



**McGRAW-HILL**  
**ENCYCLOPEDIA OF**  
**SCIENCE &**  
**TECHNOLOGY**

[www.MHEST.com](http://www.MHEST.com)

**12** **NOB-PAP**

# N

## Nobelium — Nymphaeales

### Nobelium

A chemical element, No, atomic number 102. Nobelium is a synthetic element produced in the laboratory. It decays by emitting an alpha particle, that

1																	18
H																	He
3	4															10	
Li	Be															Ne	
11	12	13	14	15	16	17									18		
Na	Mg	Al	Si	P	S	Cl	Ar									Kr	
19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36
K	Ca	Sc	Ti	V	Cr	Mn	Fe	Co	Ni	Cu	Zn	Ga	Ge	As	Se	Br	Kr
37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54
Rb	Sr	Y	Zr	Nb	Mo	Tc	Ru	Rh	Pd	Ag	Cd	In	Sn	Sb	Te	I	Xe
55	56	71	72	73	74	75	76	77	78	79	80	81	82	83	84	85	86
Cs	Ba	Lu	Hf	Ta	W	Re	Os	Ir	Pt	Au	Hg	Tl	Pb	Bi	Po	At	Rn
87	88	103	104	105	106	107	108	109	110	111	112	113					
Fr	Ra	Lr	Rf	Db	Sg	Bh	Hs	Mt	Ds	Rg							

lanthanide series	57	58	59	60	61	62	63	64	65	66	67	68	69	70
	La	Ce	Pr	Nd	Pm	Sm	Eu	Gd	Tb	Dy	Ho	Er	Tm	Yb

actinide series	89	90	91	92	93	94	95	96	97	98	99	100	101	102
	Ac	Th	Pa	U	Np	Pu	Am	Cm	Bk	Cf	Es	Fm	Md	No

is, a doubly charged helium ion. Only atomic quantities of the element have been produced to date. Nobelium is the tenth element heavier than uranium to be produced synthetically. It is the thirteenth member of the actinide series, a rare-earth-like series of elements. See ACTINIDE ELEMENTS; PERIODIC TABLE; RADIOACTIVITY; RARE-EARTH ELEMENTS; TRANSURANIUM ELEMENTS. Paul R. Fields

Bibliography. S. Hofmann, *On Beyond Uranium: Journey to the End of the Periodic Table*, 2002; G. T. Seaborg and W. D. Loveland, *The Elements Beyond Uranium*, 1990.

### Noeggerathiales

An incompletely known and poorly defined group of vascular plants whose geologic range extends from Upper Carboniferous to Triassic. Their taxonomic status and position in the plant kingdom are uncertain since morphological evidence (because of the

paucity of the fossil record) does not make it possible to place the group confidently in any recognized major subdivision of the vascular plants. A rather heterogeneous assemblage of foliar and vegetative organs assignable to fairly well-defined genera have been placed in the group, thus somewhat forcing the concept that these parts constitute a natural order of vascular plants. Internal anatomy is unknown, with exception of the possible noeggerathialean genus *Sphenostrobus*. Noeggerathialean genera include *Noeggerathia*, *Noeggerathiostrobus*, *Tingia*, *Tingiostrachya*, and *Discinites*. The reproductive organs are strobiloid, with whorled organization, and vary from homosporous to heterosporous. In *Discinites* the number of megaspores may range from 1 to 16 per sporangium. Foliar organs vary from nonarticulate and fernlike to anisophyllous four-ranked fronds. The Noeggerathiales have been proposed in the evolutionary scheme for vascular plants. See PALEOBOTANY; PLANT KINGDOM. Elso S. Barghoorn

### Noise measurement

The process of quantitatively determining one or more properties of acoustic noise. In noise assessment and control studies, knowledge of the physical properties of the undesirable sound is the initial step toward understanding the situation and what should be done to reduce or eliminate a noise problem. The instruments and techniques used for measuring noise are of fundamental importance in noise assessment and control. Many measurements may be required to characterize noise adequately because generally it is possible to make a noise measurement only at a particular point in space, while a noise field often has significant spatial variations.

The most common measures of noise are of the magnitude and frequency content of the noise sound pressure, time-averaged or as a function of time. Of increasing interest are metrics of sound quality (that

## 2 Noise measurement

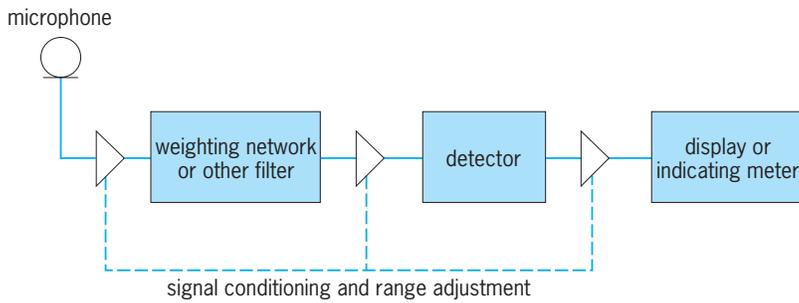


Fig. 1. Functional block diagram of a sound-level meter.

may include both physical and psychoacoustic factors), such as loudness, pitch strength, and fluctuation strength. To characterize the noise output of a source, sound power level may be determined. To locate a source, or to quantify propagation paths, sound intensity level may be measured.

Essentially all noise measurements are performed using electronic equipment. An electroacoustic transducer (a microphone in air and other gases; a hydrophone in water and other liquids) transforms sound (usually the sound pressure) at the point of observation into a corresponding electrical signal. This electrical signal is then operated on by analog or digital means with devices such as signal conditioners, filters, and detectors to determine the value (or values) of interest. This value is then given on an indicating meter or digital display. *See* ELECTRIC FILTER; HYDROPHONE; MICROPHONE; TRANSDUCER.

For noise measurements in air, a sound-level meter is the most commonly used instrument. Specifications for such instruments are given by national

and international standards. The simplest sound-level meter comprises a microphone, a frequency-weighting filter, a root-mean-square (rms) detector, and logarithmic readout of the sound pressure level in decibels relative to 20 micropascals (Figs. 1 and 2). Standard frequency weightings, designated A and C, were originally developed to approximate human response to noise at low and high levels, respectively, but now are used as specified in standards and legislation without regard to their origin (Fig. 3). [Other weightings also have been used and may be found in older or special-purpose sound-level meters.] More sophisticated sound-level meters incorporate standardized octave-band or fractional-octave-band filters, and provide additional metrics and analysis capabilities. Formerly such capability was found only in instruments called analyzers, but the differences that distinguish a sound-level meter from an analyzer and a computer are becoming size and portability rather than analysis capability. *See* DECIBEL; LOUDNESS.

**Applications.** Selection of a noise metric and measurement instrument generally depends upon the application. Customary metrics vary because, in practice, what is desired to be known varies with application. Emission metrics characterize noise source output. Immission metrics characterize noise received by a listener. Generalization is difficult, and some overlap is inevitable, but it is convenient to recognize four classes of application: (1) discrete source emission, (2) hearing conservation, (3) outdoor environmental noise, and (4) indoor room noise. Within these classes there are found additional subclasses, measurements for enforcement of legal limits often

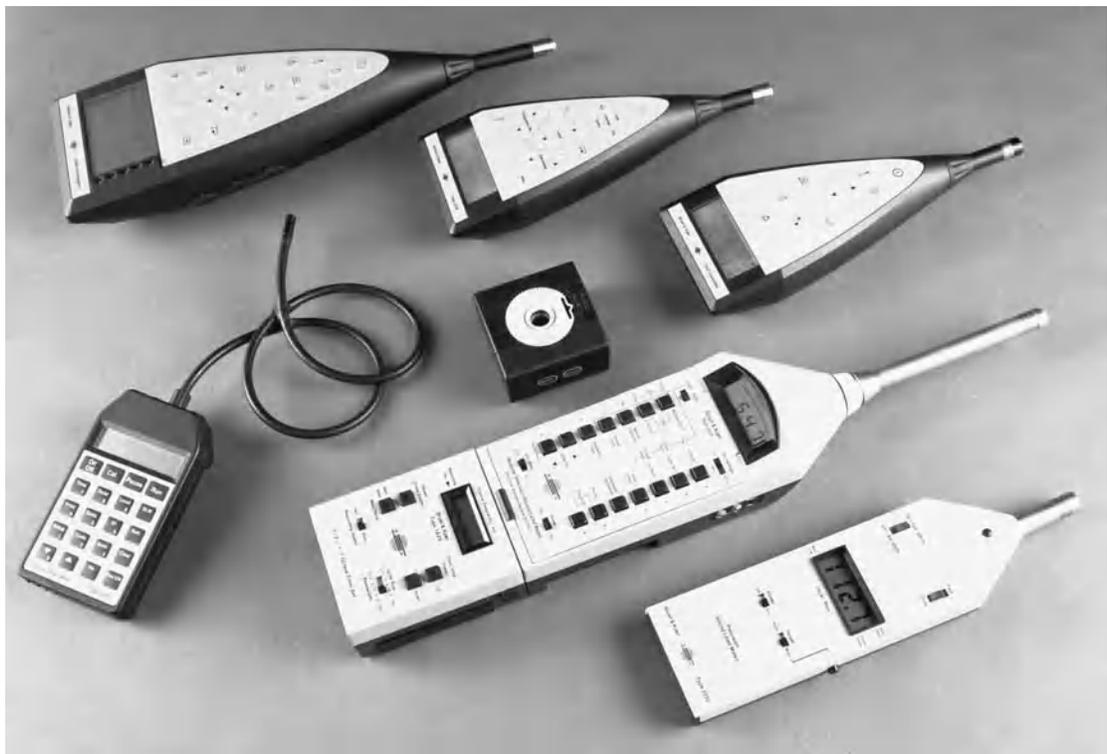


Fig. 2. Sound-level meters. (Brüel & Kjær)

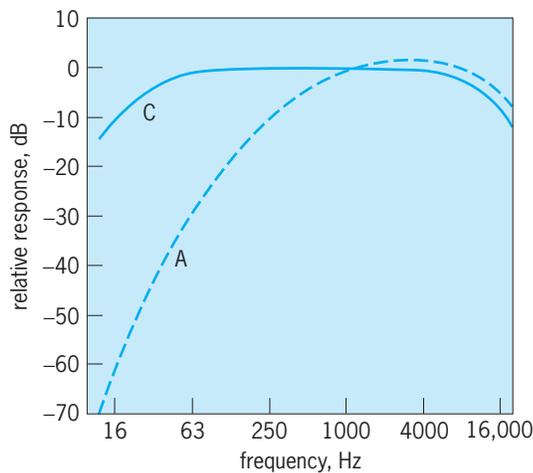


Fig. 3. Frequency weightings A and C specified for sound-level meters in standards of the International Electrotechnical Commission and the American National Standards Institute.

being done differently than those for comfort, aircraft sonic boom being measured differently than power plant noise, and so on.

Source emission is best characterized by sound power level, a quantity not directly measurable. Sound intensity or pressure measurements under controlled conditions are used to determine power. Sound power level is expressed in decibels relative to 1 picowatt. (While the decibel unit is used for sound power level as for sound pressure level, the reference quantity is different and power is a quantity fundamentally different from pressure.) Also used for the source emission description is the sound pressure level at a defined operator or bystander position. Because sound pressure level generally is a function of both the source emission and the acoustical environment, it is important that the measurement conditions be properly specified. Frequency analysis and spectral data are common as these are needed to gauge human response and to calculate the noise immission that is expected with a particular installation (in an enclosure or behind a noise barrier, for example). Transient exposures may be characterized by sound exposure level, the level of an integral of the mean-square frequency-weighted sound pressure.

Hearing conservation noise measurements generally are concerned with two quantities: time-weighted-average A-weighted sound level expressed as a noise dose, and C-weighted peak sound pressure level. The former characterizes the potential for damage due to chronic exposure, the latter the potential for immediate damage. Different jurisdictions have established various thresholds and time weightings. In the United States, for example, different schemes are used for industrial, underground mine, and military exposures.

Outdoor environmental noise may be characterized by short-term metrics, usually A-weighted sound level, as well as by longer-term metrics. Statistical metrics ( $L_N$ ) characterize the percentage of time that the sound level exceeds a stated value ( $L_{90}$ , for ex-

ample, being the level exceeded during 90% of the observation time, and  $L_{10}$  the level exceeded 10%). Such measures give an indication of both the range of levels encountered and their duration. Time-average sound level (sometimes designated  $L_{eq}$ ) disregards variations to arrive at an equivalent steady level. The period of the time average depends on the application, with averages hourly, daily, or over as much as a year being common. Modified time-average sound levels include the day-night average level and the day-evening-night average level, with penalty weightings added for nighttime (10 dB) and both night (10 dB) and evening (5 dB) noise, respectively, people being presumed more sensitive to noise at the penalty times.

Indoor room noise measures commonly incorporate both amplitude and frequency dependence. Established criteria, such as NC, RC, and NCB, are based on octave-band sound pressure levels plotted on a special chart (or the equivalent in a spreadsheet) to determine the value of the criterion. The details of these methods continue to evolve.

**Metrics.** Several common metrics are used to characterize noise. Many metrics are stated in units of decibels, and caution is necessary in comparing measurements of one metric with those of another. The effect of a given number of decibels depends as much, if not more, on the physical quantity being described by the level as on the numerical value of the level. A C-weighted peak sound level of 80 dB re 20  $\mu$ Pa with a duration of only a few milliseconds, for example, will have an impact quite different from that of a time-average sound level of 80 dB re 20  $\mu$ Pa.

The sound power level is ten times the logarithm to the base ten of the ratio of a given sound power, the sound energy radiated (by a source) per unit of time, in a stated frequency band to the standard reference sound power of 1 picowatt.

The sound intensity level is ten times the logarithm to the base ten of the ratio of the magnitude of the sound intensity vector, the vector giving the net rate of sound power transmission through a unit area normal to the vector, in a stated direction to the standard reference intensity of 1 picowatt per square meter. See SOUND INTENSITY.

The sound pressure level is ten times the logarithm to the base ten of the ratio of the square of sound pressure in a stated frequency band to the square of the standard reference sound pressure (20  $\mu$ Pa in air, 1  $\mu$ Pa in water). Root-mean-square pressure is assumed unless otherwise stated, and for root-mean square pressure the type of time average and time constant or averaging time should be stated. See SOUND PRESSURE.

Octave and one-third-octave band levels are levels of the stated kind (power, intensity, or pressure), determined in standardized frequency bands having uniform width on a logarithmic scale. Such levels usually are determined and reported without A or C weighting.

The sound level is the A-weighted exponential time-average root-mean-square sound pressure level

obtained with the F time constant (0.125 second). Alternatively C weighting or the S time constant (1.0 s) may be specified.

The time-average sound level (equivalent continuous A-weighted sound pressure level) is the A-weighted linear time-average root-mean-square sound pressure level obtained with a specified averaging time.

The sound exposure level is ten times the logarithm to the base ten of the ratio of the time integral of the squared A-weighted sound pressure, over a stated time interval or event, to the product of the squared standard reference sound pressure (20  $\mu$ Pa in air) and a reference duration of 1 second.

The sound level exceedance is the sound level exceeded a stated percent of the observation time. It is usually determined from a probability histogram of measured sound levels.

The noise dose is the standardized time-weighted average A-weighted sound level expressed as a percentage of an allowable limit. It is usually employed in assessing the potential for noise to cause hearing impairment.

Noise Criteria (NC, NCB) and Room Criteria (RC) are ratings used to assess noise in rooms.

The root-mean-square sound pressure is an element of several noise metrics. The root-mean-square is the effective average, the usual average (mean) being zero for an oscillating quantity such as sound pressure. The common mathematical definition of root-mean-square for a time-dependent signal, the square root of the time-mean of the square of the signal, presumes a signal of infinite duration and time-invariant root-mean-square value, and this assumption is not always appropriate for noise measurements. The exponential time-average root-mean-square value,  $p_{\text{rms},\tau}$ , of a signal  $p(t)$  is determined from Eq. (1), where  $\tau$  is the time constant and  $\xi$

$$p_{\text{rms},\tau}(t; \tau) = \sqrt{\frac{1}{\tau} \int_{-\infty}^t p^2(\xi) e^{-(t-\xi)/\tau} d\xi} \quad (1)$$

is a dummy variable of integration. This kind of root-mean-square value is easily implemented as a running average, being realized with the a simple RC electronic circuit or other means. Selection of a short time constant permits tracking of variations in the signal, while a long time constant effectively eliminates them. The linear time-average root-mean-square value,  $p_{\text{rms},T}$ , is determined from Eq. (2),

$$p_{\text{rms},T}(t; T) = \sqrt{\frac{1}{T} \int_{t-T}^t p^2(\xi) d\xi} \quad (2)$$

where  $T$  is the averaging (also called integration) time. Linear time averaging gives equal weight to all of the signal within the averaging interval.

**Instruments.** Sound-level meter is a general term for any instrument designed to measure sound (and noise) levels. An instrument that conforms with the requirements of a relevant national or international

standard specification should be used so that the measurements will have a known uncertainty that permits results to be compared with confidence. (Standards for sound-level meters require that an instrument in conformance be marked as such; hence, lack of marking implies that the device is not in conformance with a standard.) The minimum requirement for a standard sound-level meter is that it provide an exponential time-weighted root-mean-square sound level, F or S, with frequency-weighting A or C. Generally three classes of instrument accuracy are recognized: types 0, 1, and 2, having overall uncertainty generally considered to be  $\pm 0.5$ , 1, and 2 dB, respectively. (The uncertainty specification is frequency-dependent and for certain situations may exceed these values.) F (formerly “fast”) and S (formerly “slow”) time weighting correspond to exponential time constants of 0.125 and 1.0 second, respectively. An impulse sound-level meter includes the ability to measure the root-mean-square sound pressure of rapid transients with I (formerly “impulse”) time weighting, specified as an exponential time average with a 35-millisecond rise time followed by a peak detector with an exponential decay having a 1.5-second time constant. An integrating-averaging sound-level meter includes the capability of measuring time-average sound level and usually sound exposure level. Other metrics may be included at the option of the manufacturer. A noise dosimeter is a kind of special-purpose sound-level meter designed to be worn, providing a readout of the noise dose experienced by the wearer.

A frequent addition to the basic sound-level meter is an octave or one-third-octave band filter set that permits determination of spectrum levels. The most common approach is to measure individual frequency bands in succession, called serial analysis. Parallel analysis, measuring all bands simultaneously, previously available in analyzers intended for laboratory use, is becoming available in hand-held portable instruments.

A significant distinction between an analyzer and a sound-level meter is that, while the latter is a complete instrument, an analyzer usually requires adjunct equipment such as a transducer and input signal conditioner. Octave-band and fractional-octave-band analyzers are based on a set of filters called constant-percentage-bandwidth (cpb) filters or constant Q filters, and produce sound spectra that have uniform resolution on a logarithmic frequency scale. A logarithmic scale is convenient to display the wide range of frequencies to which human hearing is sensitive, and correlates reasonably well with the human sensation of pitch. Certain measurements are more conveniently displayed on a linear frequency scale. A filter set with constant bandwidth (cb), having uniform resolution on a linear frequency scale, is more appropriate for this purpose. A device commonly called an FFT (fast Fourier transform) analyzer performs constant-bandwidth spectrum analysis digitally, using an efficient implementation of the discrete Fourier transform. FFT analyzers usually implement a number of analyses that extend and

supplement spectrum determination. See ACOUSTIC NOISE; FOURIER SERIES AND TRANSFORMS; PITCH; SOUND.  
Joseph Pope

Bibliography. Acoustical Society of America, *Catalog of ANSI, ISO, and IEC Standards in Acoustics*, revised annually; American Society of Heating, Refrigerating, and Air-Conditioning Engineers Inc., *ASHRAE Handbook: Fundamentals*, 2001; American Society of Heating, Refrigerating, and Air-Conditioning Engineers Inc., *ASHRAE Handbook: Heating, Ventilating, and Air-Conditioning Applications*, 1999; L. L. Beranek and I. L. Ver (eds.), *Noise and Vibration Control Engineering: Principles and Applications*, John Wiley, 1992; M. J. Crocker (ed.), *Encyclopedia of Acoustics*. 4 vols., John Wiley, 1997; C. M. Harris (ed.), *Handbook of Acoustical Measurements and Noise Control*, 3d ed., McGraw-Hill, 1991, and Acoustical Society of America, 1997; A. D. Pierce, *Acoustics: An Introduction to Its Physical Principles and Applications*, McGraw-Hill, 1981, and Acoustical Society of America, 1989.

### Nomograph

A graphical relationship between a set of variables that are related by a mathematical equation or law. The fundamental principle involved in the construction of a nomographic or alignment chart consists of representing an equation containing three variables,

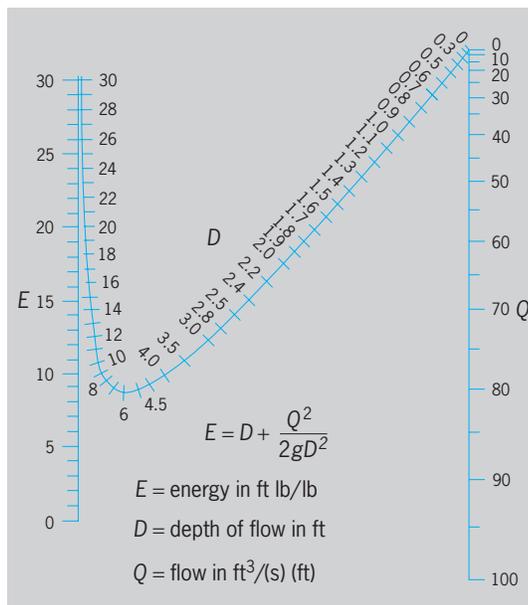


Fig. 1. Nomograph for energy content of a rectangular channel with uniform flow.

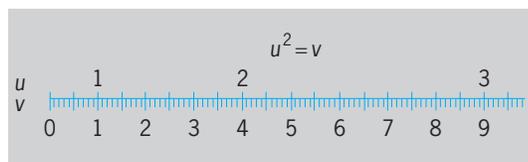


Fig. 2. Conversion scales for finding squares and square roots.

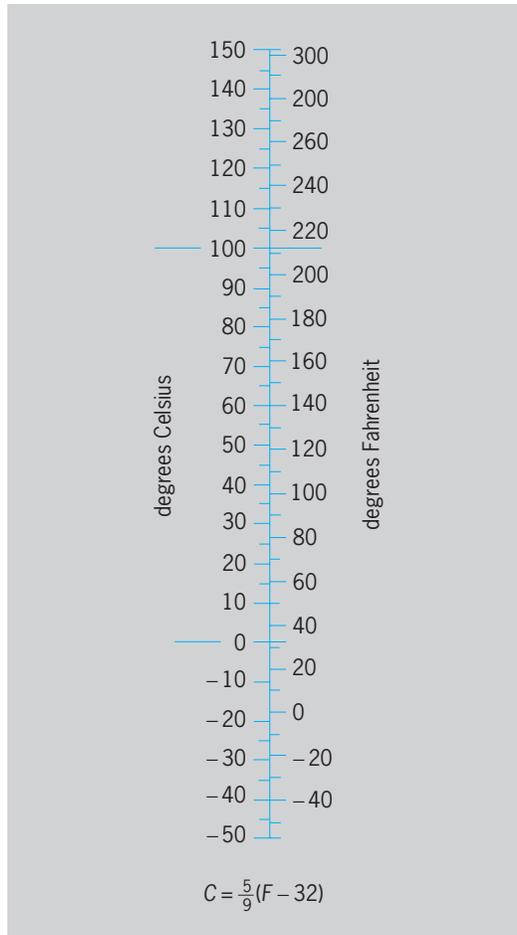


Fig. 3. Stationary scales relate Celsius and Fahrenheit temperatures.

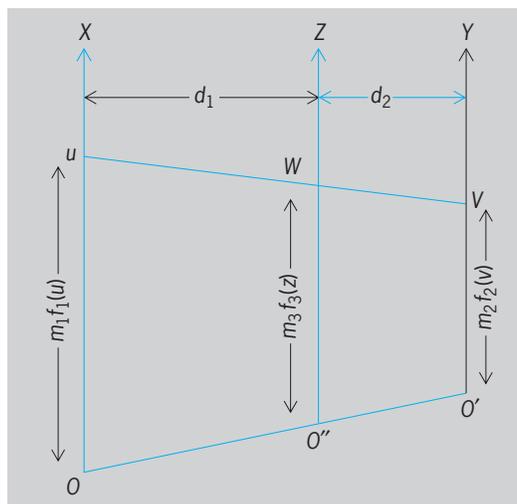


Fig. 4. Alignment chart for simple summation.

$f(u, v, w) = 0$ , by means of three scales in such a manner that a straight line cuts the three scales in values of  $u$ ,  $v$ , and  $w$ , satisfying the equation. The cutting line is called the isopleth or index line. Numbers may be quickly and easily read from the scales of such a chart even by one unfamiliar with the construction of the chart and the equation involved. **Figure 1** illustrates such an example. Assume that it is desired to find the value of  $E$  when  $D = 2$  and  $Q = 50$ . Lay a

straightedge through 50 on the  $Q$  scale and through 2 on the  $D$  scale and read 11.8 at its intersection with the  $E$  scale. As another example, it might be desired to know what value or values of  $D$  should be used if  $E$  and  $Q$  are required to be 10 and 60, respectively. A straightedge through  $E = 10$  and  $Q = 60$  cuts the  $D$  scale in two points,  $D = 2.8$  and  $9.4$ . This is equivalent to finding two positive roots of the cubic equation  $D^3 - 10D^2 + 56.25 = 0$ . It is assumed that  $g = 32 \text{ ft/s}^2$  in this equation.

**Scale.** The graphical scale is a curve or straight line, called an axis, on which is marked a series of

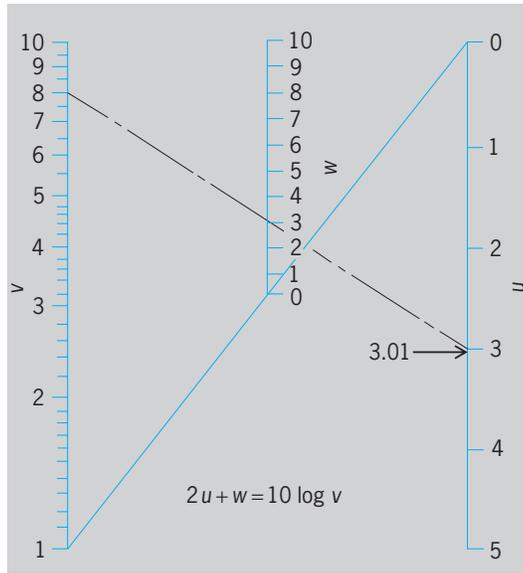


Fig. 5. Variation of alignment chart using logarithm.

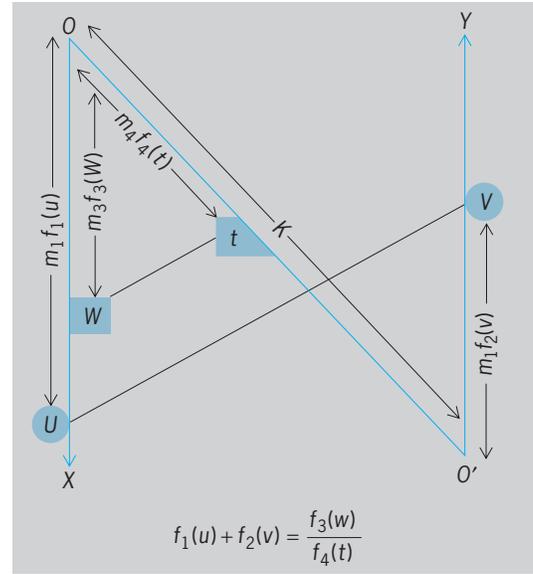


Fig. 7. Alignment chart for summation involving a ratio.

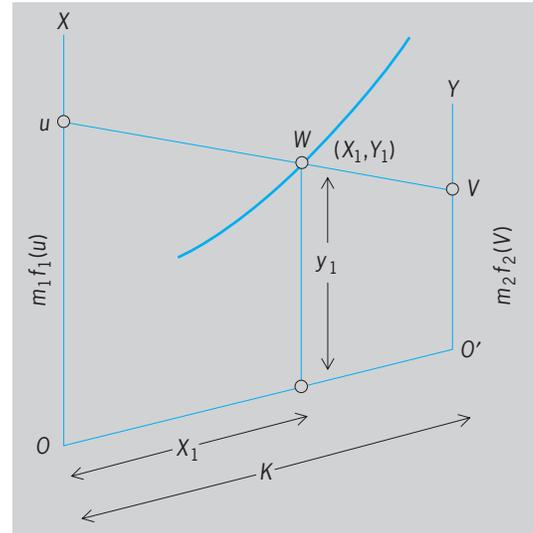


Fig. 8. Alignment chart for summation using a product.

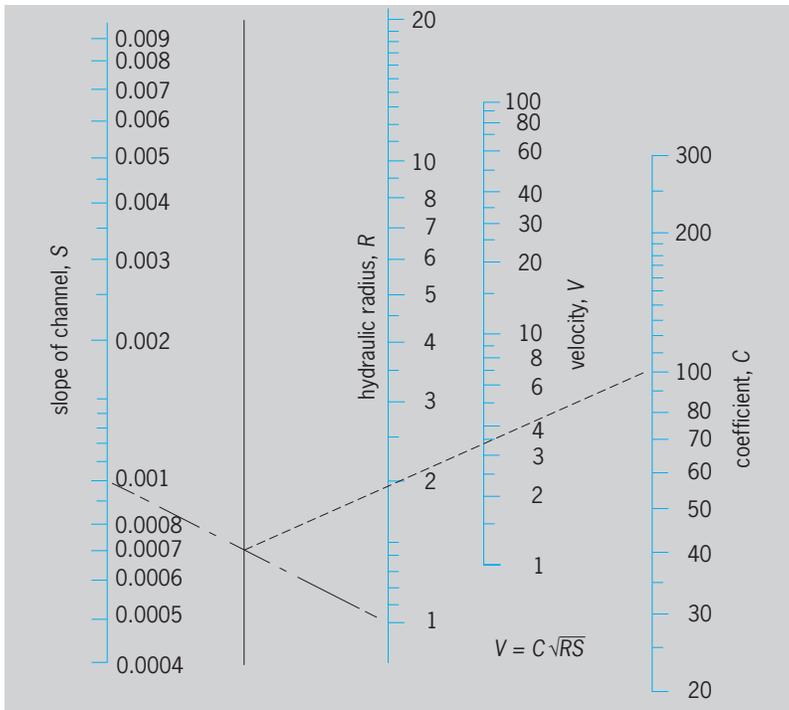


Fig. 6. Nomograph for velocity of flow of water in open channels.

points or strokes corresponding to a set of numbers arranged in order of magnitude. If the distances between successive strokes are equal, the scale is uniform; otherwise the scale is nonuniform. The scale on a yardstick or thermometer is uniform, whereas a logarithmic scale on a slide rule is nonuniform.

**Representation of a function by a scale.** Consider the function  $f(u)$ . Lay off, from a fixed point  $O$  on a straight line or curve, lengths equal to  $f(u)$  units; mark at the strokes indicating the end of each unit the corresponding value of  $u$ . If the unit of measure is an inch, the equation of the scale is  $x = f(u)$  inch. More generally, the equation of the scale is  $x = mf(u)$  units, where the constant  $m$  (modulus) regulates the length of scale used for the required values of the variable  $u$  needed.

**Stationary scales.** A relation between two variables of the form  $v = f(u)$ , or  $f(v) = f(u)$ , can be represented

by the natural scales  $x = mv$  and  $x = mf(v)$ ; or  $x = mf(v)$  and  $x = mF(u)$ , on the opposite side of the same line or axis. **Figure 2** shows the relation  $u^2 = v$  using the natural scales  $x = mu^2$  and  $x = mv$ , where in this illustration  $m = 0.43$  and the unit is an inch. By using logarithms the above equation becomes  $2 \log u = \log v$ , and the scales are  $x = m(2 \log u)$  and  $x = m \log v$ .

Adjacent stationary scales may be used to advantage in representing the relationship between the two variables in a conversion formula. **Figure 3** shows the relation between degrees Celsius and Fahrenheit. It is easy to see that  $F = 140$  when  $C = 60$ ; and when  $F = -40$ ,  $C = -40$ .

*Perpendicular scales.* A relation between two variables  $u$  and  $v$  of the form  $v = f(u)$ ,  $f(u, v) = 0$ , or  $f(u) = F(v)$  can be represented by constructing two scales  $x = mf(u)$  and  $y = mF(v)$  on perpendicular axes. Any pair of values of  $u$  and  $v$  will determine a point in the plane. The locus of all such points is a curve which represents the relationship between the variables  $u$  and  $v$ . The various types of coordinate paper, sold commercially, are constructed in this manner. Log-log, semilog, and reciprocal coordinate papers are probably the most common. These types of scales and their combinations are essential in the construction of nomographic charts, especially when the number of variables involved exceeds three.

**Types of nomographic charts.** The form of the equation serves to classify the type of chart. An equation of the form of Eqs. (1) and (2) leads to a chart

$$f_1(u) + f_2(v) = f_3(w) \quad (1)$$

$$f_1(u) + f_2(v) + \dots = f_n(t) \quad (2)$$

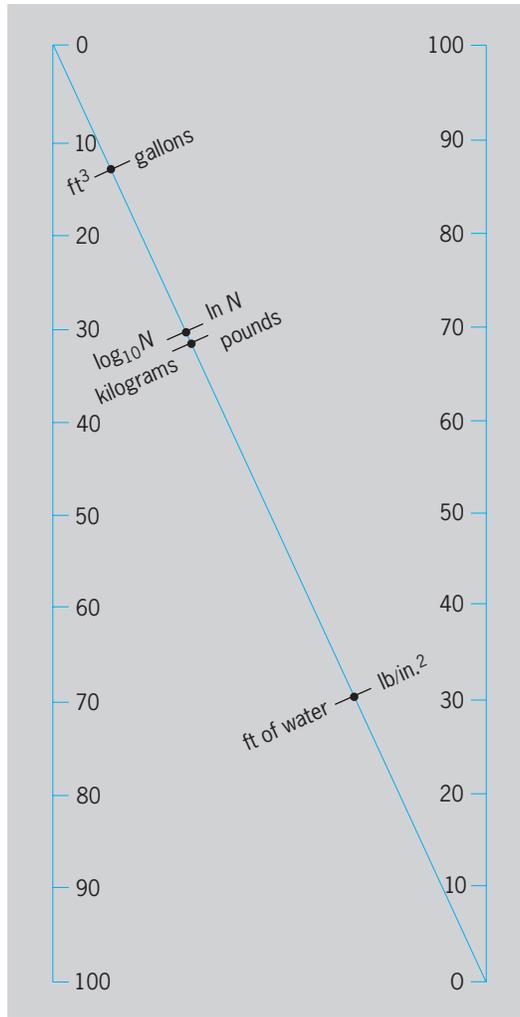
of the type shown in **Fig. 4**. The equations of the three scales are  $x = m_1 f_1(u)$ ,  $m_2 f_2(v)$ , and  $[m_1/m_2/(m_1 + m_2)] f_3(w)$ , respectively; and  $d_1/d_2 = m_1/m_2$ . The equation  $2u + w = 10 \log v$  is in this form. Taking  $m_1 = m_2 = 1$  and  $d_1 = d_2$ , the scales are  $10 \log v$ ,  $-2u$ , and  $w/2$ . The completed chart is shown in **Fig. 5**. As an example, if  $v = 8$  and  $w = 3$ , the value of  $u$  is found to be 3.01.

Another form of equation that leads to a similar chart is Eq. (3). Using logarithms, this equation takes the form of Eq. (1), which is Eq. (4). The Chezy

$$f_1(u)f_2(v) \dots = f_n(t) \quad (3)$$

$$\log f_1(u) + \log f_2(v) + \dots = \log f_n(t) \quad (4)$$

formula for velocity of the flow of water  $v = c(RS)^{1/2}$  is of this type. When logarithms are used, the equation becomes  $\log v = \log c + 1/2 \log R + 1/2 \log S$ , which may be written  $1/2 \log S + 1/2 \log R = \log v - \log c = Q$ , where  $Q$  is a dummy variable. The completed chart is shown in **Fig. 6**. To find the value of  $v$  when  $R = 1$ ,  $S = 0.001$ , and  $c = 100$ , set the straight-edge on  $S = 0.001$  and  $R = 1$ ; now connect the point of intersection on the dummy scale to  $c = 100$  and read  $v = 3$  at the point of crossing on the  $v$  scale.



**Fig. 9.** Conversion chart relates several different units of measure on linear scales.

Alternatively the equation may be of the form of Eq. (5). Using logarithms, Eq. (5) also takes the form of Eq. (1), which is Eq. (6).

$$[f_1(u)]^{f_2(v)} = f_3(w) \quad (5)$$

$$\log f_2(v) + \log \log f_1(u) = \log \log f_3(w) \quad (6)$$

A second form is Eq. (7), where  $m_1/K = m_3/m_4$ . Such a chart is shown in **Fig. 7**.

$$f_1(u) + f_2(v) = \frac{f_3(w)}{f_4(t)} \quad (7)$$

A third form is Eq. (8).

$$f_1(u) + f_2(v)f_3(w) = f_4(t) \quad (8)$$

Equations (9) and (10) apply. Such charts are

$$x_1 = \frac{m_1 K}{m_1 f_3(w) + m_2} f_3(w) \quad (9)$$

$$y_1 = \frac{m_1 m_2}{m_1 f_3(w) + m_2} f_4(w) \quad (10)$$

illustrated by Fig. 1 and Fig. 8. This is a frequently encountered type of equation. In constructing the nomograph of Fig. 1, first the equation  $D = D + Q^2/2gD^2$  was rewritten in the form  $-Q^2 + 2gD^2E = 2gD^2$ , which corresponds to the basic form of Eq. (8).

A fourth basic form is Eq. (11), where  $m_1/m_2 =$

$$\frac{f_1(u)}{f_2(v)} = \frac{f_3(w)}{f_4(t)} \quad (11)$$

$m_3/m_4$ . This can be treated as a Z-chart by using natural scales, or it takes the form of Eq. (1) by using logarithms.

A fifth form is the conversion chart, Eq. (12). This

$$f_1(u) = C f_2(v) \quad (12)$$

is a special case of Eq. (1) or (11) and is useful when several conversions are to be made. Examples are Figs. 9 and 10. In practice it may be useful or even necessary to use two or more combinations of these basic types.

**Determinant as a basis of a nomograph.** The condition that the three points  $(x_1, y_1)$ ,  $(x_2, y_2)$ , and  $(x_3, y_3)$  lie on a straight line is that the determinant equal zero, as in Eq. (13). If an equation that relates three variables  $u$ ,  $v$ , and  $w$  can be expressed in the deter-

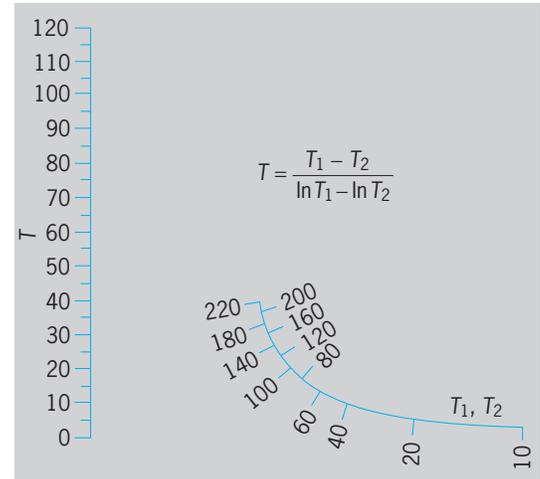


Fig. 11. Nomograph for mean temperature. To find mean temperature, lay straightedge between two given temperatures on curved scale, and then read the mean temperature on the vertical scale.

minant form shown as Eq. (14), then a nomograph can be constructed from parametric equations (15).

$$\begin{vmatrix} x_1 & y_1 & 1 \\ x_2 & y_2 & 1 \\ x_3 & y_3 & 1 \end{vmatrix} = 0 \quad (13)$$

$$\begin{vmatrix} f_1(u) & g_1(u) & 1 \\ f_2(v) & g_2(v) & 1 \\ f_3(w) & g_3(w) & 1 \end{vmatrix} = 0 \quad (14)$$

$$\begin{aligned} x &= f_1(u), & y &= g_1(u) \\ x &= f_2(v), & y &= g_2(v) \\ x &= f_3(w), & y &= g_3(w) \end{aligned} \quad (15)$$

The mean temperature equation (16) expressed as a determinant is Eq. (17). Figure 11 shows the resulting nomograph.

$$T = \frac{T_1 - T_2}{\ln T_1 - \ln T_2} \quad (16)$$

$$\begin{vmatrix} 0 & T & 1 \\ 1/\ln T_1 & T_1/\ln T_1 & 1 \\ 1/\ln T_2 & T_2/\ln T_2 & 1 \end{vmatrix} = 0 \quad (17)$$

Figure 1 could also have been constructed from the determinant form after the energy equation is written as Eq. (18).

$$\begin{vmatrix} 0 & (12.2/100^2)Q^2 & 1 \\ 9 & -E/3 & 1 \\ 9D^2 & -D^3 & 1 \\ D^2 + \frac{100^2}{73.2g} & 3D^2 + \frac{100^2}{73.2g} & 1 \end{vmatrix} = 0 \quad (18)$$

**Circular nomographs.** The general form of the basic determinant for a circular nomograph is Eq. (19), where the  $u$  and  $v$  scales lie on circular axes having the same center and with radii equal to  $1/2$ . Consider

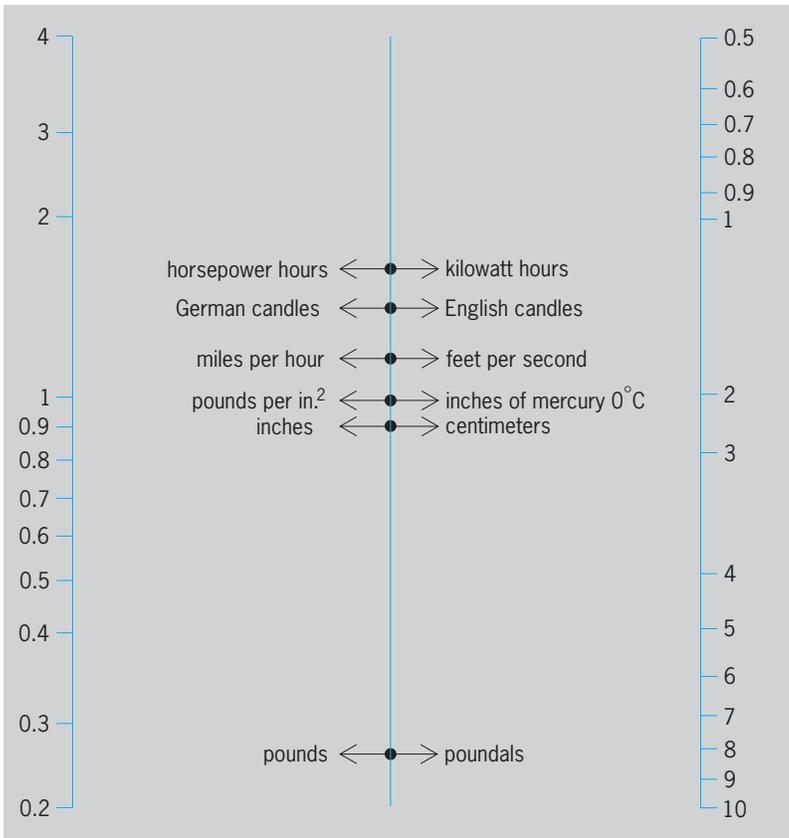


Fig. 10. Conversion chart relates several different units of measure on an exponential scale.

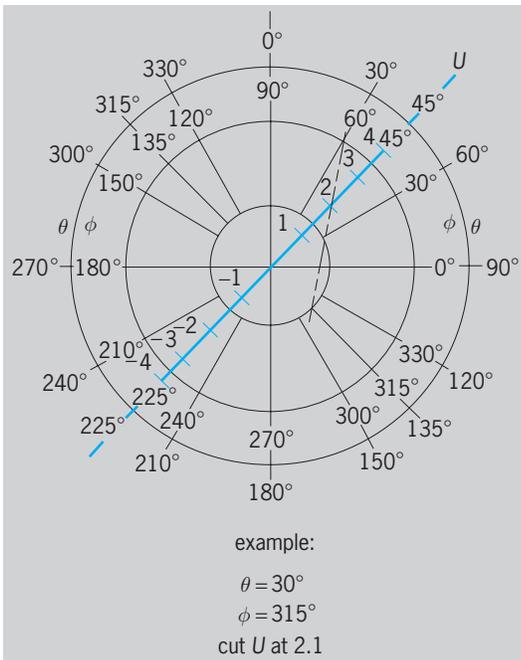


Fig. 12. Circular nomograph which results from trigonometric equation expressed in form of determinant.

the equation  $AB \cos(\theta - \mu) + BU(\sin \phi - \cos \theta) + AU(\sin \theta - \cos \phi) = 0$ . Expressed in the form of a determinant, it is Eq. (20), which leads to the nomo-

$$\begin{vmatrix} 1 & f_1(u) & 1 \\ 1 + [f_1(u)]^2 & 1 + [f_1(u)]^2 & 1 \\ 1 & f_2(v) & 1 \\ 1 + [f_2(v)]^2 & 1 + [f_2(v)]^2 & 1 \\ 1 & 0 & 1 \\ 1 + f_3(w) & 0 & 1 \end{vmatrix} = 0 \quad (19)$$

$$\begin{vmatrix} A \sin \theta & A \cos \theta & 1 \\ B \cos \phi & B \sin \phi & 1 \\ U & U & 1 \end{vmatrix} = 0 \quad (20)$$

graph of Fig. 12. See DETERMINANT; LINEAR SYSTEMS OF EQUATIONS. Raymond D. Douglass

Bibliography. L. H. Johnson, *Nomography and Empirical Equations*, 1952, reprint 1978; J. F. Kuong, *Applied Nomography*, 3 vols., 1965-1969; J. Molnar, *Nomographs: What They Are and How to Use Them*, 1981.

### Nondestructive evaluation

Nondestructive evaluation (NDE) is a technique used to probe and sense material structure and properties without causing damage. It has become an extremely diverse and multidisciplinary technology, drawing on the fields of applied physics, artificial intelligence, biomedical engineering, computer science, electrical engineering, electronics, materials science and engineering, mechanical engineering, and structural engineering. Historically, NDE techniques have been used almost exclusively for detection of macroscopic

defects (mostly cracks) in structures which have been manufactured or placed in service. Using NDE for this purpose is usually referred to as nondestructive testing (NDT).

A developing use of NDE methods is the nondestructive characterization (NDC) of materials properties (as opposed to revealing flaws and defects). Characterization typically sets out to establish absolute or relative values of material properties such as mechanical strength (elastic moduli), thermal conductivity or diffusivity, optical properties, magnetic parameters, residual strains, electrical resistivity, alloy composition, the state of cure in polymers, crystallographic orientation, and the degree of crystalline perfection. Nondestructive characterization can also be used for a variety of other specialized properties that are relevant to some aspect of materials processing in production, including determining how properties vary with the direction within the material, a property called anisotropy.

Much effort has been directed to developing techniques that are capable of monitoring and controlling (1) the materials production process; (2) materials stability during fabrication, transport, and storage; and (3) the amount and rate of degradation during the postfabrication in-service life for both components and structures. Real-time process monitoring for more efficient real-time process control, improved product quality, and increased reliability has become a practical reality. Simply put, intelligent manufacturing is impossible without integrating modern NDE into the production system. What was once a nonsystematic collection of after-the-fact techniques has become a powerful set of tools which, if used throughout the manufacturing process, can help to transform that process to make products better and more competitive. See MATERIALS SCIENCE AND ENGINEERING.

### Mature Technologies

A number of NDE technologies have been in use for some time.

**Visual inspection.** This is the oldest and most versatile NDE tool. In visual inspection, a worker examines a material using only eyesight. Visual inspection relies heavily on common sense and is vulnerable to erratic results because of human failure to comprehend the significance of what is seen.

**Liquid (or dye) penetrant.** This augmented visual method uses brightly colored liquid dye to penetrate and remain in very fine surface cracks after the surface is cleaned of residual dye.

**Magnetic particle.** This augmented visual method requires that a magnetic field be generated inside a ferromagnetic test object. Flux leakage occurs where there are discontinuities on the surface. Magnetic particles (dry powder or a liquid suspension) are captured at the leakage location and can be readily seen with proper illumination.

**Eddy current.** This method uses a probe held close to the surface of a conducting test object. The probe is a coil of fine wire on a flat spool energized by an ac oscillator, creating an ac magnetic field

orthogonal to the spool. The oscillating magnetic field penetrating the test object induces a circular oscillating image current in the plane of the surface. If the coil is moved across the surface, discontinuities in or near the surface perturb the flow of the induced current which can be detected as a change in the impedance that the original coil presents to the ac oscillator. See EDDY CURRENT.

**X-rays.** These provide a varied and powerful insight into material, but they are somewhat limited for use in the field. Radiography is similar to the familiar chest x-ray, except that it is used to image materials rather than people. Computed tomography is the nonpeople version of medical CAT scanning. Flash or pulse x-ray systems can provide x-ray images of dynamic events, much as a camera with a flash can “freeze” an athlete in midair or a horse at the finish line. (Some of these systems are suitcase size and battery-operated.) X-ray diffraction can provide information about crystallographic defects in such things as single-crystal blades for jet engines (and ground-based gas turbines) and semiconductor single crystals. **Figure 1** shows a diffraction image of a nickel-based alloy single crystal blade. Real-time radioscopy and real-time x-ray diffraction use charge-coupled-device (CCD) cameras, image intensifiers, and phosphor screens to record x-ray images at video rates and faster. The images go directly to computers where digital image processing can enhance their usefulness. See CHARGE-COUPLED DEVICES; RADIOGRAPHY; X-RAY DIFFRACTION.

**Acoustic emission.** This technique typically uses a broadband piezoelectric transducer to listen for acoustic noise. At first glance, acoustic emission (AE) appears as an attractive technology because its sensors may be configured to listen all the time, thereby capturing an acoustic emission event no matter



Fig. 1. X-ray diffraction images from a turbine blade.

when it occurs. In contrast, ultrasonic methods acquire information only when active probing is taking place. Much effort has been expended trying to use acoustic emission to detect plastic deformation and crack growth in structures as a means of monitoring system health. Except for some special cases, attempts to make broader use of acoustic emission have generally been disappointing. Crack growth certainly generates acoustic emissions, but so do a lot of other mechanisms. A highway bridge loaded with live traffic, for example, generates much acoustic noise from many different but benign sources. The difficulty is separating the signature, if any, due to crack growth. Before acoustic emission technology can be used to its full potential, identification of source mechanisms and the associated structural alteration of materials and structures must be better understood. See ACOUSTIC EMISSION; PIEZOELECTRICITY; PLASTIC DEFORMATION OF METAL.

**Thermography.** This technique uses a real-time “infrared camera,” much like a home camcorder, except that it forms images using infrared photons instead of visible ones. Imaging rates can vary from a few frames per second to hundreds of frames per second. The choice of camera depends on trade-offs between requirements for frame rate, brightness sensitivity, and spatial resolution. A representative commercial infrared camera can easily image the thermal trail written on paper at room temperature by a moving finger. A breakthrough has been the replacement of camera sensors that see only one pixel at a time with sensors that have a full field of pixels detected simultaneously. A consequence of this development is that an entire infrared image can be plucked out of a sequence, like a freeze-frame, at a predetermined time after an event. If the image has random noise (like “snow” in TV images), and if the event can be repeated over and over, then images can be plucked out over and over after exactly the same time delay. Adding such identical images together will increase the clarity of the image and reduce the snow.

**Contact ultrasonics.** This technique is the workhorse of traditional and mature NDE technology. It uses a transducer held in contact with a test object to launch ultrasonic pulses and receive echoes. These pulses are mechanical vibrational waves, which commonly contain frequencies in the megahertz range. The transducers are usually piezoelectric elements tuned to a narrow band (their resonant frequency), and often require a couplant (such as special greases) to ensure efficient transmission of the vibrations into the test object. In some cases a water path is used to couple the transducer to the test object. In looking for internal flaws, the same transducer is often used to first launch a pulse and subsequently receive its echoes. Comparing echoes from the other side of the object with echoes that return earlier determines the location of flaws. It is often advantageous to use a second transducer to function solely as a receiver. Velocity and attenuation of the waves can be determined, and in turn yield information on material properties such as elastic moduli, grain size, and porosity. See TRANSDUCER.

### Emerging Technologies

Many noncontact measurements have been developed that enhance the mature technologies.

**Noncontact ultrasonic transducers.** These involve laser ultrasonics, electromagnetic acoustic transducers, and air- or gas-coupled transducers.

Laser ultrasonics affords the opportunity to make truly noncontact measurements in materials at elevated temperatures, in hostile environments, and in locations geometrically difficult to reach. One major drawback of laser-based ultrasonic systems is that laser interferometers are generally poor receivers, particularly for shear-wave displacements. Improving received signal levels by using larger generation pulses is limited by the need to leave the surface of the material undamaged. Employing temporal or spatial modulation of the generation process has been effective in gaining increases in the signal-to-noise ratios of detected signals. See LASER; ULTRASONICS.

Acoustic vibrations in a conducting material in the presence of a dc magnetic field induce small ac currents. Electromagnetic acoustic transducers (EMATs) both provide the dc magnetic field and sense the induced currents. They also can operate in the inverse fashion, but EMATs are much better at detecting than generating. Although noncontact, EMAT-based systems are not truly remote because their efficiency rapidly decreases with the lift-off distance between the transducer face and the test object surface. Typically lift-off needs to be no larger than perhaps a millimeter (0.04 in.). Another drawback is that EMAT-based systems work only on electrically conducting materials.

Air- or gas-coupled transducers operate with much larger lift-off distances than EMATs. Their principal advantages are that they do not require mirror surface finishes (as do laser interferometers), nor do they require the test object to be conductive. The main challenge in their implementation comes from the large acoustic impedance mismatch between air and the solid objects under study; however, existing systems have succeeded in overcoming this challenge effectively. For example, a C scan system has been developed that uses air coupling for both generation and detection (Fig. 2).

**Hybrid laser ultrasonics.** Since lasers are very effective at generating "clean" ultrasonic excitation to satisfy a variety of needs, hybrid systems generally use laser generation with something other than a laser interferometer for ultrasonic detection. For example, EMATs can make excellent detectors in place of lasers. Air-coupled transducers can also provide many advantages over laser interferometers in particular applications.

Historically, making engineered monolithic structures of composite materials has been a highly labor-intensive manufacturing operation because of the tedious process of hand-layup of the prepreg material. In addition, curing the resulting structure required autoclave processing of the entire structure all at once. These practices have contributed to high-

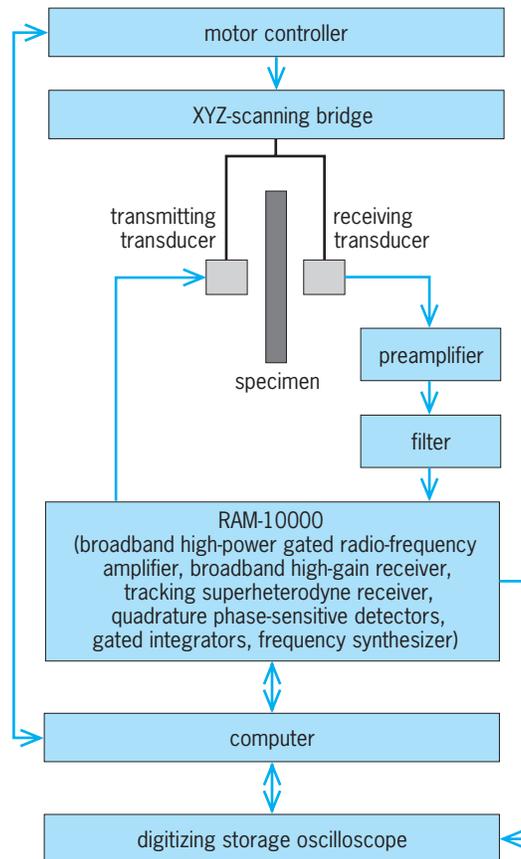


Fig. 2. Air-coupled ultrasonic C scan system.

cost, time-consuming, labor-intensive procedures, and a limit on structural size and shape to fit into an autoclave. Quality control and reproducibility of the hand-layup depended on the skill and dedication of the specialists doing the layup. To address these processing problems, the industry evolved a technology using a robot to perform precision placement of prepreg tape and to consolidate and cure the tape as it is laid (eliminating the need for an autoclave). The problem has been the need for a closed-loop real-time capability to provide quality inspection for voids, porosity, and state of cure as the tape goes down. A hybrid ultrasonic remote inspection system has been developed for this purpose (Fig. 3). It provides real-time data "on the fly" as the tape is placed in position by robot, consolidated by roller, and cured by localized thermal deposition. Narrow-band generation of surface waves is accomplished by a spatial modulation procedure that employs a periodic transmission mask. Detection in this system is by air-coupled transduction. An improvement in this same system has been the deployment of a fiber-optic light delivery system. The fiber-optic umbilical is able to deliver pulsed light from a neodymium:yttrium-aluminum-garnet (Nd:YAG) laser at an average energy as high as 55 millijoules per pulse for a pulse duration as short as 10 nanoseconds, at a 20-hertz repetition rate, without signs of fiber damage. The center frequency of the generated toneburst signal can be controlled easily by changing the periodicity

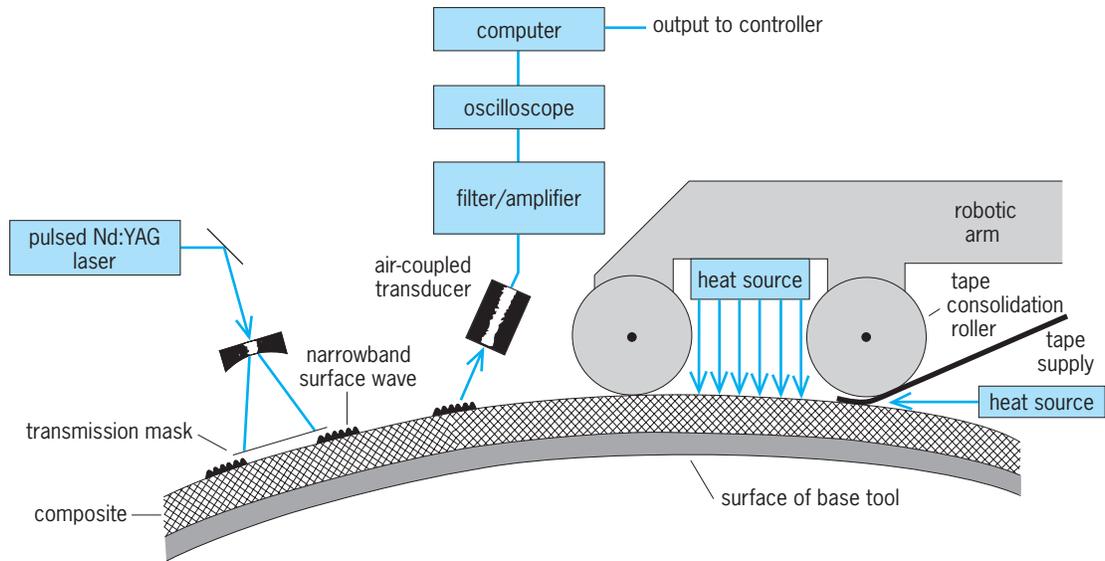


Fig. 3. Hybrid ultrasonic inspection for automated “on the fly” composite fabrication.

of the transmission mask. This step, in turn, allows control of the surface-wave penetration depth, which is determined by the wave frequency content. In addition, compared to conventional single-pulse illumination, the spatial modulation technique allows more light to illuminate the test surface while retaining thermoelastic generation conditions, thus increasing the signal detectability without material damage and ablation. See COMPOSITE MATERIAL; ROBOTICS.

**Nonlinear ultrasonics.** This has become a promising field. Generation of ultrasound in a selected range

of frequencies with detection in a different range can be sensitive to changes in materials not detectable with the common practice of generating and detecting in the same range. The coupling into other ranges is a consequence of higher-order nonlinearities in the elastic moduli (commonly presumed linear) which govern the propagation of ultrasound.

**Thermal wave imaging.** This uses a main laser beam to scan the surface of the object to be examined. The amount of laser energy absorbed by the surface is directly dependent on the surface’s thermal properties. Thermal expansion resulting from this localized energy absorption creates very slight vertical surface displacements which vary depending on the amount of laser energy absorbed (Fig. 4). In addition, the air close above the surface is heated slightly, changing its density, and therefore its index of refraction. A secondary “probe” laser beam can detect either the changes in the refractive index of the air (skim probe) or the resultant surface displacements (reflection probe). The infrared emission from the surface can be imaged by thermography as described previously, or an ultrasonic detector can be used to examine the interior using transmitted elastic waves generated by the sudden thermoelastic expansions from the heating beam pulses. If these various means of detecting are scanned systematically over the surface, high-resolution images can be generated that reveal very slight differences in the rate at which heat flows away from the surface; or they can provide ultrasonic image data about the internal structure. These differences map either inhomogeneities in the thermal properties of the material immediately below the surface, or subsurface structural flaws such as voids, cracks, or delaminations. An example of this technique is detection of areas of lack of adhesion of protective turbine blade coatings used on nickel-based superalloy blades for aviation gas turbines (jet engines).

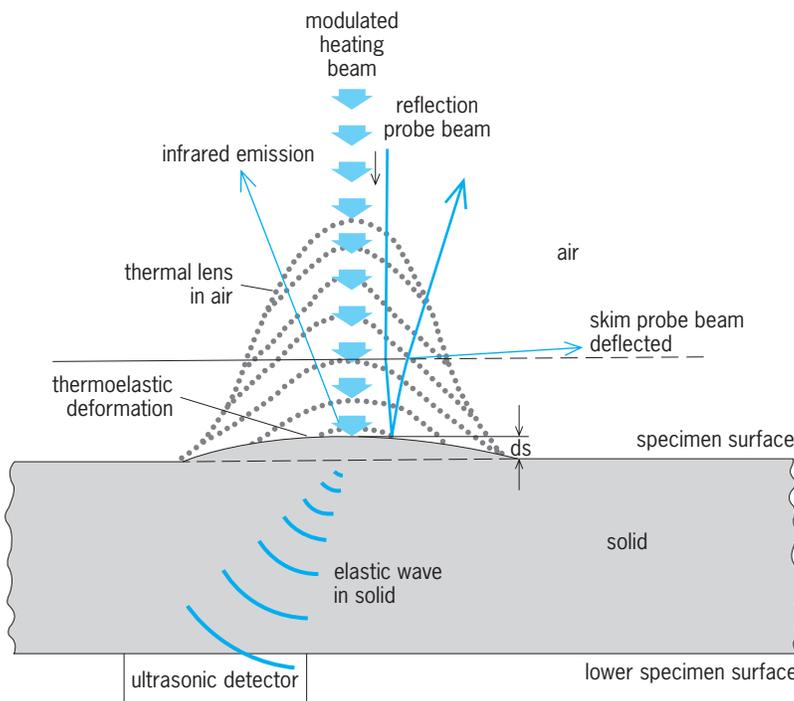


Fig. 4. Principles of thermal wave imaging. Skim probe beam detects changes in the refractive index of air; reflection probe beam detects resultant surface displacements; infrared emission is measured by thermography; and ultrasonic detector examines the interior of the specimen.

**Speckle-based interferometry.** Electronic speckle pattern interferometry is a noncontact, full-field

optical technique for high-sensitivity measurement of extremely small displacements in an object's surface. "Speckle" refers to the grainy appearance of an optically rough surface illuminated by a laser. It is actually a very complex interference pattern resulting from diffuse scattering from the rough surface. The pattern is unique for a given surface, like a fingerprint. If the surface is deformed, the pattern changes. By mathematically comparing two speckled images of the surface before and after deformation, a fringe pattern depicting the displacement distribution is generated. "Full-field" refers to the fact that an area of the test specimen is imaged, not just one point. The key concept behind this approach is that flaws or damage in the test object create local perturbations of the surface displacement profile when the object is loaded (for example, stretched). The speckle comparison patterns serve to indicate the presence, location, and other details of the flaws. A variation, called electronic shearography, measures displacement gradients instead of absolute displacements. Shearography needs only a very simple optical setup since it is based on a self-referencing coherent optical system. Consequently, vibration isolation and laser coherence length considerations may be satisfied more easily than for other coherent optical techniques such as holographic interferometry. High-sensitivity, high-resolution, ease of use, and full-field capabilities make speckle-based techniques a valuable tool for industrial nondestructive testing. See INTERFEROMETRY; NONLINEAR OPTICAL DEVICES.

**Microwave techniques.** Development activities are under way in such diverse applications as ground-penetrating radar for land-mine detection, locating delaminations in highway bridge decks, and monitoring the curing process in polymers. See MICROWAVE.

John M. Winter, Jr.

**Bibliography.** H. L. M. dos Reis, *Nondestructive Testing and Evaluation for Manufacturing and Construction*, Hemisphere Publishing, New York, 1990; J. Holbrook and J. Bussiere (eds.), *Nondestructive Monitoring of Materials Properties*, Materials Research Society, Pittsburgh, 1989; J. Krautkramer and H. Krautkramer, *Ultrasonic Testing of Materials*, 3d ed., Springer-Verlag, New York, 1983; P. K. Liaw et al. (eds.), *Nondestructive Evaluation and Materials Properties III*, The Minerals, Metals, and Materials Society, Warrendale, PA, 1996.

## Noneuclidean geometry

A system of geometry based upon a set of axioms different from those of the euclidean geometry based on the three-dimensional space of experience. Noneuclidean geometries, especially riemannian geometry, are useful in mathematical physics. This article will describe some of the basic concepts and axioms of noneuclidean geometries and then draw some comparisons with euclidean concepts. See DIFFERENTIAL GEOMETRY; EUCLIDEAN GEOMETRY; PROJECTIVE GEOMETRY; RIEMANNIAN GEOMETRY.

The famous names that are related to hyperbolic

(noneuclidean) geometry are J. Bolyai and N. I. Lobachevski. In addition, A. Cayley, E. Beltrami, K. F. Gauss, and G. F. B. Riemann have made many outstanding contributions to the subject of noneuclidean geometry.

**Projective space.** Let  $K$  denote a given field of elements, called the ground field  $K$ . In actual applications,  $K$  will be simply the field of real numbers. Consider the class  $\Sigma$  of ordered sets of  $(n + 1)$ -tuples  $x = (x^0, x^1, \dots, x^n)$ , where each  $x^i$  is an element of the ground field  $K$ . The index  $i$  is not an exponent but is merely a superscript. In this article, both superscripts and subscripts will be used. If  $x = (x^0, x^1, \dots, x^n)$  and  $y = (y^0, y^1, \dots, y^n)$  are two elements in the class  $\Sigma$ , then the sum  $(x + y)$  is defined to be the element  $x + y = (x^0 + y^0, x^1 + y^1, \dots, x^n + y^n)$  in  $\Sigma$ . Also, if  $x = (x^0, x^1, \dots, x^n)$  is an element of  $\Sigma$  and if  $\rho$  is an element of the ground field  $K$ , then the scalar product  $\rho x$  is defined to be the element  $\rho x = (\rho x^0, \rho x^1, \dots, \rho x^n)$  in  $\Sigma$ .

If  $x_1, x_2, \dots, x_m$  are  $m$  elements in  $\Sigma$  and if  $\rho_1, \rho_2, \dots, \rho_m$  is a set of  $m$  elements from the ground field  $K$ , then the element  $x = \rho_1 x_1 + \rho_2 x_2 + \dots + \rho_m x_m$  of  $\Sigma$  is called a linear combination of the elements  $x_1, x_2, \dots, x_m$  of  $\Sigma$ .

The class  $\Sigma^1$  is a proper subset of  $\Sigma$  which is composed of all ordered sets of  $(n + 1)$ -tuples  $x = (x^0, x^1, \dots, x^n)$  of  $\Sigma$  such that at least one  $x^i$  of the  $(n + 1)$  elements  $x^0, x^1, \dots, x^n$  of the ground field  $K$  is not zero. Two elements  $x = (x^0, x^1, \dots, x^n)$  and  $y = (y^0, y^1, \dots, y^n)$ , of the class  $\Sigma^1$ , are said to be equivalent if, and only if, there exists an element  $\rho \neq 0$ , of the ground field  $K$ , such that  $y = \rho x$ . This establishes an equivalence relation among the elements  $x = (x^0, x^1, \dots, x^n)$  of the class  $\Sigma^1$ . The equivalence classes in  $\Sigma^1$ , formed relative to this equivalence relation, are called points  $P$ . The collection of all these points  $P$  forms a projective space  $S_n$  of  $n$  dimensions over the ground field  $K$ .

A point  $P$  in  $S_n$  is a class in  $\Sigma^1$  of all ordered sets of  $(n + 1)$ -tuples equivalent to a given set of  $(n + 1)$ -tuples, namely,  $x = (x^0, x^1, \dots, x^n)$ . Each of these ordered sets of  $(n + 1)$ -tuples is in the equivalence class called the point  $P$  and is said to be a set of homogeneous point coordinates for the point  $P$ . Thus, if  $x = (x^0, x^1, \dots, x^n)$  is a set of homogeneous point coordinates of a point  $P$  in  $S_n$ , any other set of homogeneous point coordinates for the same point  $P$  is  $\rho x = (\rho x^0, \rho x^1, \dots, \rho x^n)$ , where  $\rho \neq 0$  is an arbitrary element of the ground field  $K$ .

A collection of  $(p + 1)$  points  $x_0, x_1, \dots, x_p$  in  $S_n$  is said to be linearly dependent if, and only if, at least one of them is a linear combination of the remaining ones such that at least one of the scalar coefficients is not zero. Otherwise they are said to be linearly independent. Clearly, if  $(p + 1)$  points  $x_0, x_1, \dots, x_p$  in  $S_n$  are linearly independent, then  $0 \leq p \leq n$ .

Let a collection of  $(p + 1)$  linearly independent points  $x_0, x_1, \dots, x_p$  in  $S_n$  be given. Then  $0 \leq p \leq n$ . The  $S_p$  determined by them is composed of all points  $x$  in  $S_n$  such that Eq. (1) holds, where at least one of

$$x = y^0 x_0 + y^1 x_1 + \dots + y^p x_p \quad (1)$$

the scalar multiples  $y^0, y^1, \dots, y^p$  is not zero. This  $S_p$  is a projective space of  $p$  dimensions over the ground field  $K$ , and is a projective subspace of  $S_n$ . A set of homogeneous point coordinates for an arbitrary point  $P$  in this  $S_p$  is  $(y^0, y^1, \dots, y^p)$ .

Define  $S_{-1}$  to be a vacuous collection of points  $P$  in  $S_n$ . Such an  $S_{-1}$  is called a projective subspace of  $S_n$  of dimension  $-1$ . Evidently every  $S_0$  is a point  $P$  and conversely. A line is an  $S_1$ , a plane is an  $S_2$ , a three-space is an  $S_3$ , and a  $p$ -space is an  $S_p$ . In particular, a hyperplane is an  $S_{n-1}$ .

**Dual coordinates.** An  $S_p$  is a locus of points  $x$  whose homogeneous point coordinates  $x = (x^0, x^1, \dots, x^n)$  satisfy a system of  $(n - p)$  linear homogeneous equations (2), such that the rank of the coefficient matrix  $(u_{ij})$  is  $(n - p)$ .

$$\sum_{j=0}^n u_{ij}x^j = 0 \quad (i = 1, 2, \dots, n - p) \quad (2)$$

In particular, a hyperplane  $S_{n-1}$  is a locus of points  $x$  whose homogeneous point coordinates  $x = (x^0, x^1, \dots, x^n)$  satisfy a single linear homogeneous equation (3), where at least one of the  $(n + 1)$  elements  $u_0, u_1, \dots, u_n$ , of the ground field  $K$ , is not zero.

$$\begin{aligned} ux &= \sum_{i=0}^n u_i x^i \\ &= u_0 x^0 + u_1 x^1 + \dots + u_n x^n = 0 \end{aligned} \quad (3)$$

Because an ordered set of  $(n + 1)$ -tuples  $u = (u_0, u_1, \dots, u_n)$  of this type uniquely defines an  $S_{n-1}$ , it is called a set of homogeneous hyperplane coordinates, for the particular  $S_{n-1}$ . If  $u = (u_0, u_1, \dots, u_n)$  is a set of homogeneous hyperplane coordinates of an  $S_{n-1}$ , any other set of homogeneous hyperplane coordinates for the same  $S_{n-1}$  is  $\sigma u = (\sigma u_0, \sigma u_1, \dots, \sigma u_n)$ , where  $\sigma \neq 0$  is an arbitrary element of the ground field  $K$ .

In the projective space  $S_n$  of  $n$  dimensions, a point  $P$  and a hyperplane  $S_{n-1}$  are termed dual objects. Thus, homogeneous point coordinates  $x = (x^0, x^1, \dots, x^n)$  and homogeneous hyperplane coordinates  $u = (u_0, u_1, \dots, u_n)$  are said to be dual systems of coordinates.

A point  $P$  with the homogeneous point coordinates  $x = (x^0, x^1, \dots, x^n)$  is on the hyperplane  $S_{n-1}$  with the homogeneous hyperplane coordinates  $u = (u_0, u_1, \dots, u_n)$  if, and only if, the condition of Eq. (3) is satisfied.

By mathematical induction, it follows that the dual of a projective subspace  $S_p$  of  $p$  dimensions is a projective subspace  $S_{n-p-1}$  of  $(n - p - 1)$  dimensions. For example, the dual of  $S_{-1}$  is  $S_n$ , the dual of an  $S_0$  is an  $S_{n-1}$ , and the dual of an  $S_1$  is an  $S_{n-2}$ .

The projective subspace  $S_p$  is said to be contained in the projective subspace  $S_q$ , or the projective subspace  $S_q$  contains the projective subspace  $S_p$ , if, and only if, every point of  $S_p$  is a point of  $S_q$ . In particular,  $S_{-1}$  is contained in every  $S_q$ , and  $S_n$  contains every  $S_p$ . This relation is written as  $S_p \subset S_q$ , or  $S_q \supset S_p$ .

The dual of the relation  $S_p \subset S_q$  is the dual relation  $S_{n-p-1} \supset S_{n-q-1}$ .

**Principle of duality.** For every theorem concerning the  $S_p$ 's and the relations  $S_p \subset S_q$ , there is obtained a dual theorem wherein each  $S_p$  is replaced by the dual object  $S_{n-p-1}$ , and each relation  $S_p \subset S_q$  is replaced by the dual relation  $S_{n-p-1} \supset S_{n-q-1}$ .

If  $S_p$  and  $S_q$  are any two projective subspaces of the projective space  $S_n$ , the largest projective subspace contained in both  $S_p$  and  $S_q$  is denoted by  $S_p \cap S_q$ , and the smallest projective subspace containing both  $S_p$  and  $S_q$  is denoted by  $S_p \cup S_q$ . Clearly,  $S_p \cap S_q$  is the intersection or meet, and  $S_p \cup S_q$  is the union or join of  $S_p$  and  $S_q$ . See SET THEORY.

The two relations  $S_p \cap S_q$  and  $S_p \cup S_q$  are dual. Concerning the inclusion relation  $S_p \subset S_q$ , the projective subspaces of the projective space  $S_n$  form a complemented modular lattice.

Finally, let  $\dim(S_p)$  denote the dimension  $p$  of the projective subspace  $S_p$  of the projective space  $S_n$ . The dimension theorem may be stated in the form shown in Eq. (4).

$$\begin{aligned} \dim(S_p \cap S_q) + \dim(S_p \cup S_q) \\ = \dim(S_p) + \dim(S_q) \end{aligned} \quad (4)$$

**Projective group.** A collineation  $T$  of the projective space  $S_n$  is a one-to-one correspondence between the points of  $S_n$ , of the form shown in Eq. (5), where

$$\rho \bar{x}^i = \sum_{j=0}^n a_j^i x^j \quad (i = 0, 1, \dots, n) \quad (5)$$

the rank of the coefficient matrix  $(a_j^i)$  is  $(n + 1)$ , and  $\rho \neq 0$  is an arbitrary constant of proportionality. Of course, all elements are from the original ground field  $K$ .

A collineation  $T$  is not only a one-to-one correspondence between the points  $S_0$  of  $S_n$ , but also a one-to-one correspondence between the  $S_p$ 's of  $S_n$ ; that is, under  $T$ , any  $S_p$  is carried into one, and only one,  $\bar{S}_p$ , and conversely.

The set of collineations  $T$  of  $S_n$  forms the collineation group  $G$  of  $S_n$ . It is composed of  $n(n + 2)$  essential parameters.

The fundamental theorem of projective geometry may be stated in the following form: There is a collineation  $T$  which carries a given set of  $(p + 1)$  linearly independent points into a prescribed set of  $(p + 1)$  linearly independent points. Moreover, if  $p = n$ , then  $T$  is uniquely determined.

A correlation  $\Gamma$  of the projective space  $S_n$  is a one-to-one correspondence between the points  $S_0$  and the hyperplanes  $S_{n-1}$  of  $S_n$ , of the form shown in Eq. (6), where the rank of the coefficient matrix  $(b_{ij})$

$$\sigma \bar{u}_i = \sum_{j=0}^n b_{ij} x^j \quad (i = 0, 1, \dots, n) \quad (6)$$

is  $(n + 1)$ , and  $\sigma \neq 0$  is an arbitrary constant of proportionality.

The set  $C$  of correlations  $\Gamma$  of  $S_n$  is composed of  $n(n + 2)$  essential parameters. In general,  $C$  is not a group.

The total projective group  $G^*$  of the projective space  $S_n$  is composed of collineations  $T$  and correlations  $\Gamma$ . It is a mixed group  $G^*$  of  $n(n+2)$  essential parameters. Obviously, the collineation group  $G$  is a subgroup of the total projective group  $G^*$ .

Projective geometry consists of the qualitative and quantitative invariants of the total projective group  $G^*$ .

**Cross ratio.** Let  $S_{r-1}$  and  $S_{r+1}$  be two fixed projective subspaces of  $S_n$  of dimensions  $(r-1)$  and  $(r+1)$ , respectively, such that  $S_{r-1}$  is contained in  $S_{r+1}$ . Evidently  $0 \leq r \leq n-1$ . A pencil is composed of all the  $S_r$ 's that contain the given  $S_{r-1}$  and that are contained in the given  $S_{r+1}$ .

For example, where  $r=0$ , one obtains a pencil of points, all of which are in a fixed line  $S_1$ . Similarly when  $r=n-1$ , there is defined a pencil of hyperplanes  $S_{n-1}$ , all of which contain a fixed projective subspace  $S_{n-2}$  of  $(n-2)$  dimensions.

A pencil of elements  $P$  is a projective space of one dimension. Therefore, the elements  $P$  of a pencil are defined by a system of homogeneous coordinates  $(\rho, \tau)$ , where at least one of the elements  $\rho$  and  $\tau$  of the ground field  $K$  is not zero. If it is assumed that  $\tau$  is always one, then  $\rho$  is said to be the nonhomogeneous coordinate of the element  $P$  which is not the element at infinity  $(1,0)$ . The element of infinity  $(1,0)$  in this system of coordinates is denoted by the symbol  $\infty$ .

If  $(\rho_1, \tau_1), (\rho_2, \tau_2), (\rho_3, \tau_3), (\rho_4, \tau_4)$  are the homogeneous coordinates of four distinct elements  $P_1, P_2, P_3, P_4$ , their cross ratio  $\mathfrak{R}$  is defined to be the numerical invariant given in Eq. (7). This is the fun-

$$\mathfrak{R}(P_1P_2, P_3P_4) = \frac{(\rho_3\tau_1 - \rho_1\tau_3)(\rho_2\tau_4 - \rho_4\tau_2)}{(\rho_2\tau_3 - \rho_3\tau_2)(\rho_4\tau_1 - \rho_1\tau_4)} \quad (7)$$

damental projective invariant. In nonhomogeneous coordinates, this is Eq. (8).

$$\mathfrak{R}(P_1P_2, P_3P_4) = \frac{(\rho_3 - \rho_1)(\rho_2 - \rho_4)}{(\rho_2 - \rho_3)(\rho_4 - \rho_1)} \quad (8)$$

In particular, Eqs. (9) and (10) hold.

$$\mathfrak{R}(\infty P_3, P_1P_2) = \frac{\rho_3 - \rho_2}{\rho_3 - \rho_1} \quad (9)$$

$$\mathfrak{R}(\infty 0, P_1P_2) = \frac{\rho_2}{\rho_1} \quad (10)$$

Of importance in projective geometry is the concept of a harmonic set of elements of a pencil provided that the ground field  $K$  is not of characteristic two. Four elements  $P_1, P_2, P_3, P_4$  of a pencil are said to form a harmonic set of elements if, and only if,  $\mathfrak{R}(P_1P_2, P_3P_4) = -1$ .

**Quadrics.** Consider a correspondence  $\Gamma$  of  $S_n$  in which every point  $P$  is converted into a single hyperplane  $S_{n-1}$ , or  $S_n$ . It is assumed that  $\Gamma$  carries every line  $S_1$  into a single  $S_{n-2}$ , or  $S_{n-1}$ , or  $S_n$ . Finally, it is supposed that if  $\Gamma$  carries a point  $P$  into an  $S_p$  and if  $\bar{P}$  is any point of this  $S_p$ , then  $\Gamma$  converts the point  $\bar{P}$  into an  $S_q$  which passes through the original point  $P$ . Such a correspondence  $\Gamma$  is called a polarity  $\Gamma$ .

In homogeneous coordinates, a polarity  $\Gamma$  is given by Eq. (11), where the matrix  $(g_{ij})$  is symmetric, that

$$\sigma \bar{u}_i = \sum_{j=0}^n g_{ij} x^j \quad (i = 0, 1, \dots, n) \quad (11)$$

is,  $g_{ij} = g_{ji}$ , and where its rank is  $(r+1)$  for which  $0 \leq r \leq n$ . If  $0 \leq r < n$ , the polarity  $\Gamma$  is said to be singular. Otherwise  $\Gamma$  is said to be nonsingular.

In the homogeneous point coordinates, the polarity  $\Gamma$  is given by Eq. (12).

$$\sum_{i,j=0}^n g_{ij} x^i \bar{x}^j = 0 \quad (12)$$

The dual of a polarity  $\Gamma$  is also a polarity  $\Gamma^*$ . Such a polarity  $\Gamma^*$  is given in homogeneous coordinates by Eq. (13), where the matrix  $(g^{ij})$  is symmetric, that is,

$$\rho \bar{x}^i = \sum_{j=0}^n g^{ij} u_j \quad (i = 0, 1, \dots, n) \quad (13)$$

$g^{ij} = g^{ji}$ , and where its rank is  $(r+1)$  for which  $0 \leq r \leq n$ .

In homogeneous hyperplane coordinates, the polarity  $\Gamma^*$  is given by Eq. (14).

$$\sum_{i,j=0}^n g^{ij} \bar{u}_i \bar{u}_j = 0 \quad (14)$$

If the polarity  $\Gamma$  is nonsingular, then the dual  $\Gamma^*$  is the original polarity  $\Gamma$ . In that event, the two matrices  $(g_{ij})$  and  $(g^{ij})$  can be considered to be inverse matrices. Thus a nondegenerate polarity can be given by Eqs. (11), (12), (13), or (14).

A hyperplane element  $(P, \pi)$  is composed of a point  $P$  and a hyperplane  $\pi$  of dimension  $(n-1)$  which passes through the point  $P$ . A quadric  $Q$  is a locus of hyperplane elements  $(P, \pi)$  such that under a given polarity  $\Gamma$  or  $\Gamma^*$ , the point  $P$  is transformed into the hyperplane  $\pi$ , or the hyperplane  $\pi$  is carried into the point  $P$ .

The polarity  $\Gamma$  or  $\Gamma^*$  is said to be a polarity relative to the corresponding quadric  $Q$ .

In homogeneous point coordinates, the equation of a quadric  $Q$  is (15).

$$\sum_{i,j=0}^n g_{ij} x^i x^j = 0 \quad (15)$$

In homogeneous hyperplane coordinates, the equation of a quadric  $Q$  is (16).

$$\sum_{i,j=0}^n g^{ij} u_i u_j = 0 \quad (16)$$

**Euclidean and noneuclidean geometries.** Consider a real projective space  $S_n$ . That is,  $S_n$  is defined over the real number system  $K$ . Sometimes it is convenient to consider that  $S_n$  is immersed in a complex projective space  $S_n^*$ , which is defined over the complex number system  $K^*$ .

Let  $k \neq 0$  be either a positive real number or infinite (that is,  $1/k = 0$ ), or else a pure imaginary number of the form  $k = il$ , where  $i^2 = -1$  and  $l$

is a positive real number. Consider the fundamental quadric  $\Sigma$  whose homogeneous point equation is (17), where the superscripts exterior to the paren-

$$f(x, x) = (kx^0)^2 + (x^1)^2 + (x^2)^2 + \dots + (x^n)^2 = 0 \quad (17)$$

theses denote exponents. The homogeneous hyperplane equation of this fundamental quadratic  $\Sigma$  is (18).

$$F(u, u) = \left(\frac{u_0}{k}\right)^2 + u_1^2 + u_2^2 + \dots + u_k^2 = 0 \quad (18)$$

The set of all collineations  $T$  of  $S_n$  which carry this fundamental quadric  $\Sigma$  into itself is a group  $G(k)$  of  $n(n+1)/2$  essential parameters. When  $k$  is a positive real number, this is the elliptic group  $G_E$  of elliptic geometry. When  $1/k = 0$ , this is the euclidean group  $G_P$  of euclidean geometry. Finally when  $k = il$  where  $i^2 = -1$  and  $l$  is a positive real number, this is the hyperbolic group  $G_H$  of hyperbolic geometry.

The study of the qualitative and quantitative invariants of these three groups  $G_E$ ,  $G_P$ , and  $G_H$  constitutes the three subjects of elliptic, euclidean, and hyperbolic geometries. The elliptic and hyperbolic geometries are usually referred to as the noneuclidean geometries.

In elliptic geometry, two distinct lines contained in a single plane always meet in a single point. Therefore, if  $L$  is a fixed line and if  $P$  is a given point not in  $L$ , there cannot be a line  $M$  passing through this point  $P$  which is parallel to the given line  $L$ .

In euclidean geometry, the ideal hyperplane  $\pi_\infty$  or the hyperplane  $\pi_\infty$  at infinity is the one whose point equation is  $x^0 = 0$ , or whose hyperplane equations are  $u_1 = 0, u_2 = 0, \dots, u_n = 0$ . The proper  $S_p$ 's for  $p = -1, 0, 1, \dots, n$  are those which are not contained in the ideal hyperplane  $\pi_\infty$ . The improper  $S_p$ 's for  $p = 0, 1, \dots, n-1$  are those which are contained in the ideal hyperplane  $\pi_\infty$ . In euclidean geometry, only proper  $S_p$ 's are studied.

If a proper  $S_p$  and a distinct proper  $S_q$  intersect in an improper  $S_r$ , then  $S_p$  and  $S_q$  are said to be parallel. Thus, two distinct lines in euclidean space are said to be parallel if, and only if, they intersect in an ideal point.

From the preceding discussion, Euclid's fifth parallel postulate is an easy consequence. That is, if in euclidean space  $L$  is a fixed line and if  $P$  is a fixed point not on  $L$ , there is one, and only one, line  $M$  parallel to  $L$  and passing through  $P$ .

In hyperbolic geometry, the points and the tangent  $S_p$ 's for  $p = 0, 1, 2, \dots, n-1$  of the fundamental quadric  $\Sigma$  given by Eq. (17) or (18) are said to be ordinary improper or ordinary infinite. The  $S_p$ 's for  $p = 0, 1, 2, \dots, n-1$ , which are in the exterior of this quadric  $\Sigma$ , are said to be ultraimproper or ultrainfinite. The proper points in hyperbolic geometry are those which are in the interior of this quadric  $\Sigma$ . The proper  $S_p$ 's for  $p = 1, 2, \dots, n-1, n$  are those which contain proper points and are considered to be sets of these proper points.

If a proper  $S_p$  and a distinct proper  $S_q$  intersect in an ordinary improper  $S_r$  or in an ultraimproper  $S_r$ , then  $S_p$  and  $S_q$  are said to be ordinary parallel or ultraparallel.

In hyperbolic geometry, if  $L$  is a fixed proper line and if  $P$  is a given proper point not on this line  $L$ , then there are two distinct proper lines  $M_1$  and  $M_2$  passing through  $P$  which are ordinary lines parallel to  $L$ .

Also passing through  $P$ , there is an infinite number of proper lines  $M$  which are ultraparallel to  $L$ . These lines  $M$  belong to the flat pencil with vertex at  $P$  and determined by the lines  $M_1$  and  $M_2$ .

**Distance.** Let  $P$  and  $Q$  be two distinct points given by the homogeneous point coordinates  $x = (x^0, x^1, \dots, x^n)$  and  $y = (y^0, y^1, \dots, y^n)$ . In euclidean and hyperbolic geometries, it is understood that  $P$  and  $Q$  are proper points. The line  $L$  determined by the two points  $P$  and  $Q$  intersects the fundamental quadric  $\Sigma$  in two distinct points  $P_\infty$  and  $Q_\infty$ . The distance  $s = s(P, Q)$  between these two points  $P$  and  $Q$  is defined by formula (19). It is understood that  $s = s(P, Q)$

$$s = s(P, Q) = \frac{k}{2i} \log \Re(PQ, P_\infty Q_\infty) \quad (19)$$

is a real nonnegative number. Also it is assumed in elliptic geometry that  $0 \leq s/k \leq \pi$ .

Let  $f(x, y)$  be expressed by Eq. (20). The two points

$$f(x, y) = k^2 x^0 y^0 + x^1 y^1 + x^2 y^2 + \dots + x^n y^n \quad (20)$$

$P$  and  $Q$  are polar reciprocal or orthogonal relative to the fundamental quadric  $\Sigma$  if and only if  $f(x, y) = 0$ . The point  $P$  is on  $\Sigma$  if, and only if,  $f(x, x) = 0$ .

A point  $R$  whose homogeneous point coordinates are  $z = (z^0, z^1, \dots, z^n)$  is on the line  $L$  determined by the two points  $P$  and  $Q$  if, and only if, a number  $\rho$  exists such that Eqs. (21) hold.

$$\begin{aligned} z^0 &= x^0 + \rho y^0 \\ z^1 &= x^1 + \rho y^1, \dots, z^n = x^n + \rho y^n \end{aligned} \quad (21)$$

This point  $R$  is on the fundamental quadric  $\Sigma$  whose point equation is Eq. (17) if, and only if,  $\rho$  satisfies the quadratic equation (22). Because the two

$$f(x, x) + 2\rho f(x, y) + \rho^2 f(y, y) = 0 \quad (22)$$

points  $P$  and  $Q$  are distinct, this will have two distinct roots  $\rho_1$  and  $\rho_2$ . The two points  $P_\infty$  and  $Q_\infty$  are on the line  $L$  whose homogeneous point equations are Eqs. (21), corresponding to the two distinct roots  $\rho_1$  and  $\rho_2$ .

In elliptic and euclidean geometries, these two points  $P_\infty$  and  $Q_\infty$  are conjugate-imaginary. In hyperbolic geometry, they are real.

From Eqs. (21) and (22), it is seen that Eq. (23) holds.

$$\begin{aligned} \Re(PQ, P_\infty Q_\infty) &= \frac{\rho_1}{\rho_2} \\ &= \frac{-f(x, y) - \sqrt{f^2(x, y) - f(x, x)f(y, y)}}{-f(x, y) + \sqrt{f^2(x, y) - f(x, x)f(y, y)}} \end{aligned} \quad (23)$$

Then, from Eq. (23), Eq. (24) applies.

$$\mathbb{R}(PQ, P_\infty Q_\infty) = \frac{\left[ f(x, y) + \sqrt{f^2(x, y) - f(x, x)f(y, y)} \right]^2}{f(x, x)f(y, y)} \quad (24)$$

Consequently the distance  $s(P, Q)$  is given by formula (25).

$$s(P, Q) = \frac{k}{i} \log \frac{f(x, y) + \sqrt{f^2(x, y) - f(x, x)f(y, y)}}{\sqrt{f(x, x)}\sqrt{f(y, y)}} \quad (25)$$

Set  $s = s(P, Q)$ . Then  $s$  is given by Eqs. (26).

$$\begin{aligned} \cos \frac{s}{k} &= \frac{f(x, y)}{\sqrt{f(x, x)}\sqrt{f(y, y)}} \\ \sin \frac{s}{k} &= \frac{\sqrt{f^2(x, y) - f(x, x)f(y, y)}}{i\sqrt{f(x, x)}\sqrt{f(y, y)}} \end{aligned} \quad (26)$$

Of course, from Eq. (19), this distance  $s = s(P, Q)$  is invariant under each of the groups  $G_E, G_P, G_H$  of elliptic, euclidean, and hyperbolic geometries.

By Eq. (20), Eqs. (26) may be written in the forms shown in Eqs. (27), in which  $\epsilon = +1$  or  $\epsilon = -1$ ,

$$\begin{aligned} \cos \frac{s}{k} &= \frac{\epsilon(k^2 x^0 y^0 + x^1 y^1 + x^2 y^2 + \dots + x^n y^n)}{\sqrt{[(kx^0)^2 + (x^1)^2 + \dots + (x^n)^2]} \cdot \sqrt{[(ky^0)^2 + (y^1)^2 + \dots + (y^n)^2]}} \\ \sin \frac{s}{k} &= \sqrt{\frac{\left[ k^2 \sum_{i=1}^n \left| \begin{smallmatrix} x^0 & x^i \\ y^0 & y^i \end{smallmatrix} \right|^2 + \frac{1}{2} \sum_{i,j=1}^n \left| \begin{smallmatrix} x^i & x^j \\ y^i & y^j \end{smallmatrix} \right|^2 \right]}{[(kx^0)^2 + (x^1)^2 + \dots + (x^n)^2]} \cdot [(ky^0)^2 + (y^1)^2 + \dots + (y^n)^2]} \quad (27) \end{aligned}$$

according to whether the geometry is elliptic (including the euclidean case) or hyperbolic.

In hyperbolic geometry,  $k = il$  where  $i^2 = -1$  and  $l$  is a positive real number. Equations (27) can be written in the forms shown in Eqs. (28).

$$\begin{aligned} \cosh \frac{s}{l} &= \frac{l^2 x^0 y^0 - x^1 y^1 - x^2 y^2 - \dots - x^n y^n}{\sqrt{[(lx^0)^2 - (x^1)^2 - \dots - (x^n)^2]} \cdot \sqrt{[(ly^0)^2 - (y^1)^2 - \dots - (y^n)^2]}} \\ \sinh \frac{s}{l} &= \sqrt{\frac{\left[ l^2 \sum_{i=1}^n \left| \begin{smallmatrix} x^0 & x^i \\ y^0 & y^i \end{smallmatrix} \right|^2 - \frac{1}{2} \sum_{i,j=1}^n \left| \begin{smallmatrix} x^i & x^j \\ y^i & y^j \end{smallmatrix} \right|^2 \right]}{[(lx^0)^2 - (x^1)^2 - \dots - (x^n)^2]} \cdot [(ly^0)^2 - (y^1)^2 - \dots - (y^n)^2]} \quad (28) \end{aligned}$$

Return to the general case of Eqs. (27). Let  $x = (x^0, x^1, \dots, x^n)$  denote affine coordinates in a euclidean space of  $(n + 1)$  dimensions. Define the special quadric  $\Sigma^*$  by Eq. (29). This is a central quadric

$$f(x, x) = (kx^0)^2 + (x^1)^2 + \dots + (x^n)^2 = k^2 \quad (29)$$

$\Sigma^*$ . Each of the two noneuclidean geometries can be visualized as the geometry on this quadric  $\Sigma^*$  in

which diametrically opposite points are identified. In particular, for elliptic geometry,  $\Sigma^*$  can be considered a sphere.

On this quadric  $\Sigma^*$ , Eqs. (27) can be written in the forms labeled Eqs. (30). Here  $k$  is a positive real

$$\begin{aligned} \cos \frac{s}{k} &= x^0 y^0 + \frac{1}{k^2} (x^1 y^1 + x^2 y^2 + \dots + x^n y^n) \\ k \sin \frac{s}{k} &= \sqrt{\sum_{i=1}^n \left| \begin{smallmatrix} x^0 & x^i \\ y^0 & y^i \end{smallmatrix} \right|^2 + \frac{1}{2k^2} \sum_{i,j=1}^n \left| \begin{smallmatrix} x^i & x^j \\ y^i & y^j \end{smallmatrix} \right|^2} \quad (30) \end{aligned}$$

number and the distance  $s$ , such that  $0 \leq s/k \pi$ , is given by the preceding equations.

Euclidean geometry can be considered to be the limiting case of either elliptic or hyperbolic geometry as  $k$  becomes infinite. In this case, one can always regard  $x^0$  as unity. Then from Eqs. (30), the distance formula  $s = s(P, Q)$  for euclidean geometry is given by Eq. (31). In this case, coordinates  $x = (x^1, x^2, \dots, x^n)$

$$\begin{aligned} s &= s(P, Q) \\ &= \sqrt{(x^1 - y^1)^2 + (x^2 - y^2)^2 + \dots + (x^n - y^n)^2} \quad (31) \end{aligned}$$

of a point  $P$  are said to be rectangular or cartesian.

The final case is that of hyperbolic geometry. Here  $k = il$ , where  $i^2 = -1$  and  $l$  is a positive real number. From Eqs. (30), the distance formula  $s = s(P, Q)$  is given by Eqs. (32).

$$\begin{aligned} \cosh \frac{s}{l} &= x^0 y^0 - \frac{1}{l^2} (x^1 y^1 + x^2 y^2 + \dots + x^n y^n) \\ l \sinh \frac{s}{l} &= \sqrt{\sum_{i=1}^n \left| \begin{smallmatrix} x^0 & x^i \\ y^0 & y^i \end{smallmatrix} \right|^2 - \frac{1}{2l^2} \sum_{i,j=1}^n \left| \begin{smallmatrix} x^i & x^j \\ y^i & y^j \end{smallmatrix} \right|^2} \quad (32) \end{aligned}$$

In each of the three geometries it is assumed that the relationship  $s = s(P, Q)$  is zero if, and only if, the two points  $P$  and  $Q$  are identical.

Each of the three geometries is an abstract metric space. That is,

- i.  $s(P, Q) \geq 0$ , and  $s(P, Q) = 0$  if, and only if,  $P = Q$
- ii.  $s(P, Q) = s(Q, P)$
- iii.  $s(P, Q) + s(Q, R) \geq s(P, R)$

The condition (i) is that of positive definiteness; the condition (ii) is that of symmetry; the condition (iii) is that of the well-known triangular inequality.

**Angle.** By dualizing the concept of distance  $s = s(P, Q)$ , the concept of angle  $\theta = \theta(\pi, \sigma)$  between two hyperplanes  $\pi$  and  $\sigma$  is obtained. Let  $\pi$  and  $\sigma$  be two distinct hyperplanes which are given by the homogeneous hyperplane coordinates  $u = (u_0, u_1, \dots, u_n)$  and  $v = (v_0, v_1, \dots, v_n)$ . In euclidean and hyperbolic geometries, it is understood that  $\pi$  and  $\sigma$  are proper hyperplanes which determine a pencil  $\lambda$ . In this pencil  $\lambda$  there are two distinct hyperplanes  $\pi_\infty$  and  $\sigma_\infty$  which are tangent to the fundamental quadric  $\Sigma$ . The angle  $\theta = \theta(\pi, \sigma)$ , where  $0 \leq \theta \leq \pi$ , between these two hyperplanes  $\pi$  and  $\sigma$  is defined by formula (33). This angle  $\theta = \theta(\pi, \sigma)$  is given by

Eqs. (34), in which  $\epsilon = +1$  or  $\epsilon = -1$  according

$$\begin{aligned} \theta &= \theta(\pi, \sigma) = \frac{l}{2i} \log \Re(\pi\sigma, \pi_\infty\sigma_\infty) \quad (33) \\ \cos \theta &= \frac{\epsilon \left( \frac{u_0 v_0}{k^2} + u_1 v_1 + u_2 v_2 + \cdots + u_n v_n \right)}{\sqrt{\left( \frac{u_0^2}{k^2} + u_1^2 + \cdots + u_n^2 \right) \cdot \left( \frac{v_0^2}{k^2} + v_1^2 + \cdots + v_n^2 \right)}} \\ \sin \theta &= \frac{\epsilon \left[ \frac{1}{k^2} \sum_{i=1}^n \begin{vmatrix} u_0 & u_i \\ v_0 & v_i \end{vmatrix} + \frac{1}{2} \sum_{i,j=1}^n \begin{vmatrix} u_i & u_j \\ v_i & v_j \end{vmatrix} \right]}{\sqrt{\left( \frac{u_0^2}{k^2} + u_1^2 + \cdots + u_n^2 \right) \cdot \left( \frac{v_0^2}{k^2} + v_1^2 + \cdots + v_n^2 \right)}} \quad (34) \end{aligned}$$

to whether the geometry is elliptic (including the euclidean case) or hyperbolic.

It is evident that when  $\epsilon = +1$  and  $k$  becomes infinite, the preceding formulas become those for the angle  $\theta$  between the two hyperplanes  $\pi$  and  $\sigma$  in a euclidean space of  $n$  dimensions.

The two hyperplanes  $\pi$  and  $\sigma$  are said to be orthogonal if, and only if,  $\Re(\pi\sigma, \pi_\infty\sigma_\infty) = -1$ . Let  $\sigma_1$  and  $\sigma_2$  be two distinct hyperplanes which intersect in a line  $L$ . Then  $L$  is said to be orthogonal to a hyperplane  $\pi$  if  $\sigma_1$  and  $\sigma_2$  are each orthogonal to  $\pi$ . If  $L$  and  $M$  are two distinct lines which pass through a point  $P$  and if  $M$  is contained in a hyperplane  $\pi$  orthogonal to  $L$ , then  $L$  and  $M$  are said to be orthogonal.

If  $\pi$  is a fixed hyperplane and if  $P$  is a point not in  $\pi$ , then there is one and only one line  $L$  passing through the point  $P$  orthogonal to the hyperplane  $\pi$ .

**Differential of arc length.** If  $P$  and  $Q$  are two nearby points on the quadric  $\Sigma^*$  defined by the two sets of coordinates  $(\bar{x}^0, \bar{x}^1, \dots, \bar{x}^n)$  and  $(\bar{x}^0 + d\bar{x}^0, \bar{x}^1 + d\bar{x}^1, \dots, \bar{x}^n + d\bar{x}^n)$ , then by Eqs. (30), the square of the differential  $ds$  of arc length  $s$  between the points  $P$  and  $Q$  is Eq. (35).

$$\begin{aligned} ds^2 &= \sum_{i=1}^n (\bar{x}^0 d\bar{x}^i - \bar{x}^i d\bar{x}^0)^2 \\ &+ \frac{1}{2k^2} \sum_{i,j=1}^n (\bar{x}^i d\bar{x}^j - \bar{x}^j d\bar{x}^i)^2 \quad (35) \end{aligned}$$

For the point  $P$  on this quadric  $\Sigma^*$  introduce a new set of coordinates  $(x^1, x^2, \dots, x_n)$  defined by Eq. (36), where it is understood that  $\bar{x}^0 \neq -1$ .

$$x^i = \frac{2\bar{x}^i}{1 + \bar{x}^0} \quad (i = 1, 2, \dots, n) \quad (36)$$

In this new set of coordinates, the quadric  $\Sigma^*$  is given by Eq. (37).

$$(x^1)^2 + (x^2)^2 + \cdots + (x^n)^2 + 4k^2 = \frac{8k^2}{1 + \bar{x}^0} \quad (37)$$

In the euclidean and noneuclidean geometries, the square of the differential  $ds$  of arc length  $s$  is given

by Eq. (38).

$$ds^2 = \frac{(dx^1)^2 + (dx^2)^2 + \cdots + (dx^n)^2}{\left[ 1 + \frac{(x^1)^2 + (x^2)^2 + \cdots + (x^n)^2}{4k^2} \right]^2} \quad (38)$$

As  $k$  approaches infinity, the differential  $d\sigma$  of arc length  $\sigma$  of euclidean geometry is obtained. If  $ds$  represents the differential of arc length  $s$  in elliptic or hyperbolic geometry, then Eq. (39) holds, in which the scale  $\rho$  is given by Eq. (40).

$$ds = \rho d\sigma \quad (39)$$

$$\rho = \frac{1}{1 + \frac{(x^1)^2 + (x^2)^2 + \cdots + (x^n)^2}{4k^2}} \quad (40)$$

Thus, each of the noneuclidean geometries may be visualized as a conformal image of euclidean geometry.

Each of these three geometries is a special case of riemannian geometry. Let the  $g_{ij}(x)$  be  $n(n+1)/2$  real functions which are continuous and possess continuous partial derivatives of as high an order as is necessary in a certain region of real  $n$ -dimensional space for which the coordinates of a point  $P$  are  $x = (x^1, x^2, \dots, x^n)$ . The  $g_{ij}$ 's are symmetric; that is,  $g_{ij} = g_{ji}$ , for all  $i, j = 1, 2, \dots, n$ . The quadratic form (41)

$$\sum_{i,j=1}^n g_{ij} \lambda^i \lambda^j \quad (41)$$

is assumed to be positive definite, that is, expression (41) is nonnegative; it is zero if and only if  $\lambda^i = 0$ , for all  $i = 1, 2, \dots, n$ . Then the riemannian space  $V_n$  is one for which the square of the differential  $ds$  of arc length  $s$  between two nearby points  $P$  and  $Q$  is given by Eq. (42).

$$ds^2 = \sum_{i,j=1}^n g_{ij} dx^i dx^j \quad (42)$$

For this riemannian space  $V_n$ , the Christoffel symbols of the first kind are shown as Eq. (43), for  $i, j$ ,

$$\Gamma_{ijk} = \frac{1}{2} \left( \frac{\partial g_{ik}}{\partial x^j} + \frac{\partial g_{jk}}{\partial x^i} - \frac{\partial g_{ij}}{\partial x^k} \right) \quad (43)$$

$k = 1, 2, \dots, n$ . The Christoffel symbols of the second kind are shown as Eq. (44), for  $i, j, k = 1, 2,$

$$\Gamma_{jk}^i = \sum_{l=1}^n g^{il} \Gamma_{jkl} \quad (44)$$

$\dots, n$ . Of course,  $(g^{ij})$  is the inverse matrix of  $(g_{ij})$ . The Christoffel symbols of the second kind are also called the affine connections of  $V_n$ .

The vector geodesic curvature  $\kappa^i$  of a curve  $C$  in  $V_n$  is Eq. (45). A curve  $C$  of  $V_n$  is said to be a geodesic

$$\kappa^i = \frac{d^2 x^i}{ds^2} + \sum_{j,k=1}^n \Gamma_{jk}^i \frac{dx^j}{ds} \frac{dx^k}{ds} \quad (45)$$

if, and only if,  $\kappa^i = 0$  for  $i = 1, 2, \dots, n$  at every point  $P$  of  $C$ .

Each of the three spaces already discussed is a riemannian space. The geodesics of any one such space are the lines of the space.

A riemannian space of constant curvature is applicable to (that is, can be mapped isometrically into) elliptic, euclidean, or hyperbolic space. In this case, this constant curvature is equal to the gaussian curvature  $G$ . From Eq. (35) or (38), this constant gaussian curvature  $G$  is given by Eq. (46). For elliptic space,

$$G = \frac{1}{k^2} \quad (46)$$

$G$  is of positive constant curvature. For euclidean space,  $G$  is identically zero. Finally for hyperbolic space,  $G$  is of negative constant curvature.

In each of these three geometries, consider a geodesic triangle. This is formed by three points  $P, Q, R$ , not all on one geodesic, and the three geodesics passing through every two of the three points  $P, Q, R$ . If  $A, B, C$  are the angles of this geodesic triangle, and if  $T$  denotes its area, then Eq. (47) holds.

$$A + B + C - \pi = \frac{T}{k^2} \quad (47)$$

Thus, according to whether the geometry is elliptic, euclidean, or hyperbolic, the sum of the angles of a geodesic triangle is greater than  $\pi$ , equal to  $\pi$ , or less than  $\pi$ .

Elliptic geometry of two dimensions may be represented upon a sphere in euclidean space of three dimensions. On the other hand, hyperbolic geometry of two dimensions can be depicted on a pseudosphere in euclidean space of three dimensions. The pseudosphere is obtained by revolving the tractrix about its asymptote.

John De Cicco

Bibliography. H. S. M. Coxeter, *Introduction to Geometry*, 2d ed., 1969, paper 1989; J. Gray, *Ideas of Space: Euclidean, Non-Euclidean, and Relativistic*, 2d ed., 1989; M. J. Greenberg, *Euclidean and Non-Euclidean Geometries: Development and History*, 3d ed., 1995; E. Kasner and J. R. Newman, *Mathematics and the Imagination*, 1940, reprint 1989; G. E. Martin, *Foundations of Geometry and the Non-Euclidean Plane*, 1986; B. A. Rosenfeld, *The History of Non-Euclidean Geometry*, 1988; P. J. Ryan, *Euclidean and Non-Euclidean Geometry: An Analytic Approach*, 1986.

## Nonlinear acoustics

The study of amplitude-dependent acoustical phenomena. The amplitude dependence is due to the nonlinear response of the medium in which the sound propagates, and not to the nonlinear behavior of the sound source. According to the linear theory of acoustics, increasing the level of a source by 10 dB results in precisely the same sound field as before, just 10 dB more intense. Linear theory also predicts that only frequency components radiated directly by the source can be present in the sound field. These principles do not hold in nonlinear acoustics. See

AMPLITUDE (WAVE MOTION); LINEARITY; NONLINEAR PHYSICS.

The extent to which nonlinear acoustical effects are strong or even significant depends on the competing influences of energy loss, frequency dispersion, geometric spreading, and diffraction. When conditions are such that nonlinear effects are strong, acoustic signals may experience substantial waveform distortion and changes in frequency content as they propagate, and shock waves may be present. Nonlinear acoustical effects occur in gases, liquids, and solids, and they are observed over a broad range of frequencies. Shock waves present in sonic booms and thunder claps are in the audio frequency range. Principles of nonlinear acoustics form the basis for procedures at megahertz frequencies used in medical ultrasound and nondestructive evaluation of materials. Nonlinearity can also induce changes in non-fluctuating properties of the medium. These include acoustic streaming, which is the steady fluid flow produced by sound, and radiation pressure, which results in a steady force exerted by sound on its surroundings. See ACOUSTIC RADIATION PRESSURE; BIOMEDICAL ULTRASONICS; NONDESTRUCTIVE EVALUATION; SHOCK WAVE; SONIC BOOM; THUNDER.

The study of nonlinear acoustics dates back to the mideighteenth century, when L. Euler derived the first exact nonlinear wave equation for a plane wave in a gas, revealing the nonlinear distortion of the acoustic waveform prior to shock formation, was derived by S. D. Poisson at the turn of the nineteenth century. The understanding of shock waves was provided by G. G. Stokes, W. J. M. Rankine, and P. H. Hugoniot in the latter half of the nineteenth century. Solutions for harmonic generation in sound waves, both before and after shock formation, were derived in the first half of the twentieth century. Foundations of the modern theory of nonlinear acoustics were developed in the second half of the twentieth century, motivated by engineering advances in the transmission and reception of high-intensity sound.

**Waveform distortion.** The principal feature that distinguishes nonlinear acoustics from nonlinear optics is that most acoustical media exhibit only weak dispersion, whereas media in which nonlinear optical effects arise exhibit strong dispersion. Dispersion is the dependence of propagation speed on frequency. In optical media, strong nonlinear wave interactions require that phase-matching conditions be satisfied, which can be accomplished only for several frequency components at one time. In contrast, all frequency components in a sound wave propagate at the same speed and are automatically phase-matched, which permits strong nonlinear interactions to occur among all components in the frequency spectrum. See NONLINEAR OPTICS.

Nonlinear coupling due to phase matching of the components in the frequency spectrum of a traveling sound wave manifests itself as waveform distortion that accumulates with distance. The main effect that opposes nonlinear distortion of the waveform is

energy absorption due to viscosity and heat conduction, and especially in air, relaxation due to molecular vibration. For sound radiated at a single frequency  $f$ , the parameter that characterizes the balance of nonlinearity and energy loss is the Gol'dberg number, given by Eq. (1). Here  $\ell_a = 1/\alpha$  is the absorption

$$\Gamma = \ell_a/\bar{x} \quad (1)$$

length associated with the exponential attenuation coefficient  $\alpha$  predicted for the wave amplitude by linear theory, and  $\bar{x}$  is the shock formation distance, given by Eq. (2). The coefficient of nonlinearity  $\beta$

$$\bar{x} = \frac{1}{\beta \epsilon k} \quad (2)$$

depends on the state equation for the propagation medium (discussed below),  $\epsilon = p_0/(\rho_0 c_0^2)$  is the peak acoustic Mach number associated with the source waveform,  $p_0$  is the corresponding peak sound pressure,  $\rho_0$  is the ambient density of the medium,  $c_0$  is the sound speed predicted by linear theory, and  $k = 2\pi f/c_0$ . See SOUND; SOUND ABSORPTION; SOUND PRESSURE.

For a plane wave, the criterion  $\Gamma < 1$  indicates that

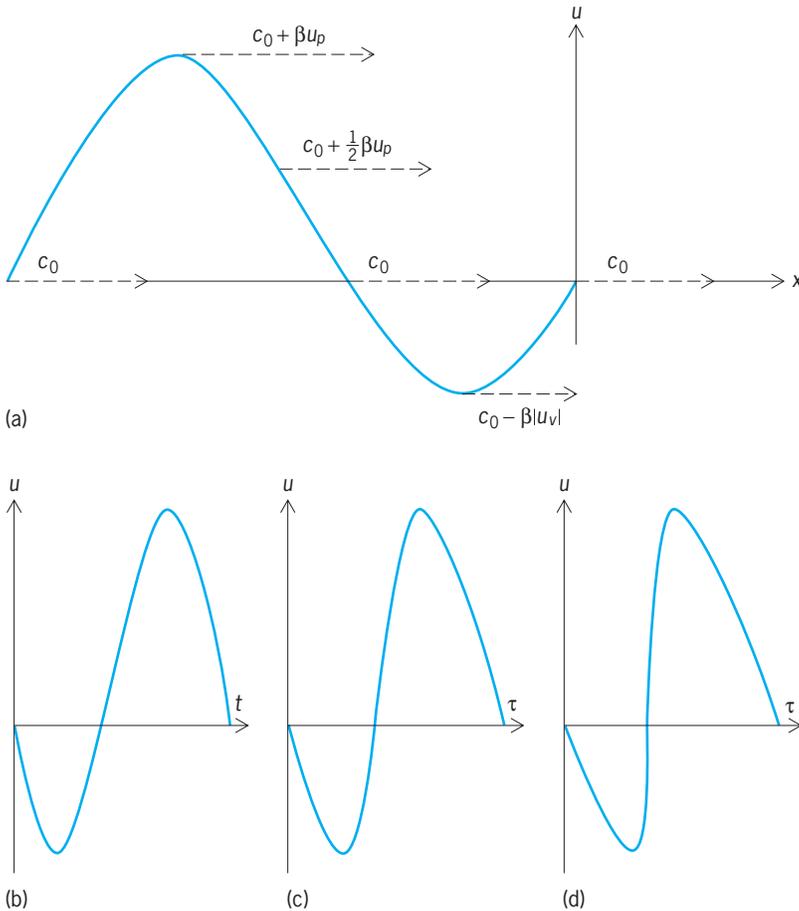


Fig. 1. Nonlinear distortion of an intense acoustic waveform. (a) Spatial plot (amplitude versus distance) of the initial acoustic waveform near  $x = 0$ . (b) Time waveform (plot of amplitude versus time) at distance  $x = 0$ . (c) Time waveform at  $x > 0$ .  $\tau = t - x/c_0$ . (d) Time waveform at  $x = \bar{x}$ , where a shock is formed. (After D. T. Blackstock, *Nonlinear acoustics (theoretical)*, in D. E. Gray, ed., *American Institute of Physics Handbook*, 3d ed. pp. 3-183-3-205, McGraw-Hill, 1972)

energy loss due to absorption overwhelms the tendency of the waveform to distort, and linear theory provides an accurate description of the wave propagation. For  $\Gamma > 1$ , the initially sinusoidal waveform generates harmonics of the source frequency as it propagates. These harmonics cannot be described by linear theory. For  $\Gamma \gg 1$ , the waveform distortion is strong and results in the development of a shock front. Under this condition, a shock wave forms at distance  $\bar{x}$ . See HARMONIC (PERIODIC PHENOMENA).

Waveform distortion accumulates with distance because points on a waveform having different amplitudes propagate at different speeds. The relation for the propagation speed of individual points on a traveling sound wave is given by Eq. (3), where  $u$

$$\frac{dx}{dt} = c_0 + \beta u \quad (3)$$

is the local particle velocity within the wave. This distortion mechanism is illustrated in Fig. 1a. The peak in the sound wave has particle velocity  $u_p$  and travels faster than the zero crossings; the valley has a lower (negative) particle velocity  $u_v$  and travels slower. Time waveforms in Fig. 1b-d illustrate the waveform distortion that results from the amplitude-dependent propagation speed. Peaks advance on the zero crossings and valleys recede, until at  $x = \bar{x}$  a shock is formed. Shock formation is said to occur when a vertical tangent first appears in the waveform.

The entire propagation history of a sinusoidal plane wave radiated by a source under the condition  $\Gamma \gg 1$  is depicted in Fig. 2. The source waveform is illustrated in Fig. 2a, at the shock formation distance ( $x = \bar{x}$ ) in Fig. 2c, and at the location of maximum shock amplitude [ $x = (\pi/2)\bar{x}$ ] in Fig. 2d. Figure 2e and f reveal the propagation of a stable sawtooth waveform ( $3\bar{x} < x \ll \ell_a$ ). Eventually, for  $x \sim \ell_a$  (Fig. 2g), energy losses reduce the wave amplitude to such an extent that nonlinear effects no longer sustain a well-defined shock front. Finally, as depicted in Fig. 2h, for  $x \gg \ell_a$ , the old-age region is reached. Here, the waveform resembles its original sinusoidal shape, although substantially reduced in amplitude.

An example helps to place the above discussion in context. If a source radiates a plane wave in air at  $f = 5$  kHz (at which frequency the absorption length is  $\ell_a \sim 100$  m or 330 ft) with peak acoustic Mach number  $\epsilon = 0.01$  (that is, a sound pressure level of 154 dB re 20  $\mu$ Pa), the Gol'dberg number is  $\Gamma \simeq 100 \gg 1$ , and the shock formation distance is  $\bar{x} \simeq 1$  m (3.3 ft). A fully developed sawtooth wave exists beyond  $3\bar{x} \simeq 3$  m (10 ft).

Plane waves, however, are idealizations often used to approximate propagation in ducts and the near-fields of sound beams. When propagation takes place in unbounded media, spreading of the wave reduces its amplitude as a function of distance from the source. For a wave radiated by a spherical source of radius  $r_0$ , the shock formation distance is given

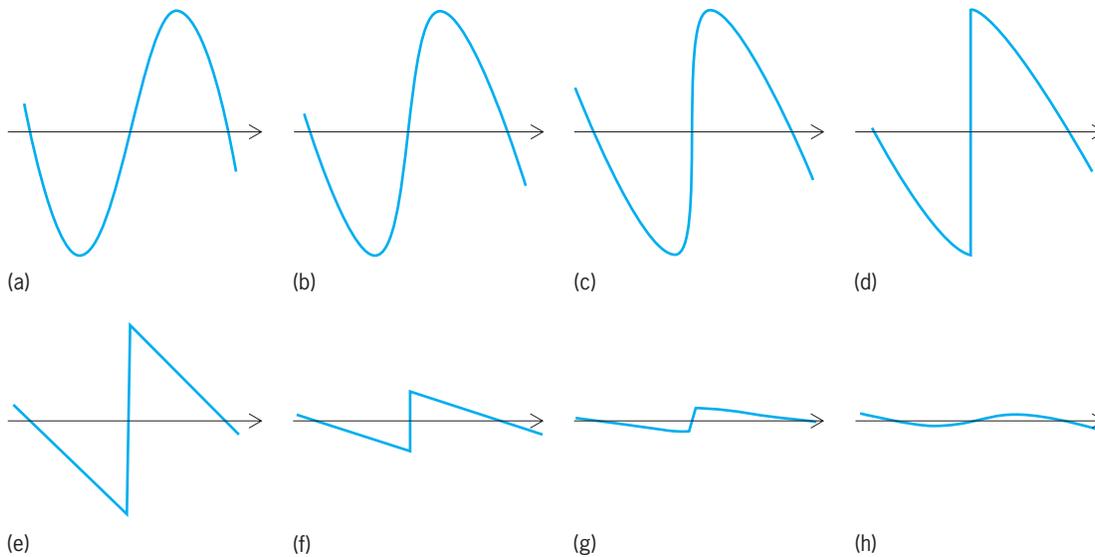


Fig. 2. Propagation history of an intense acoustic waveform that is sinusoidal at the source. (a) Source waveform,  $x = 0$ . (b) Distortion becoming noticeable. (c) Shock formation,  $x = \bar{x}$ . (d) Maximum shock amplitude,  $x = (\pi/2) \bar{x}$ . (e) Full sawtooth shape,  $x = 3 \bar{x}$ . (f) Decaying sawtooth. (g) Shock beginning to disperse. (h) Old age. (After J. A. Shooter et al., *Acoustic saturation of spherical waves in water*, *J. Acous. Soc. Amer.*, 55:54–62, 1974)

by Eq. (4) instead of Eq. (2). The shock formation

$$\bar{r} = r_0 e^{1/\beta \epsilon k r_0} \quad (4)$$

distance for a spherical wave can be substantially greater than that for a plane wave. For the same frequency and source level as in the previous example, the shock formation distance for a spherical wave radiated by a source of radius  $r_0 = 0.5$  m (1.6 ft) is  $\bar{r} \simeq 3$  m (10 ft), but for  $r_0 = 0.1$  m (0.3 ft) it increases to  $\bar{r} \simeq 900$  m (3000 ft). For the smaller source the relation  $\bar{r} \gg l_a$  is obtained, shock formation cannot occur, and even second-harmonic generation is very weak. Spherical spreading reduces the amplitude of the wave so rapidly that shock formation cannot occur before the wave falls victim to strong energy losses.

**Acoustic saturation.** Energy losses at shock fronts impose an upper bound on how much sound power can be transmitted beyond a certain range. As the source level is increased, the shock formation distance is reduced, which allows more energy to be lost at the shocks before the wave arrives at a given location. For plane waves radiated by a monofrequency source for which  $\Gamma \gg 1$ , the peak sound pressure that can be produced at distance  $x$  is given by Eq. (5), which corresponds to the peak pressure

$$p_{\text{sat}} = \frac{\pi \rho_0 c_0^2}{\beta k x} \quad (5)$$

in the sawtooth waveforms depicted in Fig. 2e and f. Equation (5) is independent of the source pressure amplitude  $p_0$ . All additional energy pumped into the wave is lost at the shock fronts, and acoustic saturation is said to have occurred. The saturation curve for such a wave is shown in Fig. 3, where the extra attenuation indicates the energy loss not predicted by linear theory, designated by the diagonal line.

Acoustic saturation is not restricted to plane

waves—saturation is frequently observed in sound beams—but not all waveforms saturate. In principle, any periodic acoustic wave will saturate provided sufficient sound power is available from the source. In contrast, short pulses do not experience saturation. The N wave (the acoustic waveform resembles the letter N) is typical of sonic booms and also waveforms generated by spark sources in air, and it illustrates the point. For sufficiently high source levels, the peak amplitude of the N wave becomes proportional to the square root of the source amplitude, but it never becomes independent of source amplitude. There is nevertheless excess attenuation, because under these circumstances a 10-dB increase in source level results in only a 5-dB increase in the level of the received signal.

**Frequency generation.** The waveform distortion in Fig. 2 is accompanied by the generation of harmonics of the source frequency  $f$ . For  $\Gamma \gg 1$ , the amplitudes

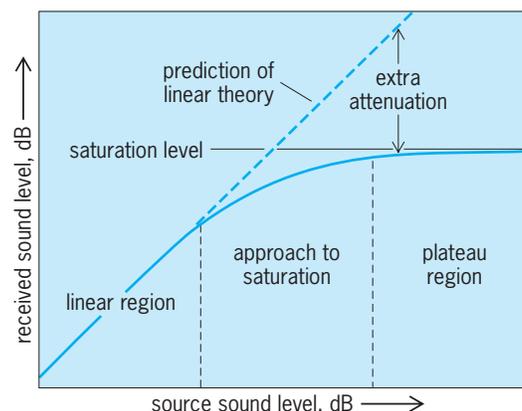
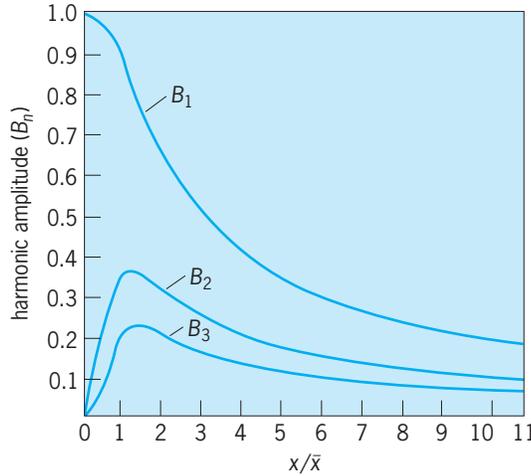


Fig. 3. Response curve illustrating the development of acoustic saturation. (After D. A. Webster and D. T. Blackstock, *Finite-amplitude saturation of plane sound waves in air*, *J. Acous. Soc. Amer.*, 62:518–523, 1977)



**Fig. 4.** Amplitudes of frequency components in an intense sound wave that is sinusoidal at the source. (After D. T. Blackstock, *Connection between the Fay and Fubini solutions for plane sound waves of finite amplitude*, *J. Acous. Soc. Amer.*, 39:1019–1026, 1966)

of the harmonic components at frequencies  $nf$  ( $n = 1, 2, \dots$ ), normalized by the source pressure  $p_0$ , are given at different distances by Eqs. (6) and (7), where

$$B_n(x) = \frac{2J_n(nx/\bar{x})}{nx/\bar{x}} \quad (6)$$

$$x/\bar{x} \leq 1$$

$$B_n(x) = \frac{2}{n(1 + x/\bar{x})} \quad (7)$$

$$3 \lesssim x/\bar{x} \lesssim \Gamma$$

$J_n$  is the Bessel function of the first kind, of order  $n$ . The first several amplitudes are plotted in **Fig. 4** [including the intermediate range  $1 < x/\bar{x} < 3$ , not covered by Eqs. (6) and (7)]. As the wave propagates in the preshock region,  $x/\bar{x} < 1$ , the amplitude of the source frequency component  $B_1$  decreases as energy is expended to generate higher harmonics. In the region  $x/\bar{x} > 3$ , corresponding to **Fig. 2e** and **f**, the amplitudes of all spectral components decrease with distance from the source and are related in the ratio  $1/n$ , as is required by the Fourier series for an ideal sawtooth waveform. See BESSEL FUNCTIONS; FOURIER SERIES AND TRANSFORMS.

In a sound beam radiated from a monofrequency source, the beamwidths of the harmonic components become smaller approximately in proportion to  $1/\sqrt{n}$ , and the relative levels of the side lobes are also reduced as  $n$  increases. These effects improve the directional properties of the higher harmonics and are used to enhance resolution in acoustic imaging. In a sound beam so intense that acoustic saturation is approached along its axis, nonlinear losses tend to reduce these trends. Extra attenuation is most pronounced in the center of the beam, where the intensity is highest and shocks develop first. Erosion of the main lobe due to the increased energy loss near the axis causes both the beamwidth and side-lobe levels to increase. See DIRECTIVITY.

**Coefficient of nonlinearity.** For a given source amplitude and frequency, Eq. (2) for the shock formation distance indicates that the coefficient of nonlinearity  $\beta$  is a critical parameter in determining the rate of nonlinear wave distortion. For compressional waves in gases, liquids, and isotropic solids, standard expressions for this coefficient are given by Eqs. (8)–(10). In Eq. (8),  $\gamma$  is the ratio of specific heats, and

$$\beta = \frac{\gamma + 1}{2} \quad \text{gas} \quad (8)$$

$$\beta = \frac{B}{1 + 2A} \quad \text{liquid} \quad (9)$$

$$\beta = -\left(\frac{3}{2} + \frac{A + 3B + C}{\rho_0 c_0^2}\right) \quad \text{isotropic solid} \quad (10)$$

therefore with  $\gamma = 1.4$  for diatomic gases  $\beta = 1.2$  for air.

For liquids, an arbitrary equation of state relating the sound pressure to the excess density  $\rho'$  at constant entropy may be expanded in the Taylor series given by Eq. (11), where  $A = \rho_0 c_0^2$  and the ratio  $B/A$  is given by Eq. (12). The subscript  $s$  indicates that

$$p = A \left(\frac{\rho'}{\rho_0}\right) + \frac{B}{2!} \left(\frac{\rho'}{\rho_0}\right)^2 + \frac{C}{3!} \left(\frac{\rho'}{\rho_0}\right)^3 + \dots \quad (11)$$

$$\frac{B}{A} = \frac{\rho_0}{c_0^2} \left(\frac{\partial^2 p}{\partial \rho'^2}\right)_{s,0} \quad (12)$$

the derivatives of pressure with respect to density are evaluated at constant entropy, with the 0 referring to ambient conditions. Although Eq. (12) is the fundamental definition, other formulations are often used for measuring  $B/A$ , for example, in terms of sound speed variation as a function of pressure, or for conditions of constant pressure and temperature rather than constant entropy. A short compilation of measurements is presented in the **table**. For pure water at room temperature,  $B/A = 5.0$ , in which case  $\beta = 3.5$ . Because of the important role of nonlinear acoustics in medical applications, considerable effort has been devoted to collecting measurements of  $B/A$  for a wide variety of biologic materials, values of which nominally range from 5 to 10. It is rarely necessary to consider the higher-order coefficients,  $C$ ,  $D$ , and so forth, in Eq. (11).

Equation (10) shows that although only one quantity must be measured to determine  $\beta$  for a liquid, three are required for an isotropic solid. The parameters  $A$ ,  $B$ , and  $C$  are referred to as third-order elastic constants because they appear at third order in an expansion of the strain energy function. Moreover, it is difficult to obtain accurate measurements of these constants. Values of  $\beta$  for common isotropic solids tend to fall within the range of values for common liquids. Whereas  $\beta$  is invariably positive for fluids (excluding anomalous behavior near critical points), it is negative for several solids, such as fused quartz and Pyrex glass. In this case, the acoustic waveforms distort in directions that are opposite those shown in **Figs. 1** and **2**. For example, the relation  $\beta > 0$  indicates that only compression shocks can form in gases

Values of $B/A$ for liquids at atmospheric pressure*		
Liquid	$T, ^\circ\text{C} (^{\circ}\text{F})$	$B/A$
Distilled water	0 (32)	4.2
	20 (68)	5.0
	40 (104)	5.4
	60 (140)	5.7
	80 (176)	6.1
	100 (212)	6.1
Seawater (3.5% salinity)	20 (68)	5.25
Alcohols		
Methanol	20 (68)	9.6
Ethanol	20 (68)	10.5
<i>n</i> -Propanol	20 (68)	10.7
<i>n</i> -Butanol	20 (68)	10.7
Organic liquids		
Acetone	20 (68)	9.2
Benzene	20 (68)	9.0
Chlorobenzene	30 (86)	9.3
Cyclohexane	30 (86)	10.1
Diethylamine	30 (86)	10.3
Ethylene glycol	30 (86)	9.7
Ethyl formate	30 (86)	9.8
Glycerol (4% H <sub>2</sub> O)	30 (86)	9.0
Heptane	30 (86)	10.0
Hexane	30 (86)	9.9
Methyl acetate	30 (86)	9.7
Methyl iodide	30 (86)	8.2
Nitrobenzene	30 (86)	9.9
Liquid metals		
Bismuth	318 (604)	7.1
Indium	160 (320)	4.6
Mercury	30 (86)	7.8
Potassium	100 (212)	2.9
Sodium	110 (230)	2.7
Tin	240 (464)	4.4

\* After R. T. Beyer, *Nonlinear Acoustics*, Navy Sea Systems Command, 1974.

and liquids, yet only expansion shocks can form for  $\beta < 0$ , as in these particular solids.

Nonlinearity in anisotropic solids depends on larger numbers of third-order elastic constants. The strength of nonlinear effects in crystals depends on the direction of propagation, but in general it is of the order encountered in isotropic solids. In contrast, microinhomogeneous features of rock yield effective values of  $\beta$  that may be larger than those of isotropic solids by several orders of magnitude. However, acoustic propagation in rock is often accompanied by strong attenuation that may offset nonlinear frequency generation. There is significant interest in nonlinear properties of rock in the oil exploration industry because seismologists can potentially deduce, by acoustic measurements, the magnitude of the local stress environment. In addition, there is the expectation that measured values of the third-order elastic constants may correlate inversely with the strength of the rock. Both of these considerations are relevant for the safe construction of productive oil wells. See GEOPHYSICAL EXPLORATION; OIL AND GAS WELL DRILLING.

**Parametric array.** The parametric array is a classical application of nonlinear acoustics developed originally for application to underwater sonar. A source radiates two signals simultaneously at the two neighboring frequencies  $f_1 \simeq f_2$ . These two sound beams interact as they propagate and generate a spectrum of intermodulation frequencies, among them the dif-

ference frequency  $f_- = |f_1 - f_2|$ , which is much less than  $f_1$  or  $f_2$  and is the lowest in the spectrum. It is the volume of fluid in which the nonlinear interaction generates the difference frequency signal that is referred to as the parametric array. When absorption terminates nonlinear interaction within the nearfield of the primary beams, the beam profile of the difference frequency signal is given by Eq. (13), where

$$D(\theta) = \frac{1}{\sqrt{1 + (k_- \ell_a)^2 \sin^4(\theta/2)}} \quad (13)$$

$\theta$  is angle from the central axis,  $\ell_a$  is the absorption length at the primary frequencies, and  $k_- = 2\pi f_- / c_0$ .

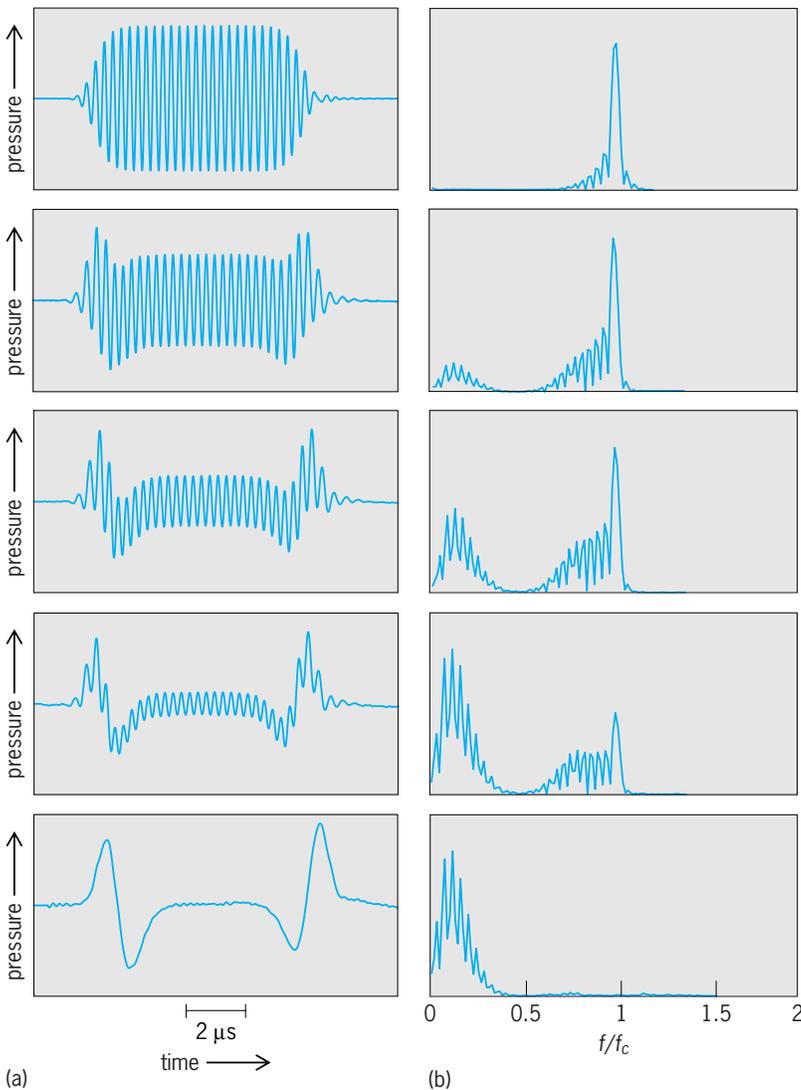
The most useful feature of the parametric array is its beamwidth. The beamwidth is inversely proportional to  $\sqrt{k_- \ell_a}$ , which depends on the absorption coefficient of the primary waves, but not on the size of the source. In contrast, direct radiation by a circular source of radius  $a$  at frequency  $f_-$  produces a beamwidth that is inversely proportional to  $k_- a$ , which depends critically on source size. The principal benefit of the parametric array is that very narrow beams of low-frequency sound can be generated by the nonlinear interaction of high-frequency primary beams radiated from a fairly small source. Because  $\ell_a \gg a$ , the beamwidth can be far narrower, by an order of magnitude, than can be obtained by direct radiation from the source of the primary waves, but at frequency  $f_-$ .

The absorption length  $\ell_a$  is the effective aperture length of the array because it defines the volume of fluid in front of the source where the difference frequency is generated. Since the difference frequency component propagates at the same speed as the primary waves that generate it, the nonlinear interaction volume behaves as an end-fire array. In underwater applications, the parametric array is used for long-range propagation of highly directional, low-frequency sound. It also forms the basis for audio loudspeakers currently under development, which depend on a related phenomenon called self-demodulation.

**Self-demodulation.** The parametric array may be used to generate low-frequency sound possessing a beam pattern described by Eq. (13), yet having a bandwidth not restricted to a single difference frequency component. A carrier wave of frequency  $f_c$  is amplitude-modulated by an envelope  $E(t)$  and radiated by a directional sound source. In the farfield along the axis of the beam, the time dependence is given by Eq. (14). There is no remnant of the car-

$$F(t) = \frac{d^2 E^2(t)}{dt^2} \quad (14)$$

rier frequency  $f_c$ , because it is much greater than the frequencies associated with the modulation envelope  $E(t)$ . The carrier is therefore absorbed much more rapidly by the fluid, in which attenuation increases close to quadratically with frequency. All that remains is a squared and differentiated version of the envelope. Evolution of the waveform toward this farfield result is called self-demodulation.



**Fig. 5.** Measurements of self-demodulation of an acoustic pulse in glycerin at increasing distances from the source, from 5 cm (2.0 in.) in the first row to 26 cm (10.2 in.) in the bottom row. (a) Acoustic waveforms. (b) Corresponding frequency spectra.  $f$  = frequency;  $f_c$  = carrier frequency = 3.5 MHz. (After M. A. Averkiou et al., *Self-demodulation of amplitude and frequency modulated pulses in a thermoviscous fluid*, *J. Acous. Soc. Amer.*, 94:2876–2883, 1993)

Self-demodulation in glycerin of an acoustic pulse with carrier frequency  $f_c = 3.5$  MHz is depicted in **Fig. 5**, with measured acoustic waveforms and the corresponding frequency spectra. Each row down is at a distance farther from the source, from 5 cm (2.0 in.) for the first row to 26 cm (10.2 in.) for the last. In the first row, there is no observable effect of nonlinearity, either in the time waveform or in the frequency spectrum. As the high-frequency carrier wave is absorbed by the fluid, a self-demodulated signal emerges in the time waveforms, and a spectral shift from high to low frequencies is observed. Equation (14) is an accurate description of the waveform in the bottom row, where no energy is apparent at frequencies  $f \simeq f_c$ .

In air, parametric arrays are under development that are driven at ultrasonic frequencies and generate low-frequency sound in the audio range. This application is sometimes referred to as the audio

spotlight, because it is capable of transmitting sound (with some distortion) in beams that are remarkably narrow and can selectively insonify very localized regions in space. Electrical predistortion of the signal input to the loudspeaker is needed to compensate for the squaring of the modulation envelope  $E(t)$  in Eq. (14).

**Acoustic streaming.** Acoustic streaming is the steady flow of fluid produced by the absorption of sound. It is a nonlinear effect because the velocity of the flow depends quadratically on the amplitude of the sound, and the flow is not predicted by linear theory. Absorption due to viscosity and heat conduction results in a transfer of momentum from the sound field to the fluid. This momentum transfer manifests itself as steady fluid flow.

Acoustic streaming produced in sound beams is enhanced considerably when shocks develop. Shock formation generates a frequency spectrum rich in higher harmonics. Because thermoviscous absorption increases quadratically with frequency, attenuation of the wave, and therefore the streaming velocity, increases markedly following shock formation. Streaming is also generated in acoustic boundary layers formed by standing waves in contact with surfaces.

Measurements of acoustic streaming have been used to determine the bulk viscosity coefficients of fluids. Acoustic streaming affects the performance of thermoacoustic engines and refrigerators, which exploit the heat transfer induced by the temperature and velocity fluctuations in a sound wave in contact with a surface possessing large heat capacity. These devices are adversely affected by heat transport associated with streaming. In thermoacoustic devices operating at audio frequencies in gases, the streaming velocities are of order 10 cm/s (4 in./s). At the megahertz frequencies and acoustic intensities encountered in biomedical applications, streaming velocities attained in liquids are also of order 10 cm/s.

**Phase conjugation.** Phase conjugation refers to wavefront reversal, also called time reversal, at a single frequency. The latter terminologies more clearly describe this procedure. A waveform is captured by a phase conjugation device and reversed in such a way that it propagates back toward the source in the same way that it propagated toward the conjugator. Under ideal circumstances, it is as though a movie of the wave propagation from source to conjugator were played backward in time. Sound that is radiated from a point source and propagates through an inhomogeneous medium that introduces phase distortion in the wave field is thus retransmitted by the conjugator in such a way as to compensate for the phase distortion and to focus the wave back on the point source.

Phase conjugation is used to compensate for phase distortion in applications involving imaging and re-targeting of waves on sources. Although it is a mature technology in nonlinear optics, practical methods of phase conjugation in nonlinear acoustics were not developed until the 1990s. The most successful techniques for acoustical phase conjugation are based on

modulation of acoustical properties of a material that captures the incident sound wave. The modulation is twice the frequency of the incident sound wave, and it is induced by an electric field applied to piezoelectric material, or a magnetic field applied to magnetostrictive material. Often the modulated property of interest is the sound speed in the material. When the incident wave at frequency  $f$  propagates through a medium in which the sound speed fluctuates at frequency  $2f$ , parametric interaction generates a wave at the difference frequency  $f$  that propagates backward as though reversed in time. See MAGNETOSTRICTION; OPTICAL PHASE CONJUGATION; PIEZOELECTRICITY.

The use of phase conjugation to reduce phase distortion in the acoustic imaging of a perforated steel plate embedded in an aberrating layer is demonstrated in Fig. 6. The plate perforations are in the shapes of various letters. The plate was embedded in gel having a corrugated surface (Fig. 6a) that in-

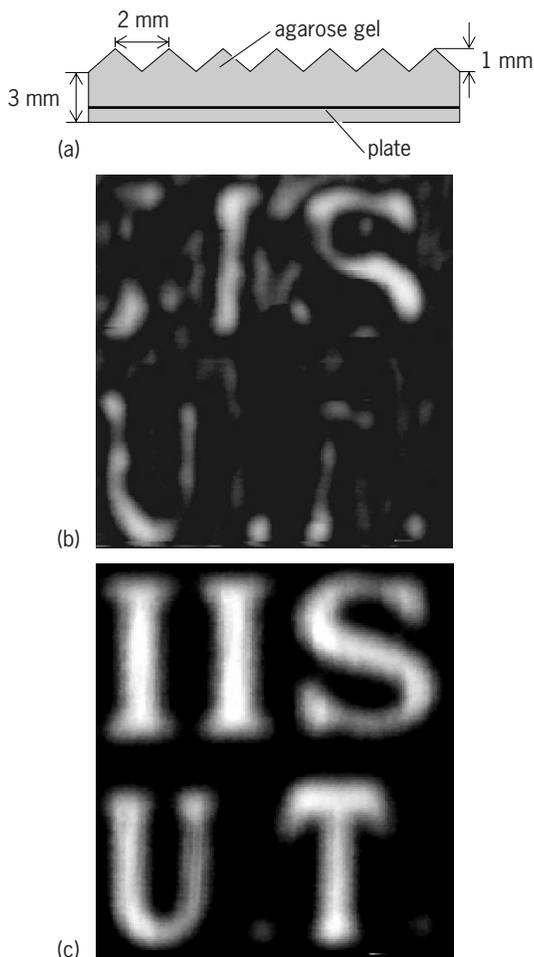


Fig. 6. Use of phase conjugation to improve acoustic imaging of a perforated steel plate embedded in an aberrating layer of gel. The perforations have the shapes of letters. (a) Geometry and dimensions of the sample, consisting of a plate embedded in gel. (b) Conventional acoustic image of the plate within the gel. (c) Acoustic image of the plate within the gel when phase conjugation is used. (From K. Yamamoto et al., *Acoustic phase conjugation by nonlinear piezoelectricity, II. Visualization and application to imaging systems*, *J. Acous. Soc. Amer.*, 106:1339–1345, 1999)

roduced phase distortion of acoustic wavefronts propagating through the layer. The letters are barely discernible in an image constructed from a conventional acoustic scan of the sample at 10 MHz (Fig. 6b). When phase conjugation was introduced by applying a dc electric bias field and a fluctuating electric field of frequency 20 MHz to a piezoelectric material, an image was obtained with the letters considerably reduced in distortion (Fig. 6c). Phase conjugation based on magnetostriction, also being developed for acoustic imaging, permits substantial amplification of the conjugate wave. The amplification is sufficiently large that harmonics can be generated in the conjugate wave field, which provides additional opportunities for image enhancement.

**Biomedical applications.** Phenomena associated with nonlinear acoustics have proved useful in both diagnostic and therapeutic applications of biomedical ultrasound. A very significant breakthrough in diagnostic imaging, especially for echocardiography and abdominal ultrasound imaging, is based on second-harmonic generation. Medical ultrasound imaging is performed at frequencies of several megahertz. Images constructed from the backscattered second-harmonic component have substantially reduced clutter and haze associated with the propagation of ultrasound through the outer layers of skin, which is the primary cause of phase aberrations. In another technique, microbubbles are injected into the bloodstream to enhance echoes backscattered from blood flow. The microbubbles are fabricated to make them resonant at diagnostic imaging frequencies, and they become strongly nonlinear oscillators when excited by ultrasound. Imaging is based on echoes at harmonics of the transmitted signal. Frequencies backscattered from the microbubbles differ from those in echoes coming from the surrounding tissue, which highlights the locations of the microbubbles and therefore of the blood flow itself.

A notable therapeutic application is lithotripsy, which refers to the noninvasive disintegration of kidney stones and gallstones with focused shock waves. Nonlinear acoustical effects in lithotripsy are associated not only with propagation of the shock wave but also with the generation of cavitation activity near the stones. Radiation of shock waves due to the collapse of cavitation bubbles is believed to be the dominant cause of stone breakup. An emerging therapeutic application, high-intensity focused ultrasound (HIFU), utilizes the heat dissipated by shock waves that develop in beams of focused ultrasound. The heating is so intense and localized that the potential exists for noninvasive cauterization of internal wounds and removal of tumors and scar tissue. See CAVITATION; ULTRASONICS. Mark F. Hamilton

Bibliography. R. T. Beyer, *Nonlinear Acoustics*, rev. ed., Acoustical Society of America, 1997; M. F. Hamilton and D. T. Blackstock, *Nonlinear Acoustics*, Academic Press, 1998; K. Naugolnykh and L. Ostrovsky, *Nonlinear Wave Processes in Acoustics*, Cambridge University Press, 1998; O. V. Rudenko and S. I. Soluyan, *Theoretical Foundations of Nonlinear Acoustics*, Plenum Press, 1977.

### Nonlinear control theory

A control system involves a plant and a controller. Plants are objects as diverse as a satellite, a distillation column, a robot arm, and a colony of bacteria. After measuring actual outputs of the plant, the controller computes signals that are applied at the inputs to the plant to achieve desired outputs. The design of controllers must be based upon mathematical models of plants, which are in most realistic situations composed of nonlinear differential and difference equations. A standard approach is to linearize the equations and use the powerful methods available for the design of linear control systems. See DIFFERENTIAL EQUATION; LINEAR SYSTEM ANALYSIS; OPTIMAL CONTROL (LINEAR SYSTEMS).

When the controlled outputs are allowed to have large deviations from the desired steady-state values, a linearized model will cease to describe the plant accurately, thereby causing erroneous results in the design. Linearized design models also fail in those important situations where nonlinearities are introduced into a controller to achieve a desired performance, generally at reduced cost. Typical examples of nonlinear controllers are on-off relays for temperature regulation in heating and cooling systems, switching elements in robot manipulators, and jet thrusters for the attitude control of space vehicles.

Unlike linear systems, there is no general theory for nonlinear systems. Nonlinear control theory is fragmented into areas centered around those classes of systems that are most prominent in applications. The basic ideas and concepts underlying a few of the areas will be described, with an emphasis on some fundamental mathematical notions and techniques.

**Nonlinear systems.** A simple nonlinear control system is the inverted pendulum (Fig. 1a), where it is required to keep the pendulum in an upright position by applying the torque  $u(t)$  at its base. The pendulum can be considered as a generic model for a rocket booster balanced on top of gimbaled thruster engine (Fig. 1b), and a controlled robot arm (Fig. 1c).

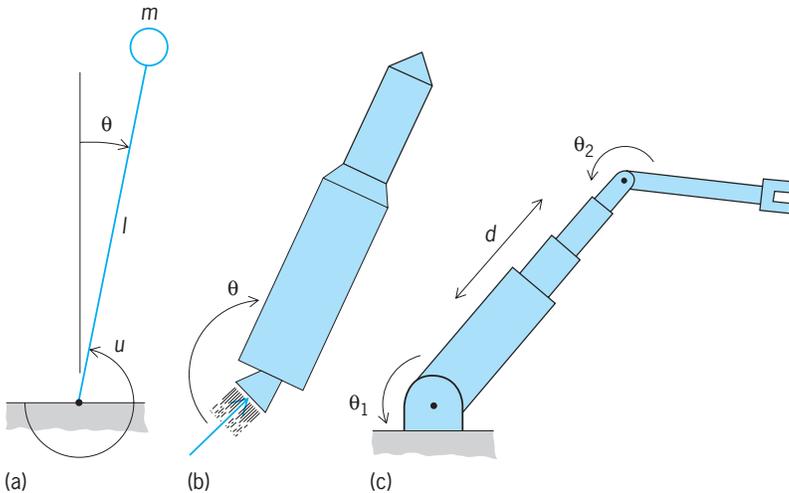


Fig. 1. Nonlinear systems. (a) Inverted pendulum, which serves as a model for (b) a rocket booster and (c) a robot arm.

Newton's second law of motion for the pendulum is Eq. (1), where  $m$  is the mass of the bob,  $l$  is the length of the pendulum, and  $g$  is the acceleration due to gravity. By choosing the angular position  $\theta = x_1$  and angular velocity  $\dot{\theta} = x_2$  of the bob as the state of the system, Eq. (1) can be rewritten as two first-order state equations (2). See PENDULUM.

$$m\dot{\theta}(t) = \frac{mg}{l} \sin \theta(t) + u(t) \quad (1)$$

$$\begin{aligned} \dot{x}_1 &= x_2 \\ \dot{x}_2 &= \frac{g}{l} \sin x_1 + u \end{aligned} \quad (2)$$

If the torque is absent ( $u = 0$ ), the upright position of a motionless pendulum ( $\dot{x}_1 = 0, \dot{x}_2 = 0$ ) is an equilibrium state:  $x_1 = 0, x_2 = 0$ . If the pendulum is slightly perturbed at this state, it will fall down and keep oscillating around the other equilibrium point at  $x_1 = \pi$  and  $x_2 = 0$ , where again,  $\dot{x}_1 = 0$  and  $\dot{x}_2 = 0$ .

To keep the pendulum in the upright position in physically realistic situations involving perturbations, the torque  $u$  is chosen to be a suitable function  $\phi$  of the state  $x_1, x_2$ , as in Eq. (3). That  $u$  should be a

$$u = \phi(x_1, x_2) \quad (3)$$

feedback control law in terms of the state is one of the most important results of control theory.

**Linearization.** A large class of plants can be described by a vector differential equation (4), where

$$\dot{x} = f(t, x, u) \quad (4)$$

the vector  $x = (x_1, x_2, \dots, x_n)^T$  is the state, the vector  $u = (u_1, u_2, \dots, u_m)^T$  is the input of the plant at time  $t$ , and  $T$  denotes the transpose which sends row vectors into column vectors. The plant is specified by the vector-valued function  $f = (f_1, f_2, \dots, f_n)^T$ , which is the infinitesimal state transition function describing the evolution of the state  $x \rightarrow x + dx$  corresponding to the change in time  $t \rightarrow t + dt$ .

In the absence of control, the plant is described by Eq. (5). The motions  $x(t; t_0, x_0)$  of the plant are

$$\dot{x} = f(t, x) \quad (5)$$

solutions of Eq. (5) for given initial conditions  $t_0, x_0$ . Equilibrium states are constant solutions of Eq. (5), which are determined by Eq. (6). When a system

$$f(t, x) = 0 \quad \text{for all } t \quad (6)$$

starts at an equilibrium state  $x_e$ , it stays there forever; that is, it satisfies Eq. (7).

$$x(t; t_0, x_e) = x_e \quad \text{for all } t > t_0 \quad (7)$$

A common approach to control system design is to assume small deviations of the state from a desired equilibrium  $x_e$  and linearize the plant at  $x_e$ . It is assumed, for simplicity, that the plant is time-invariant, so that Eq. (4) reduces to Eq. (8), and that  $x_e = 0$ .

$$\dot{x} = f(x, u) \quad (8)$$

The function  $f(x,u)$  can then be expanded in a Taylor series at  $x = 0$  and  $u = 0$  to obtain an approximate linear model given by Eq. (9), where the constant

$$\dot{x} = Ax + Bu + b(x, u) \quad (9)$$

matrices  $A = (a_{ij})$  and  $B = (b_{ij})$  of dimensions  $n \times n$  and  $n \times m$ , respectively, have elements given by Eqs. (10). The vector-valued function  $b(x,u)$  repre-

$$\begin{aligned} a_{ij} &= \frac{\partial f_i}{\partial x_j} \\ b_{ij} &= \frac{\partial f_i}{\partial u_j} \end{aligned} \quad (10)$$

sents the higher-order terms in the series and any modeling uncertainty in the system.

A linear feedback control law given by Eq. (11) can be used in Eq. (9) to obtain the closed-loop system given by Eq. (12). Numerous methods of linear

$$u = -Kx \quad (11)$$

$$\dot{x} = (A - BK)x + b(x) \quad (12)$$

system analysis can be applied to determine an appropriate  $m \times n$  gain matrix  $K = (k_{ij})$ , so that the linear part of Eq. (12) behaves satisfactorily and, at the same time, dominates any adverse effect of the nonlinear perturbation  $b(x)$ .

The linearized model of the pendulum is given by Eq. (13). A linear control, which can be expressed by

$$\dot{x} = \begin{bmatrix} 0 & 1 \\ \frac{g}{\ell} & 0 \end{bmatrix} x + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u + b(x) \quad (13)$$

Eq. (14), can always be chosen ( $k_1 > g/\ell$ ,  $k_2 > 0$ ) to make the closed loop, given by Eq. (15), stable, such

$$u = -k^T x \quad k = (k_1, k_2) \quad (14)$$

$$\dot{x} = \begin{bmatrix} 0 & 1 \\ \frac{g}{\ell} - k_1 & -k_2 \end{bmatrix} x + b(x) \quad (15)$$

that the controlled pendulum would return to the upright position after perturbation. In fact, stabilizing gains  $k_1$  and  $k_2$  can always be found regardless of the size of perturbation  $b(x)$ , although they must often have large values; this is so-called high-gain feedback.

Nonlinear feedback can be used to achieve exact linearization. By choosing the torque given by Eq. (16), the original nonlinear system of Eqs. (2)

$$u = -\frac{g}{\ell} \sin x_1 - k_1 x_1 - k_2 x_2 \quad (16)$$

becomes the linear system of Eqs. (15) without any perturbation  $b(x)$ . However, a simple cancellation of nonlinearity by the first term in Eq. (16) is not always possible; sophisticated concepts of differential geometry are needed to accomplish an exact linear design of an otherwise nonlinear plant. See DIFFERENTIAL GEOMETRY.

**Lyapunov's method.** An equilibrium of a dynamic system is stable if, after the system is slightly perturbed from the equilibrium, any subsequent motion of the system remains close to the equilibrium. If, in addition, each motion approaches the equilibrium as time increases, that is, expression (17) is

$$x(t; t_0, x_0) \rightarrow x_e \quad \text{as } t \rightarrow \infty \quad (17)$$

valid, then the equilibrium is asymptotically stable. See CONTROL SYSTEM STABILITY.

The underlying idea of Lyapunov's direct method, applied to a given isolated system, involves finding an energylike function  $V(x)$  that is positive and decays at each  $x$  except at the equilibrium  $x_e$ ; that is,  $V(x) > 0$  and  $\dot{V}(x) < 0$  for all  $x \neq x_e$ . Then, as  $V(x)$  continually decreases toward its minimum value  $V(x_e)$ , the motion  $x(t; t_0, x_0)$  asymptotically approaches  $x_e$ .

In Eq. (15), if  $b(x)$  is neglected and the values in Eqs. (18) are chosen for  $k_1$  and  $k_2$ , then the resulting linear control system is given by Eqs. (19). If the

$$k_1 = 1 + \frac{g}{\ell} \quad k_2 = 1 \quad (18)$$

$$\begin{aligned} \dot{x}_1 &= x_2 \\ \dot{x}_2 &= -x_1 - x_2 \end{aligned} \quad (19)$$

function  $V(x)$  is chosen as given in Eq. (20), then it follows from Eqs. (19) that the derivative of  $V(x)$  is given by Eqs. (21), which proves the asymptotic sta-

$$V(x) = \frac{3}{2}x_1^2 + x_1x_2 + x_2^2 > 0 \quad (20)$$

$$\begin{aligned} \dot{V}(x) &= 3x_1\dot{x}_1 + \dot{x}_1x_2 + x_1\dot{x}_2 + 2x_2\dot{x}_2 \\ &= -(x_1^2 + x_2^2) < 0 \end{aligned} \quad (21)$$

bility of  $x_e = 0$ . The motions of the system intersect the closed constant  $V$ -contours from the outside toward the inside until they reach the equilibrium (Fig. 2).

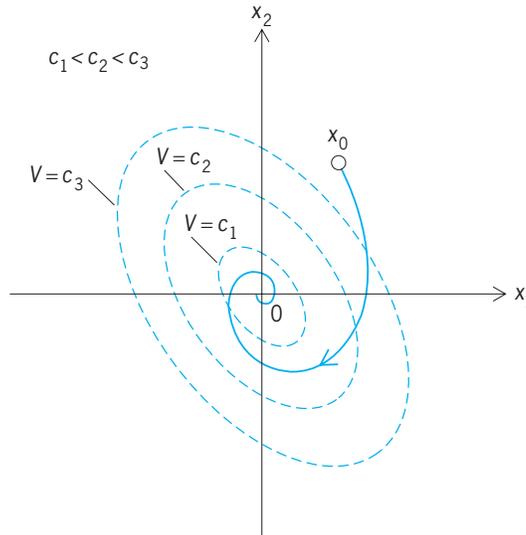


Fig. 2. Phase-plane ( $x_1, x_2$ -plane) portrait of the system given by Eqs. (19), showing contours of constant value of the Lyapunov function  $V$  given in Eq. (20) and a typical trajectory.

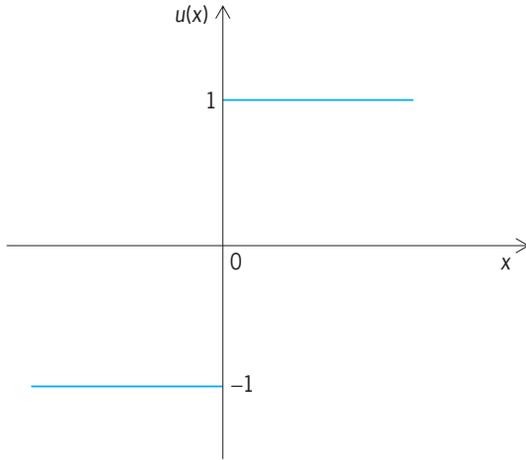


Fig. 3. Control function characteristic of a relay.

When an additional control term given by Eq. (22)

$$u = -\text{sgn}(\frac{1}{2}x_1 + x_2) \quad (22)$$

is inserted into the second of Eqs. (19), it increases the negativity of  $V(x)$  by  $-(\frac{1}{2}x_1 + x_2) \text{sgn}(\frac{1}{2}x_1 + x_2)$ . The added term produces the steepest possible descent motion of the system toward the equilibrium and, at the same time, makes the closed-loop system robust to uncertain perturbations  $b(x)$  that were neglected in Eq. (19). The control law of Eq. (22) can be implemented by a relay (Fig. 3). Controllers of this type are simple and reliable, resulting in robust variable structure systems.

In general, when a system is given as in Eq. (5), the aim of Lyapunov's method is to find a positive definite function  $V(t,x)$ , such that its total time derivative, given by Eq. (23), is negative definite along any

$$\dot{V}(t,x) = \frac{\partial V}{\partial t} + (\text{grad } V)^T f(t,x) \quad (23)$$

solution of Eq. (5). Stability, if present, is established without solving Eq. (5). Drawbacks of the method are its inconclusiveness when it fails and the difficulty in finding a suitable Lyapunov function  $V(t,x)$  that proves the stability of a given system. Generalizations of the method include vector Lyapunov functions for large-scale systems, hereditary systems, stochastic control, and robust control of uncertain systems. See LARGE SYSTEMS CONTROL THEORY; STOCHASTIC CONTROL THEORY.

**Popov's method.** A wide class of nonlinear control systems can be identified as the Lur'e type having the representation of Eqs. (24) as well as a block-

$$\dot{x} = Az + bu \quad (24a)$$

$$y = c^T x \quad (24b)$$

$$u = \phi(y) \quad (24c)$$

diagram representation (Fig. 4a). The system is an interconnection of the linear plant and a nonlinear controller. The linear part is described by the two state equations (22a) and (22b) or, equivalently, by

the transfer function of Eq. (25), where  $A$  is a matrix,

$$G(s) = c^T(A - sI)^{-1}b \quad (25)$$

$b$  and  $c$  are column vectors,  $I$  is the identity matrix (all of appropriate dimensions), and  $s = \sigma + j\omega$  is the complex variable. The nonlinear function  $\phi$  of the controller belongs to a sector  $[0,k]$  (Fig. 4b).

A Lur'e-type system is said to be absolutely stable if the equilibrium  $x_e = 0$  is stable for any initial state  $x_0$  and any  $\phi$  that belongs to the sector  $[0,k]$ . For this class of systems, a Lyapunov function is available as given in Eq. (26), which is a weighted

$$V(x) = x^T Hx + q \int_0^{c^T x} \phi(y) dy \quad (26)$$

sum of a quadratic form (with positive definite matrix  $H$ ) plus an integral of the nonlinearity. The system is absolutely stable, and  $V(x)$  is a Lyapunov function, if there is a number  $q$  so that the Popov's frequency domain condition, given by inequality (27), is satisfied.

$$k^{-1} + \text{Re}[(1 + j\omega q)G(j\omega)] > 0 \quad \text{for all } \omega \quad (27)$$

A graphical interpretation of inequality (27) can be given (Fig. 5), where the Popov locus  $G^*(j\omega)$ , given by Eq. (28), is required to lie to the right of the Popov line.

$$G^*(j\omega) = \text{Re } G(j\omega) + j\omega \text{Im } G(j\omega) \quad (28)$$

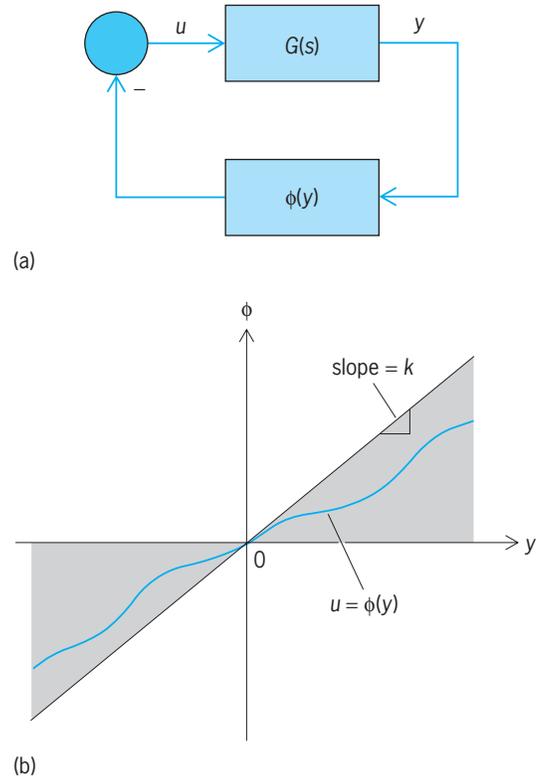


Fig. 4. Lur'e-type nonlinear control system. (a) Block diagram. (b) Nonlinearity sector, the shaded area between the  $y$  axis and the line through the origin with slope  $k$ , to which the nonlinear function of the controller,  $u = \phi(y)$ , is confined.

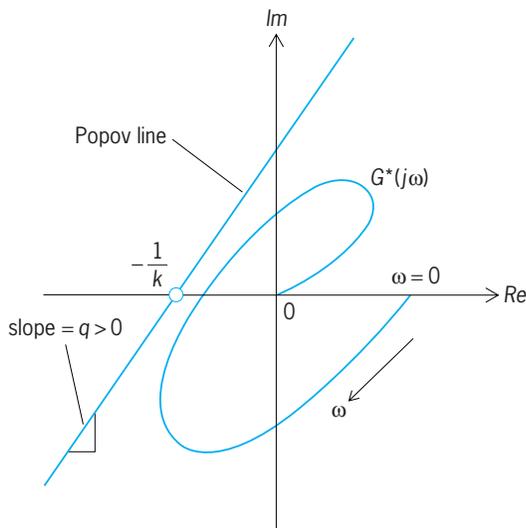


Fig. 5. Popov's graphical criterion. The Popov locus  $G^*(j\omega)$  must lie to the right of the Popov line.

The essential significance of the Popov's method is its use of the powerful frequency characteristics in the stability analysis of nonlinear control systems. In particular, it can provide a justification for the use of the describing function method, which is another widely used frequency domain technique but relies on a Fourier-series approximation of the nonlinearity. Generalizations of the Popov's method include plants with multiple inputs and outputs, hyperstability, and adaptive control systems. See ADAPTIVE CONTROL; CONTROL SYSTEM STABILITY; CONTROL SYSTEMS.

Dragoslav D. Šiljak

Bibliography. K. J. Åström and B. Wittenmark, *Adaptive Control*, 2d ed., 1994; A. Isidori, *Nonlinear Control Systems: An Introduction*, 1985; H. K. Khalil, *Nonlinear Systems*, 2d ed., 1995; V. M. Popov, *Hyperstability of Control Systems*, 1973; D. D. Šiljak, *Nonlinear Systems*, 1969; M. W. Spong and M. Vidyasagar, *Robot Dynamics and Control*, 1989; V. I. Utkin, *Sliding Modes in Control and Optimization*, 1992.

## Nonlinear optical devices

Devices that use the fact that the polarization in any real medium is a nonlinear function of the optical field strength to implement various useful functions. The nonlinearities themselves can be grouped roughly into second-order and third-order. Materials that possess inversion symmetry typically exhibit only third-order nonlinearities, whereas materials without inversion symmetry can exhibit both second- and third-order nonlinearities. See CRYSTALLOGRAPHY; ELECTRIC SUSCEPTIBILITY; ELECTROMAGNETIC RADIATION; POLARIZATION OF DIELECTRICS.

**Second-order devices.** Devices based on the second-order nonlinearity involve three-photon (or three-wave) mixing. In this process, two photons are mixed together to create a third photon, subject to energy- and momentum-conservation constraints.

Different names are ascribed to this mixing process, depending upon the relative magnitudes of the energies of the three photons. See CONSERVATION OF ENERGY; CONSERVATION OF MOMENTUM.

*Second harmonic generation and sum-frequency mixing.* When the two beginning photons are of equal energy or frequency, the mixing process gives a single photon with twice the energy or frequency of the original ones. This mixing process is called second-harmonic generation. Second-harmonic generation is used often in devices where photons of visible frequency are desired but the available underlying laser system is capable of producing only infrared photons. For example, the neodymium-doped yttrium-aluminum-garnet (Nd:YAG) laser produces photons in the infrared with a wavelength of 1.06 micrometers. These photons are then mixed in a crystal with a large second-order nonlinearity and proper momentum-conservation characteristics to yield green second-harmonic photons of 0.532- $\mu\text{m}$  wavelength. See LASER.

Under different momentum-conservation constraints, a similar interaction can take place between two photon fields of different frequency, resulting in photons whose energy or frequency is the sum of those of the original photons. This process is called sum-frequency mixing.

*Optical parametric oscillation/amplification.* This kind of mixing process occurs when one of the two initial photons has the largest energy and frequency of the three. A high-energy photon and a low-energy photon mix to give a third photon with an energy equal to the difference between the two initial photons. If initially the third field amplitude is zero, it is possible to generate a third field whose frequency is not previously present; in this case the process is called optical parametric oscillation. If the third field exists but at a low level, it can be amplified through the optical parametric amplification process. The existence of this mixing process is also subject to momentum-conservation constraints inside the second-order nonlinear crystal, in a fashion similar to the second harmonic generation process. See PARAMETRIC AMPLIFIER.

**Third-order devices.** Devices based on the third-order nonlinearity involve a process called four-photon (or four-wave) mixing. In this process, three photons are mixed together to create a fourth photon, subject to energy- and momentum-conservation constraints. The four-photon mixing nonlinearity is responsible for the existence of so-called self-action effects where the refractive index and absorption coefficient of a light field are modified by the light field's own presence, for third-harmonic generation and related processes, and for phase-conjugation processes.

*Nonlinear self-action effects.* In a medium with a third-order nonlinearity, the refractive index and absorption coefficient of a light field present in the medium are modified by the strength of the light intensity. Because the field effectively acts on itself, this interaction is termed a self-action effect. The momentum-conservation constraints are automatically satisfied

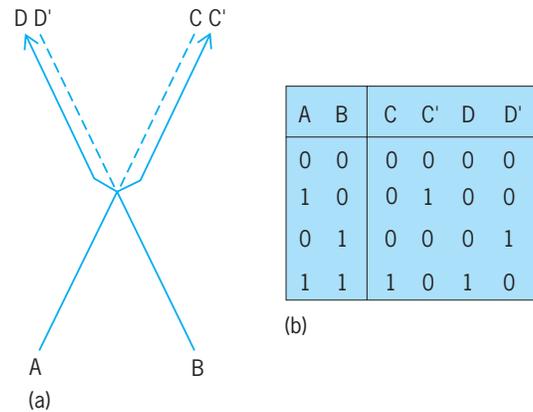
because of the degenerate frequencies involved in the interaction. Such an interaction manifests itself by changing the total absorption experienced by the light field as well as by changing the velocity of propagation of the light field. See ABSORPTION OF ELECTROMAGNETIC RADIATION; REFRACTION OF WAVES.

There are many devices based on the self-action effects. A reverse saturable absorber becomes more opaque because of the nonlinear absorption (also called two-photon adsorption) that it manifests. Refractive-index changes can be used to change the transmission characteristics of resonant cavities and other structures by modifying the effective optical path length (the product of actual structure length times the effective refractive index for the structure) and shifting the cavity resonances to other frequencies. Several nonlinear optical switches have been proposed based upon this resonance-shifting phenomenon. See CAVITY RESONATOR; OPTICAL INSTABILITY.

Finally, the self-action effects can be used in a nonlinear waveguide under certain conditions to cancel out the deleterious effects of either group velocity dispersion for short light pulses or diffraction for continuous beams. The special light pulses based upon this cancellation behave somewhat like particles, and are named optical solitons. This is a consequence of the fact that Maxwell's equations, which govern the propagation of light, can be put into a special form called the nonlinear Schrödinger (NLS) equation when the nonlinearity is of self-action form. The NLS equation resembles the Schrödinger equation of quantum mechanics, with the potential term in the latter equation replaced by a nonlinear term proportional to the local intensity of the light field, and possesses soliton solutions. See MAXWELL'S EQUATIONS; OPTICAL PULSES; QUANTUM MECHANICS; SCHRÖDINGER'S WAVE EQUATION.

Solitons are interesting types of waves because of the particlelike characteristics they exhibit. They are robust in the presence of large-scale perturbations; even strong interactions (collisions with other solitons) do not modify their shape but only shift their center of gravity by a small amount. These robust characteristics suggest that solitons should have application in telecommunication transmission systems where pulses of light are required to propagate great distances in optical fibers characterized by relatively large values of dispersion and nonlinearity. Optical solitons can, indeed, propagate the thousands of miles necessary for intercontinental fiber links. Various other proposals have also been made for schemes utilizing the interaction properties of solitons in logic operations. These devices are of two primary types, soliton dragging gates and soliton collision gates (see *illus.*). See LOGIC CIRCUITS; OPTICAL COMMUNICATIONS; OPTICAL FIBERS; SOLITON.

*Third-harmonic generation.* In a third-harmonic generation process, three photons of like energy and frequency are mixed to yield a single photon with three times the energy and frequency of the initial photons. As in second-harmonic generation, this process requires a set of momentum-conservation con-



**Soliton interaction gate.** (a) Solitons are initiated at A and B, and detectors are placed at C, C', D, D'. If only soliton A or soliton B is present, the soliton trajectories will follow the broken lines; if both A and B are present, the solitons will follow the solid lines. (b) Truth table. 0 represents the absence of a soliton; 1 represents the presence of a soliton.

straints to be met. Applications of third-harmonic generation are typically in the areas of frequency up-conversion. For example, the fundamental Nd:YAG laser emission can be converted to ultraviolet light of  $0.355\text{-}\mu\text{m}$  wavelength with the appropriate nonlinear crystal.

*Phase conjugation.* Phase-conjugation devices make use of a property that third-order media possess whereby energy- and frequency-degenerate photons from two counterpropagating fields are mixed with an incoming photon to yield a photon with exactly the opposite propagation direction and conjugate phase. This phase-conjugate field will pass out of the nonlinear optical device in exactly the direction opposite to the incoming field. Such devices are used in phase-conjugate mirrors, mirrors which have the ability to cancel phase variation in a beam due to, for example, atmospheric turbulence. See ADAPTIVE OPTICS; OPTICAL PHASE CONJUGATION.

**Materials.** The suitability of available nonlinear optical materials is a critical factor in the development of nonlinear optical devices. For certain applications, silica glass fibers may be used. Because of the long propagation distances involved in intercontinental transmission systems, the small size of the optical nonlinearity in silica is not a drawback, although the fibers used in the soliton transmission systems have been especially engineered for reduced dispersion at the transmission wavelength of  $1.55\ \mu\text{m}$ . However, generally the nonlinearities available are either below or right at the figure-of-merit thresholds for particular devices. Key materials are semiconductors [such as gallium arsenide (GaAs), zinc selenide (ZnSe), and indium gallium arsenide phosphide (InGaAsP)], certain organic polymeric films, hybrid materials such as semiconductor-doped glasses, and liquid crystals. However, there are several formidable fundamental limits to the development of materials due to the intimate relationship between the absorption and refraction processes that occur in all materials. See NONLINEAR OPTICS; OPTICAL MATERIALS.

David R. Andersen

Bibliography. R. W. Boyd, *Nonlinear Optics*, 1992; P. N. Butcher and D. Cotter, *The Elements of Nonlinear Optics*, 1990; M. Schubert and B. Wilhelm, *Nonlinear Optics and Quantum Electronics*, 1986.

## Nonlinear optics

A field of study concerned with the interaction of electromagnetic radiation and matter in which the matter responds in a nonlinear manner to the incident radiation fields. The nonlinear response can result in intensity-dependent variation of the propagation characteristics of the radiation fields or in the creation of radiation fields that propagate at new frequencies or in new directions. Nonlinear effects can take place in solids, liquids, gases, and plasmas, and may involve one or more electromagnetic fields as well as internal excitations of the medium. Most of the work done in the field has made use of the high powers available from lasers. The wavelength range of interest generally extends from the far-infrared to the vacuum ultraviolet, but some nonlinear interactions have been observed at wavelengths extending from the microwave to the x-ray ranges. See LASER.

**Nonlinear materials.** Nonlinear effects of various types are observed at sufficiently high light intensities in all materials. It is convenient to characterize the response of the medium mathematically by expanding it in a power series in the electric and magnetic fields of the incident optical waves. The linear terms in such an expansion give rise to the linear index of refraction, linear absorption, and the magnetic permeability of the medium, while the higher-order terms give rise to nonlinear effects. See ABSORPTION OF ELECTROMAGNETIC RADIATION; REFRACTION OF WAVES.

In general, nonlinear effects associated with the electric field of the incident radiation dominate over magnetic interactions. The even-order dipole susceptibilities are zero except in media which lack a center of symmetry, such as certain classes of crystals, certain symmetric media to which external forces have been applied, or at boundaries between certain dissimilar materials. Odd-order terms can be nonzero in all materials regardless of symmetry. Generally the magnitudes of the nonlinear susceptibilities decrease rapidly as the order of the interaction increases. Second- and third-order effects have been the most extensively studied of the nonlinear interactions, although effects up to order 30 have been observed in a single process. In some situations, multiple low-order interactions occur, resulting in a very high effective order for the overall nonlinear process. For example, ionization through absorption of effectively 100 photons has been observed. In other situations, such as dielectric breakdown or saturation of absorption, effects of different order cannot be separated, and all orders must be included in the response. See ELECTRIC SUSCEPTIBILITY; POLARIZATION OF DIELECTRICS.

**Second-order effects.** Second-order effects involve

a polarization with the dependence  $P_m^{(2)} = dE^2$ , where  $E$  is the electric field of the optical waves and  $d$  is a nonlinear susceptibility. The second-order polarization has components that oscillate at sum and difference combinations of the incident frequencies, and also a component that does not oscillate. The oscillating components produce a propagating polarization wave in the medium with a propagation vector that is the appropriate sum or difference of the propagation vectors of the incident waves. The nonlinear polarization wave serves as a source for an optical wave at the corresponding frequency in a process that is termed three-wave parametric mixing.

**Phase matching.** The strongest interaction occurs when the phase velocity of the polarization wave is the same as that of a freely propagating wave of the same frequency. The process is then said to be phase-matched. Dispersion in the refractive indices, which occurs in all materials, usually prevents phase matching from occurring unless special steps are taken. Phase matching in crystals can be achieved by using noncollinear beams, materials with periodic structures, anomalous dispersion near an absorption edge, or compensation using free carriers in a magnetic field, or by using the birefringence possessed by some crystals. In the birefringence technique, the one used most commonly for second-order interactions, one or two of the interacting waves propagate as an extraordinary wave in the crystal. The phase-matching conditions are achieved by choosing the proper temperature and propagation direction. For a given material these conditions depend on the wavelengths and direction of polarization of the individual waves. The conditions for phase-matched three-wave parametric mixing can be summarized by Eqs. (1) and (2), where the  $\nu$ 's are the frequencies

$$\nu_3 = \nu_1 \pm \nu_2 \quad (1)$$

$$\mathbf{k}_3 = \mathbf{k}_1 \pm \mathbf{k}_2 \quad (2)$$

of the waves and the  $\mathbf{k}$ 's are the propagation constants. Phase matching increases the power in the generated wave by many orders of magnitude. See CRYSTAL OPTICS.

**Harmonic generation and frequency mixing.** In three-wave sum- and difference-frequency mixing, two incident waves at  $\nu_1$  and  $\nu_2$  are converted to a third wave at  $\nu_3$  according to Eq. (1). The simplest interaction of this type is second-harmonic generation in which  $\nu_3 = 2\nu_1$ . For this interaction the phase-matching condition reduces to Eq. (3), where  $n(\nu_3)$  is the refrac-

$$n(\nu_3) = \frac{n_1(\nu_1)}{2} + \frac{n_2(\nu_1)}{2} \quad (3)$$

tive index at the harmonic wavelength, and  $n_1(\nu_1)$  and  $n_2(\nu_1)$  are the refractive indices of the incident waves at the fundamental frequency.

Second-harmonic generation and second-order frequency mixing have been demonstrated at wavelengths ranging from the infrared to the ultraviolet, generally coinciding with the transparency range

of the nonlinear crystals. Second-harmonic conversion has been used with pump radiation extending from 10  $\mu\text{m}$  (from carbon dioxide lasers) to 217 nm, the phase-matching limit of currently available crystals. Sum-frequency mixing has been used to generate wavelengths as short as 172 nm. Difference-frequency mixing can be used to generate both visible and infrared radiation out to wavelengths of about 2 mm. If one or more of the incident wavelengths is tunable, the generated wavelength will also be tunable, providing a useful source of radiation for high-resolution spectroscopy. Radiation can be converted from longer-wavelength ranges, such as the infrared, to shorter-wavelength regions, such as the visible, where more sensitive and more convenient detectors are available. *See* LASER SPECTROSCOPY.

In principle, 100% of the energy in the incident radiation can be converted to the generated wave in frequency-mixing interactions. In most situations, however, additional linear and nonlinear interactions prevent the conversion efficiency from reaching its theoretical maximum. For second-order interactions the most important limiting effects are absorption of the radiation at either incident or generated frequencies (generally important for generation in the mid- to far-infrared) or failure to achieve or maintain phase matching throughout the interaction (important at shorter wavelengths in the infrared, or in the visible or ultraviolet). Divergent angles, a spread of frequencies, or even minor amounts of phase structure (multiple frequency components) in the incident beams can limit phase matching. Phase matching can also be disturbed by heating of the crystal following absorption of a small amount of the pump radiation. In practice, second-harmonic conversion efficiency of over 90% has been achieved with radiation from high-power pulsed lasers. Conversion of radiation from continuous-wave lasers of over 30% has been obtained when the nonlinear crystal is placed in the laser cavity or in an external resonator.

*Parametric generation.* In parametric generation, which is the reverse process of sum-frequency mixing, a single input wave at  $\nu_3$  is converted to two lower-frequency waves according to the relation  $\nu_3 = \nu_1 + \nu_2$ . The individual values of  $\nu_1$  and  $\nu_2$  are determined by the simultaneous satisfaction of the phase-matching condition in Eq. (2). Generally this condition can be satisfied for only one pair of frequencies at a time for a given propagation direction. By changing the phase-matching conditions, the individual longer wavelengths can be tuned. Parametric generation can be used to amplify waves at lower frequencies than the pump wave or, in an oscillator, as a source of tunable radiation in the visible or infrared over a wavelength range similar to that covered in difference-frequency mixing.

*Optical rectification.* The component of the second-order polarization that does not oscillate produces an electrical voltage. The effect is called optical rectification and is a direct analog of the rectification that occurs in electrical circuits at much lower fre-

quencies. It has been used in conjunction with ultrashort mode-locked laser pulses to produce some of the shortest electrical pulses generated, with durations of less than 250 femtoseconds. *See* OPTICAL PULSES.

**Third-order interactions.** Third-order interactions give rise to several types of nonlinear effects. Four-wave parametric mixing involves interactions which generate waves at sum- and difference-frequency combinations of the form of Eq. (4) with the corresponding phase-matching condition of Eq. (5). Phase

$$\nu_4 = \nu_1 \pm \nu_2 \pm \nu_3 \quad (4)$$

$$\mathbf{k}_4 = \mathbf{k}_1 \pm \mathbf{k}_2 \pm \mathbf{k}_3 \quad (5)$$

matching in liquids is usually accomplished by using noncollinear pump beams. Phase matching in gases can also be accomplished by using anomalous dispersion near absorption resonances or by using mixtures of gases, one of which exhibits anomalous dispersion. The use of gases, which are usually transparent to longer and shorter wavelengths than are the solids used for second-order mixing, allows the range of wavelengths covered by parametric mixing interactions to be extended considerably. Four-wave mixing processes of the type  $\nu_4 = \nu_1 - \nu_2 - \nu_3$  have been used to generate far-infrared radiation out to wavelengths of the order of 25  $\mu\text{m}$ . Sum-frequency mixing and third-harmonic generation have been used to generate radiation extending to 57 nm.

The nonlinear susceptibility is greatly increased, sometimes by four to eight orders of magnitude, through resonant enhancement that occurs when the input frequencies or their multiples or sum or difference combinations coincide with appropriate energy levels in the nonlinear medium. Two-photon resonances are of particular importance since they do not involve strong absorption of the incident or generated waves. Resonant enhancement in gases has allowed tunable dye lasers to be used for nonlinear interactions, providing a source of tunable radiation in the vacuum ultraviolet and in the far-infrared. Such radiation is useful in spectroscopic studies of atoms and molecules.

Just as with three-wave mixing, four-wave sum-frequency generation can be used to convert infrared radiation to the visible, where it is more easily detected. These interactions have been used for infrared spectroscopic studies and for infrared image conversion. *See* INFRARED IMAGING DEVICES; INFRARED SPECTROSCOPY.

Conversion efficiency of the third- and higher-order interactions is generally lower than for the second-order processes because the competing effects are more important. Third- and higher-order interactions are generally limited by linear and nonlinear absorption, Stark shifting of resonant levels, and multiphoton ionization that lead to reduction of the nonlinear susceptibility, population redistribution, absorption of radiation, or breaking of phase matching. In third-order effects, the competing mechanisms are of the same order as the frequency-mixing interaction, leading to a maximum efficiency that de-

depends only on the material constants and detunings from resonance. The maximum possible conversion efficiency for these interactions is of the order of 30%, while the maximum reported is 10%.

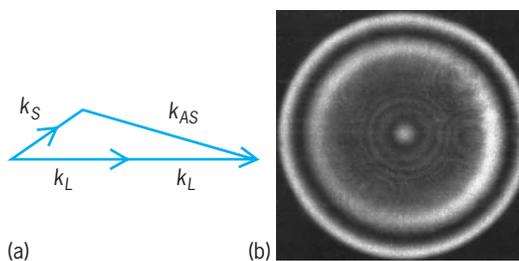
Conversion efficiencies for the higher-order interactions are usually lower than for the lower-order ones. The higher-order interactions are thus most useful in extending coherent radiation into a range of wavelengths not reachable with lower-order interactions. Conversion to harmonics of very high order has been achieved with extremely intense optical pulses with durations of the order of several hundred femtoseconds or less, where the energy in the various higher harmonic orders is relatively constant, rather than decreasing with harmonic order. Conversion to harmonic orders of up to 111 and generation of radiation down to 7.26 nm have been observed. At the intensities used in these interactions, generally greater than  $10^{15}$  W/cm<sup>2</sup>, perturbation theory breaks down and a more complete theory of the interaction is needed.

Two-wave mixing is an interaction that occurs in photorefractive materials (discussed below) between two waves of the same frequency, in which power is transferred from one wave to another. It can be used to amplify a weaker beam with a stronger one. When the stronger beam has spatial aberrations, it transfers the power to the weaker beam, but not the aberrations, effectively reducing the distortion of the optical beam.

**Stimulated scattering.** Light can scatter inelastically from fundamental excitations in the medium, resulting in the production of radiation at a frequency that is shifted from that of the incident light by the frequency of the excitation involved. The difference in photon energy between the incident and scattered light is accounted for by excitation or deexcitation of the medium. Some examples are Brillouin scattering from acoustic vibrations; various forms of Raman scattering involving molecular rotations or vibrations, electronic states in atoms or molecules, lattice vibrations or spin waves in solids, spin flips in semiconductors, and electron plasma waves in plasmas; Rayleigh scattering involving density or entropy fluctuations; and scattering from concentration fluctuations in gases. See SCATTERING OF ELECTROMAGNETIC RADIATION.

At the power levels available from pulsed lasers, the scattered light experiences exponential gain, and the process is then termed stimulated, in analogy to the process of stimulated emission in lasers. In stimulated scattering, the incident light can be almost completely converted to the scattered radiation. Stimulated scattering has been observed for all of the internal excitations listed above. The most widely used of these processes are stimulated Raman scattering and stimulated Brillouin scattering.

*Stimulated Raman scattering.* In its simplest form, stimulated Raman scattering involves transfer of energy from a laser, or pump, wave to a longer-length wave termed the Stokes wave. If the Stokes wave becomes intense enough, it can produce its own Stokes wave, termed a second Stokes wave, shifted to longer



**Fig. 1. Anti-Stokes generation by four-wave mixing.** (a) Phase-matching diagrams showing the direction of propagation of the waves involved. (b) Characteristic ring structure of anti-Stokes generation in hydrogen gas. The appearance of two rings is caused by an interaction between stimulated Raman scattering and four-wave mixing that reduces the anti-Stokes generation at exact phase matching. (From M. D. Duncan et al., *Parametric Raman gain suppression in D<sub>2</sub> and H<sub>2</sub>*, *Opt. Lett.*, 11:803-805, 1986)

wavelengths from the first Stokes frequency by the material excitation frequency. Light at wavelengths shorter than that of the original laser, termed anti-Stokes waves, can be produced by four-wave mixing processes of the form of Eq. (6), where  $\nu_L$

$$\nu_{AS} = 2\nu_L - \nu_S \quad (6)$$

is the laser frequency and  $\nu_S$  and  $\nu_{AS}$  are the frequencies of the Stokes and anti-Stokes waves, respectively. These waves are usually emitted in cones because of the need for phase matching (Fig. 1). At sufficiently high intensities the stimulated Raman scattering spectrum can consist of many Stokes and anti-Stokes waves.

Stimulated Raman scattering can occur in the forward or backward direction. In the forward direction the most common uses are amplifying weak Stokes signals, generating multiple-wavelength or tunable sources for spectroscopic applications, removing laser-beam aberrations, and combining radiation from multiple lasers. Backward-stimulated Raman scattering is commonly used for pulse compression. See RAMAN EFFECT.

*Stimulated Brillouin scattering.* Stimulated Brillouin scattering involves scattering from acoustic waves and is typically done in liquids or gases. The Stokes wave is usually generated in the backward direction, for which the gain is the highest. Frequency shifts are typically of the order of hundreds of megahertz to tens of gigahertz, depending on the material and the wavelength of the original laser. The most common use of stimulated Brillouin scattering is in phase conjugation. Stimulated Brillouin scattering is also a source of damage in many solids that limits the intensity that can be used.

**Self-action and related effects.** Nonlinear polarization components at the same frequencies as those in the incident waves can result in effects that change the index of refraction or the absorption coefficient, quantities that are constants in linear optical theory.

*Multiphoton absorption and ionization.* Materials that are transparent at low optical intensities can undergo an increase in absorption at high intensities. This effect involves simultaneous absorption of two or more photons from one or more incident waves. When the

process involves transitions to discrete upper levels in gases or to conduction bands in solids that obey certain quantum-mechanical selection rules, it is usually termed multiphoton absorption, or, more specifically,  $n$ -photon absorption, where  $n$  is the number of photons involved. When transitions to the continuum of gases are involved, the process is termed multiphoton ionization.

*Saturable absorption.* In materials which have a strong linear absorption at the incident frequency, the absorption can decrease at high intensities, an effect termed saturable absorption. Saturable absorbers are useful in operating Q-switched and mode-locked lasers.

*Self-focusing and self-defocusing.* Intensity-dependent changes in refractive index can affect the propagation characteristics of a laser beam. For many materials the index of refraction increases with increasing optical intensity. If the laser beam has a profile that is more intense in the center than at the edges, the profile of the refractive index corresponds to that of a positive lens, causing the beam to focus. This effect, termed self-focusing, can cause an initially uniform laser beam to break up into many smaller spots with diameters of the order of a few micrometers and intensity levels that are high enough to damage many solids. This mechanism limits the maximum intensities that can be obtained from some high-power pulsed solid-state lasers. See KERR EFFECT.

In other materials the refractive index decreases as the optical intensity increases. The resulting nonlinear lens causes the beam to defocus, an effect termed self-defocusing. When encountered in media

that are weakly absorbing, the effect is termed thermal blooming. It is prominent, for example, in the propagation of high-power infrared laser beams through the atmosphere.

*Broadening of spectrum.* The nonlinear refractive index can lead to a broadening of the spectrum of pulsed laser fields. The broadened spectrum may extend from the ultraviolet to the infrared and has been used with picosecond-duration pulses for time-resolved spectroscopic studies. It can also be combined with dispersive delay lines to shorten the pulse duration.

*Degenerate four-wave mixing.* The same interactions that give rise to the self-action effects can also cause nonlinear interactions between waves that are at the same frequency but are otherwise distinguishable, for example, by their direction of polarization or propagation. Termed degenerate four-wave mixing, this interaction gives rise to a number of effects such as amplification of a weak probe wave and phase conjugation (discussed below).

*Optical fibers.* Propagation through optical fibers can involve several nonlinear optical interactions. Self-phase modulation resulting from the nonlinear index can be used to spread the spectrum, and subsequent compression with diffraction gratings and prisms can be used to reduce the pulse duration. The shortest optical pulses, with durations of the order of 6 femtoseconds, have been produced in this manner.

Linear dispersion in fibers causes pulses to spread in duration and is one of the major limitations on data transmission through fibers. Dispersive pulse spreading can be minimized with solitons, which are specially shaped pulses that propagate long distances without spreading. They are formed by a combined interaction of spectral broadening due to the nonlinear refractive index and anomalous dispersion found in certain parts of the spectrum. See SOLITON.

Stimulated Raman scattering from vibrational excitations in the fiber can limit the intensity that can be transmitted in the laser wave. Because of the long distances involved, this can happen at intensities accessible by diode lasers. If two pulses with different wavelengths are propagated, stimulated Raman scattering can be used to amplify the longer-wavelength pulse in the fiber continuously, reducing the need for electronic repeaters in long-distance systems. See OPTICAL COMMUNICATIONS; OPTICAL FIBERS.

*Control by low-frequency fields.* Low-frequency electric, magnetic, or acoustic fields can be used to control the polarization or propagation of an optical wave. These effects, termed electro-, magneto-, or acousto-optic effects, are useful in the modulation and deflection of light waves and are used in information-handling systems. See ACOUSTOOPTICS; ELECTROOPTICS; MAGNETOOPTICS; OPTICAL INFORMATION SYSTEMS; OPTICAL MODULATORS.

**Coherent effects.** Another class of effects involves a coherent interaction between the optical field and an atom in which the phase of the atomic wave functions is preserved during the interaction. These interactions involve the transfer of a significant fraction of the atomic population to an excited state. As

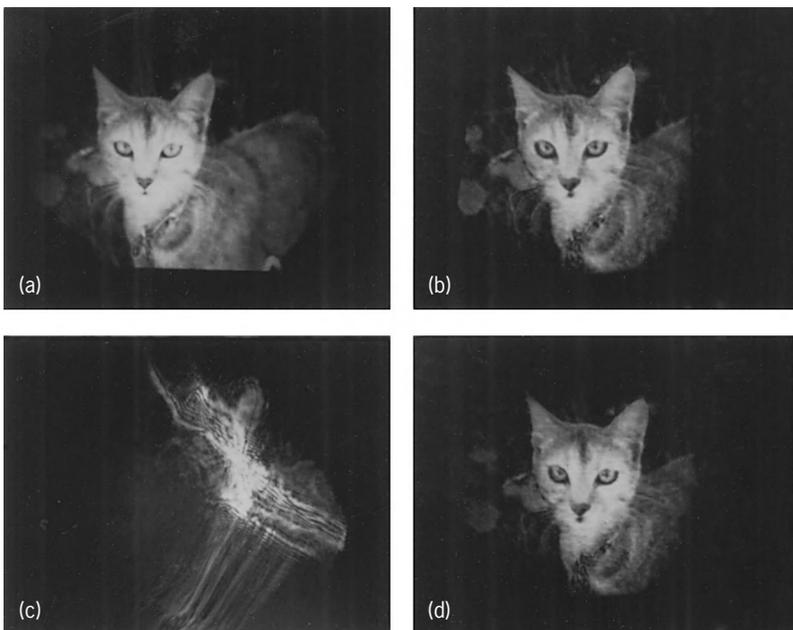


Fig. 2. Use of nonlinear optical phase conjugation for correction of distortions on an image. (a) Image obtained with a normal mirror and no distortion. (b) Image obtained with an optical conjugator and no distortion. (c) Image obtained with a normal mirror when the object is viewed through a distorting piece of glass. (d) Image obtained with an optical conjugator when the object is viewed through a distorting piece of glass. (From J. Feinberg, *Self-pumped, continuous-wave phase conjugator using internal reflection*, *Opt. Lett.*, 7:486-488, 1982)

a result, they cannot be described with the simple perturbation expansion used for the other nonlinear optical effects. Rather they require that the response be described by using all powers of the incident fields. These effects are generally observed only for short light pulses, of the order of several nanoseconds or less. In one interaction, termed self-induced transparency, a pulse of light of the proper shape, magnitude, and duration can propagate unattenuated in a medium which is otherwise absorbing.

Other coherent effects involve changes of the propagation speed of a light pulse or production of a coherent pulse of light, termed a photon echo, at a characteristic time after two pulses of light spaced apart by a time interval have entered the medium. Still other coherent interactions involve oscillations of the atomic polarization, giving rise to effects known as optical nutation and free induction decay. Two-photon coherent effects are also possible.

**Other applications.** In addition to the applications noted above in the discussions of the various interactions, many applications can be accomplished with more than one type of interaction.

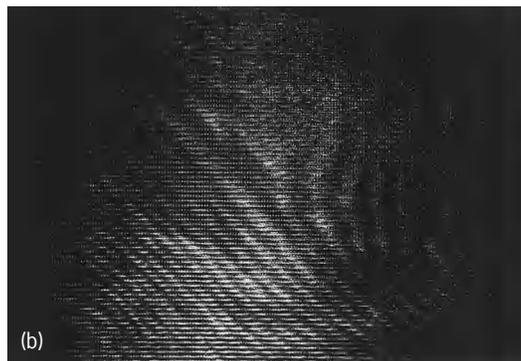
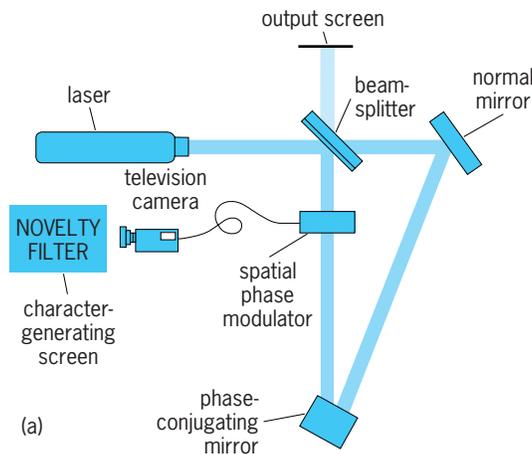
*Nonlinear spectroscopy.* The variation of the nonlinear susceptibility near the resonances that correspond to sum- and difference-frequency combinations of the input frequencies forms the basis for various types of nonlinear spectroscopy which allow study of energy levels that are not normally accessible with linear optical spectroscopy.

Nonlinear spectroscopy can be performed with many of the interactions discussed earlier. Multiphoton absorption spectroscopy can be performed by using two strong laser beams, or a strong laser beam and a weak broadband light source. If two counterpropagating laser beams are used, spectroscopic studies can be made of energy levels in gases with spectral resolutions much smaller than the Doppler limit. Nonlinear optical spectroscopy has been used to identify many new energy levels with principal quantum numbers as high as 150 in several elements. See RESONANCE IONIZATION SPECTROSCOPY; RYDBERG ATOM.

Many types of four-wave mixing interactions can also be used in nonlinear spectroscopy. The most widespread of these processes, termed coherent anti-Stokes Raman spectroscopy (CARS), offers the advantage of greatly increased signal levels over linear Raman spectroscopy for the study of certain classes of materials.

*Phase conjugation.* Optical phase conjugation is an interaction that generates a wave that propagates in the direction opposite to a reference, or signal, wave, and has the same spatial variations in intensity and phase as the original signal wave, but with the sense of the phase variations reversed. Several nonlinear interactions are used to produce phase conjugation.

Optical phase conjugation allows correction of optical distortions that occur because of propagation through a distorting medium (Fig. 2). This process can be used for improvement of laser-beam quality, optical beam combining, correction of distortion because of mode dispersion in fibers, and



**Fig. 3.** Novelty filter that uses phase conjugation in a photorefractive material to select changes in a pattern. (a) Configuration of components. (b) Dark output of the novelty filter that occurs when the filter has adapted to the pattern in the spatial light modulator. (c) Output of the filter immediately after a new pattern has been supplied to the spatial light modulator. If the pattern is left unchanged, the output will gradually change to b. (From D. Z. Anderson and M. C. Erie, *Resonator memories and optical novelty filters*, *Opt. Eng.*, 26:434-444, 1987)

stabilized aiming. It can also be used for neural networks that exhibit learning properties (Fig. 3). See NEURAL NETWORK; OPTICAL PHASE CONJUGATION.

*Optical bistability.* Optical elements whose intensity transmission is bistable can be produced with Fabry-Perot etalons that contain materials that exhibit either saturable absorption or an intensity-dependent index of refraction. These devices can be used to perform many of the functions associated with transistors, such as differential gain, switching, and limiting. Hybrid forms, which involve electrooptical

elements, are compatible with integrated optical devices. See OPTICAL BISTABILITY.

*Time measurements.* Various nonlinear interactions, such as second harmonic generation, multiphoton absorption followed by fluorescence, or coherent Raman scattering, have provided a means for measurement of pulse duration and excited-state lifetimes in the picosecond and subpicosecond time regimes.

*Self-organizing patterns.* Some types of systems, for example those involving nonlinear interactions in an optical cavity, produce optical beams with self-organizing patterns. Such patterns occur as a result of the competition among modes of the cavity in the presence of the nonlinear element. They have potential applications in areas such as information processing and optical displays. John F. Reintjes

**Photorefractive effect.** The photorefractive effect occurs in many electrooptic materials. A change in the index of refraction in a photorefractive medium arises from the redistribution of charge that is induced by the presence of light. Charge carriers that are trapped in impurity sites in a photorefractive medium are excited into the material's conduction band when exposed to light. The charges migrate in the conduction band until they become retrapped at other sites. The charge redistribution produces an electric field that in turn produces a spatially varying index change through the electrooptic effect in the material. Unlike most other nonlinear effects, the index change of the photorefractive effect is retained for a time in the absence of the light and thus may be used as an optical storage mechanism. The memory time depends on the dark electrical conductivity of the medium. Storage times range from milliseconds to months or years, depending upon the material and the methods employed. See TRAPS IN SOLIDS.

*Applications.* Photorefractive materials are often used for holographic storage. In this case, the index change mimics the intensity interference pattern of two beams of light. Over 500 holograms have been stored in the volume of a single crystal of iron-doped lithium niobate. See HOLOGRAPHY.

Photorefractive materials are typically sensitive to very low light levels. Wave-mixing interactions in photorefractive materials are analogous to conventional third-order interactions. Excellent phase-conjugate wave fidelity has been achieved in several photorefractive materials, for example. The photorefractive effect is, however, extremely slow by the standards of optical nonlinearity. Because of their sensitivity, photorefractive materials are increasingly used for image and optical-signal processing applications. See IMAGE PROCESSING; NONLINEAR OPTICAL DEVICES. Dana Z. Anderson

*Photorefractive polymers.* Photorefractive polymers are among the most sensitive nonlinear optical recording materials. They exhibit large refractive index changes when exposed to low-power laser beams. When the optical excitation consists of two interfering coherent beams, the periodic light distribution produces a periodic refractive index modulation. The resulting index change produces a hologram in

the volume of the polymer film. The hologram can be reconstructed by diffracting a third laser beam on the periodic index modulation. In contrast to many physical processes that can be used to generate a refractive index change, the photorefractive effect is fully reversible, meaning that the recorded holograms can be erased with a spatially uniform light beam. This reversibility makes photorefractive polymers suitable for real-time optical processing applications.

The mechanism that leads to the formation of a photorefractive index modulation involves the formation of an internal electric field through the absorption of light, the generation of carriers, and the transport and trapping of the carriers over macroscopic distances. The resulting electric field produces a refractive index change through orientational or nonlinear optical effects. While a refractive index change is obtained in crystals through the Pockels effect, most photorefractive polymers exhibit index changes due to orientational Kerr effects. In current polymers, the highest index modulation amplitude is of the order of 1% and the fastest materials have a response time of a few milliseconds.

Due to the transport process, the index modulation amplitude is phase-shifted with respect to the periodic light distribution produced by the interfering optical beams that generate the hologram. This phase shift enables the coherent energy transfer between two beams propagating in a thick photorefractive material. This property, referred to as two-beam coupling, is used to build optical amplifiers.

The photorefractive effect was discovered in inorganic crystals, then studied for several decades mainly in inorganic crystals and semiconductors. The effect was discovered in organic materials in the 1990s. Photorefractive materials are used in holographic storage, nondestructive testing, holographic time gating for image filtering, novelty filtering, phase conjugation, optical correlation, and reconfigurable interconnects. While crystals and semiconductors require advanced fabrication technologies, organic photorefractive polymers have the potential to be mass-produced by techniques such as injection molding. Hence, they can be fabricated into objects of various shapes at low cost. This property together with advances in fabricating integrated laser sources at lower cost can provide great momentum to the development of new optical processing technologies. As media for real-time optical recording and processing, photorefractive polymers are expected to play a major role in these technologies. Bernard Kippelen

*Bibliography.* N. Bloembergen, *Nonlinear Optics*, 4th ed., World Scientific, 1996; R. W. Boyd, *Nonlinear Optics*, Academic Press, 1992; P. Günter and J. P. Huignard (eds.), *Photorefractive Materials and Their Applications II*, Springer-Verlag, Berlin, 1988; B. Kippelen et al., Infrared photorefractive polymers and their applications for imaging, *Science*, 279:54–57, 1998; K. Meerholz et al., A photorefractive polymer with high optical gain and diffraction efficiency near 100%, *Nature*, 371:497–500, 1994; J. V. Moloney and A. C. Newell, *Nonlinear Optics*,

Addison Wesley Longman, 1992; H. S. Nalwa and S. Miyata (eds.), *Nonlinear Optics of Organic Molecules and Polymers*, CRC Press, Boca Raton, 1997; D. D. Nolte (ed.), *Photorefractive Effects and Materials*, Kluwer Academic, Boston, 1995; J. Reintjes, *Nonlinear Optical Parametric Processes in Liquids and Gases*, 1984; J.-I. Saki, *Phase Conjugate Optics*, McGraw-Hill, 1992; P. Yeh, *Introduction to Photorefractive Nonlinear Optics*, Wiley, 1993; J. Zyss (ed.), *Molecular Nonlinear Optics*, Academic Press, San Diego, 1994.

## Nonlinear physics

The study of situations where, in a general sense, cause and effect are not proportional to each other; or more precisely, if the measure of what is considered to be the cause is doubled, the measure of its effect is not simply twice as large. Many examples have been known in physics for a long time, and they seemed well understood. Over the last few decades, however, physicists have noticed that this lack of proportionality in some of the basic laws of physics often leads to unexpected complications, if not to outright contradictions. Thus, the term nonlinear physics refers more narrowly to these developments in the understanding of physical reality.

**Linear versus nonlinear behavior.** The foundations of physics are often expressed in what are called laws, in a misleading analogy to the rules that govern an orderly society. These laws proclaim a mathematical relation between two different kinds of quantities, such as the force on a spring and its length (Hooke's law) or the force on a body and its acceleration (Newton's second law). Each of the two quantities has to be measured in a separate procedure, and their mathematical relation becomes a purely empirical result. Although the two quantities in each of these two laws are proportional to one another, there is an important difference. Newton's second law is correct on a very deep level; that is, the proportionality does not depend on any further assumptions. By contrast, Hooke's law can be verified by experiment only in special circumstances and for special materials. See HOOKE'S LAW; NEWTON'S LAWS OF MOTION; PHYSICAL LAW.

The combination of Hooke's law with Newton's second law leads to the linear (or harmonic) oscillator, a common idealization that pervades modern physics. If a body of fixed mass is suspended by a spring and the body is pulled away from its equilibrium position, the spring and the mass perform simple oscillations. The validity of Hooke's law guarantees that the motion is represented by a pure sine curve and that the period of the oscillation does not depend on its amplitude. Such is not the case for an ordinary pendulum, however: Its motion is not purely sinusoidal, and its period increases with the amplitude of the swing. The force that tends to restore the pendulum to its vertical position is not proportional to the angle with the vertical. This oscillator is, therefore, called nonlinear (or anharmonic).

See ANHARMONIC OSCILLATOR; HARMONIC MOTION; HARMONIC OSCILLATOR; PENDULUM.

The combination of Faraday's law with Ampere's law, as extended by J. C. Maxwell to include so-called electric displacement currents, again gives rise to the phenomenon of a harmonic oscillation inside a cavity. But if the electromagnetic field (or light) is allowed to interact with a collection of excited atoms, a profoundly nonlinear process occurs: The atoms are stimulated to give up their extra energy at a rate that is proportional to the square of the electromagnetic field. The result is a (nonlinear) laser oscillation whose amplitude increases as long as the cavity remains supplied with excited atoms, either from the outside or by the oscillation itself. See CAVITY RESONATOR; ELECTROMAGNETIC FIELD; LASER; MAXWELL'S EQUATIONS.

Newton's law of universal gravitation, the foundation of celestial mechanics, is nonlinear: Two bodies attract one another with a force that is proportional to the product of their masses and inversely proportional to the square of their distance (rather than proportional to the distance). Coulomb's law, the base for all atomic, molecular, and mesoscopic physics, is almost identical except that the product of the masses is replaced by the product of the electric charges. The product of the masses (or charges) in the numerator as well as the square of the distance in the denominator can be traced directly to general relativity in the case of gravitation, and to Maxwell's equations for Coulomb's law. The nonlinearity is basic and inescapable. See COULOMB'S LAW; GRAVITATION; RELATIVITY.

**Linearity in nonlinear systems.** Nevertheless, the gravitational or the electrostatic attraction acting between just two bodies leads to simple results, such as Kepler's laws in celestial mechanics and the Balmer-Bohr spectrum of the hydrogen atom in quantum mechanics. See ATOMIC STRUCTURE AND SPECTRA; CELESTIAL MECHANICS.

When a large number of particles starts out in a condition of stable equilibrium, the result of small external forces is well-coordinated vibrations of the whole collection, for example, the vibrations of a violin string, or of the electric current in an antenna. Each collective motion acts like an independent oscillator, each with its own frequency. In more complicated systems, many vibrational modes can be active simultaneously without mutual interference. A large dynamical system is, therefore, described in terms of its significant degrees of freedom, thought to be only loosely coupled. The motion of any part of the whole becomes multiperiodic; for example, a water molecule has bending and stretching vibrations with different frequencies, and both are coupled with the rotational motion of the whole molecule. See ANTENNA (ELECTROMAGNETISM); DEGREE OF FREEDOM (MECHANICS); MOLECULAR STRUCTURE AND SPECTRA; VIBRATION.

Traditionally, the important degrees of freedom are first identified, their periods are found, and then the extent of their coupling is determined. The mathematical procedures for this type of analysis have

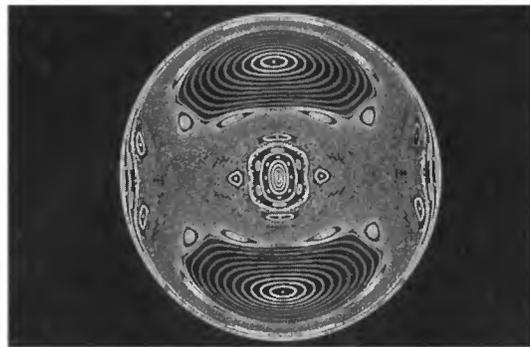
been developed into a fine art, whose most sophisticated and often quite successful products are found particularly in modern statistical mechanics and in high-energy physics. See ELEMENTARY PARTICLE; PERTURBATION (ASTRONOMY); PERTURBATION (MATHEMATICS); STATISTICAL MECHANICS.

**Failure of perturbation theory.** H. Poincaré discovered at the end of the nineteenth century that for many problems, this so-called perturbation theory is not entirely satisfactory. He showed, in the case of the Moon's motion around the Earth, that the disturbance by the Sun is strong enough that this standard mathematical procedure fails. The main culprits are resonances, which occur when the frequencies of different degrees of freedom are combined through their nonlinear coupling. A key nonperturbative phenomenon is known to engineers as phase lock: When different frequencies arise in simple multiples of one another, the whole dynamical system falls into a dynamical trap; and for a continuous range of initial conditions, the interaction changes the frequencies of the individual degrees of freedom sufficiently so as to "lock" the motion into the resonance. See ASTEROID; PHASE-LOCKED LOOPS; RESONANCE (ACOUSTICS AND MECHANICS).

**Structure of motion in phase space.** An example of these effects is provided by a particle of mass  $m$  moving in the  $(x, y)$  plane. Its state of motion at any time  $t$  is uniquely determined by its position  $(x, y)$  and its momentum  $(p_x, p_y)$ ; the four-dimensional space with coordinates  $(x, y, p_x, p_y)$  is called the phase space. If no work is done by outside forces, the value of the total energy, that is, the sum of the kinetic energy  $(p_x^2 + p_y^2)/2m$  and the potential energy  $V(x, y)$ , stays constant at some value  $E$ . The motion at a fixed energy  $E$  is best understood if the values of  $x$  and  $p_x$  are monitored every time the particle crosses the  $x$  axis, that is,  $y = 0$ . The corresponding points are represented in a  $p_x$ -versus- $x$  plot, called a surface of section because the flow of the trajectories in phase space at the energy  $E$  is intersected by the surface  $y = 0$ .

If the motion is multiperiodic, these points form a smooth closed curve, and different motions at the same energy  $E$  yield a sequence of nested curves. But this type of regular structure in phase space is found only for certain special potentials  $V(x, y)$  with a high degree of symmetry. As soon as a small coupling potential  $W(x, y)$  is added to  $V(x, y)$ , many smooth curves dissolve into a disorderly scatter of individual points that cannot be connected in any sensible manner; the corresponding motions are called chaotic. In other parts of the surface of section, an originally smooth curve decays into a sequence of small localized pieces, like a chain of islands (see *illus.*). See CHAOS.

**KAM theorem.** In the 1950s, A. N. Kolmogoroff provided a first account of how the addition of the coupling  $W(x, y)$  generates the chaotic regions and islands. This problem was later worked out in detail by V. Arnold and J. Moser to yield the KAM theorem. Each smooth curve in the surface-of-section without the coupling is characterized by two frequencies



Poincaré surface-of-section, representing motion of an electron in a hydrogen atom in the presence of a strong magnetic field. The section is a slice of four-dimensional phase space representing the electron's position and momentum. Filled regions indicate chaotic behavior. (Image by Dieter Wintgen)

whose ratio varies smoothly from curve to curve. The coupling causes phase lock wherever this ratio is a simple rational number, like  $1/1$ ,  $1/2$ ,  $1/3$ ,  $2/3$ , and so forth; the smooth curve breaks up into islands with chaotic regions in between. But the original curves with an irrational frequency ratio, like  $\sqrt{2}$ ,  $\sqrt{3}$ , and so forth, remain smooth in spite of the coupling  $W$ . The total area in the surface-of-section, outside the remaining smooth curves and the islands, increases continuously with the strength of the coupling; this chaotic area has a complicated fractal structure. See FRACTALS.

**Implications of nonlinearity.** The KAM theorem gives detailed information about the loss of the regular structure as the strength of the coupling increases. It does not say anything, however, about the trajectories in the newly created areas of chaotic behavior. These further investigations are the main goal of such fields as chaos or complexity. The impact of Poincaré's general arguments and the KAM theorem reaches into every area of nonlinear physics. The oldest among them is hydrodynamics, where the phenomenon of turbulent flow has so far resisted any effective control. This is what makes weather prediction so difficult. Signal propagation along the nerves and transmission of pulses through synaptic connections are other well-known nonlinear processes. See NEUROBIOLOGY; TURBULENT FLOW.

The two pillars of modern physics, electromagnetism and quantum mechanics, are described, each separately, by linear equations. They become effectively nonlinear, however, in the practical solutions concerned with large atoms or molecules. Moreover, when the two fields are combined into quantum electrodynamics (QED), the coupling between them is nonlinear. Fortunately, this coupling is weak enough for perturbation theory to work very well, and QED has become the inspiration for all the theoretical models in nuclear and high-energy physics. Unfortunately, the coupling in these mathematical models is very strong, so that perturbation theory is useless. At the deepest level of physical reality, there is no escape from nonlinear physics. The manifestations of the nonlinear features in the most basic problems

have barely begun to be accounted for. See NONLINEAR ACOUSTICS; NONLINEAR OPTICS; QUANTUM CHROMODYNAMICS; QUANTUM ELECTRODYNAMICS; QUANTUM FIELD THEORY; QUANTUM MECHANICS.

Martin C. Gutzwiller

Bibliography. F. Gross, *Relativistic Quantum Mechanics and Field Theory*, 1993; R. C. Hilborn, *Chaos and Nonlinear Dynamics*, 2d ed., 2001; L. Lam (ed.), *Nonlinear Physics for Beginners*, 1998; R. Z. Sagdeev, D. A. Usikov, and G. M. Zaslavsky, *Nonlinear Physics from the Pendulum to Turbulence and Chaos*, 1988; S. H. Strogatz, *Nonlinear Dynamics and Chaos*, 1994; M. Tabor, *Chaos and Integrability in Nonlinear Dynamics*, 1989.

## Nonlinear programming

The area of applied mathematics and operations research concerned with finding the largest or smallest value of a function subject to constraints or restrictions on the variables of the function. Nonlinear programming is sometimes referred to as nonlinear optimization.

A useful example concerns a power plant that uses the water from a reservoir to cool the plant. The heated water is then piped into a lake. For efficiency, the plant should be run at the highest possible temperature consistent with safety considerations, but there are also limits on the amount of water that can be pumped through the plant, and there are ecological constraints on how much the lake temperature can be raised. The optimization problem is to maximize the temperature of the plant subject to the safety constraints, the limit on the rate at which water can be pumped into the plant, and the bound on the increase in lake temperature.

The nonlinear programming problem refers specifically to the situation in which the function to be minimized or maximized, called the objective function, and the functions that describe the constraints are nonlinear functions. Typically, the variables are continuous; this article is restricted to this case.

The general topic of mathematical programming includes many important special cases, each with its own methods and techniques. If the objective function and the constraint functions are linear, the problem is a linear programming one; if the objective function is quadratic and the constraints are linear, the problem is a quadratic programming one. If there are no constraints, the problem is unconstrained. If the variables are restricted to a set of discrete or integer values, the problem is referred to as an integer programming problem. See LINEAR PROGRAMMING.

Researchers in nonlinear programming consider both the theoretical and practical aspects of these problems. Theoretical issues include the study of algebraic and geometric conditions that characterize a solution, as well as general notions of convexity that determine the existence and uniqueness of solutions. Among the practical questions that are addressed are the mathematical formulation of a specific problem and the development and analysis of algorithms for finding the solution of such problems.

**General theory.** The general nonlinear programming problem can be stated as that of minimizing a scalar-valued objective function  $f(\mathbf{x})$  over all vectors  $\mathbf{x}$  satisfying a set of constraints. The constraints are in the form of general nonlinear equations and inequalities. Mathematically, the nonlinear programming problem may be expressed as notation (1),

$$\begin{aligned} &\text{minimize } f(\mathbf{x}) \text{ with respect to } \mathbf{x} \\ &\text{subject to: } g_i(\mathbf{x}) \leq 0, \quad i = 1, 2, \dots, m \quad (1) \\ &\quad \quad \quad b_j(\mathbf{x}) = 0, \quad j = 1, 2, \dots, p \end{aligned}$$

where  $\mathbf{x} = (x_1, x_2, \dots, x_n)$  are the variables of the problem,  $f$  is the objective function,  $g_i(\mathbf{x})$  are the inequality constraints, and  $b_j(\mathbf{x})$  are the equality constraints. This formulation is general in that the problem of maximizing  $f(\mathbf{x})$  is equivalent to minimizing  $-f(\mathbf{x})$  and a constraint  $g_i(\mathbf{x}) \geq 0$  is equivalent to the constraint  $-g_i(\mathbf{x}) \leq 0$ .

Many modern approaches for solving (1) rely on exploiting the equations that characterize a solution. The most familiar characterization is for the case when the functions  $f$ ,  $g_i$ , and  $b_j$  are all differentiable. This case can be described more easily by considering the problem in which there are only equality constraints, that is, when  $m = 0$ . The following notation is employed. If  $s(\mathbf{x})$  is a scalar function of  $\mathbf{x}$ , then  $\nabla s(\mathbf{x})$  denotes the gradient vector of  $s(\mathbf{x})$ , that is, the vector of partial derivatives of  $s$  with respect to the components of  $\mathbf{x}$ . Thus  $(\nabla s(\mathbf{x}))_i = \partial s(\mathbf{x})/\partial x_i$ . Likewise, if  $\mathbf{h}(\mathbf{x})$  denotes the vector function whose  $i$ th component is  $b_i(\mathbf{x})$ , then  $\nabla \mathbf{h}(\mathbf{x})$  denotes the  $(n \times p)$  matrix whose  $i$ th column is  $\nabla b_i(\mathbf{x})$ . For any two vectors  $\mathbf{a}$  and  $\mathbf{b}$ ,  $\mathbf{a}^t$  denotes the transpose of the vector  $\mathbf{a}$  and  $\mathbf{a}^t \mathbf{b}$  denotes the standard dot product (or inner product). See CALCULUS OF VECTORS.

The characterization is provided in terms of the Lagrangian function defined by Eq. (2),

$$L(\mathbf{x}, \boldsymbol{\lambda}) = f(\mathbf{x}) + \sum_{i=1}^p \lambda_i b_i(\mathbf{x}) = f(\mathbf{x}) + \mathbf{h}(\mathbf{x})^t \boldsymbol{\lambda} \quad (2)$$

where  $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_p)$  is the vector of Lagrange multipliers. It follows that Eq. (3)

$$\nabla L(\mathbf{x}^{(*)}, \boldsymbol{\lambda}) = \nabla f(\mathbf{x}^{(*)}) + \nabla \mathbf{h}(\mathbf{x}^{(*)}) \boldsymbol{\lambda} \quad (3)$$

is satisfied. If the vector  $\mathbf{x}^{(*)}$  solves the nonlinear programming problem (1) when  $m = 0$ , then there exist values of the Lagrange multipliers such that Eqs. (4) are satisfied.

$$\begin{aligned} \nabla L(\mathbf{x}^{(*)}, \boldsymbol{\lambda}) &= 0 \\ b_j(\mathbf{x}^{(*)}) &= 0, \quad j = 1, \dots, p \end{aligned} \quad (4)$$

Other conditions are needed to complete the theory, but Eq. (4) provides the basis for an important class of methods of solving the nonlinear programming problem. (This class is described below.) In the unconstrained case, Eq. (4) becomes the condition that if  $\mathbf{x}^{(*)}$  is a solution then  $\nabla f(\mathbf{x}^{(*)}) = 0$ . There may be no points satisfying all of the constraints, in which case the problem is said to be inconsistent.

The Lagrange multipliers  $\lambda$ , also called dual variables, have an important interpretation in the nonlinear programming problem: They give the sensitivity of the objective function value to the constraints.

**Computational methods.** Since general nonlinear equations cannot be solved in closed form, iterative methods must be used. Such methods generate a sequence of approximations, or iterates, that will converge to a solution under specified conditions. Newton's method is one of the best-known methods and is the basis for many of the fastest methods for solving the nonlinear programming problem.

If  $F(\mathbf{x}) = 0$  is one equation in one unknown, Newton's method for solving this problem produces the next iterate,  $\mathbf{x}^{(k+1)}$ , from the current iterate,  $\mathbf{x}^{(k)}$ , by formula (5),

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \frac{F(\mathbf{x}^{(k)})}{F'(\mathbf{x}^{(k)})} \quad (5)$$

where  $F'(\mathbf{x}^{(k)})$  is the derivative of  $F$  with respect to  $\mathbf{x}$  evaluated at  $\mathbf{x}^{(k)}$ . If  $\mathbf{F}(\mathbf{x}) = 0$  is now  $n$  equations in  $n$  unknowns, then  $F'(\mathbf{x}^{(k)})$  becomes  $\nabla\mathbf{F}(\mathbf{x}^{(k)})$ , the  $(n \times n)$  matrix of partial derivatives defined by Eq. (6),

$$\nabla\mathbf{F}(\mathbf{x})_{ij} = \frac{\partial F_i(\mathbf{x})}{\partial x_j} \quad (6)$$

evaluated at  $\mathbf{x}^{(k)}$ . That is, the  $ij$  component of  $\nabla\mathbf{F}(\mathbf{x})$  is the partial derivative of the  $i$ th component of  $\mathbf{F}$  with respect to  $j$ th element of  $\mathbf{x}$ . The analog to the division by  $F'(\mathbf{x}^{(k)})$  in Eq. (5) is the multiplication by the inverse of  $\nabla\mathbf{F}(\mathbf{x}^{(k)})$  [if its determinant is not zero], so that Eq. (5) becomes Eq. (7).

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - [\nabla\mathbf{F}(\mathbf{x}^{(k)})]^{-1} \mathbf{F}(\mathbf{x}^{(k)}) \quad (7)$$

If the initial approximation,  $\mathbf{x}^{(0)}$ , is sufficiently good and other reasonable conditions hold, then Newton's method converges rapidly to a solution of the system of equations. Modifications need to be made, however, to create an efficient and robust method that can be used as a general tool for solving the nonlinear programming problem. These modifications involve the development of tests that assess  $\mathbf{x}^{(k+1)}$  to determine if it is an improvement over  $\mathbf{x}^{(k)}$ , and procedures to adjust the steps to ensure that improvement occurs. Since the computation of  $\nabla\mathbf{F}(\mathbf{x}^{(k)})$  is often the most expensive part of the calculation, other modifications seek to approximate this matrix by computationally inexpensive alternatives. Research continues on these and other ideas, especially as they relate to large-scale problems.

To solve the nonlinear programming problem when there are no inequality constraints, the nonlinear system of equations (4) can be attacked by Newton's method as just described. The nonlinear system is in terms of the variables  $\mathbf{x}$  and the dual variables  $\lambda$ , and is therefore  $(n + p)$  equations in  $(n + p)$  unknowns, namely, Eqs. (8).

$$\mathbf{F}(\mathbf{x}, \lambda) = \begin{cases} \nabla L(\mathbf{x}, \lambda) = 0 \\ \mathbf{h}(\mathbf{x}) = 0 \end{cases} \quad (8)$$

The formulas involved in the computation of the matrix  $\nabla\mathbf{F}(\mathbf{x}^{(k)}, \lambda^{(k)})$  are straightforward and require

the evaluation of the matrix of second partial derivatives of  $L(\mathbf{x}, \lambda)$  with respect to  $\mathbf{x}$ . This matrix is denoted by  $\nabla^2 L(\mathbf{x}, \lambda)$ . Thus, Eq. (9) holds,

$$\nabla\mathbf{F}(\mathbf{x}^{(k)}, \lambda^{(k)}) = \begin{bmatrix} \nabla^2 L(\mathbf{x}^{(k)}, \lambda^{(k)}) & \nabla\mathbf{h}(\mathbf{x}^{(k)}) \\ \nabla\mathbf{h}(\mathbf{x}^{(k)})^t & 0 \end{bmatrix} \quad (9)$$

and the Newton iteration for Eq. (8) is then given by Eq. (10).

$$\begin{bmatrix} \mathbf{x}^{(k+1)} \\ \lambda^{(k+1)} \end{bmatrix} = \begin{bmatrix} \mathbf{x}^{(k)} \\ \lambda^{(k)} \end{bmatrix} - [\nabla\mathbf{F}(\mathbf{x}^{(k)}, \lambda^{(k)})]^{-1} \begin{bmatrix} \nabla L(\mathbf{x}^{(k)}, \lambda^{(k)}) \\ \mathbf{h}(\mathbf{x}^{(k)}) \end{bmatrix} \quad (10)$$

A means of handling inequality constraints completes this description. First, it can be shown that Eq. (10) is equivalent to solving the quadratic programming problem, expressed by notation (11),

$$\begin{aligned} &\text{minimize } \nabla f(\mathbf{x}^{(k)})^t \delta + \frac{1}{2} \delta^t \nabla^2 L(\mathbf{x}^{(k)}, \lambda^{(k)}) \delta \\ &\quad \text{with respect to } \delta \quad (11) \\ &\text{subject to: } \nabla\mathbf{h}(\mathbf{x}^{(k)})^t \delta + \mathbf{h}(\mathbf{x}^{(k)}) = 0 \end{aligned}$$

for  $\delta$  and setting  $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \delta$  and  $\lambda^{(k+1)}$  to the Lagrange multiplier associated with problem (11). Formulated in this way, the method is known as sequential quadratic programming (SQP). At each step of an SQP algorithm, one formulates and solves a quadratic programming approximation to the nonlinear programming problem and uses its solution to construct the next iterate. To incorporate inequality constraints into this algorithm, the constraints of the quadratic programming problem are augmented with Eqs. (12).

$$\nabla\mathbf{g}(\mathbf{x}^{(k)})^t \delta + \mathbf{g}(\mathbf{x}^{(k)}) \leq 0 \quad (12)$$

Solving inequality-constrained quadratic programs is harder than solving quadratic programs with equality constraints only, but effective methods exist.

Sequential quadratic programming continues to be one of the most powerful and successful general approaches for solving the nonlinear programming problem with both equality and inequality constraints. Algorithms based on sequential quadratic programming have been used to solve a wide variety of science, engineering, and management problems. Research is continuing on sequential quadratic programming and related methods with an emphasis on large-scale problems. Research is also continuing on methods for problems in which the functions are not differentiable. See OPERATIONS RESEARCH; OPTIMIZATION. Paul T. Boggs

**Bibliography.** S. G. Nash and A. Sofer, *Linear and Nonlinear Programming*, 1995; J. Nocedal and S. J. Wright, *Numerical Optimization*, 1999.

## Nonmetal

The elements are conveniently, but arbitrarily, divided into metals and nonmetals. The nonmetals do not conduct electricity readily, are not ductile, do

not have a complex refractive index, and in general have high ionization potentials. The nonmetals vary widely in physical properties. Hydrogen is a colorless permanent gas; bromine is a dark-red, volatile liquid; and carbon, as diamond, is a solid of great hardness and high refractive index. If the periodic table is divided diagonally from upper left to lower right, all the nonmetals are on the right-hand side of the diagonal. Examples of elements which do not fit neatly into this useful but arbitrary classification are tin, which exists in two allotropic modifications, one definitely metallic and the other with many properties of a nonmetal, and tellurium and antimony. Such elements are called metalloids. See IONIZATION POTENTIAL; METAL; METALLOID; PERIODIC TABLE. Thomas C. Waddington

## Non-newtonian fluid

A fluid that departs from the classic linear newtonian relation between stress and shear rate. In a strict sense, a fluid is any state of matter that is not a solid, and a solid is a state of matter that has a unique stress-free state. A conceptually simpler definition is that a fluid is capable of attaining the shape of its container and retaining that shape for all time in the absence of external forces. Therefore, fluids encompass a wide variety of states of matter including gases and liquids as well as many more esoteric states (for example, plasmas, liquid crystals, and foams). See FLUIDS; FOAM; GAS; LIQUID; LIQUID CRYSTALS; PLASMA (PHYSICS).

A newtonian fluid is one whose mechanical behavior is characterized by a single function of temperature, the viscosity, a measure of the "slipperiness" of the fluid. For the example of Fig. 1, where a fluid is sheared between a fixed plate and a moving plate, the viscosity is given by Eq. (1). Thus, as the viscosity

$$\text{Viscosity} = \frac{\text{force/area}}{\text{velocity/height}} \quad (1)$$

of a fluid increases, it requires a larger force to move the top plate at a given velocity. For simple, newtonian fluids, the viscosity is a constant dependent on only temperature; but for non-newtonian fluids, the viscosity can change by many orders of magnitude as the shear rate (velocity/height in Fig. 1) changes. See NEWTONIAN FLUID; VISCOSITY.

Many of the fluids encountered in everyday life (such as water, air, gasoline, and honey) are ade-

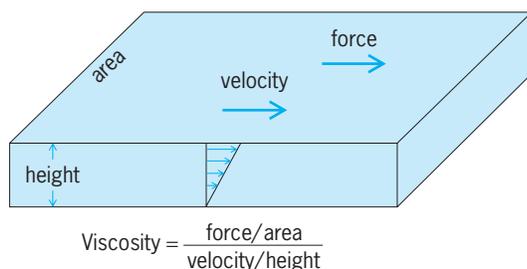


Fig. 1. Steady shear flow of a fluid between a fixed plate and a parallel moving plate, illustrating the concept of viscosity.

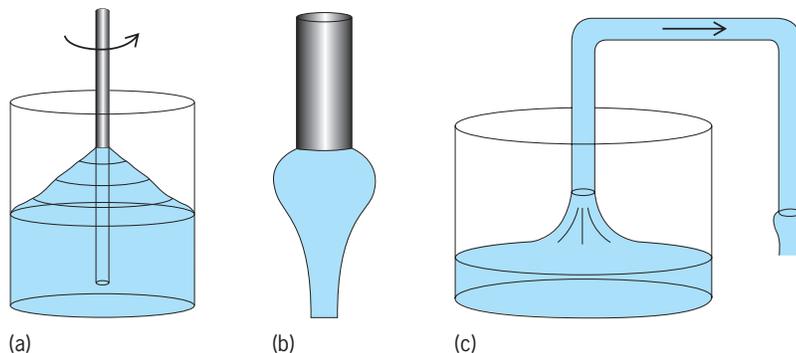


Fig. 2. Examples of the counterintuitive behavior display by non-newtonian fluids. (a) Weissenberg effect. (b) Die swell. (c) Tubeless siphon.

quately described as being newtonian, but there are even more that are not. Common examples include mayonnaise, peanut butter, toothpaste, egg whites, liquid soaps, and multigrade engine oils. Other examples such as molten polymers and slurries are of considerable technological importance. A distinguishing feature of many non-newtonian fluids is that they have microscopic or molecular-level structures that can be rearranged substantially in flow. See PARTICLE FLOW; POLYMER.

**Counterintuitive behavior.** Our intuitive understanding of how fluids behave and flow is built primarily from observations and experiences with newtonian fluids. However, non-newtonian fluids display a rich variety of behavior that is often in dramatic contrast to these expectations. For example, an intuitive feel for the slipperiness of fluids can be gained from rubbing them between the fingers. Furthermore, the slipperiness of water, experienced in this way, is expected to be the same as the slipperiness of automobile tires on a wet road. However, the slipperiness (viscosity) of many non-newtonian fluids changes a great deal depending on how fast they move or the forces applied to them.

Intuitive expectations for how the surface of a fluid will deform when the fluid is stirred (with the fluid bunching up at the wall of the container) are also in marked contrast to the behavior of non-newtonian fluids. When a cylindrical rod is rotated inside a container of a newtonian fluid, centrifugal forces cause the fluid to be higher at the wall. However, for non-newtonian fluids, the normal stress differences cause the fluid to climb the rod; this is called the Weissenberg effect (Fig. 2a). Intuitive understanding about the motion of material when the flow of a fluid is suddenly stopped, for example by turning off a water tap, is also notably at odds with the behavior of non-newtonian fluids. See CENTRIFUGAL FORCE.

A non-newtonian fluid also displays counterintuitive behavior when it is extruded from an opening. A newtonian fluid tapers to a smaller cross section as it leaves the opening, but the cross section for a non-newtonian fluid first increases before it eventually tapers. This phenomenon is called die swell (Fig. 2b). See NOZZLE.

When a newtonian fluid is siphoned and the fluid level goes below the entrance to the siphon tube, the

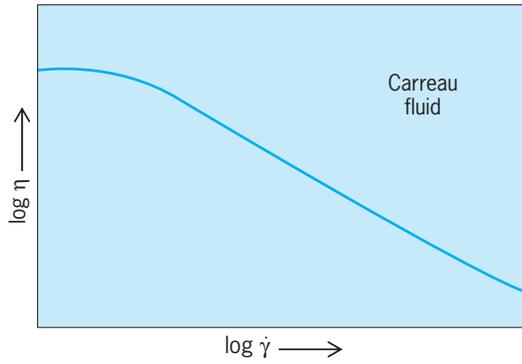


Fig. 3. Typical dependence of the viscosity ( $\eta$ ) on shear rate ( $\dot{\gamma}$ ) for a non-newtonian fluid. The functional form shown is called the Carreau model.

siphoning action stops. For many non-newtonian fluids, however, the siphoning action continues as the fluid climbs from the surface and continues to enter the tube. This phenomenon is called the tubeless siphon (Fig. 2c).

**Non-newtonian viscosity.** For newtonian fluids, the viscosity is independent of the speed at which the fluid is caused to move. However, for non-newtonian fluids, the viscosity can change by many orders of magnitude as the speed (velocity gradient) changes. Typically, the viscosity ( $\eta$ ) of these fluids is given as a function of the shear rate ( $\dot{\gamma}$ ), which is a generalization of the ratio velocity/height in Fig. 1. A common dependence for this function is given in Fig. 3. In this particular model (called the Carreau model), the viscosity as a function of shear rate is given by Eq. (2).

$$\eta = \eta_{\infty} + (\eta_0 - \eta_{\infty}) [1 + (\lambda \dot{\gamma})^2]^{(n-1)/2} \quad (2)$$

This model involves four parameters:  $\eta_0$ , the viscosity at very small shear rates;  $\eta_{\infty}$ , the viscosity at very large shear rates;  $\lambda$ , which describes the shear rate at which the curve begins to decrease; and  $n$ , which determines the rate of viscosity decrease. Figure 3 shows the viscosity for a fluid that shear-thins (that is, the viscosity decreases as the shear rate increases,  $n$  is less than one, and  $\eta_0$  is greater than  $\eta_{\infty}$ ). For other non-newtonian fluids, the viscosity might increase as the shear rate increases (shear-thickening fluids, for which  $n$  is greater than one and  $\eta_0$  is less than  $\eta_{\infty}$ ).

**Nonlinear effects.** Although shear-rate-dependent viscosity is the most important non-newtonian effect for many engineering applications of these fluids, it is not by itself capable of describing any of the phenomena depicted in Fig. 2. Both the Weissenberg effect (rod climbing) and die swell are generally attributed to nonlinear effects, by which is meant stresses that are present in non-newtonian fluids and not present in newtonian fluids. The stress is a generalization of the concept of pressure; at any point in the fluid there are nine components of the stress,  $\tau_{ij}$ , where  $i$  and  $j$  take on the three coordinate directions. Then  $\tau_{ij}$  is the force per unit area in the  $j$  direction on a surface of constant  $i$  (for example,  $\tau_{xy}$  is the force per unit area in the  $y$  direction on a surface of constant

$x$ ). Most significant among the anomalous stresses in non-newtonian fluids is the first normal stress difference, which causes a tension in the direction of fluid flow. (There is also a second normal stress difference, which tends to be smaller and less significant than the first.) These stresses are generally caused by the flow trying to deform the structure of the material and the microstructure of the material resisting this deformation. The most common and pronounced result in a shear flow is a tension in the direction of the flow trying to resist flow. The Weissenberg effect is easily understood in terms of the first normal stress difference. Here the flow direction is circular around the rod. The tension pulls the fluid inward toward the rod, and the fluid must rise because it cannot go downward through the bottom of the container. See STRESS AND STRAIN.

**Viscoelasticity.** Perhaps the most striking behavior of non-newtonian fluids is a consequence of their viscoelasticity. Solids can be thought of as having perfect memory. If they are deformed through the action of a force, they return to their original shape when the force is removed. This happens when a rubber ball bounces; the ball is deformed as it hits a surface, but the rubber remembers its undeformed spherical shape. Recovery of the shape causes the ball to bounce back. In contrast, newtonian fluids have no memory; when a force is removed, they retain their condition at the time the force is removed (or continue moving as the result of inertia). When a newtonian fluid is dropped onto a surface, it does not bounce. Non-newtonian fluids are viscoelastic in the sense that they have fading memory. If a force is removed shortly after it is applied, the fluid will remember its undeformed shape and return toward it. However, if the force is applied on the fluid for a long time, the fluid will eventually forget its undeformed shape. If a sample of a non-newtonian fluid is dropped onto a surface, it will bounce like a ball. However, if the fluid is simply placed on the surface, it will flow smoothly. Viscoelasticity is frequently the cause of many of the secondary flows that are observed for non-newtonian fluids. These are fluid motions that are small for newtonian fluids (for example, swirling motions) but can become dominant for non-newtonian fluids. See ELASTICITY.

**Constitutive relations.** Analysis of fluid flow operations is typically performed by examining local conservation relations—conservation of mass, momentum (Newton's second law), and energy. This analysis requires material-specific information (for example, the relation between density, pressure, and temperature) that is collectively known as constitutive relations. The science devoted to obtaining suitable constitutive equations for description of the behavior of non-newtonian fluids is called rheology. The most important constitutive equation for fluid mechanics is that relating the stress in the fluid to the kinematics of the motion (that is, the velocity, the derivatives of the velocity with respect to position, and the time history of the velocity).

Two general methods exist for formulation of appropriate constitutive equations: (1) empiricism

combined with flow measurements, and (2) molecular or structural theories. A wide variety of standard experiments have been designed to probe the non-newtonian behavior of fluids and to test the predictions of constitutive equations. Examples include steady shear flow as in Fig. 1, with the upper plate moving at constant velocity; small-amplitude oscillatory shear flow, where the velocity of the upper plate depicted in Fig. 1 is caused to oscillate sinusoidally in time; and the sudden inception of shear flow, where the upper plate in Fig. 1 instantaneously begins moving at constant velocity. In the latter two experiments, the stress (force) is recorded as a function of time. These experiments provide useful probes of the viscoelasticity of the fluid. Another category of flow experiment involves elongation, wherein the fluid is caused to move so that fluid elements are stretched. In addition, other experiments have been used where the fluid is caused to move in more complicated ways. All of these experiments are used to define material functions that characterize the response of the fluid (for example, the viscosity of Fig. 3).

*Generalized newtonian fluid.* The simplest constitutive equation (obtained empirically from examination of experimental viscosity data) for a non-newtonian fluid is that for the generalized newtonian fluid, given by Eq. (3), where  $\dot{\gamma}_{ij}$ , given by Eq. (4), is called the

$$\tau_{ij} = -\eta(\dot{\gamma})\dot{\gamma}_{ij} \quad (3)$$

$$\dot{\gamma}_{ij} = \frac{\partial v_j}{\partial x_i} + \frac{\partial v_i}{\partial x_j} \quad (4)$$

rate-of-strain tensor, and  $\eta(\dot{\gamma})$  is the viscosity as a function of shear rate. Here,  $v_i$  is the  $i$ th component of the velocity vector, and the  $x_i$  are the position coordinates. For the Carreau fluid,  $\eta(\dot{\gamma})$  is given by Eq. (2).

The generalized newtonian fluid constitutive equation is not capable of describing viscoelasticity or any effects of normal stress differences. The simplest constitutive equation that is capable of describing viscoelasticity is the Maxwell model (named after James C. Maxwell). In the Maxwell model, Eq. (5)

$$\tau_{ij} = -G \int_{-\infty}^t e^{-(t-t')/\lambda} \dot{\gamma}_{ij}(t') dt' \quad (5)$$

expresses the stresses as time integrals over the history of the velocity derivatives, where  $G$  (the shear modulus) and  $\lambda$  (the relaxation time) are material-dependent constants, and the time integral extends over all past time. It is straightforward to see how the Maxwell model incorporates the memory of the fluid; the stresses at the current time depend on the velocity derivatives at past times, and the memory fades because of the exponential that becomes smaller at more distant past times. For smaller values of  $\lambda$  the memory fades more rapidly, and for larger values of  $\lambda$  the fluid behaves more elastically (remembers better). The Maxwell model has been modified to describe fluid elasticity more accurately by writ-

ing the stress as a sum of integrals each with different values for  $G$  and  $\lambda$ , but the model still suffers from the serious deficit in predicting that the viscosity is constant (that is, independent of shear rate).

*Requirements and generalizations.* Constitutive equations, whether formulated empirically or as the result of a structural theory, are generally required to obey several fundamental postulates, including frame invariance and objectivity. Frame invariance specifies that the stress in a material should not depend on the measuring system (for example, the coordinate system) that is used to quantify it. From a practical point of view, this means that the mathematical language of tensor calculus must be used for the formulation. Objectivity is a more confusing concept, but it means loosely that the stress in a material must be independent of any rigid-body motion of the material. *See TENSOR ANALYSIS.*

Unfortunately, the Maxwell model as formulated above does not satisfy the objectivity postulate. However, it has been modified in a variety of ways by using more complicated descriptions of the fluid motion to make it objective. It has also been modified in other ways to incorporate shear-rate-dependent viscosity and normal stress effects. In addition, a large class of constitutive equations has been developed that expresses the stress as the solution of a differential equation (as opposed to a time integral), and these are often easier to use in analyzing specific flows.

Although the non-newtonian behavior of many fluids has been recognized for a long time, the science of rheology is, in many respects, still in its infancy, and new phenomena are constantly being discovered and new theories proposed. Advancements in computational techniques are making possible much more detailed analyses of complex flows and more sophisticated simulations of the structural and molecular behavior that gives rise to non-newtonian behavior. Engineers, chemists, physicists, and mathematicians are actively pursuing research in rheology, particularly as more technologically important materials are found to display non-newtonian behavior. *See FLUID FLOW; FLUID-FLOW PRINCIPLES; RHEOLOGY.* John M. Wiest

## Nonrelativistic quantum theory

The modern theory of matter and its interaction with electromagnetic radiation, applicable to systems of material particles which move slowly compared to the velocity of light and which are neither created nor destroyed.

This article details the formal structure of quantum theory, which can be summarized in the form of postulates as unequivocal as are Newton's laws of classical mechanics; a less logically rigorous presentation is adopted here, however. Even so, the reader unfamiliar with quantum theory is advised to read first another article (*see* QUANTUM MECHANICS) which more qualitatively discusses the novel (from the standpoint of classical physics) features of nonrelativistic quantum theory.

That article and this one attempt to make convincing the thesis that the formalism of nonrelativistic quantum theory is an unarbitrary and physically reasonable extension of the older classical theories. Belief in quantum theory stems as much from acceptance of this thesis as from the broad range, barely hinted at in these articles, of successful application of the theory. For added details concerning special formal topics see MATRIX MECHANICS; PERTURBATION (QUANTUM MECHANICS).

For generalizations of nonrelativistic quantum theory to relativistic particles (particles with speeds close to the velocity of light), or to systems in which particle creation and destruction can occur, see QUANTUM ELECTRODYNAMICS; QUANTUM FIELD THEORY; RELATIVISTIC QUANTUM THEORY.

### Wave Function and Probability Density

Basic to quantum mechanics is the belief that the wave properties of matter are associated with a function, the wave function, obeying an equation called a wave equation. The simplest possible wave in three-dimensional space is a so-called scalar wave, in which the wave disturbance is wholly characterized by a single function  $\psi(x,y,z,t) \equiv \psi(\mathbf{r},t)$ . It is natural, therefore, to postulate that a wave function  $\psi(x,y,z,t)$  provides a complete description of the simplest possible physical system, namely, a single particle moving in a force field specified by a potential  $V(\mathbf{r})$ . It is further postulated that  $|\psi(\mathbf{r},t)|^2$ , which classically would be proportional to the wave intensity, is the probability density; that is,  $|\psi(\mathbf{r},t)|^2 d\mathbf{r}$  is the probability, at time  $t$ , of finding the particle in the volume  $dxdydz \equiv d\mathbf{r}$  of space lying between  $x$  and  $x + dx$ ,  $y$  and  $y + dy$ ,  $z$  and  $z + dz$ .

There is the obvious generalization that a wave function  $\psi(\mathbf{r}_1, \dots, \mathbf{r}_g, t)$  will completely describe a system of  $g$  particles, with  $|\psi(\mathbf{r}_1, \dots, \mathbf{r}_g, t)|^2 d\mathbf{r}_1 \dots d\mathbf{r}_g$  the probability at time  $t$  of simultaneously finding particle 1 in the volume element  $d\mathbf{r}_1 = dx_1 dy_1 dz_1, \dots$ , particle  $g$  in  $d\mathbf{r}_g$ . Moreover, for a system of  $g$  distinguishable particles, the probability  $P_j(\mathbf{r}) d\mathbf{r}$  of finding particle  $j$  in the volume element  $d\mathbf{r}$  at  $\mathbf{r} \equiv x,y,z$  of physical space is given by Eq. (1), where

$$P_j(\mathbf{r}) d\mathbf{r} = d\mathbf{r} \int \{d\mathbf{r}_1 \dots d\mathbf{r}_{j-1}, d\mathbf{r}_{j+1} \dots d\mathbf{r}_g \times |\psi(\mathbf{r}_1, \dots, \mathbf{r}_{j-1}, \mathbf{r}, \mathbf{r}_{j+1}, \dots, \mathbf{r}_g, t)|^2\} \quad (1)$$

$|\psi(\mathbf{r}^1, \dots, \mathbf{r}_g, t)|^2$  is integrated over all positions of particles 1 to  $j - 1$  and  $j + 1$  to  $g$ , with  $\mathbf{r}_j$  put equal to  $\mathbf{r}$ .

**Normalization.** Because each of the particles  $1, \dots, g$  must be somewhere in physical space, Eq. (1) demands that Eq. (2a) be integrated over all positions of all  $g$  particles. When Eq. (2a) is satisfied,  $\psi$  is said

$$\int d\mathbf{r} P_j(\mathbf{r}) = \int d\mathbf{r}_1 \dots d\mathbf{r}_g |\psi(\mathbf{r}_1, \dots, \mathbf{r}_g, t)|^2 = 1 \quad (2a)$$

$$\int d\mathbf{r}_1 \dots d\mathbf{r}_g |\psi'|^2 = C \neq 1 \quad (2b)$$

to be normalized. If  $\psi'(\mathbf{r}_1, \dots, \mathbf{r}_g, t)$  is a proposed wave function which satisfies Eq. (2b), the probabilities specified by  $\psi'$  are found from Eq. (1) using the normalized  $\psi = C^{-1/2} \psi' \exp(i\eta)$ , provided  $C$  is not infinite, that is, provided  $\psi'$  is quadratically integrable; the phase factor  $\exp(i\eta)$ ,  $\eta$  being real, can be chosen arbitrarily. Though the absolute probabilities  $P_j(\mathbf{r})$  are not defined when  $\psi'(\mathbf{r}_1, \dots, \mathbf{r}_g)$  is not quadratically integrable,  $|\psi'|^2$  may remain a useful measure of the relative probability of finding particles  $1, \dots, g$  at  $\mathbf{r}_1, \dots, \mathbf{r}_g$ .

One need be concerned only with wave functions which can represent actual physical systems. It is postulated that an admissible (physically possible) wave function  $\psi'(\mathbf{r}_1, \dots, \mathbf{r}_g)$  is quadratically integrable (type 1), or fails to be quadratically integrable (type 2) only because  $\psi'$  vanishes too slowly or at worst remains finite as infinity is approached in the  $3g$ -dimensional space of  $\mathbf{r}_1, \dots, \mathbf{r}_g$ . Convincing physical and mathematical reasons can be found for excluding wave functions other than these types. One-particle wave functions  $\psi'(\mathbf{r})$  of type 2 correspond to classically unconfined systems, for example, an electron ionized from a hydrogen atom; for these systems  $\psi(\mathbf{r}) = \infty^{-1/2} \psi' = 0$  has the obvious interpretation that an unconfined particle is sure to be found outside any given finite volume. Such a  $\psi'(\mathbf{r})$  also can represent a very large (effectively infinite) number of independently moving identical particles in a very large (effectively infinite) volume, for example, a beam of free electrons issuing from an electron gun; in this event  $|\psi'(\mathbf{r})|^2$ , although it continues to describe the likelihood of observing an electron, can be thought to equal the number density of electrons at any point  $\mathbf{r}$ , with  $C = \infty$  in Eq. (2b) indicating that the number of particles in all of space is infinite. These considerations can be extended to quadratically nonintegrable many-particle wave functions  $\psi'(\mathbf{r}_1, \dots, \mathbf{r}_g)$ .

**Spin.** The preceding formalism accepts the presumption of classical physics that a particle is a structureless entity, about which "everything" is known when its position  $\mathbf{r}$  is known. This presumption is inaccurate (and therefore the formalism is not wholly adequate) for systems of electrons, protons, and neutrons, these being the fundamental particles which compose what usually is termed stable matter. In particular an electron, proton, or neutron cannot be described completely by a single wave function  $\psi(\mathbf{r}, t)$ . For each of these particles two wave functions  $\psi_1(\mathbf{r}, t)$  and  $\psi_2(\mathbf{r}, t)$  are required, which may be regarded as components of an overall two-component wave function  $\psi(\mathbf{r}, t)$ . The need for a multicomponent wave function has the immediate classical interpretation that the particle has internal degrees of freedom, that is, that knowledge of the position of the particle is not "everything." It can be shown that these internal degrees of freedom are associated with so-called intrinsic angular momentum or spin. See SPIN (QUANTUM MECHANICS).

For electrons, protons, or neutrons the spin is  $1/2\hbar$ , where  $\hbar = h/2\pi$ , and  $h$  is Planck's constant; the only allowed values of  $s_z$ , the  $z$  component of the spin, are

$\pm 1/2$  (in units of  $\hbar$ ). Thus when the system contains a single particle of spin  $1/2$ , an electron, say,  $|\psi_1(\mathbf{r}, t)|^2 d\mathbf{r}$  can be interpreted as the probability of finding, in the volume  $d\mathbf{r}$ , an electron with  $s_z = +1/2$ ;  $|\psi_2(\mathbf{r}, t)|^2$  is the probability density for finding an electron with  $s_z = -1/2$ . The normalization condition replacing Eq. (2a) is given by Eq. (3).

$$\int d\mathbf{r} [|\psi_1(\mathbf{r}, t)|^2 + |\psi_2(\mathbf{r}, t)|^2] = 1 \quad (3)$$

This formalism is readily extended to many-particle systems. For example, when the system contains  $g$  particles of spin  $1/2$ , the overall wave function  $\psi$  has  $2^g$  components  $\psi_j$ ;  $|\psi_j(\mathbf{r}_1, \dots, \mathbf{r}_g)|^2$  is the probability density for finding each of particles 1 to  $g$  with spin oriented along  $+z$ ; and the normalization condition is given by Eq. (4), summed from  $j = 1$  to  $2^g$ . The

$$\sum_j \int d\mathbf{r}_1 \cdots d\mathbf{r}_g |\psi_j(\mathbf{r}_1, \dots, \mathbf{r}_g)|^2 = 1 \quad (4)$$

appropriate reinterpretations when the wave function is not quadratically integrable are obvious. Complications which arise from particle indistinguishability are discussed later in the article.

### Operators

Whereas in classical mechanics particle coordinates  $\mathbf{r}$  and momenta  $\mathbf{p}$  are numbers which can be specified independently at any instant of time, in quantum mechanics the components of  $\mathbf{r}$  and  $\mathbf{p}$  are linear operators, as also are functions  $f(\mathbf{r}, \mathbf{p})$  of the coordinates and momenta. It is postulated that (i) the operator  $\mathbf{x}$  (here distinguished by boldface from the  $x$  coordinate to which  $\mathbf{x}$  corresponds) simply multiplies a wave function  $\psi(x)$  by  $x$ , that is,  $\mathbf{x}\psi = x\psi$ ; (ii) the operator corresponding to the canonically conjugate  $x$  component of momentum of that particle is  $p_x = (\hbar/i)\partial/\partial x$ , that is,  $p_x\psi = (\hbar/i)\partial\psi/\partial x$ . Thus  $A\psi$  denotes the new wave function  $\psi' = A\psi$ , resulting from the linear operation  $A$  on a given wave function  $\psi$ . When  $A, B, C$  are linear operators, and  $\psi, \xi$  are any two functions, Eqs. (5) hold. Moreover, if  $\xi$

$$\begin{aligned} A(\psi + \xi) &= A\psi + A\xi \\ (A + B)\psi &= A\psi + B\psi \\ AB\psi &= A(B\psi) \end{aligned} \quad (5)$$

and  $\psi$  can be added or equated, then  $\xi$  and  $\psi$  must have the same number of components and depend upon the same space and spin coordinates; if  $\xi = \psi$ , corresponding components of  $\xi$  and  $\psi$  are equal. A spin-independent operator performs the same operation on each component of a many-component wave function, for example,  $(p_x\psi)_j = p_x\psi_j = (\hbar/i)\partial\psi_j/\partial x$ . Spin-dependent operators are more complicated; for instance, in a one-particle system of spin  $1/2$  the components of  $\psi' = s_z\psi$ , where  $s_z$  denotes the  $z$  component of the spin operator, are given by Eqs. (6), using

$$\psi'_1 = 1/2\hbar\psi_1 \quad \psi'_2 = -1/2\hbar\psi_2 \quad (6)$$

the notation adopted previously in the discussion of spin.

The operators  $A$  and  $B$  are said to commute when, for any  $\psi$ ,  $A(B\psi) = B(A\psi)$ , implying that Eq. (7) is

$$AB - BA = 0 \quad (7)$$

valid. The operator  $(AB - BA)$  is termed the commutator of  $A$  and  $B$ . Any operator  $f(A)$  expressible as a power series in the operator  $A$  commutes with  $A$ . By performing the indicated operations, Eq. (8) is

$$(xp_x - p_x x)\psi = i\hbar\psi \quad (8)$$

obtained, which shows that pairs of operators need not commute. In a  $g$ -particle system, all particle coordinates  $x_1, y_1, z_1, \dots, x_g, y_g, z_g$  commute with each other; all momentum coordinates  $\mathbf{p}_1, \dots, \mathbf{p}_g$  commute with each other; any component of  $\mathbf{r}_1$  or  $\mathbf{p}_1$  commutes with all components of  $\mathbf{r}_2, \dots, \mathbf{r}_g$  and of  $\mathbf{p}_2, \dots, \mathbf{p}_g$ ; the  $x$  coordinate of any particle commutes with  $p_y$  and  $p_z$  of that particle, and so on.

**Hermitian operators.** An operator  $A$  relevant to a given system of  $g$  particles is termed hermitian if Eq. (9) holds for all pairs of sufficiently well-behaved

$$\begin{aligned} \sum_j \int d\mathbf{r}_1 \cdots d\mathbf{r}_g \xi_j^*(A\psi)_j \\ = \sum_j \int d\mathbf{r}_1 \cdots d\mathbf{r}_g (A\xi)_j^* \psi_j \end{aligned} \quad (9)$$

quadratically integrable wave functions,  $\xi(\mathbf{r}_1, \dots, \mathbf{r}_g)$ ,  $\psi(\mathbf{r}_1, \dots, \mathbf{r}_g)$ .

In Eq. (9) the asterisk denotes the complex conjugate, and the sum is over all components  $j$  of  $\xi$ ,  $A\psi$ ,  $A\xi$ ,  $\psi$ . Evidently every particle coordinate  $x_1, y_1, z_1, \dots, x_g, y_g, z_g$  is a hermitian operator, as is any reasonably well-behaved  $f(\mathbf{r}_1, \dots, \mathbf{r}_g)$ . Recalling that quadratically integrable functions vanish at infinity, integration by parts shows that every component of  $\mathbf{p}_1, \dots, \mathbf{p}_g$  is hermitian, as is any  $f(\mathbf{p}_1, \dots, \mathbf{p}_g)$  expressible as a power series in components of  $\mathbf{p}_1, \dots, \mathbf{p}_g$ ; for example, in the simple one-dimensional spinless case, Eq. (10) holds. It is implied that  $\xi$  and  $\psi$

$$\begin{aligned} \int_{-\infty}^{\infty} dx \xi^* \frac{\hbar}{i} \frac{\partial \psi}{\partial x} &= -\frac{\hbar}{i} \int_{-\infty}^{\infty} dx \frac{\partial \xi^*}{\partial x} \psi \\ &= \int_{-\infty}^{\infty} dx \left( \frac{\hbar}{i} \frac{\partial \xi}{\partial x} \right)^* \psi \end{aligned} \quad (10)$$

are continuous; otherwise the integration by parts yields extra terms on the right side of Eq. (10). Similarly,  $p_x^2$  has the desired hermitian property, defined by Eq. (9), only when  $\xi$  and  $\psi$  are continuous and have continuous first derivatives at all points.

When  $A, B, C, \dots$ , are individually hermitian Eq. (11) holds. For simplicity the integration vari-

$$\int \xi^* [(ABC \cdots)\psi] = \int [(\cdots CBA)\xi]^* \psi \quad (11)$$

ables and the summation over components are not indicated explicitly in Eq. (11). When  $A$  and  $B$  are hermitian, Eq. (11) implies (i)  $1/2(AB + BA)$  is hermitian;

(ii)  $AB$  and  $BA$  are not hermitian unless  $A$  and  $B$  commute. For example,  $xp_x$  and  $p_x x$  are not hermitian, but  $1/2(xp_x + p_x x)$  is; classically, of course, there is no distinction between  $xp_x$ ,  $p_x x$ , or  $1/2(xp_x + p_x x)$ . By appropriately symmetrizing, taking note of Eq. (11), one can construct the quantum-mechanical hermitian operator corresponding to any classical  $f(\mathbf{r}_1, \dots, \mathbf{r}_g; \mathbf{p}_1, \dots, \mathbf{p}_g)$  expressible as a power series in components of coordinates and momenta.

If one supposes that the forces acting on a quantum-mechanical system of  $g$  particles are precisely the same as in the classical case, the energy operator is the classical hamiltonian in Eq. (12), where

$$H(\mathbf{r}_1, \dots, \mathbf{r}_g; \mathbf{p}_1, \dots, \mathbf{p}_g) = T + V \quad (12)$$

the kinetic energy  $T = p_1^2/2m_1 + \dots + p_g^2/2m_g$ ;  $V(\mathbf{r}_1, \dots, \mathbf{r}_g)$  is the potential energy; and  $m_i$  is the mass of particle  $i$ . See HAMILTON'S EQUATIONS OF MOTION.

Quantum-mechanical nonclassical forces, with associated potential energy operators more complicated than  $V(\mathbf{r}_1, \dots, \mathbf{r}_g)$ , are not uncommon, however. For example, the interaction between two neutrons is believed to include space exchange as in Eq. (13), where  $J(r_{12})$  is an ordinary function of the

$$\begin{aligned} V\psi_j(\mathbf{r}_1, \mathbf{r}_2) &\equiv J(r_{12})P_{12}\psi_j(\mathbf{r}_1, \mathbf{r}_2) \\ &= J(r_{12})\psi(\mathbf{r}_2, \mathbf{r}_1) \end{aligned} \quad (13)$$

distance  $r_{12}$  between the particles, and the space exchange operator  $P_{12}$  interchanges the space coordinates of particles 1 and 2 in each component of  $\psi$ . See NUCLEAR STRUCTURE.

**Real eigenvalues.** The eigenvalue equation for an operator  $A$  is given by Eq. (14), where the number  $\alpha$

$$Au(\alpha) = \alpha u(\alpha) \quad (14)$$

is the eigenvalue, and the corresponding eigenfunction  $u(\alpha)$  is a not identically zero wave function solving Eq. (14) for that value of  $\alpha$ . The eigenvalue equation  $Hu = Eu$  for the energy operator  $H$  has special importance and is known as the time-independent Schrödinger equation. Since the eigenvalues  $\alpha$  are identified with the results of measurement, it is desirable that (i) the eigenvalues  $\alpha$  all be real numbers and not complex numbers; (ii) the corresponding eigenfunctions form a complete set, the meaning and importance of which is explained subsequently. Property (i) is important because actual measurements yield real numbers only, for example, lengths, meter readings, and the like. If the eigenvalue of  $A$  were complex, it could not be maintained that each value of  $\alpha$  represented a possible result of exact measurement of  $A$ . This assertion is not negated by the fact that it is formally possible to combine real quantities into complex expressions; for example, the coordinates of  $\mathbf{r}$  in the  $xy$  plane form the complex vector  $x + iy$ . See EIGENFUNCTION; EIGENVALUE (QUANTUM MECHANICS).

The eigenvalues  $\alpha_n$  belonging to the quadratically integrable eigenfunctions  $u(\alpha_n) \equiv u_n$  of a hermitian operator  $A$  are necessarily real. In the notation of

Eq. (11), letting  $\xi = \psi = u_n$  in Eq. (9) and employing Eq. (14), one obtains Eq. (15).

$$\begin{aligned} \alpha_n \int u_n^* u_n &= \int u_n^* (\alpha_n u_n) = \int (Au_n) u_n^* \\ &= \int (Au_n)^* u_n = \int (\alpha_n u_n)^* u_n = \alpha_n^* \int u_n^* u_n \end{aligned} \quad (15)$$

The equality of the first and last terms of Eq. (15) demonstrates that  $\alpha_n = \alpha_n^*$ . For this reason it is postulated that all "observable" operators are hermitian operators, and conversely that all hermitian operators represent observable quantities; henceforth all operators are supposed hermitian. In addition, it is necessary to require that the allowed eigenfunctions preserve the hermiticity property of Eq. (9); otherwise Eq. (15) would not hold. For the important class of hamiltonian operators  $H = T + V$ , except for highly singular (discontinuous) potentials, the boundary conditions that  $u$  and  $\partial u/\partial x$  must be continuous guarantee reality of the eigenvalue corresponding to a quadratically integrable  $u$  solving Eq. (14). Admissible wave functions, defined following Eq. (2b), may be quadratically nonintegrable, however. It always is assumed that the physically desirable properties (i) and (ii) that were discussed above follow from the equally physically desirable simple requirement that the eigenfunctions  $u(\alpha)$  of  $A$  must be admissible, provided that  $u$ ,  $\partial u/\partial x$ , etc., satisfy the continuity conditions which make Eq. (15) correct for quadratically integrable  $u_n$ . In systems containing a single spinless particle this assumption has been justified rigorously for operators of interest; the widespread quantitative successes of quantum theory support the belief that the assumption is equally valid in more complicated systems.

**Orthogonality.** The eigenvalues  $\alpha$  corresponding to quadratically integrable eigenfunctions typically form a denumerable (countable), though possibly infinite, set and compose the discrete spectrum of  $A$ . The admissible nonquadratically integrable eigenfunctions typically correspond to a continuous set of real eigenvalues composing the continuous spectrum of  $A$ . An eigenvalue  $\alpha$  is degenerate, with order of degeneracy  $d \geq 2$ , if there exist  $d$  independent eigenfunctions  $u_1, \dots, u_d$  corresponding to the same value of  $\alpha$ , whereas every  $d + 1$  such eigenfunctions are dependent. The  $d$  functions  $\psi_1, \dots, \psi_d$  are (linearly) dependent if the relation in Eq. (16) can be

$$c_1 \psi_1 + \dots + c_d \psi_d = 0 \quad (16)$$

true with not all the constants  $c_1, \dots, c_d$  equal to zero. Eigenvalues which are either discrete or continuous or both may be degenerate. While so-called accidental degeneracy can occur, degeneracy of the eigenvalues  $\alpha$  of  $A$  ordinarily is associated with the existence of one or more operators which commute with  $A$ . In the absence of degeneracy the eigenfunctions  $u(\alpha)$  are uniquely indexed by  $\alpha$ , and can be chosen to satisfy the orthonormal (orthogonality and normalizing) relations [compare Eq. (4)], as shown

in Eq. (17).

$$\sum_j \int d\mathbf{r}_1 \cdots d\mathbf{r}_g u_j^*(\alpha) u_j(\alpha') = \delta_{\alpha\alpha'} \text{ or } \delta(\alpha - \alpha') \quad (17)$$

In Eq. (17) the Kronecker symbol  $\delta_{\alpha\alpha'}$  is employed when  $\alpha$  lies in the discrete spectrum;  $\delta_{\alpha\alpha'} = 0$  for  $\alpha \neq \alpha'$ ,  $\delta_{\alpha\alpha'} = 1$  for  $\alpha = \alpha'$ . The Dirac delta function  $\delta(\alpha - \alpha')$  is employed when  $\alpha$  lies in the continuous spectrum;  $\delta(\alpha - \alpha') = 0$  when  $\alpha \neq \alpha'$ , but has a finite integral defined by Eq. (18a). For a wide variety of functions  $f(\alpha)$  these properties of  $\delta(\alpha - \alpha')$  imply that Eq. (18b) is valid.

$$\int_{-\infty}^{\infty} d\alpha \delta(\alpha - \alpha') = \int_{-\infty}^{\infty} d\alpha' \delta(\alpha - \alpha') = 1 \quad (18a)$$

$$\int_{-\infty}^{\infty} d\alpha f(\alpha) \delta(\alpha - \alpha') = \int_{-\infty}^{\infty} d\alpha f(\alpha) \delta(\alpha' - \alpha) = f(\alpha') \quad (18b)$$

When Eq. (17) holds,  $u(\alpha)$  are said to be normalized on the  $\alpha$ -scale. Evidently  $\delta(x)$  is highly singular at  $x = 0$ , and is an even function of  $x$ .

Equation (17) asserts that eigenfunctions corresponding to different eigenvalues always are orthogonal, that is, that the integral in Eq. (17) equals zero whenever  $\alpha \neq \alpha'$ . For discrete  $\alpha$  (or  $\alpha'$ ) this assertion is readily proved by an argument similar to Eq. (17); in fact, the orthogonality for  $\alpha \neq \alpha'$  holds whether or not the eigenvalues are degenerate. When  $\alpha$  and  $\alpha'$  both lie in the continuous spectrum, however, the integral of Eq. (17) does not converge and therefore actually is not defined. Thus the delta function  $\delta(\alpha - \alpha')$  primarily is a useful formal device; here and elsewhere in the theory, delta functions always are eventually integrated over their arguments, as in Eqs. (18); such integration makes expressions like the left side of Eq. (17) effectively convergent. A mathematically rigorous justification of the use of the delta function in quantum theory, or a rigorous justification of Eq. (17), encounters difficulties. These are related to the difficulties in establishing rigorously the aforementioned properties (i) and (ii). For operators and eigenfunctions of physical interest, experience suggests no reason to doubt the basic correctness of the mathematical procedures of nonrelativistic quantum theory. See OPERATOR THEORY.

### Illustrative Applications

The immediately following subheadings apply the preceding formalism to some representative problems involving a one-dimensional spinless particle free to move in the  $x$  direction only; in every case, one begins by seeking the appropriate solution to Eq. (14). As subsequent discussion makes clear, results for this simplest of systems are pertinent to more complicated systems.

**Momentum.** Equation (14) is written as Eq. (19),

$$p_x u = \frac{\hbar}{i} \frac{\partial u}{\partial x} = \hbar k u \quad (19)$$

where  $\hbar k$  is the eigenvalue; conventionally the eigenfunctions are indexed by the wave number  $k$ . The solutions to Eq. (19) have the form  $u(x, k) = C(k) \exp[ikx]$ , where  $C(k)$  is a normalizing constant. When  $k$  is real,  $u(x, k)$  is finite for all  $x$ ,  $-\infty \leq x \leq \infty$ . Consequently, the spectrum is continuous and includes all real  $k$ ,  $-\infty \leq k \leq \infty$ . If  $k$  has an imaginary part, that is, if  $k = k_1 + ik_2$ ,  $k_2 \neq 0$ , then  $u(x, k_1 + ik_2)$  is not admissible since it becomes infinite at either  $x = +\infty$  or  $x = -\infty$ . If the eigenfunction could vanish identically for  $|x| \geq a$ , it would be quadratically integrable for complex  $k$ ; in this event the eigenfunction would be discontinuous at  $x = a$ . However, Eq. (20) holds

$$|C(k) \exp[ikx]| \neq 0 \quad (20)$$

unless  $C(k) = 0$ . Thus the requirements that  $u$  be admissible and continuous ensure that the eigenvalues of  $p_x$  are real, as asserted previously in the discussion of real eigenvalues. The  $u(x, k)$  are quadratically nonintegrable, as expected for the continuous spectrum; each  $u(x, k)$  can be regarded as representing a beam of particles, all moving with the same velocity. Because  $\exp[ikx]$  is periodic with wavelength  $\lambda = 2\pi/k$ , such a beam will demonstrate wavelike properties; in fact,  $p_x = \hbar k = h/\lambda$ , in agreement with the de Broglie relations. This agreement is not trivial; although the conclusion that the quantum-mechanical momentum operator is  $p_x = (\hbar/i) \partial/\partial x$  can be argued starting from the de Broglie relations, the form of  $p_x$  also can be inferred directly from Eq. (8), which in turn can be argued from the formal analogy between the properties of the commutator and the Poisson bracket, without any reference to the de Broglie relations. See CANONICAL TRANSFORMATIONS.

Normalized on the  $k$  scale, the eigenfunctions are given by Eq. (21a), for which corresponding to Eq. (17) Eq. (21b) may be written. Equation (21b)

$$u(x, k) = \frac{1}{\sqrt{2\pi}} e^{ikx} \quad (21a)$$

$$\int_{-\infty}^{\infty} dx u^*(k) u(k') = \frac{1}{2\pi} \int_{-\infty}^{\infty} dx e^{i(k' - k)x} = \delta(k - k') \quad (21b)$$

amounts to a nonrigorous statement of the Fourier integral transform theorem, as shown by Eqs. (22).

$$\psi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} dk e^{ikx} c(k) \quad (22a)$$

$$c(k) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} dx e^{-ikx} \psi(x) \quad (22b)$$

See FOURIER SERIES AND TRANSFORMS.

Equation (22b) can be derived by mathematically rigorous procedures. The quantum-theoretic derivation multiplies Eq. (22a) by  $(2\pi)^{-1/2} \exp(-ik'x)$  and

integrates over all  $x$ . There is obtained, after interchanging the orders of  $x$  and  $k$  integration, Eq. (23).

$$\begin{aligned} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} dx e^{-ik'x} \psi(x) \\ = \frac{1}{2\pi} \int_{-\infty}^{\infty} dk c(k) \int_{-\infty}^{\infty} dx e^{i(k-k')x} \\ = \int_{-\infty}^{\infty} dk c(k) \delta(k-k') = c(k') \end{aligned} \quad (23)$$

**Kinetic energy.** For the kinetic energy  $T_x = p_x^2/2m$ , Eq. (14) is written as Eq. (24). The eigenvalue is  $E$ .

$$T_x u = -\frac{\hbar^2}{2m} \frac{\partial^2 u}{\partial x^2} = Eu \quad (24)$$

Admissible solutions are given by Eq. (25), where

$$u(x, E) = C(E) \exp\left(\frac{i}{\hbar} x \sqrt{2mE}\right) \quad (25)$$

$\sqrt{E}$  must be real. Thus the spectrum is continuous and runs from  $E = 0$  to  $E = +\infty$ . The square root may be either positive or negative in Eq. (25) so that there are two independent eigenfunctions at each value of  $E$ ; that is, the spectrum is degenerate. The eigenfunctions may be labeled  $u_+(x, E)$  and  $u_-(x, E)$ ; normalized on the  $E$  scale, Eq. (26) obtains, in which

$$u_{\pm}(x, E) = \left(\frac{m}{2b^2 E}\right)^{1/4} \exp\left(\pm \frac{i}{\hbar} x \sqrt{2mE}\right) \quad (26)$$

the square root always is positive,  $u_+$  and  $u_-$  individually satisfy Eq. (17), but the sets  $u_+$  and  $u_-$  are orthogonal to each other. Introducing  $k = \sqrt{2mE}/\hbar$ , the eigenfunctions can be labeled by the single parameter  $k$ ,  $-\infty \leq k \leq \infty$ , instead of by  $E$ ,  $0 \leq E \leq \infty$ , and the subscripts + or -. Evidently the  $u(x, k)$  of Eq. (21a) are eigenfunctions not only of  $p_x$  but also of  $T_x$ ; each  $u(x, k)$  corresponds to a different eigenvalue of  $p_x$ , but  $u(x, k)$  and  $u(x, -k)$  correspond to the same eigenvalue of  $T_x$ . Interpreted physically, these results mean that the energy  $E$  of a free particle is known exactly when its momentum  $p_x$  is known exactly, and that  $E = p_x^2/2m = (-p_x)^2/2m$  just as in classical mechanics. The normalizations in Eqs. (21a) and (26) are different because  $dE = (\hbar^2 k/m) dk$ . The eigen-

functions in Eqs. (27) also are normalized on the  $E$

$$\begin{aligned} u'_+(x, E) &= \frac{1}{\sqrt{2}} [u_+(x, E) + u_-(x, E)] \\ &= \left(\frac{2m}{b^2 E}\right)^{1/4} \cos \frac{x}{\hbar} \sqrt{2mE} \end{aligned} \quad (27a)$$

$$\begin{aligned} u'_-(x, E) &= \frac{1}{\sqrt{2}} [u_+(x, E) - u_-(x, E)] \\ &= \left(\frac{2m}{b^2 E}\right)^{1/4} \sin \frac{x}{\hbar} \sqrt{2mE} \end{aligned} \quad (27b)$$

scale, and are an alternative set to  $u_{\pm}(x, E)$  of Eq. (25).

**Typical energy operator.** Equation (14) is written as Eq. (28), where the operators in the brackets operate

$$\begin{aligned} Hu &= (T_x + V)u \\ &= \left[ \frac{-\hbar^2}{2m} \frac{\partial^2}{\partial x^2} + V(x) \right] u = Eu \end{aligned} \quad (28)$$

on  $u$ . Typically, but not always (see the two subheadings immediately following),  $V(x)$  is an everywhere smooth finite function which approaches zero at  $x = \pm\infty$ . **Figure 1** is a plot of such a  $V(x)$ , having attractive force centers (negative potential energy) near  $x = \pm a$ , and a repulsive region (positive potential energy) near  $x = 0$ ; in this case the minimum potential at  $x = \pm a$  is  $V_0 < 0$ .

Equation (28) is rewritten in the form of Eqs. (29).

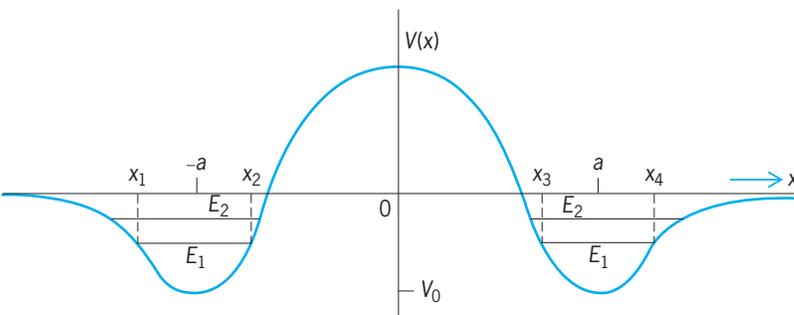
$$\frac{\partial^2 u}{\partial x^2} = Ku \quad (29a)$$

$$K(x, E) = \frac{2m}{\hbar^2} [V(x) - E] \quad (29b)$$

When  $K > 0$ , Eq. (29b) implies that  $(\partial/\partial x)(\partial u/\partial x)$  has the same sign as  $u$ ; for example,  $\partial u/\partial x$  increases with increasing  $x$  if  $u > 0$ . In other words, at points  $x$  where  $K(x, E) > 0$ , solutions of Eq. (29a), are convex toward the  $x$  axis; where  $K(x, E) < 0$ ,  $u(x, E)$  is concave toward the  $x$  axis. Provided  $V(x)$  decreases to zero sufficiently rapidly,  $K(\pm\infty, E) < 0$  when  $E > 0$ , so that at  $x = \pm\infty$  solutions to Eq. (28) are approximated by linear combinations of the oscillatory functions of Eqs. (27): when  $E < 0$ , however,  $K(\pm\infty, E) > 0$ , so that at  $x = \pm\infty$  solutions to Eq. (28) are approximated by linear combinations of the convex functions [compare Eq. (25)] in Eqs. (30).

$$\begin{aligned} u_1(x, E) &= \exp [x(2m|E|)^{1/2}/\hbar] \\ u_2(x, E) &= \exp [-x(2m|E|)^{1/2}/\hbar] \end{aligned} \quad (30)$$

Function  $u_1$  is infinite at  $x = +\infty$ ;  $u_2$  is infinite at  $x = -\infty$ . Thus, for  $E < 0$ , every admissible solution of Eq. (28) must behave like  $u_1$  at  $x = -\infty$  and like  $u_2$  at  $x = +\infty$ , that is, must be asymptotic to the  $x$  axis at both  $x = +\infty$  and  $x = -\infty$ . When  $V_0 < E < 0$  there are values of  $x$  where the solution is concave, so that a smooth (satisfying the boundary condition that  $u$  and  $\partial u/\partial x$  must be continuous) transition from a curve like  $G$  or  $L$  (**Fig. 2**) on the left to  $F$  on the right may be possible. When  $E < V_0$  solutions to Eq. (28) are everywhere convex and, like  $F$ ,  $G$ ,  $L$  of



**Fig. 1.** A potential  $V(x)$ , capable of trapping particles near  $x = \pm a$ .

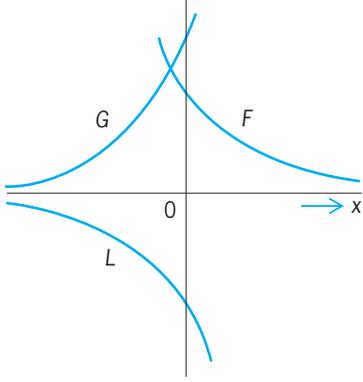


Fig. 2. Everywhere convex solutions asymptotic to  $x$  axis.

Fig. 1, never are asymptotic to the  $x$  axis at both  $x = \pm\infty$ .

It is concluded that (i) the continuous spectrum includes all positive values of  $E$ ,  $0 \leq E \leq \infty$ , and at each such  $E$  there are two independent, nonquadratically integrable (oscillatory at  $x = \pm\infty$ ) eigenfunctions, as in the previously discussed example of kinetic energy; (ii) there are no eigenvalues  $E < V_0$ ; (iii) there may but need not be discrete eigenvalues  $V_0 < E < 0$  corresponding to quadratically integrable eigenfunctions. The lowest eigenfunction, for  $E = E_1$ , has the least region of concavity and therefore joins  $F$  of Fig. 2 to  $G$  without crossing the  $x$  axis, as in Fig. 3. The eigenfunction  $u(x, E_2)$  corresponding to the next higher eigenvalue  $E = E_2 > E_1$  has one node (one zero), decreases less rapidly at  $x = \pm\infty$  than  $u(x, E_1)$ , and links  $F$  to  $L$ ; the next higher eigenfunction corresponding to  $E_3 > E_2$  has two nodes, again links  $F$  to  $G$ , and so on.

The horizontal lines in the neighborhood of  $x = \pm a$  in Fig. 1 show the energy levels of the two lowest eigenvalues  $E_1, E_2$ . The points  $x_1, x_2$ , where  $E = V_1$  and where the curvature of  $u(x, E_1)$  changes sign (Fig. 3), are the turning points (velocity equals zero) of a classical particle oscillating with total energy  $E_1$  in the attractive potential well at  $x = -a$ ;  $x_3, x_4$  are similar turning points near  $x = a$ . For  $E_1 < E < E_2$  a solution  $u(x, E)$  starting out as does  $F$  at  $x = +\infty$  crosses the  $x$  axis but becomes negatively infinite at  $x = -\infty$ , because it is insufficiently concave to join smoothly with  $L$ ; this makes understandable the general result that eigenvalues corresponding to quadratically integrable eigenfunctions are discrete.

**Potential barrier.** Consider Eq. (28) with  $V(x) = 0$  for  $x < 0$ ;  $V(x) = V_1 > 0$  for  $x > 0$ , and  $0 < E < V_1$ , as shown in Fig. 4a. As previously explained,  $u$  and  $\partial u/\partial x$  must be everywhere continuous, even though  $V(x)$  is discontinuous at  $x = 0$ . Recollecting Eqs. (26), (29), and (30) for  $0 < E < V_1$ , there must be one and only one independent eigenfunction, having the form shown in Eqs. (31), where  $R$  and  $S$  are con-

$$u(x, E) = \exp[ix(2mE)^{1/2}/\hbar] + R \exp[-ix(2mE)^{1/2}/\hbar] \quad x < 0 \quad (31a)$$

$$u(x, E) = S \exp[-x(2m|E - V_1|)^{1/2}/\hbar] \quad x > 0 \quad (31b)$$

stants. It is reasonable and customary to interpret the first exponential in Eq. (31a) as a beam of particles moving with momentum  $p_x = (2mE)^{1/2}$  toward the classical barrier or turning point  $x = 0$ ; the second exponential in Eq. (31a) represents a reflected beam. Since the amplitude of the incident beam has been normalized to unity [not a normalization consistent with Eq. (17)],  $|R|^2$  must be the reflection coefficient of the barrier. The continuity requirements at  $x = 0$  yield Eqs. (32), showing that  $|R|^2 = 1$  in agreement

$$R = \frac{iE^{1/2} + (V_1 - E)^{1/2}}{iE^{1/2} - (V_1 - E)^{1/2}} \quad (32)$$

$$S = \frac{2iE^{1/2}}{iE^{1/2} - (V_1 - E)^{1/2}}$$

with classical expectation for  $E < V_1$ . See REFLECTION AND TRANSMISSION COEFFICIENTS.

The eigenfunction  $u(x, E)$  is sketched in Fig. 4a. Since  $|u|^2 \neq 0$  at  $x > 0$ , particles penetrate the classically inaccessible region to the right of the barrier;

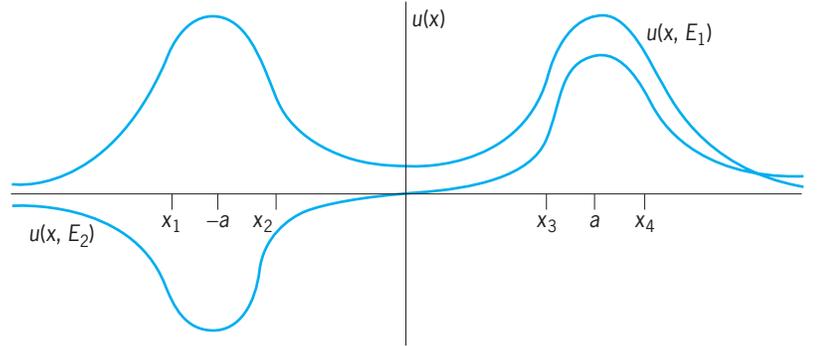


Fig. 3. Eigenfunctions indicated for the two lowest eigenvalues of the hamiltonian with the potential of Fig. 1.

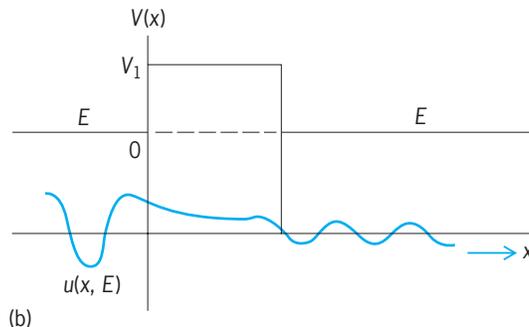
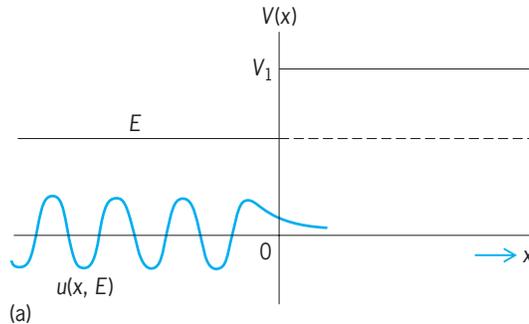


Fig. 4. Reflections (a) from a barrier of infinite thickness, and (b) from one of finite thickness.

because  $|u|^2 = 0$  at  $x = +\infty$ , all particles eventually are turned back, however, consistent with  $|R|^2 = 1$ . This penetration, and Eqs. (31) and (32), closely resembles the penetration (and corresponding classical optics expressions) of a totally reflected light wave into a medium with smaller index of refraction. Hence Eqs. (31) and (32), which are unforced results of solving Eq. (28) for  $V(x)$  of Fig. 4a, manifest the wave-particle duality inherent in the quantum theoretic formalism.

For the barrier of finite thickness  $a$  (Fig. 4b), there are two solutions at every  $0 < E < V_1$ . The solution representing a beam incident from the left, transmitted with coefficient  $|T|^2$  and reflected with coefficient  $|R|^2$ , is found from Eq. (31a) together with Eqs. (33). Equations (33) mean that, except for the

$$u(x, E) = C \exp[-x(2m|E - V_1|)^{1/2}/\hbar] + D \exp[x(2m|E - V_1|)^{1/2}/\hbar] \quad 0 < x < a \quad (33a)$$

$$u(x, E) = T \exp[ix(2mE)^{1/2}/\hbar] \quad x > a \quad (33b)$$

incident  $\exp[ix(2mE)^{1/2}]$  at  $x = -\infty$ , the solution at  $x = \pm\infty$  must consist of waves traveling out toward infinity; a similar outgoing boundary condition specifies the continuum solution  $E > 0$  representing more complicated collisions, for example, the scattering of a beam of particles by target particles in a foil. Outgoing boundary conditions are employed in classical wave theories as well. See SCATTERING EXPERIMENTS (ATOMS AND MOLECULES); SCATTERING EXPERIMENTS (NUCLEI); SCATTERING OF ELECTROMAGNETIC RADIATION.

The continuity requirements lead to Eq. (34a), and

$$|T|^2 = \left\{ 1 + \frac{V_1^2 \sinh^2 [(a/\hbar)\sqrt{2m(V_1 - E)}]}{4E(V_1 - E)} \right\}^{-1} \quad (34a)$$

$$|T|^2 = \frac{16E(V_1 - E)}{V_1^2} \times \exp[-2a(2m|E - V_1|)^{1/2}/\hbar] \quad (34b)$$

$|R|^2 = 1 - |T|^2$ , as is necessary if  $|R|^2$  and  $|T|^2$  are to represent, respectively, reflection and transmission coefficients. Equation (34a), and the corresponding expression when  $E \gg V_1$ , resemble the equations for transmission of light through thin films. When  $(a/\hbar)\sqrt{2m(V_1 - E)} \gg 1$ , Eq. (34a) is closely approximated by Eq. (34b), where the exponential factor, sometimes termed the barrier penetrability, is  $\ll 1$ . The transmission through a less simple barrier  $V(x)$  than Fig. 4b is measured by the penetrability factor defined by Eq. (35), where  $x_1, x_2$  are the turning

$$P = \exp \frac{-2}{\hbar} \int_{x_1}^{x_2} dx (2m|E - V(x)|)^{1/2} \quad (35)$$

points  $E = V(x_1) = V(x_2)$ . The barrier penetrability governs the rates at which (i) an incident proton, whose kinetic energy is less than the height of the repulsive Coulomb potential barrier surrounding a nucleus, is nonetheless able to penetrate the nucleus to produce nuclear reactions; (ii) alpha particles can escape from a radioactive nucleus; (iii) electrons can

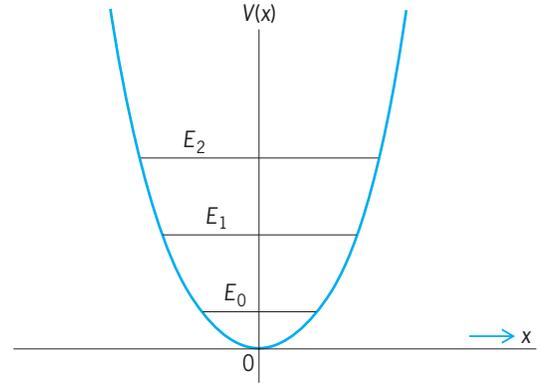


Fig. 5. Harmonic oscillator potential.

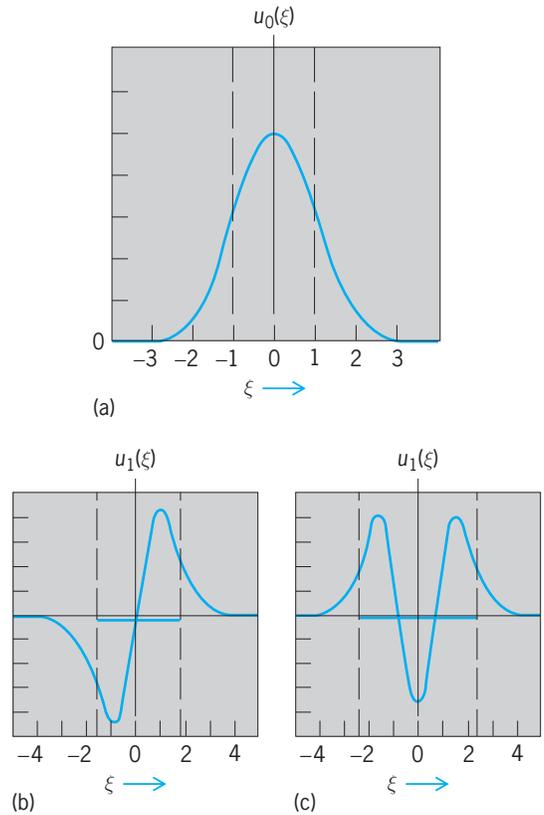


Fig. 6. Harmonic oscillator eigenfunctions. (a)  $n = 0$ . (b)  $n = 1$ . (c)  $n = 2$ . (After L. Pauling and E. B. Wilson, Jr., *Introduction to Quantum Mechanics*, McGraw-Hill, 1935)

be pulled out of a metal by a strong electric field. See FIELD EMISSION; NUCLEAR FUSION; RADIOACTIVITY.

**Harmonic oscillator.** The potential  $V(x) = 1/2Kx^2$  (Fig. 5) describes a classical harmonic oscillator whose equilibrium position is  $x = 0$ ; the classical oscillator frequency is  $\nu = (2\pi)^{-1}\sqrt{K/m}$ . Since  $V(x)$  becomes infinite at  $x = \pm\infty$ , there is no continuous spectrum, but there must be an infinite number of discrete eigenvalues. Various sophisticated methods exist for finding the eigenfunctions and eigenvalues. See HARMONIC OSCILLATOR.

The energy levels turn out to be as in Eq. (36),

$$E_n = (n + 1/2)h\nu \quad (36)$$

where  $n = 0, 1, 2, \dots$  (Fig. 5). The corresponding first three eigenfunctions  $u_0(\xi)$ ,  $u_1$ ,  $u_2$  are sketched in Fig. 6, with  $\xi = x(mK)^{1/4}/\hbar^{1/2}$  a convenient dimensionless variable; the broken vertical lines indicate the turning points at  $\xi = \pm\sqrt{2n+1}$ . Figure 7 plots the probability density  $|u(\xi)|^2$  for  $n = 10$ . The classical probability of finding a particle in the  $x$  interval  $dx$  is proportional to the time spent in  $dx$ . The curved dashed line in Fig. 7 plots the classical probability density for a classical oscillator whose energy is  $E_{10} = 2^{1/2}bv$ , Eq. (36). See DIMENSIONLESS GROUPS; ENERGY LEVEL (QUANTUM MECHANICS).

The agreement between the classical probability density and the average over oscillations of  $|u_{10}(\xi)|^2$  illustrates the connection between classical particle mechanics and the more fundamental dual wave-particle interpretation of quantum theory. With increasing  $n$  the oscillations of  $|u_n(\xi)|^2$  become more rapid, and the agreement with the classical probability density improves, in accordance with the correspondence principle. These harmonic oscillator results are a good first approximation to the energy levels and eigenfunctions of vibrating atoms in isolated molecules and in solids. See LATTICE VIBRATIONS; MOLECULAR STRUCTURE AND SPECTRA; SPECIFIC HEAT OF SOLIDS.

### Expectation Values

Suppose  $B$  is an operator that commutes with  $A$ . If  $Au_n = \alpha_n u_n$ ,  $BAu_n = B(\alpha_n u_n)$ , so that, using Eq. (7), one obtains Eq. (37), which means that  $Bu_n$  also is

$$A(Bu_n) = \alpha_n(Bu_n) \quad (37)$$

an eigenfunction of  $A$  corresponding to the eigenvalue  $\alpha_n$ . When  $\alpha_n$  is not degenerate, Eq. (16) means  $Bu_n$  must be a multiple of  $u_n$ , that is,  $Bu_n = \beta u_n$ ,  $\beta$  being a constant; thus  $u_n$  is simultaneously an eigenfunction of the pair of commuting operators  $A$ ,  $B$ . When the order of degeneracy of  $\alpha_n$  is  $d$ ,  $Bu_n$  need not be a multiple of  $u_n$ , but  $d$  independent eigenfunctions  $u_{n1}(\alpha_n, \beta_1), \dots, U_{nd}(\alpha_n, \beta_d)$  always can be

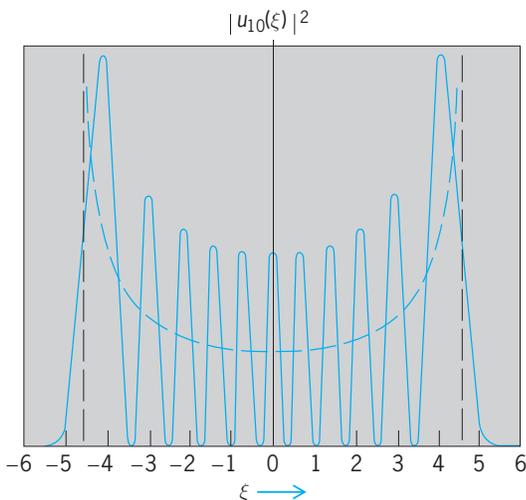


Fig. 7. Probability density for harmonic oscillator in state  $n = 10$ . (After L. Pauling and E. B. Wilson, Jr., *Introduction to Quantum Mechanics*, McGraw-Hill, 1935)

found, such that  $u_{ns}(\alpha_n, \beta_s)$  is an eigenfunction of  $B$  corresponding to the eigenvalue  $\beta_s$ , as well as an eigenfunction of  $A$  corresponding to  $\alpha_n$ . If all the  $\beta_s$  are different,  $u_{n1}, \dots, u_{nd}$ , being eigenfunctions corresponding to different eigenvalues of  $B$ , are mutually orthogonal (see the earlier discussion on orthogonality). If all the  $\beta_s$  are not different, a third operator  $C$ , simultaneously commuting with  $A$  and  $B$ , is sought, and so on. For the system of  $g$  particles of spin  $1/2$ , one determines in this fashion a complete set of simultaneously commuting observables,  $A, B, C, \dots$ , whose corresponding eigenvalues  $\alpha, \beta, \gamma, \dots$  (except for accidental degeneracy) uniquely index an orthonormal set of  $2^g$ -component eigenfunctions as in Eq. (38), which satisfies Eq. (39).

$$u_j(\mathbf{r}_1, \dots, \mathbf{r}_g; \alpha, \beta, \gamma, \dots), \quad j = 1 \text{ to } 2^g \quad (38)$$

$$\int u^*(\alpha, \beta, \gamma, \dots) u(\alpha', \beta', \gamma', \dots) \\ = \delta(\alpha - \alpha') \delta(\beta - \beta') \delta(\gamma - \gamma') \dots \quad (39)$$

In Eq. (39), and henceforth unless specifically indicated otherwise, the integral is in the simplified notation of Eq. (11). Equation (39) generalizes Eq. (17);  $\delta(\alpha - \alpha')$  is replaced by a Kronecker symbol  $\delta_{\alpha\alpha'}$  when  $\alpha, \alpha'$  are discrete, and so on.

The meaning of property (ii) mentioned in the earlier discussion of real eigenvalues can now be explained. The eigenfunctions of an operator  $A$  having been indexed as in Eq. (39), any sufficiently smooth quadratically integrable  $2^g$ -component wave function  $\psi(\mathbf{r}_1, \dots, \mathbf{r}_g)$  can be expressed in the form of Eq. (40), where each eigenvalue is integrated over its

$$\psi(\mathbf{r}_1, \dots, \mathbf{r}_g) = \int d\alpha \int d\beta \int d\gamma \dots \\ \times c(\alpha, \beta, \gamma, \dots) u(\mathbf{r}_1, \dots, \mathbf{r}_g; \alpha, \beta, \gamma, \dots) \quad (40)$$

entire spectrum, with the understanding (further to simplify the notation) that in the discrete spectrum integration is replaced by summation. Employing Eq. (39), the constants  $c(\alpha, \beta, \gamma, \dots)$  are found to be as shown in Eq. (41).

$$c(\alpha, \beta, \gamma, \dots) = \int u_j^*(\alpha, \beta, \gamma, \dots) \psi \quad (41)$$

Equations (40) and (41) are consistent with the often instructive interpretation that  $c(\alpha, \beta, \gamma, \dots)$  are the projections of the vector  $\psi$  on a complete set of orthogonal unit vectors  $u(\alpha, \beta, \gamma, \dots)$  in an infinite-dimensional vector space.

**Probability considerations.** When  $\psi$  is normalized, Eqs. (4), (18), (39), and (40) imply Eq. (42), in the notation of Eq. (40). Equations (40) and (42) make

$$\int d\alpha \int d\beta \int d\gamma \dots |c(\alpha, \beta, \gamma, \dots)|^2 = 1 \quad (42)$$

it reasonable to postulate that whenever a system is instantaneously described by the normalized

quadratically integrable  $\psi(\mathbf{r}_1, \dots, \mathbf{r}_g)$ , the instantaneous probability of finding the system in the state corresponding to  $A = \alpha$  is given by Eq. (43a).

$$P(\alpha) = \int d\beta \int d\gamma \dots |c(\alpha, \beta, \gamma, \dots)|^2 \quad (43a)$$

$$P(\alpha, \beta) = \int d\gamma \dots |c(\alpha, \beta, \gamma, \dots)|^2 \quad (43b)$$

Similarly, the simultaneous probability of finding the system in the state corresponding to  $A = \alpha$  and  $B = \beta$  is given, by Eq. (43b), and so on. Clearly Eqs. (1) and (2) are a special case of Eqs. (42) and (43), in which the coordinate operators,  $\mathbf{x}_1, \mathbf{y}_1, \mathbf{z}_1, \dots, \mathbf{x}_g, \mathbf{y}_g, \mathbf{z}_g$  of the  $g$  spinless particles form the complete set of simultaneously commuting observables, and  $\psi(\mathbf{x}_1, \mathbf{y}_1, \mathbf{z}_1, \dots, \mathbf{x}_g, \mathbf{y}_g, \mathbf{z}_g)$  is the projection of  $\psi$  on the eigenfunction simultaneously corresponding to  $\mathbf{x}_1 = x_1, \mathbf{y}_1 = y_1, \dots, \mathbf{z}_g = z_g$ .

The total probability  $\int d\alpha P(\alpha)$  is unity, Eq. (42), and therefore any measurement of the observable  $A$  must yield a number equal to one of its eigenvalues. Moreover, by the very meaning of probability, the average or expectation value of the observable  $A$  must be as given in Eq. (44a), which shows that the same

$$\langle A \rangle = \int d\alpha \alpha P(\alpha) \quad (44a)$$

$$\langle A \rangle = \int \psi^* A \psi \quad (44b)$$

operator may have different forms in different representations. For instance, in the momentum representation discussed earlier where  $\alpha \equiv k, p_x c(k)$  must equal simply  $\hbar k c(k)$ ; consequently, to satisfy Eq. (8),  $\mathbf{x}c(k)$  must equal  $i(\partial c / \partial k)$ . When  $\psi$  is a wave function for which Eq. (9) holds with  $\xi \equiv \psi$ , Eqs. (5), (14), (39), and (40) imply that Eq. (44b) is equivalent to Eq. (44a). Equation (40b), usually more convenient than Eq. (44a), predicts the expectation value of any given observable in an arbitrary physical situation described by a reasonably well-behaved quadratically integrable wave function; expectation values are not defined for nonquadratically integrable  $\psi$ . Equation (9) guarantees that  $\langle A \rangle$  computed from Eq. (44b) is a real number, necessary if  $\langle A \rangle$  is to represent the results of measurement (see the earlier discussion on real eigenvalues). See PROBABILITY.

**Uncertainty principle.** A precise measure of the spread or uncertainty in the value of  $A$  is  $\Delta A$  defined by Eq. (45). Here  $(\Delta A)^2$  is the average square

$$(\Delta A)^2 = \langle (A - \langle A \rangle)^2 \rangle = \langle A^2 \rangle - \langle \langle A \rangle \rangle^2 \quad (45)$$

deviation of  $A$  from its average  $\langle A \rangle$ . The quantity  $\Delta A = 0$  when and only when  $\psi$  is an eigenfunction of  $A$ , that is,  $A\psi = \alpha\psi$ , in which event  $\langle A \rangle = \alpha, \langle A^2 \rangle = \alpha^2$ . The discussion following Eq. (37) implies  $\Delta A$  and  $\Delta B$  simultaneously can be zero; that is,  $A$  and  $B$  are simultaneously exactly measurable, whenever the commutator  $AB - BA$  is zero. If  $AB - BA \neq 0$ , introduce  $A' = A - \langle A \rangle, B' = B - \langle B \rangle$ ; use Eqs. (11) and

(44b); and employ the so-called Schwartz inequality. Then Eq. (46) holds.

$$\begin{aligned} (\Delta A)^2 (\Delta B)^2 &= \int (A'\psi)^* (A'\psi) \int (B'\psi)^* (B'\psi) \\ &\geq \left| \int (A'\psi)^* (B'\psi) \right|^2 = \left| \int \psi^* (A'B'\psi) \right|^2 \end{aligned} \quad (46)$$

But since Eq. (47) is valid, it leads, after further manipulation, to Eq. (48), which is the rigorous quan-

$$A'B' = \frac{1}{2}(A'B' - B'A') + \frac{1}{2}(A'B' + B'A') \quad (47)$$

$$(\Delta A)^2 (\Delta B)^2 \geq \frac{1}{4} |\langle AB - BA \rangle|^2 \quad (48)$$

tum theoretic formulation of the uncertainty principle. When  $A = x, B = p_x$ , Eq. (8) yields Eq. (49).

$$(\Delta x)(\Delta p_x) \geq \hbar/2 \quad (49)$$

This simple derivation of the uncertainty principle demonstrates anew the necessity for a dual wave-particle interpretation of the operator formalism.

### Time Dependence

The procedures developed thus far predict the results of measurement at any given instant of time  $t_0$  that the wave function  $\psi(x, t_0)$  is known. Given  $\psi$  at  $t_0$ , to predict the results of measurement at some future instant  $t$ , it is necessary to know the time evolution of  $\psi$  from  $t_0$  to  $t$ . It is postulated that this evolution is determined by the time-dependent Schrödinger equation (50), where  $H$  is the hamiltonian or energy operator;  $\psi$  is supposed to be a continuous function of  $t$ . If  $H$  were a number, the solution to Eq. (50) would be as shown in Eq. (51). Equation

$$H\psi = i\hbar \frac{\partial \psi}{\partial t} \quad (50)$$

(51) remains valid when  $H$  is an operator, provided Eq. (52) holds, with  $I$  the unit operator,  $I\psi = \psi$ ;

$$\psi(t) = \exp \left[ \frac{-iH(t - t_0)}{\hbar} \right] \psi(t_0) \quad (51)$$

(51) remains valid when  $H$  is an operator, provided Eq. (52) holds, with  $I$  the unit operator,  $I\psi = \psi$ ;

$$\begin{aligned} \exp \left[ \frac{-iH(t - t_0)}{\hbar} \right] &\equiv I - \frac{i(t - t_0)}{\hbar} H \\ &+ \frac{[-i(t - t_0)]^2}{2\hbar^2} H^2 + \dots \end{aligned} \quad (52)$$

the right side of Eq. (52) is the usual series expansion of the exponential on the left side. These operational methods, which manipulate operators like numbers, are widely used in quantum theory; they must be used cautiously, but usually lead rapidly to the same results as more conventional mathematical techniques.

When  $H\psi(t_0) = E\psi(t_0)$ , that is, when  $\psi(t_0)$  is known to be an eigenfunction  $u(E, t_0)$  of the energy operator, Eq. (51) shows that  $\psi(t) \equiv u(E, t) = u(E, t_0) \exp[-iE(t - t_0)/\hbar]$ , so that  $|u(E, t)|^2 = |u(E, t_0)|^2$ ; similarly, the expectation values  $\langle A \rangle$  of allopertors  $A$

are time-independent when the system is in a stationary state  $u(E, t)$ . In an arbitrary state  $\psi(t)$ , Eqs. (5), (11), (44b) and (50) imply [assuming that  $A$  is not explicitly time-dependent, for example,  $A = A(\mathbf{r}, \mathbf{p})$  but not  $A(\mathbf{r}, \mathbf{p}, t)$ ] Eq. (53), which uses the notation

$$\begin{aligned} \frac{d}{dt}\langle A \rangle &= \int \left[ \psi^* \left( A \frac{\partial \psi}{\partial t} \right) + \frac{\partial \psi^*}{\partial t} (A \psi) \right] \\ &= \frac{1}{i\hbar} \int [\psi^* (A H \psi) - (H \psi)^* A \psi] \\ &= \frac{1}{i\hbar} \int [\psi^* (A H - H A) \psi] \\ &= \frac{1}{i\hbar} \langle (A H - H A) \rangle \end{aligned} \quad (53)$$

of Eq. (11). Of course  $\langle (A H - H A) \rangle = 0$  whenever  $\psi(t)$  is a stationary state  $u(E, t)$ . Equation (53) shows, however, that if  $A$  commutes with the hamiltonian, then  $\langle A \rangle$  is independent of time whether or not the system is in a stationary state. Consequently, operators commuting with the hamiltonian are termed constants of the motion; a system initially described by an eigenfunction  $u(E, \beta, \gamma, \dots)$  of the simultaneously commuting observables  $H, B, C, \dots$ , remains in an eigenstate of  $H, B, C, \dots$ , as the wave function evolves in time.

Equation (53) is closely analogous to the classical mechanics expression for  $dA(\mathbf{r}, \mathbf{p})/dt$ , where  $A(\mathbf{r}, \mathbf{p})$  is the classical quantity corresponding to the quantum-mechanical operator  $A$ . If  $A$  is put equal to  $I$  in Eq. (53), one obtains Eq. (54).

$$\frac{d}{dt} \int \psi^* \psi = \frac{1}{i\hbar} \int [\psi^* (H \psi) - (H \psi)^* \psi] \quad (54)$$

Therefore the requirement that  $H$  be hermitian, Eq. (9), which has been justified on the grounds that the eigenvalues of  $H$  must be real, has the further important consequence that the right side of Eq. (54) is zero, that is, that Eq. (2a) is obeyed at all times  $t > t_0$  if it obeyed at  $t = t_0$ . This result is necessary for the consistency of the formalism: otherwise it could not be claimed that  $|\psi(t)|^2$  from Eq. (50) is the probability density at  $t > t_0$ . For a single spinless three-dimensional particle, with  $H = p^2/2m + V(x, y, z)$ , it follows directly from Eq. (50) that one obtains Eq. (55a), where Eq. (55b) holds, and so on. In

$$\frac{\partial}{\partial t} (\psi^* \psi) + \frac{\partial}{\partial x} S_x + \frac{\partial}{\partial y} S_y + \frac{\partial}{\partial z} S_z = 0 \quad (55a)$$

$$S_x = \frac{\hbar}{2mi} \left( \psi^* \frac{\partial \psi}{\partial x} - \psi \frac{\partial \psi^*}{\partial x} \right) \quad (55b)$$

Eq. (55a)  $S_x, S_y, S_z$  can be interpreted as the components of a probability current vector  $\mathbf{S}$ , whose flow across any surface enclosing a volume  $\tau$  accounts for the change in the probability of finding the particle inside  $\tau$ . See EQUATION OF CONTINUITY; MAXWELL'S EQUATIONS.

For a nonquadratically integrable  $\psi$  in the one-particle case where Eq. (55b) is applicable, the probability current at infinity generally has the value  $|\psi|^2 \mathbf{v}$ , where  $\mathbf{v}$  is the classical particle velocity at infinity; the one-dimensional plane waves of Eq. (21a)

trivially illustrate this assertion. Consequently (see the preceding discussion of normalization),  $\mathbf{S}$  of Eq. (55b) is interpretable as particle current density when  $|\psi|^2$  is nonvanishing at infinity. These considerations may be generalized to more complicated systems and are important in collision problems, where the incoming and outgoing current at infinity determine the cross section.

**Invariance.** Extremely general arguments support the view that the form of the Schrödinger equation (50) for any  $g$ -particle system isolated from the rest of the universe must be (i) translation invariant, that is, independent of the origin of coordinates; (ii) rotation invariant, that is, independent of orientation of the coordinate axes; and (iii) reflection invariant, that is, independent of whether one chooses to use a left-handed or right-handed coordinate system.

The only known failures of these general requirements occur for reflections, in a domain outside the scope of nonrelativistic quantum theory, namely, in phenomena, such as beta decay, that are connected with the so-called weak interactions. Correspondingly, it can be inferred that the hamiltonian operator  $H$  for any such isolated system must commute with (i) the total linear momentum operator  $\mathbf{p}_R = \mathbf{p}_1 + \dots + \mathbf{p}_g$ ; (ii) the total angular momentum operator  $\mathbf{J}$ ; (iii) the parity operator  $P$ , which reflects every particle through the origin, that is, changes  $\mathbf{r}_1$  to  $-\mathbf{r}_1, \dots, \mathbf{r}_g$  to  $-\mathbf{r}_g$ . For additional information see PARITY (QUANTUM MECHANICS); SYMMETRY LAWS (PHYSICS).

In quantum mechanics as in classical mechanics, therefore, linear momentum and total angular momentum are conserved, that is, are constants of the motion. Since for an infinitesimal displacement  $\epsilon$  in the  $x$  direction Eq. (56) holds, the connection be-

$$\begin{aligned} \psi(x_1 + \epsilon, y_1, z_1, x_2 + \epsilon, y_2, z_2, \dots, x_g + \epsilon, y_g, z_g) \\ &= \psi(x_1, y_1, z_1, \dots, x_g, y_g, z_g) \\ &+ \epsilon \left( \frac{\partial \psi}{\partial x_1} + \frac{\partial \psi}{\partial x_2} + \dots + \frac{\partial \psi}{\partial x_g} \right) \\ &= \psi + \epsilon p_{Rx} \psi \end{aligned} \quad (56)$$

tween  $p_{1x} + \dots + p_{gx}$  and translation in the  $x$  direction can be understood; the connection between  $\mathbf{J}$  and rotation is understood similarly. Because a discontinuous change in position, from  $\mathbf{r}$  to  $-\mathbf{r}$ , is inconceivable classically, the conservation of parity concept has no relevance to classical mechanics.

**Transition probability.** Frequently the hamiltonian  $H$  of Eq. (50) has the form  $H_0 + V(t)$ , where the time-dependent potential energy  $V(t)$  represents an externally imposed interaction, for example, with measuring equipment; it is supposed that  $V(t) = 0$  for  $t < 0$  and  $t > t_1$ . Usually the system is in a stationary state  $u(E_i)$  of  $H_0$  at time  $t < 0$ , and one wishes to compute the probability of finding the system in some other stationary state  $u(E_f)$  of  $H_0$  at times  $t > t_1$ . From Eq. (40) and the discussion preceding Eq. (53), Eq. (57) is obtained, which, when substituted in

Eq. (50) for times  $0 \leq t \leq t_1$ , yields Eq. (58).

$$\begin{aligned} \psi(t) &= \int dE d\beta c(E, \beta, t) \exp(-iEt/\hbar) u(E, \beta) \quad (57) \\ i\hbar \int dE d\beta \frac{dc(E, \beta, t)}{dt} \exp(-iEt/\hbar) u(E, \beta) \\ &= \int dE d\beta c(E, \beta, t) \\ &\quad \times \exp(-iEt/\hbar) V'(t) u(E, \beta) \quad (58) \end{aligned}$$

In Eqs. (57) and (58),  $E$  is  $\alpha$  of Eq. (40) and  $\beta$  stands for all other indices  $\beta, \gamma, \dots$ , necessary to make  $u(E)$ , a complete orthonormal set; if  $V'(t)$  were zero, the projections  $c(E, \beta)$  would equal  $\delta(E - E_i)\delta(\beta - \beta_i)$  independent of time.

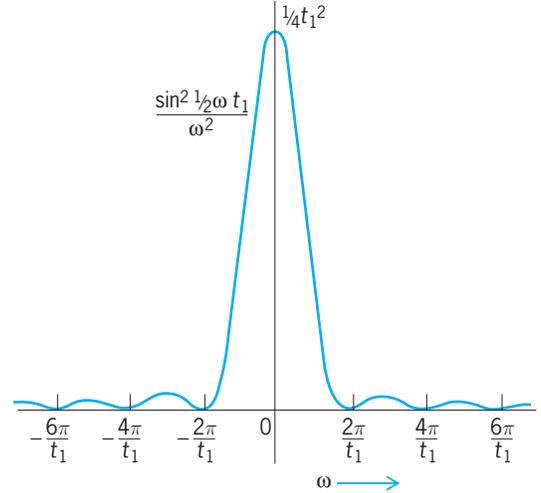
Most problems in quantum theory cannot be solved exactly; therefore some approximate treatment using perturbations must be devised. In the present case it is assumed that  $c(E, \beta, t)$  do not change appreciably from their initial values at  $t = 0$ , so that it is a reasonable approximation to replace  $c(E, \beta, t)$  by  $c(E, \beta, 0)$  on the right side of Eq. (58). With the further approximation that  $V'$  is constant during the interval  $0 \leq t \leq t_1$ , one finds, using the notation of Eq. (11), that Eq. (59) obtains, where  $\hbar\omega = E_f - E_i$ .

$$\begin{aligned} |c(E_f, \beta_f, t_1)|^2 &\equiv \frac{4}{\hbar^2} \left| \int u^*(E_f, \beta_f) V' u(E_i, \beta_i) \right|^2 \\ &\quad \times \frac{\sin^2 1/2\omega t_1}{\omega^2} \\ &\equiv \frac{4}{\hbar^2} |V'_{fi}|^2 \frac{\sin^2 1/2\omega t_1}{\omega^2} \quad (59) \end{aligned}$$

Equation (59) shows that the probability of finding the system in some new stationary state  $u(E_f, \beta_f)$  after the system is no longer perturbed is proportional to (i)  $|V'_{fi}|^2$ , the square of the matrix element of the perturbation between initial and final states; (ii) an oscillating factor which for given  $t_1$  has a peak  $t_1^2/4$  at  $\omega = 0$ .

$\sin^2(1/2\omega t_1)/\omega^2$  is plotted as a function of  $\omega$  in **Fig. 8**; evidently  $|c(E_f, \beta_f)|^2$  is relatively small for energies such that  $|\omega| > \sim \pi/t_1$ . The most likely occupied states after the perturbation conserve energy ( $E_f = E_i$ ), and the spread in energy of the final states is  $\Delta E = \hbar\Delta\omega = \sim 2\pi\hbar/t_1 \equiv b/\Delta t$ , where  $t_1 \equiv \Delta t$  is, for example, the duration of the measurement. Thus Eq. (59) provides a version of the uncertainty principle between energy and time. As  $t_1$  approaches infinity, the area under the curve in Fig. 8 becomes proportional to  $t_1$ , since the main peak has a height proportional to  $t_1^2$  and a width proportional to  $1/t_1$ . Therewith one obtains a widely employed formula giving the approximate transition probability  $w$  per unit time for making transitions, under the influence of a steady perturbation  $V'$ , from an initial stationary  $u(E_i, \beta_i)$  to a set of final states of the same energy, namely, Eq. (60), where  $\rho(E_f) = d\beta d\gamma \dots$  is the den-

$$w = \frac{2\pi}{\hbar} \rho(E_f) |V'_{fi}|^2 \quad (60)$$



**Fig. 8.** Plot of  $\sin^2 1/2 \omega t_1 / \omega^2$  versus  $\omega = (E_f - E_i)/\hbar$ . (After L. I. Schiff, *Quantum Mechanics*, 2d ed., McGraw-Hill, 1955)

sity of independent final states in the neighborhood of  $E = E_f = E_i$ ,  $\beta = \beta_f$ ,  $\gamma = \gamma_f$ , and so on. For instance, with  $u_i$  a plane wave  $e^{ikz}$  moving in the  $z$  direction, Eq. (21a), and  $u_f$  a plane wave in some other direction, Eq. (60) yields the Born approximation to the cross section for elastic scattering by a potential. Equation (60) is also applicable to problems outside the domain of nonrelativistic quantum theory, for example, to the theory of beta decay wherein new particles are created.

The preceding considerations are important for understanding how a measurement of an operator  $A$  not commuting with the hamiltonian  $H$  can cause an initially stationary state  $u(E)$  to evolve into an eigenstate  $u(\alpha)$  of  $A$ . Equations (50)–(53), with  $H = T + V = H_0$ , the unperturbed hamiltonian, hold only in the intervals between measurements; during the measurement  $u(E)$ , though an eigenfunction of  $H_0$ , is not a stationary state of the complete hamiltonian. But this paragraph does not do justice to the subtle questions involved in the quantum theory of measurements.

### Further Illustrative Applications

When  $g$  particles are noninteracting, the hamiltonian has the trivially separable form  $H \equiv H_1 + H_2 \dots + H_g$ , and Eq. (61) holds, where  $u_i(\mathbf{r}_i, E_i, \alpha_i)$  are a complete

$$\begin{aligned} (H_1 + \dots + H_g) [u_1(\mathbf{r}_1, E_1, \alpha_1) \dots u_g(\mathbf{r}_g, E_g, \alpha_g)] \\ = (E_1 + \dots + E_g) \\ \times [u_1(\mathbf{r}_1, E_1, \alpha_1) \dots u_g(\mathbf{r}_g, E_g, \alpha_g)] \quad (61) \end{aligned}$$

orthonormal set of eigenfunctions of  $H_1$  for particle 1, and so on. Thus, because an operation performed solely on particle 1 always commutes with an operation performed solely on particle 2, the products  $u_1(\mathbf{r}_1, E_1, \alpha_1) \dots u_g(\mathbf{r}_g, E_g, \alpha_g)$  are a complete orthonormal set of eigenfunctions of  $H = H_1 + \dots + H_g$ ; also the energy levels of  $H$  are the set of possible sums of individual particle energies,  $E = E_1 + \dots + E_g$ .

Similarly, if  $H(x) = T_x + V(x)$ , with eigenfunctions  $u(x, E_x)$ , is a hamiltonian for a one-dimensional spinless particle, then any quadratically integrable  $\psi(x, y, z)$  describing a three-dimensional spinless particle can be expanded in a series of products  $u(x, E_x) u(y, E_y) u(z, E_z)$ . This paragraph explains the relevance, to three-dimensional many-particle systems of the results obtained for the illustrative applications previously discussed. The immediately following subheadings continue to illustrate the general formalism. The reader is cautioned that the remaining contents of this article, although very important, especially for applications to atomic and nuclear structure, are for the most part admittedly more condensed than the material presented heretofore.

**Parity.** Because  $P^2\psi(\mathbf{r}) = P\psi(-\mathbf{r}) = \psi(\mathbf{r})$ , the parity operator has but two eigenvalues, namely,  $+1$  and  $-1$ ; the corresponding eigenfunctions are said to have even or odd parity. Evidently  $P$  commutes with the harmonic oscillator hamiltonian  $p_x^2/2m + (1/2)Kx^2$ . The harmonic oscillator eigenvalues are nondegenerate, and therefore every harmonic oscillator eigenfunction (Fig. 6) has either even or odd parity. Similarly the eigenfunctions  $u'_+$ ,  $u'_-$  of  $T_x$ , Eqs. (27), have even and odd parity, respectively; eigenfunctions of  $T_x$  which do not have definite parity also exist, however, Eq. (26), because the eigenvalues  $E$  are degenerate. The eigenfunctions of  $p_x$  do not have definite parity, Eq. (21a), because  $p_x P + P p_x = 0$ , that is,  $p_x$  does not commute, but instead anticommutes, with  $P$ .

**Time evolution of packet.** A wave function  $\psi$  representing a single (spinless) particle localized in the neighborhood of a point is termed a wave packet. Assuming  $H = p^2/2m + V(\mathbf{r})$ , Eq. (53) yields Eq. (62a)

$$\frac{d}{dt}\langle \mathbf{x} \rangle = \frac{1}{i\hbar}\langle \mathbf{x}H - H\mathbf{x} \rangle = \left\langle \frac{p_x}{m} \right\rangle \quad (62a)$$

$$\frac{d^2}{dt^2}\langle \mathbf{x} \rangle = \frac{d}{dt}\left\langle \frac{p_x}{m} \right\rangle = \frac{-1}{m}\left\langle \frac{\partial V}{\partial x} \right\rangle \quad (62b)$$

by employing Eq. (8); similarly, one obtains Eq. (62a) by employing Eq. (8); similarly, one obtains Eq. (62b). Equations (62) mean (i) the average position of the particle, that is, the center of the packet, moves with a velocity given by the expectation value of the momentum; (ii) the acceleration of the center of the packet is found from the expectation value of the classical force  $-\partial V/\partial x$ . Equations (62) illustrate the correspondence between quantum and classical mechanics and show that the classical description of particle motion is valid when the spread of the packet about its mean position can be ignored. When the particle is free,  $H = p^2/2m$ , Eqs. (8), (45), and (63), where the subscripts  $t, 0$  refer, respectively, to

$$\begin{aligned} (\Delta x)_t^2 &= (\Delta x)_0^2 + \left\{ \frac{t}{m}\langle xp + px \rangle_0 - 2\langle x \rangle_0\langle p \rangle_0 \right\} \\ &\quad + \frac{t^2}{m^2}(\Delta p_x)_0^2 \quad (63) \end{aligned}$$

expectation values at  $t$  and at initial time zero. Equa-

tion (63) shows that, although an unconfined free wave packet may contract for a while, ultimately it will spread over all space; when the minimum spread happens to occur at  $t = 0$ , the term linear in  $t$  vanishes in Eq. (63).

**Orbital angular momentum.** The quantum mechanical operators representing the components of orbital angular momentum  $\mathbf{L}$  have the same form as in classical mechanics, namely, for each particle  $L_x = yp_z - zp_y$ , and so on. By using Eq. (8), one obtains Eqs. (64),

$$L^2 L_z - L_z L^2 = 0 \quad (64a)$$

$$L_x L_y - L_y L_x = i\hbar L_z \quad (64b)$$

and so on, where  $L^2 \equiv L_x^2 + L_y^2 + L_z^2$ . According to Eq. (48), therefore (i)  $L^2$  and  $L_z$  are simultaneously exactly measurable; (ii) once the values of  $L^2$  and  $L_z$  are specified, the values of  $L_x$  and  $L_y$  must be uncertain. In spherical coordinates  $z = r \cos \theta$ ,  $x = r \sin \theta \cos \phi$ ,  $y = r \sin \theta \sin \phi$ ,  $L_x$  and  $L^2$  becomes Eqs. (65).

$$L_x = \frac{\hbar}{i} \frac{\partial}{\partial \phi} \quad (65a)$$

$$L^2 = -\hbar^2 \left( \frac{1}{\sin \theta} \frac{\partial}{\partial \theta} \sin \theta \frac{\partial}{\partial \theta} + \frac{1}{\sin^2 \theta} \frac{\partial^2}{\partial \phi^2} \right) \quad (65b)$$

Equation (14) for  $L_x$  is solved by  $u(\phi, m) = \exp(im\phi)$ , where  $m\hbar$  is the eigenvalue. It can be argued that  $u(\phi, m)$  must have a unique value at any point  $x, y, z$ , meaning  $u(\phi + 2\pi, m) = u(\phi, m)$ , so that  $m$  must be a positive or negative integer, or zero. With  $\partial^2/\partial \phi^2 = -m^2$  in Eq. (65b), the eigenvalues of  $L^2$  turn out to be  $l(l+1)\hbar^2$ , where  $l = 0, 1, \dots$ , independent of  $m$ , except that for each  $l$  the allowed values of the magnetic quantum number  $m$  are  $m = -l, -l+1, \dots, l-1, l$ ; thus each  $l$  has order of degeneracy  $2l+1$ . Because  $L^2$  and  $L_x$  commute with  $P$ , the eigenfunctions  $u(l, m)$  have definite parity; in fact, Eq. (66) holds.

$$P_u(l, m) = (-1)^l u(l, m) \quad (66)$$

In a two-particle system, the components of the total orbital angular momentum  $\mathbf{L} = \mathbf{L}_1 + \mathbf{L}_2$  obey the same commutation, Eqs. (64); as a result the eigenvalues of  $L^2$  and  $L_z$  (but not the eigenfunctions) are the same as in the one-particle case.  $L^2$  and  $L_z$  commute with  $L_1^2$  and  $L_2^2$ , but  $L^2$  does not commute with  $L_{1z}$  or  $L_{2z}$ . Consequently the total orbital angular momentum eigenfunctions are labeled by  $l, m, l_1, l_2$ . For given  $l_1, l_2$ , the possible values of  $l$  are the positive integers from  $l = l_1 + l_2$  down to  $l = |l_1 - l_2|$ ; the corresponding eigenfunctions have parity  $(-1)^{l_1+l_2}$  independent of  $l, m$ . These rules for combining angular momenta are readily generalized to more complicated systems including spin, are well established, and form the basis for the vector model of the atom. See ATOMIC STRUCTURE AND SPECTRA.

**Coulomb potential.** The hamiltonian for an (assumed spinless) electron of mass  $m_e$  in the field of a fixed charge  $Ze$  is  $H = \mathbf{p}^2/2m_e + V(r)$  where  $V(r) = -Ze^2/r$ . In spherical coordinates, Eq. (14) for

the eigenfunctions  $u(r, \theta, \phi)$  is written as Eq. (67),

$$Hu = \left[ \frac{-\hbar^2}{2m_e} \frac{1}{r^2} \frac{\partial}{\partial r} \left( r^2 \frac{\partial}{\partial r} \right) + \frac{1}{2m_e r^2} L^2 + V(r) \right] u = Eu \quad (67)$$

with  $L^2$  defined by Eq. (65b). Now  $H$  commutes with  $L^2$  and  $L_z$  and  $r^2 H$  is separable; that is, Eq. (68)

$$r^2 H(r, \theta, \phi) = H_1(r) + (2m_e)^{-1} L^2(\theta, \phi) \quad (68)$$

holds [compare Eq. (61)]. Thus  $u(r)u(l, m)$  are a complete set of eigenfunctions;  $L^2 u(r)u(l, m) = l(l+1) \cdot \hbar^2 u(r)u(l, m)$ , and therefore the radial eigenfunctions  $u(r) \equiv u(E, l)$  must satisfy Eq. (69).

$$\left[ \frac{-\hbar^2}{2m_e} \frac{1}{r^2} \frac{\partial}{\partial r} \left( r^2 \frac{\partial}{\partial r} \right) + \frac{l(l+1)\hbar^2}{2m_e r^2} + V(r) \right] u(r) = Eu(r) \quad (69)$$

The positive term  $l(l+1)\hbar^2/2m_e r^2$  acts as an added repulsive potential; it can be understood in terms of the classical centripetal force needed to maintain the angular momentum. For  $E < 0$ , admissible solutions to Eq. (69) must be exponentially decreasing at  $r = \infty$ ; moreover, because of the  $r^{-1}$  and  $r^{-2}$  terms in Eq. (69), an eigenfunction  $u(E, l)$  which behaves properly at  $r = \infty$  becomes infinite at  $r = 0$  unless  $E$  is specially chosen. Thus (as always) the quadratically integrable eigenfunctions form a discrete set. The corresponding bound state energies  $E < 0$  are given by Eq. (70).

$$E = -\frac{m_e Z^2 e^4}{2\hbar^2 n^2} \quad (70)$$

In Eq. (67) the principal quantum number  $n = 1, 2, 3, \dots$ ; for given  $l$  and  $n$  the number  $n_r \geq 0$  of zeros (between  $r = 0$  and  $r = \infty$ ) of the corresponding  $u(E, l)$  is  $n_r = n - l - 1$ . Because  $dx dy dz = r^2 \sin \theta dr d\theta d\phi$ , the radial probability density is  $r^2 |u(r, E, l)|^2$ . **Figure 9** shows the radial probability density plotted versus  $r$  (in units of the Bohr radius  $a_0 = \hbar^2/m_e e^2 \cong 5 \times 10^{-9}$  cm) for several low-lying stationary states of atomic hydrogen,  $Z = 1$ . The notation for the eigenfunctions is standard in atomic physics: the principal quantum number is supplemented by a lowercase letter  $s, p, d, \dots$  corresponding to  $l = 0, 1, 2, \dots$ ; for example, a  $3d$  state has  $l = 2$  and therefore  $n_r = 0$  or no radial nodes, as in Fig. 9. The eigenfunction  $u(l = 0, m = 0)$  is a constant; that is, an  $s$  state is spherically symmetric,  $|u(r, \theta, \phi)|^2$  is proportional to  $\cos^2 \theta$  or to  $\sin^2 \theta$  in  $p$  states, and so on. The eigenfunctions, although they are spread over all space, have their maxima at about the radii expected on the older Bohr theory of Eq. (70).

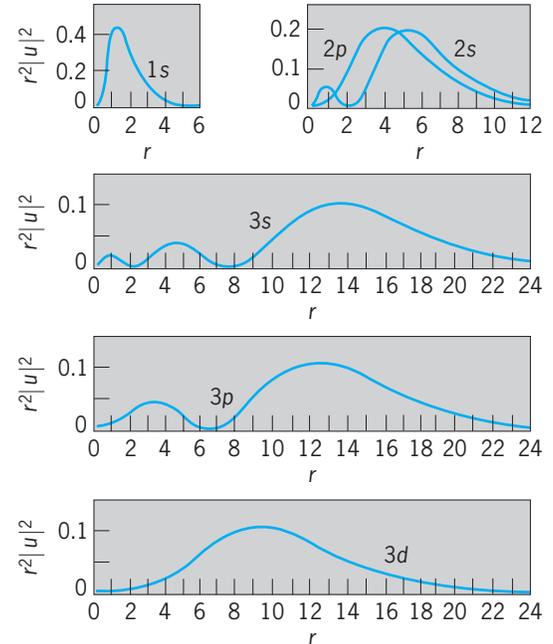
In the actual hydrogen atom the nucleus, of mass  $M$ , is not fixed. The hamiltonian is Eq. (71a), where

$$H = \frac{p_1^2}{2m_e} + \frac{p_2^2}{2M} - \frac{Ze^2}{|r_1 - r_2|} \quad (71a)$$

$$H = \frac{p_R^2}{2(M + m_e)} + \frac{p^2}{2\mu} - \frac{Ze^2}{r} \quad (71b)$$

the subscripts 1 and 2 refer to the electron and the nucleus, respectively. Introducing the center of mass  $(X, Y, Z) \equiv \mathbf{R} = (M + m_e)^{-1} [m_e \mathbf{r}_1 + M \mathbf{r}_2]$ ,  $(x, y, z) \equiv \mathbf{r} = \mathbf{r}_1 - \mathbf{r}_2$ , Eq. (71a) takes the separable form of Eq. (71b), where  $(\mathbf{p}_R)_x \equiv p_X = (\hbar/i) \partial/\partial X$ , etc., are the components of the total momentum  $\mathbf{p}_R = \mathbf{p}_1 + \mathbf{p}_2$ ;  $\mathbf{p}_x = (\hbar/i) \partial/\partial x$ , etc.; and the reduced mass  $\mu = m_e M / (M + m_e)$ . Therefore  $\mathbf{p}_R$  is a constant of the motion, as asserted in connection with Eq. (56); moreover, Eq. (71b) is separable in  $\mathbf{R}$  and  $\mathbf{r}$ . In other words, the center of mass moves like a free particle, completely independent of the internal  $\mathbf{r}$  motion. Comparing Eqs. (67) and (71b), and recalling Eq. (61), the eigenvalues of Eq. (71a), after the kinetic energy of the center of mass is subtracted, are given by Eq. (70) with  $\mu$  (depending on  $m_e/M$ ) replacing  $m_e$ . The independence of internal and center-of-mass motion means that the temperature broadening of spectral lines can be explained quantum-mechanically in terms of the classical Doppler effect for a moving fixed-frequency source. This paragraph illustrates the correspondence between the classical and quantum theories. See CENTER OF MASS.

The eigenvalues  $E$  of Eq. (70) have order of degeneracy  $n^2$ ; this degeneracy stems from (i) the fact that  $V(r)$  is spherically symmetric, which permits  $H$  to commute with  $\mathbf{L}$ ; (ii) a special symmetry of Eq. (67) for the specially simple Coulomb potential, causing solutions of Eq. (69) for different  $l$  to have the same energy. For an arbitrary spherically symmetric  $V(r)$ , the bound-state eigenvalues  $E(n_r, l)$  of Eq. (69) do not coincide for different  $l$ , and each bound state has degeneracy  $2l + 1$ , corresponding to the  $2l + 1$  possible values  $m = -l$  to  $l$  of the magnetic quantum number



**Fig. 9.** Radial probability density in atomic hydrogen. (After F. K. Richtmyer, E. H. Kennard, and T. Lauritsen, *Introduction to Modern Physics*, 5th ed., McGraw-Hill, 1955)

$m$ ; an energy level associated with orbital angular momentum  $l$  has parity  $(-)^l$ . For any such  $V(r)$  the bound-state energies (i) increase with increasing  $n_r$  for a constant value of  $l$ , for the same reasons that were discussed in connection with the eigenvalues of Eq. (28); (ii) increase with increasing  $l$  for constant  $n_r$  because the rotational kinetic energy  $(2m_e r^2)^{-1} \cdot l(l+1)\hbar^2$  increases. Except for these regularities, the order and spacing of the levels depends on the details of  $V(r)$ . For potentials  $V(r)$  decreasing more rapidly than  $1/r$ , the total number of bound states generally is finite, whereas this number is infinite for the Coulomb  $V(r) = -Ze^2/r$ .

**Removal of degeneracy.** The hamiltonian of Eq. (67) is spin-independent, which is why Eq. (67) has been treated as if the wave function had only one component; compare the remarks following Eqs. (5). Corresponding to any one-component solution  $u$  of Eq. (67), there are two independent two-component eigenfunctions; (i)  $\psi_1 = u, \psi_2 = 0$  and (ii)  $\psi_1 = 0, \psi_2 = u$ ; compare the earlier discussion of spin. Thus for an electron the degeneracy of the energy levels in a Coulomb field is  $2n^2$ ; similarly the degeneracy for neutrons, protons, or electrons in an arbitrary spherically symmetric potential is  $2(2l+1)$ . The energy operator for an electron in an actual atom, for example, hydrogen, is not spin-independent, however. Relativistic effects add, to the central  $V(r)$  of Eq. (67), noncentral spin-orbit potentials  $V'(r)[L_x s_x + L_y s_y + L_z s_z] \equiv V'(r) \mathbf{L} \cdot \mathbf{s}$ ; here  $\mathbf{s}$  is the spin operator and obeys the same commutation Eqs. (64) as  $\mathbf{L}$ .

Equations (64) show that  $V'(r)\mathbf{L} \cdot \mathbf{s}$  commutes with  $L^2, s^2, (\mathbf{L} + \mathbf{s})^2$ , and  $L_z + s_z$ , but not with  $L_x$  or  $s_x$ , illustrating the principle of conservation of total angular momentum  $\mathbf{J} = \mathbf{L} + \mathbf{s}$ ; compare the remarks preceding Eq. (56). Consequently, referring to the final part of the discussion of orbital angular momentum (i)  $J^2 = j(j+1)\hbar^2$ ; (ii) for given  $l \neq 0$  (and  $s = 1/2$ );  $j$  has but two possible values  $l \pm 1/2$ ; (iii) the energy levels are labeled by  $j$  and have a  $(2j+1)$ -fold degeneracy corresponding to the  $2j+1$  possible orientations of  $J_z = -j$  to  $+j$ ; (iv) because  $2\mathbf{L} \cdot \mathbf{s} = (\mathbf{L} + \mathbf{s})^2 - L^2 - s^2$ , levels of different  $j$  have different energies, and the splitting of the energies depends predictably on  $l$ ; (v)  $L_x$  (and  $s_x$ ) no longer are constants of the motion, although  $L^2$  and  $s^2 = 1/2 \cdot 3/2 \hbar^2$  still are. Moreover, because  $2j+1 = 2l+2$  for  $j = l + 1/2$ , and  $= 2l$  for  $j = l - 1/2$ , the total number of independent eigenfunctions associated with given  $l$  (and  $n_r$ ) remains  $2(2l+1) = (2l+2) + (2l)$ .

In the independent particle model of atoms and nuclei, one assumes that to a first approximation each particle, particle  $i$ , say, moves in a potential  $V(r_i)$  which has been averaged over the coordinates of all the other particles, so that  $V(r_i)$  depends only on the distance of  $i$  from the atomic or nuclear center. To a first approximation, therefore, the energy levels are associated with configurations of one-particle eigenfunctions; for example, the ground state of atomic beryllium is  $1s^2 2s^2$ . In higher approximation one introduces two-body interactions  $V(r_i, r_j)$ , which may be said to mix different configurations. The considerations of this and the paragraphs just preceding,

together with the exclusion principle discussed subsequently, account for the periodic system of the elements and are the basis for the highly successful nuclear shell model.

The observation that splitting the levels does not change the number of independent eigenfunctions illustrates a general principle and justifies the postulate that the statistical weight of a discrete level equals its order of degeneracy. This principle can be understood on the basis that the number of bound-state eigenfunctions should be a continuous function of the parameters in a reasonably well-behaved hamiltonian; because this number by definition is an integer, it must change discontinuously and, therefore, except under unusual mathematical circumstances, cannot change at all. See BOLTZMANN STATISTICS; STATISTICAL MECHANICS.

In an external magnetic field  $B$  the hamiltonian of a many-electron atom (i) no longer is independent of the orientation of the coordinate axes, so that the degeneracy associated with this symmetry is removed; (ii) retains symmetry with respect to rotation about the magnetic field, so that (with the  $z$  axis along  $B$ )  $J_z$  commutes with the hamiltonian. Thus in a magnetic field a level associated with the total angular momentum quantum number  $j$  should split into  $2j+1$  levels, each of which is associated with one of the magnetic quantum numbers  $-j$  to  $+j$ . This prediction is thoroughly confirmed in the Zeeman effect and in the Stern-Gerlach experiment. See ZEEMAN EFFECT.

**Radiation.** The classical hamiltonian for a charged particle in an electromagnetic field has the form given by Eq. (72), where  $\mathbf{A}$  and  $\phi$  are the scalar

$$H = \frac{1}{2m} \left( \mathbf{p} - \frac{e\mathbf{A}}{c} \right)^2 + e\phi \quad (72)$$

and vector potentials, respectively. See ELECTRON MOTION IN VACUUM.

It is postulated that when properly symmetrized, that is, when  $\mathbf{A} \cdot \mathbf{p} + p \cdot \mathbf{A}$  replaces the classical  $2\mathbf{A} \cdot \mathbf{p}$  (see the earlier discussion on hermitian operators),  $H$  of Eq. (72) is the quantum-mechanical energy operator. The presence of terms linear in  $\mathbf{p}$  modifies some of the formulas which have been given, for example, Eq. (55a). When plane waves of light (frequency  $f$ , wavelength  $\lambda$ , and moving in the  $z$  direction) fall on a hydrogen atom,  $\mathbf{A}$  is proportional to  $\cos[2\pi(z/\lambda - ft)]$ , and  $e\phi$  is the Coulomb potential  $V(r)$  of Eq. (67).

Proceeding as in Eqs. (57)-(60), noting that  $\mathbf{A}$  contains terms that are proportional to  $\exp(2\pi ift)$  and  $\exp(-2\pi ift)$ , and neglecting the small (as can be shown)  $A^2$  terms, one obtains an expression similar to Eq. (59), except that  $\omega \pm 2\pi f$  replaces  $\omega \equiv \hbar^{-1} \cdot (E_f - E_i)$ . In other words, after a long time there are appreciable transition probabilities only to final states  $f$  whose energies satisfy  $E_f - E_i \pm hf = 0$ , in agreement with the notion that a quantum of energy  $hf$  has been emitted or absorbed; the  $+$  sign corresponds to emission, the  $-$  sign to absorption. The corresponding transition probabilities are given by Eq. (60) with  $V'$  proportional to  $\exp[2\pi iz/\lambda]$ ,

and  $u_i, u_f$  stationary-state atomic wave functions satisfying the radiation-unperturbed Eq. (67). The final expressions are analogous to the classical formulas for emission or absorption of radiation; for instance, when the wavelength  $\lambda$  is large compared to atomic dimensions, expanding  $\exp [2\pi iz/\lambda]$  in powers of  $z/\lambda$  shows that the leading term in the transition probability is the matrix element, between initial and final states, of the dipole moment  $ez$  corresponding to classical electric dipole emission or absorption. Because  $z$  changes sign on reflection through the origin, the dipole matrix element vanishes unless  $u_i$  and  $u_f$  have opposite parities. This is one of the selection rules for electric dipole radiation; other selection rules, connected with angular momentum conservation, are obtained similarly. See ELECTROMAGNETIC RADIATION; SELECTION RULES (PHYSICS).

The theory starting with Eq. (72) is termed semiclassical, because it does not replace the classical  $\mathbf{A}, \phi$  by a quantum mechanical operator description of the electromagnetic field. This semiclassical theory has led to the induced emission and absorption probabilities in the presence of external radiation, but not to the spontaneous probability of emission of a photon in the absence of external radiation. The spontaneous transition probabilities can be inferred from the induced probabilities by thermodynamic arguments. The spontaneous transition probability is deduced directly, however, without appeal to the arguments of thermodynamics, when the radiation field is quantized.

### Particle Indistinguishability

For systems of  $g$  identical, and therefore indistinguishable, particles, the formalism is further complicated because the probability  $P$  of finding a given particle in a specified volume  $dx dy dz$  of space must be  $P_1 + P_2 + \dots + P_g$ , where  $P_1, \dots, P_g$  are the probabilities given previously for distinguishable particles; expectation values and normalizations must be reinterpreted accordingly. Moreover, the Pauli exclusion principle asserts that the only physically permissible wave functions must change sign when the space and spin coordinates of any pair of indistinguishable particles of spin  $1/2$  are interchanged. See EXCLUSION PRINCIPLE.

To amplify this assertion, consider the four-component wave function of the two electrons in atomic helium in Eq. (73), where  $|\psi_{++}(\mathbf{r}_1, \mathbf{r}_2)|^2$  is the

$$\psi = \psi_{++}(\mathbf{r}_1, \mathbf{r}_2), \psi_{+-}(\mathbf{r}_1, \mathbf{r}_2), \\ \psi_{-+}(\mathbf{r}_1, \mathbf{r}_2), \psi_{--}(\mathbf{r}_1, \mathbf{r}_2) \quad (73)$$

probability density for finding both electrons with spin along  $+z$ , and so forth. The exclusion principle requires validity of Eqs. (74).

$$\psi_{++}(\mathbf{r}_1, \mathbf{r}_2) = -\psi_{++}(\mathbf{r}_2, \mathbf{r}_1) \\ \psi_{--}(\mathbf{r}_1, \mathbf{r}_2) = -\psi_{--}(\mathbf{r}_2, \mathbf{r}_1) \quad (74) \\ \psi_{+-}(\mathbf{r}_1, \mathbf{r}_2) = -\psi_{-+}(\mathbf{r}_2, \mathbf{r}_1)$$

In the independent particle approximation, all

components of the ground-state eigenfunctions of atomic He are composed of products  $u(\mathbf{r}_1)u(\mathbf{r}_2)$ , where  $u(\mathbf{r})$  is the lowest  $1s$  eigenfunction solving Eq. (67) for the spherically symmetric  $V(r)$  appropriate to He. In this approximation, therefore,  $\psi_{++} = \psi_{--} = 0$ ; if  $\psi_{+-} = u(\mathbf{r}_1)u(\mathbf{r}_2)$ ,  $\psi_{-+}$  is necessarily equal to  $-u(\mathbf{r}_1)u(\mathbf{r}_2)$ . Of the four independent  $1s^2$  eigenfunctions originally possible, only one remains, which can be shown to be a total spin zero eigenfunction; the exclusion principle literally has excluded the three eigenfunctions corresponding to total spin one.

These results are summarized by the rule that at most two electrons,  $sz = \pm 1/2$ , can occupy one-particle states with the same quantum numbers  $n, l$ , and  $m = -l$  to  $l$ . By the general principle explained previously in the discussion of removal of degeneracy, the introduction of two-particle interactions  $V(\mathbf{r}_1, \mathbf{r}_2)$  does not change the number of independent eigenfunctions consistent with the exclusion principle. In the next higher  $1s2s$  configuration of He,  $\psi_{++}$  will equal  $u_{1s}(\mathbf{r}_1)u_{2s}(\mathbf{r}_2) - u_{1s}(\mathbf{r}_2)u_{2s}(\mathbf{r}_1)$ . This antisymmetrized wave function makes nonclassical exchange energy contributions, of the form given by expression (75), interaction to the expectation

$$\int u_{1s}^*(\mathbf{r}_1)u_{2s}^*(\mathbf{r}_2)V(\mathbf{r}_1, \mathbf{r}_2)u_{1s}(\mathbf{r}_2)u_{2s}(\mathbf{r}_1) \quad (75)$$

value  $\langle V(\mathbf{r}_1, \mathbf{r}_2) \rangle$  of the energy. Exchange energies are important for understanding chemical binding.

Except for the complication of spin, this article presupposes that electrons, neutrons, and protons are immutable structureless mass points. Actually, modern theories of nuclear forces and high-energy scattering experiments indicate that this assumption is untrue. When creation, destruction, or other alterations of fundamental particle structure are improbable, however, the general separability of internal and center-of-mass motion [see the discussion following Eq. (71b)] implies that each fundamental particle is sufficiently described by the position of its center of mass and by its spin orientation, that is, by the many-component wave functions used here.

In circumstances wherein more obviously composite systems undergo no changes in internal structure, they too can be treated as particles. For instance, in the slow collisions of two atoms the slowly changing potentials acting on the electrons do not induce transitions to new configurations, and the collision can be described by solving an equation of the form (67) for the relative motion; in rapid collisions electron transitions occur, and the many-electron Schrödinger equation must be employed. Similarly, since a deuteron is a neutron-proton bound state, with total angular momentum unity, in a deuterium molecule (i) each deuterium can be treated as if it were a fundamental particle of spin 1; (ii) the wave function of a deuterium molecule must be symmetric under interchange of the space and spin coordinates of the two deuterons, which interchange involves successive antisymmetric interchanges of the two neutrons and of the two protons. In other words

(when they can be treated as particles) deuterons and other composite systems of integral spin obey Bose-Einstein statistics; composite systems of half-integral spin obey Fermi-Dirac statistics.

When the particles composing a many-particle system can be represented by nonoverlapping wave packets which spread an amount  $\ll \Delta x$  as the center packet moves a distance equal to its width  $\Delta x$ , individual classical particle trajectories can be distinguished; under these circumstances, the particles, whether or not identical, are in effect distinguishable classical particles, and one expects Bose-Einstein and Fermi-Dirac statistics to reduce to the classical Maxwell-Boltzmann statistics. The well-known condition for the validity of classical statistics in an electron gas,  $Nb^3(2\pi mkT)^{-3/2} \ll 1$ , implies that such packets can be constructed; here  $N$  is the electron density,  $k$  is Boltzmann's constant, and  $T$  is the absolute temperature. In the lower vibrational states of a molecule such packets cannot be constructed for the vibrating nuclei, however (compare the discussion of the harmonic oscillator presented earlier), so that quantum statistics cannot be ignored, for example, in the specific heat of  $H_2$  at low temperatures. See QUANTUM STATISTICS. Edward Gerjuoy

Bibliography. L. E. Ballentine, *Quantum Mechanics*, 2d ed., 1998; A. Bohm, *Quantum Mechanics: Foundations and Applications*, 3d ed., 1994; B. H. Bransden and C. J. Joachim, *Introduction to Quantum Physics*, 1989; P. A. M. Dirac, *Principles of Quantum Mechanics*, 4th ed., 1958; H. Ohanian, *Principles of Quantum Mechanics*, 1989; D. A. Park, *Introduction to the Quantum Theory*, 3d ed., 1992; R. Shankar, *Principles of Quantum Mechanics*, 2d ed., 1994.

## Nonsinusoidal waveform

The representation of a wave that does not vary in a sinusoidal manner. Electric circuits containing nonlinear elements, such as iron-core magnetic devices, rectifying devices, and transistors, commonly produce nonsinusoidal currents and voltages. When these are repetitive functions of time, they are called nonsinusoidal periodic electric waves. Oscillograms, tabulated data, and sometimes mathematical functions for segments of such waves are often used to describe the variation through one cycle. The term cycle corresponds to  $2\pi$  electrical radians and covers the period, which is the time interval  $T$  in seconds in which the wave repeats itself.

These electric waves can be represented by a constant term, the average or dc component, plus a series of harmonic terms in which the frequencies of the harmonics are integral multiples of the fundamental frequency. The fundamental frequency  $f_1$ , if it does exist, has the time span  $T = 1/f_1$  s for its cycle. The second-harmonic frequency  $f_2$  then will have two of its cycles within  $T$  seconds, and so on.

**Fourier series representation.** The series of terms stated above is known as a Fourier series and can be

expressed in the form of Eq. (1), where  $y(t)$ , plot-

$$\begin{aligned} y(t) &= B_0 + A_1 \sin \omega t + A_2 \sin 2\omega t + \cdots \\ &\quad + A_n \sin n\omega t + B_1 \cos \omega t + B_2 \cos 2\omega t \\ &\quad + \cdots + B_n \cos n\omega t \\ &= B_0 + C_1 \sin(\omega t + \phi_1) + \cdots \\ &\quad + C_n \sin(n\omega t + \phi_n) \\ &= \sum_{n=0}^{\infty} C_n \sin(n\omega t + \phi_n) \end{aligned} \quad (1)$$

ted over a cycle of the fundamental, gives the shape of the nonsinusoidal wave. The terms on the right-hand side show the Fourier series representation of the wave where Eqs. (2) and (3) apply. Here  $A_0$  is

$$C_n = \sqrt{A_n^2 + B_n^2} \quad (2)$$

$$\phi_n = \arctan \frac{B_n}{A_n} \quad (3)$$

identically zero, and  $B_0 = C_0$  in Eq. (1). The radian frequency of the fundamental is  $\omega = 2\pi f_1$ , and  $n$  is either zero or an integer.  $C_1$  is the amplitude of the fundamental ( $n = 1$ ), and succeeding  $C_n$ 's are the amplitudes of the respective harmonics having frequencies corresponding to  $n = 2, 3, 4$ , and so on, with respect to the fundamental. The phase angle of the fundamental with respect to a chosen time reference axis is  $\phi_1$ , and the succeeding  $\phi_n$ 's are the phase angles of the respective harmonics. See FOURIER SERIES AND TRANSFORMS.

The equation for  $y(t)$  in general, includes an infinite number of terms. In order to represent a given nonsinusoidal wave by a Fourier series, it is necessary to evaluate each term, that is,  $B_0$ , and all  $A_n$ 's and all  $B_n$ 's. In practice, the first several terms usually yield an approximate result sufficiently accurate for portrayal of the actual wave. The degree of accuracy desired in representing faithfully the actual wave determines the number of terms that must be used in any computation.

The constant term  $B_0$  is found by computing the average amplitude of the actual wave over one cycle. Assuming any reference time  $t = 0$  on the wave,  $B_0$  is given by Eq. (4), where the angle  $\omega t$  is replaced

$$\begin{aligned} B_0 &= \frac{1}{T} \int_0^T y(t) dt = \frac{1}{2\pi} \int_0^{2\pi} y(\omega t) d\omega t \\ &= \frac{1}{2\pi} \int_0^{2\pi} y(\theta) d\theta \end{aligned} \quad (4)$$

by  $\theta$ . Since  $B_0$  is a constant, or dc, term, it merely raises or lowers the entire wave and does not affect its shape.

The coefficients of the sine series are obtained by multiplying the wave  $y(\theta)$  by  $\sin n\theta$ , integrating this product over a full cycle of the fundamental, and dividing the result by  $\pi$ . Thus,  $A_n$  is given by Eq. (5).

$$A_n = \frac{1}{\pi} \int_0^{2\pi} y(\theta) \sin n\theta d\theta \quad (5)$$

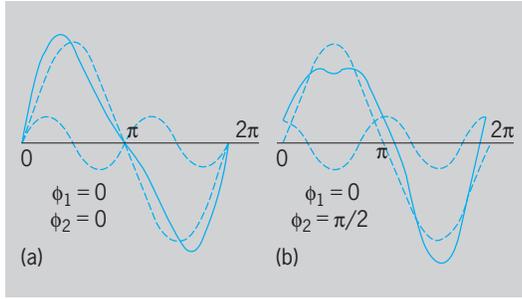


Fig. 1. Addition of a fundamental and a second harmonic. (a)  $\phi_1 = 0, \phi_2 = 0$ . (b)  $\phi_1 = 0, \phi_2 = +\pi/2$  radians.

The coefficients of the cosine terms are obtained in like manner, except that  $\cos n\theta$  replaces  $\sin n\theta$ . Thus,  $B_n$  is given by Eq. (6).

$$B_n = \frac{1}{\pi} \int_0^{2\pi} y(\theta) \cos n\theta \, d\theta \quad (6)$$

If mathematical expressions are required to describe  $y(\theta)$ , Eqs. (4), (5), and (6) give the coefficients of the series directly through analytical methods. If oscillograms or tabulated data describe  $y(\theta)$ , then graphical or tabular forms of integration are used.

**Effect of even harmonics.** Figure 1 shows waves composed of a fundamental and a second harmonic only. In Fig. 1a, both  $\phi_1$  and  $\phi_2$  are zero. In Fig. 1b,  $\phi_1$  is zero and  $\phi_2$  is  $+\pi/2$  radians with respect to one cycle of the second harmonic. In the example given in Fig. 1b the negative part of the overall wave is completely unlike the positive portion. Also, in general, these two portions will have different time intervals. Even harmonics give unsymmetrical waves. See HARMONIC (PERIODIC PHENOMENA).

**Effect of odd harmonics.** Figure 2 shows waves composed of a fundamental and a third harmonic. In Fig. 2a, both  $\phi_1$  and  $\phi_3$  are zero. In Fig. 2b,  $\phi_1$  is zero and  $\phi_3$  is  $+\pi$  radians with respect to one cycle of the third harmonic. In Fig. 2c,  $\theta_1$  is zero and  $\theta_3$  is  $+\pi/2$  radians on the third-harmonic time scale. In these diagrams the negative and positive parts of the overall waves are alike and both embrace  $\pi$  radians

of the fundamental. Odd harmonics lead to symmetrical waves.

**Symmetry.** To determine symmetry of nonsinusoidal waves, the constant term  $b_0$  is first removed. This means shifting the wave down or up by the value of  $B_0$ . After this, if the wave from  $\pi$  to  $2\pi$  is rotated about the horizontal axis and moved forward  $\pi$  radians, and then coincides exactly with the section from 0 to  $\pi$ , the total wave is said to have half-wave symmetry. Often the wave is said merely to be symmetrical. This means that in such cases  $y(\theta + \pi) = -y(\theta)$ . If, in turn, each half of the wave is symmetrical about the vertical axes of  $\pi/2$ , or  $3\pi/2$ , the wave is said to have quarter-wave symmetry as well. Half-wave and quarter-wave symmetry do not necessarily accompany each other.

All three waves of Fig. 2 have half-wave symmetry. The first two have quarter-wave symmetry also, but that of Fig. 2c does not. Waves having only a fundamental and odd harmonics show half-wave symmetry. Conversely, half-wave symmetry indicates that only the fundamental, if it exists, and odd harmonics are present in the total wave. Half-wave symmetry permits Eqs. (5) and (6) to be integrated over the interval  $\pi$ , with the result multiplied by two. Quarter-wave symmetry permits integration over one-quarter cycle of the fundamental, with the result multiplied by four.

Half-wave symmetry means that the fundamental and all odd harmonics may pass through their zero values at times quite distinct from each other. With quarter-wave symmetry the fundamental and the odd harmonics all pass through their zero values at the same time; therefore all phase angles  $\phi_n$  are either zero or  $180^\circ$ .

The wave of Fig. 1a has no symmetry of the kind discussed above. Waves containing only the fundamental and even harmonics, or even harmonics alone, are unsymmetrical. Although half-wave symmetry is absent, quarter-wave symmetry may exist. A wave which is the same from  $\pi$  to  $2\pi$  as it is from 0 to  $\pi$ , that is,  $y(\theta) = y(\theta + \pi)$ , is unsymmetrical. Only even harmonics are present, and the wave has no fundamental component. The output from a full-wave rectifier, for example, contains only the average term  $B_0$  and even harmonics.

Waves that do not meet any of the special conditions noted above can be expected to contain both even and odd harmonics, and probably both sines and cosines also. Any doubt arising on the harmonic content of a wave is resolved by assuming all components of the Fourier series to be present. Analysis will show exactly those that actually exist.

**Even and odd functions.** The time origin of a wave can be chosen arbitrarily. If the reference axis  $t = 0$  is such that the wave to the left is merely the wave to the right rotated about this axis, then  $y(-\theta) = y(\theta)$ , which is said to be an even function. Only cosine terms will be present in the Fourier series for the wave. On the other hand, if the wave to the left is the wave to the right rotated about the  $t = 0$  axis, and then rotated about the horizontal axis,

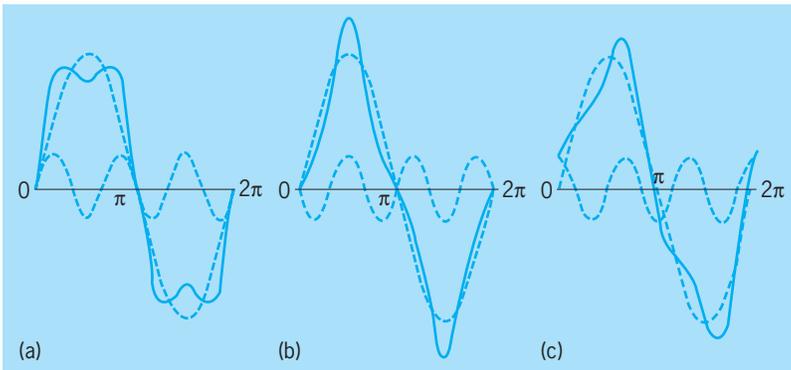


Fig. 2. Addition of a fundamental and a third harmonic. (a)  $\phi_1 = 0, \phi_3 = 0$ . (b)  $\phi_1 = 0, \phi_3 = +\pi$  radians. (c)  $\phi_1 = 0, \phi_3 = +\pi/2$  radians.

$y(-\theta) = -y(\theta)$ , which is said to be an odd function. Only sine terms will appear in the Fourier series. Neither case precludes the possibility of the presence of both even and odd harmonics.

**The rms value of nonsinusoidal wave.** A nonsinusoidal wave has an rms value obtained through the following steps:

1. Combine all terms of the same frequency so as to have a single  $A_1, A_2, \dots, A_n; B_1, B_2, \dots, B_n$ ; or a single  $C_1, C_2, \dots, C_n$ . Terms such as  $\sin(\omega t \pm \alpha)$  and  $\cos(\omega t \pm \beta)$  each have sine and cosine components which can be separated out by trigonometric expansion.

2. Form the series  $y(\theta)$  as in Eq. (1).

3. The rms value of the wave is then given by Eq. (7).

$$y_{\text{rms}} = \left( B_0^2 + \frac{A_1^2}{2} + \frac{A_2^2}{2} + \dots + \frac{A_n^2}{2} + \frac{B_1^2}{2} + \frac{B_2^2}{2} + \dots + \frac{B_n^2}{2} \right)^{1/2}$$

$$= \left( B_0^2 + \frac{C_1^2}{2} + \frac{C_2^2}{2} + \dots + \frac{C_n^2}{2} \right)^{1/2} \quad (7)$$

If  $y_{\text{rms}}$  represents a voltage or a current, this value is shown by an electro-dynamometer or iron-vane voltmeter or ammeter. The rms of the wave is merely the square root of the sum of the squares of the rms values of all of the frequency components. See ROOT-MEAN-SQUARE.

**Power.** An indicating wattmeter with a nonsinusoidal voltage impressed on its potential circuit and a nonsinusoidal current in its current coils indicates the average power taken by the circuit. Designating peak values of the component voltages and currents by  $E_n$ 's and  $I_n$ 's in place of  $C_n$ 's results in Eq. (8). Each

$$\text{Average power} = \frac{1}{2\pi} \int_0^{2\pi} ei \, d\theta$$

$$= E_0 I_0 + \frac{E_1 I_1}{2} \cos \theta_1 + \dots + \frac{E_n I_n}{2} \cos \theta_n \quad (8)$$

coefficient is simply the product of arms voltage and current. No cross-product terms involving different frequencies result from the integration. That is, no power can be contributed by a voltage of one frequency and a current of another frequency.

**Power factor.** The apparent power taken by a circuit carrying nonsinusoidal voltage and current may be defined as the product of the rms values of these quantities. A power factor for such a case may be defined only by the ratio of the average power to the apparent power. Thus, the power factor is given by Eq. (9). Power factor is hence the ratio of instrument

$$\text{Power factor (pf)} = \frac{\text{watts average power}}{\text{rms volts} \times \text{rms amperes}} \quad (9)$$

readings as stated. All circuits have a power factor, but  $\text{pf} = \cos \theta$  only for a sine wave voltage and current of the same frequency. There is no average or representative phase angle for a circuit carrying nonsinusoidal waves. See POWER FACTOR.

**Example of nonsinusoidal waves.** Assume a series circuit to have 8 ohms resistance and 15.91 millihenries inductance and that the impressed voltage is given by Eq. (10). The problem is to calculate the

$$e = 100 \sin 377t + 80 \sin 1131t \quad (10)$$

rms voltage and current, the average power, and the power factor. The voltage has a fundamental component of 60 cycles ( $f_1 = \omega_1/2\pi = 377/2\pi$ ) and a third harmonic of 180 cycles ( $f_3 = \omega_3/2\pi = 1131/2\pi$ ). At 60 cycles:

$$X_{L_1} = 377 \times 0.01591$$

$$= 6.0 \text{ ohms inductive reactance}$$

$$Z_1 = \sqrt{8^2 + 6^2} = 10 \text{ ohms impedance}$$

$$I_1 = 100/10 = 10 \text{ A max fundamental current}$$

$$\theta_1 = \arctan (6/8) = 36.87^\circ$$

At 180 cycles:

$$X_{L_3} = 3 \times 6 = 18 \text{ ohms inductive reactance}$$

$$Z_3 = \sqrt{8^2 + 18^2} = 19.70 \text{ ohms impedance}$$

$$I_3 = 80/19.7$$

$$= 4.06 \text{ A max third-harmonic current}$$

$$\theta_3 = \arctan (18/8)$$

$$= 66.06^\circ (= 22.02^\circ \text{ on fundamental scale})$$

The equation for the current is

$$i = 10 \sin (377t - 36.87^\circ)$$

$$+ 4.06 \sin (1131t - 22.02^\circ)$$

$$E_{\text{rms}} = \sqrt{\frac{100^2}{2} + \frac{80^2}{2}} = 90.06 \text{ volts}$$

$$I_{\text{rms}} = \sqrt{\frac{10^2}{2} + \frac{4.06^2}{2}} = 7.63 \text{ amperes}$$

$$\text{Apparent power} = 90.06 \times 7.63$$

$$= 687 \text{ volt-amperes}$$

$$\text{Average power} = I^2 R = 7.63^2 \times 8 = 466 \text{ watts}$$

$$\text{Power factor} = \frac{466}{682} = 0.678$$

See ALTERNATING-CURRENT CIRCUIT THEORY.

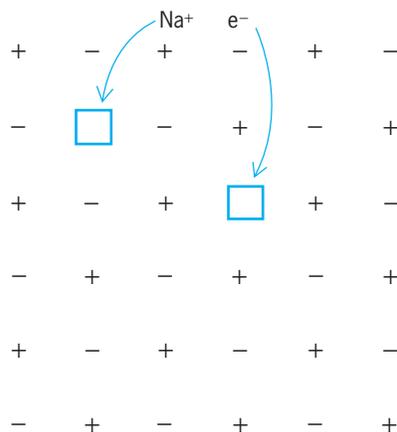
Burtis L. Robertson; W. S. Pritchett

Bibliography. C. K. Alexander and M. N. O. Sadiku, *Fundamentals of Electric Circuits*, 2d ed., McGraw-Hill, 2004; R. C. Dorf and J. A. Svoboda, *Introduction to Electric Circuits*, 6th ed., Wiley, 2003; T. L. Floyd, *Principles of Electric Circuits*, 7th ed., Prentice Hall, 2002; I. D. Mayergoyz and W. Lawson, *Basic Electric Circuit Theory*, Academic Press, 1997; J. W. Nilsson and S. A. Riedel, *Electric Circuits*, 7th ed., Prentice Hall, 2004; H. B. Tilton, *Waveforms: A Modern Guide to Nonsinusoidal Waves and Non-linear Processes*, Prentice Hall, 1986.

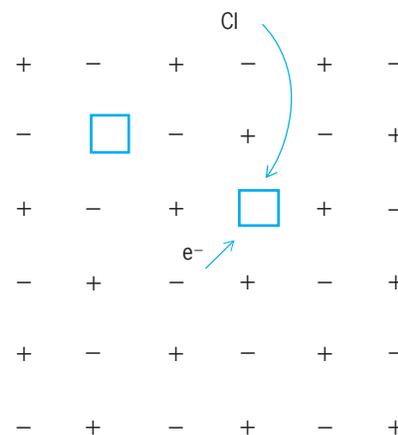
### Nonstoichiometric compounds

Chemical compounds in which the relative number of atoms is not expressible as the ratio of small whole numbers, hence compounds for which the subscripts in the chemical formula are not rational (for example,  $\text{Cu}_{1.987}\text{S}$ ). Sometimes they are called berthollide compounds to distinguish them from daltonides, in which the ratio of atoms is generally simple. Nonstoichiometry is a property of the solid state and arises because a fraction of the atoms of a given kind may be (1) missing from the regular structure (for example,  $\text{Fe}_{1-\delta}\text{O}$ ), (2) present in excess over the requirements of the structure (for example,  $\text{Zn}_{1+\delta}\text{O}$ ), or (3) substituted by atoms of another kind (for example,  $\text{Bi}_2\text{Te}_{3\pm\delta}$ ). The resulting materials are generally of variable composition, intensely colored, metallic or semiconducting, and different in chemical reactivity from the parent stoichiometric compounds from which they are derived. Nonstoichiometry is best known in the binary compounds of the transition elements, particularly the hydrides, oxides, chalcogenides, pnictides, carbides, and borides. It is also well represented in the so-called insertion or intercalation compounds, in which a metallic element or neutral molecule has been inserted in a stoichiometric host. Nonstoichiometric compounds are important in some solid-state devices (such as rectifiers, thermoelectric generators, and photodetectors) and are probably formed as chemical intermediates in many reactions involving solids (for example, heterogeneous catalysis and metal corrosion).

**Generality of the phenomenon.** Simple stoichiometry, in which a compound  $\text{A}_x\text{B}_y$  is characterized by small integral values for the composition parameters  $x$  and  $y$ , is strictly speaking a required characteristic only of molecules in the gaseous state. In the condensed state (solid or liquid), unless the simple molecular aggregate of the gas phase clearly retains its identity,  $x$  and  $y$  no longer need be small integers. Indeed, they may be as large as the number of atoms in a crystal, for example,  $10^{22}$ . Given such large



**Fig. 1.** Formation of  $\text{Na}_{1+\delta}\text{Cl}$  by incorporation of a neutral sodium atom as  $\text{Na}^+$  at a vacant cation site and an  $e^-$  at a vacant anion site. The  $+$ 's represent sodium ions, and the  $-$ 's represent chloride ions.



**Fig. 2.** Formation of  $\text{NaCl}_{1+\delta}$  by incorporation of a neutral chlorine atom at a vacant anion site and migration of an electron from some chloride ion in the normal structure of the crystal.

numbers, small but detectable departures from a simple  $x:y$  ratio can be achieved, even in cases such as  $\text{NaCl}$ , without seriously affecting the energetics of the system, provided a mechanism exists for keeping the compound electrically neutral. In the case of sodium chloride, there is no a priori requirement that the number of  $\text{Na}$  atoms and  $\text{Cl}$  atoms be exactly identical. As a matter of fact, it is relatively simple to heat a colorless stoichiometric crystal of composition  $\text{NaCl}$  in sodium vapor and convert it to yellow-brown  $\text{Na}_{1.001}\text{Cl}$ . Similarly, heating it in chlorine gas can produce  $\text{NaCl}_{1+\delta}$ . Both of these deviations from stoichiometry result from Schottky and Frenkel defects, such as those which are native in any real crystal. See CRYSTAL DEFECTS; SOLID-STATE CHEMISTRY.

Take-up of excess  $\text{Na}$  can be achieved, as shown in **Fig. 1**, by accommodating  $\text{Na}^+$  at a cation vacancy and  $e^-$  at an anion vacancy. Excess chlorine is similarly accommodated, as shown in **Fig. 2**, by incorporating the added  $\text{Cl}$  atom at an anion vacancy. The electron absence, or hole, which distinguishes the  $\text{Cl}^0$  from the normal  $\text{Cl}^-$  ions of the structure, can jump around to any of the other atoms on the anion sublattice. The amount of excess sodium or the amount of excess chlorine that can be accommodated, hence the maximum deviation from stoichiometry, is related to the number of defects in the stoichiometric crystal (intrinsic disorder). Since this intrinsic disorder is an increasing function of temperature, the deviation from stoichiometry that can be achieved is strongly dependent on the preparation conditions, specifically, the temperature at which the preparation is made and the pressure of the constituent elements. Furthermore, since nonstoichiometric compounds are in general prepared by quenching from relatively high temperatures (where atom mobility may be quite high) to room temperature (where diffusion motion is practically negligible), the materials will not usually be in thermodynamic equilibrium with their component elements at room temperature. If the quenching is not efficient, the room-temperature composition will not

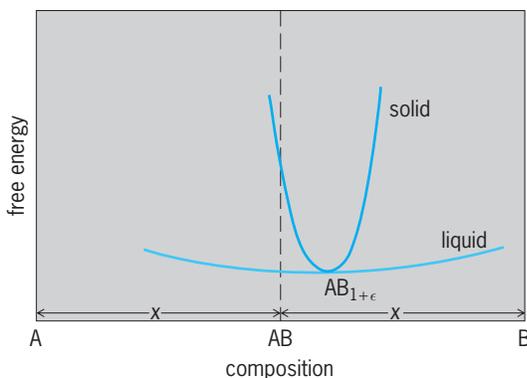


Fig. 3. Free energy versus composition for a two-element system forming the nonstoichiometric compound  $AB_{1+\epsilon}$ . Temperature and pressure are constant.

correspond to an equilibrium state at the preparation temperature, either.

**Thermodynamic considerations.** A thermodynamic description for formation of a nonstoichiometric compound at a fixed temperature is illustrated in Fig. 3, which shows possible free-energy relations for the binary system composed of elements A and B forming nonstoichiometric compound  $AB_{1+\epsilon}$ . The particular case shown corresponds to one in which the compound is completely dissociated in the liquid; that is, there are no specific interactions between A and B in the liquid. At the melting-point equilibrium the chemical potential of each component must be identical in the liquid and solid states. Furthermore, to merit the designation “compound” (as distinct from “solid solution” or “peritectic”), the chemical composition of the solid and liquid phases coexisting at the melting point must be identical. As shown in Fig. 3, this equality is fulfilled not at the stoichiometric composition AB but at  $AB_{1+\epsilon}$ . This comes about as follows: If in the liquid A and B atoms are completely independent of each other, then starting from AB replacement of some A for B or B for A will have no influence on the enthalpy  $H$  and only a slight influence in decreasing the entropy  $S$ . The free energy  $G = H - TS$  of the liquid will show a rather broad minimum as the composition is varied. In the solid the specific interactions are larger and more important. Replacement of A for B or B for A sharply raises the enthalpy, and the result is a deep minimum in the free-energy curve of the solid. Furthermore, the curve for the solid is generally not symmetric about the ideal composition AB. In the case shown, replacement of A for B raises the enthalpy more than does the replacement of B for A. As a result, the minimum lies in the B-rich region, and the equilibrium compound will be  $AB_{1+\epsilon}$ . (A similar but parallel argument can be drawn for the case where AB initially has a finite concentration of vacant lattice sites due to Schottky or Frenkel disorder.) See CHEMICAL THERMODYNAMICS.

**Phase relations.** Figure 4 shows possible phase relations involving formation of the compound  $AB_{1+\delta}$  from the melt. The solid  $AB_{1+\epsilon}$  will crystallize on cooling a melt containing A and B in the ratio  $1:1 + \epsilon$ .

On the other hand, if the melt contains, for example, stoichiometric amounts of A and B, the solidification temperature will be less than  $T_{\max}$ , and the solid first separating will not be  $AB_{1+\epsilon}$  but  $AB_{1+\epsilon'}$  ( $\epsilon' < \epsilon$ ). As crystallization proceeds, the liquid composition would change along  $L_1$  and  $\epsilon'$  would get progressively smaller, eventually even becoming negative. In practice, to keep crystallizing a solid of fixed composition other than  $AB_{1+\epsilon}$ , means would have to be provided to keep the liquid composition constant. See PHASE EQUILIBRIUM.

For simple ionic solids, particularly where the ionic polarizabilities are low and a multiplicity of oxidation states is not possible, the range of nonstoichiometry is generally small. Thus alkali or alkaline earth fluorides are notoriously hard to make as nonstoichiometric compounds. Covalent solids, on the other hand, are generally easier to deviate from stoichiometry, especially when the constituent elements are similar to each other. Bismuth telluride, for example, can be made either bismuth-rich or tellurium-rich by simply putting Bi atoms on Te sites or vice versa. (The unexpected  $p$ -type semiconductivity of  $Bi_2Te_{3-\delta}$  is attributable to the fact that the antistructure defect, Bi on a Te site, has placed a five-valent atom on a site calling for six valence electrons, thereby creating a hole.) Transition elements are particularly good at forming nonstoichiometric compounds, partly because of the capability of the ions to exist in several oxidation states and partly because there is an increased possibility of using  $d$  orbitals either for covalent bonding or for delocalizing electronic charge through  $d$ -orbital overlap. Still, many of the broad, reportedly homogeneous regions in which transition-metal-compound composition appears to be continuously variable actually contain within themselves a series of closely related compounds. As an example, the system  $TiO_x$  ( $1.7 < x < 1.9$ ) actually includes at least seven discrete oxides, each of which can be described by the common formula  $Ti_nO_{2n-1}$ , with  $n$  ranging from 4 to

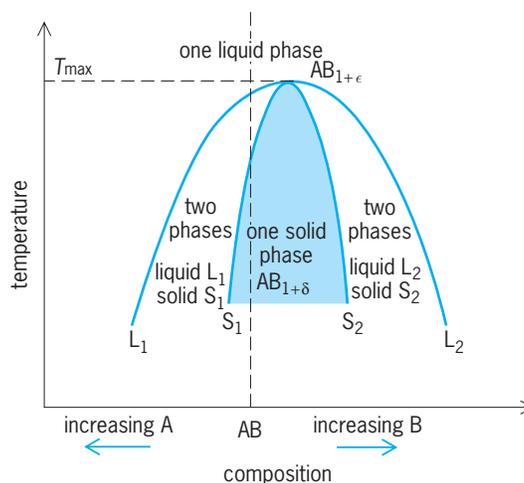


Fig. 4. Temperature-composition plot showing phase relations for formation of the nonstoichiometric compound  $AB_{1+\epsilon}$ . The tinted region corresponds to a single solid phase of variable composition.

10. Similar series have been recognized in  $W_nO_{3n-1}$ ,  $W_nO_{3n-2}$ , and  $MoO_{3n-1}$ .

Even single-crystal compositions that do not fit into such homologous series can frequently be explained as arising from the presence of clusters of defects. X-ray, neutron, and electron diffraction techniques have been supplemented by the technique of lattice imaging in transmission electron microscopy to gain information about the structural nature of defect clusters. One such defect cluster, which appears particularly often, is the crystallographic shear plane, also known as the Wadsley defect. It arises when, along a plane in the crystal, there is a change in the type of polyhedral grouping, as, for example, from corner sharing of octahedra to edge sharing. Each crystallographic shear plane introduces a departure from the stoichiometry of the parent structure; ordered sets of parallel crystallographic shear planes at fixed repetitive spacings can lead to block or column structures as found in different homologous series. In some cases, crystallographic shear planes parallel to two different crystallographic directions, for example,  $\{120\}$  and  $\{130\}$ , may coexist. Given this kind of possible complexity in structure, it is clear that the apparent ranges of homogeneity of phases may well depend upon the wavelength of the radiation used to examine them.

**Classification.** The simplest way to classify nonstoichiometric compounds is to consider which element is in excess and how this excess is brought about. A classification scheme largely based on this distinction but which also includes some examples of ternary systems is as follows.

*Binary compounds:*

- I. Metal:nonmetal ratio greater than stoichiometric
  - a. Metal in excess, for example,  $Zn_{1+\delta}O$
  - b. Missing nonmetal, for example,  $UH_{3-\delta}$ ,  $WO_{3-\delta}$
- II. Metal:nonmetal ratio less than stoichiometric
  - a. Metal-deficient, for example,  $Co_{1-\delta}O$
  - b. Nonmetal in excess, for example,  $UO_{2+\delta}$
- III. Deviations on both sides of stoichiometry, for example,  $TiO_{1\pm\delta}$

*Ternary compounds (insertion compounds):*

- IV. Oxide "bronzes," for example,  $M_\delta WO_3$ ,  $M_\delta V_2O_5$
- V. Intercalation compounds, for example,  $K_{1.5+\delta}MoO_3$ ,  $Li_\delta TiS_2$

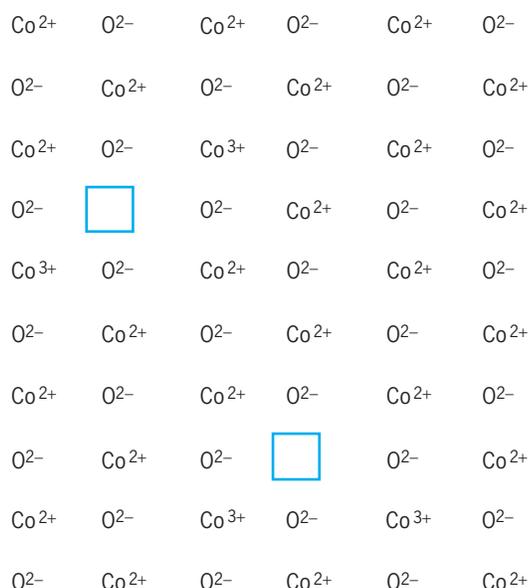
(Excluded from consideration are the recognized impurity materials, such as  $Na_{1-2x}Ca_xCl$ , which are best considered as conventional solid solutions wherein ions of one kind and perhaps vacancies have replaced an equivalent number of ions of another kind.) The specific compounds listed here were chosen to be illustrative of structures and properties commonly encountered in nonstoichiometry. They are discussed individually below.

*Zinc oxide,  $Zn_{1+\delta}O$ .* Rather small deviations from stoichiometry can be obtained by heating zinc oxide crystals in zinc vapor at temperatures of the order

of 600–1200°C (1100–2200°F). The crystals become red and their room-temperature conductivity is considerably enhanced over that of stoichiometric ZnO. The red color and the increased conductivity are attributed to interstitial zinc atoms. Since the ZnO structure is rather open, there is ample room for extra Zn atoms in the tunnels of the  $P6_3mc$  structure. The activation energy for diffusion of the Zn is only 0.55 eV, supporting the belief the nonstoichiometry arises from interstitial zinc and not from oxygen vacancy. The conductivity of  $Zn_{1+\delta}O$  is *n*-type corresponding to action of Zn as a donor. Hall measurements indicate only one free electron from each Zn atom, thus suggesting  $Zn^+$  ions. Similar properties can be produced by heating ZnO in hydrogen gas, but the cause appears to be the addition of H atoms, probably as OH.

*Uranium hydride,  $UH_{3-\delta}$ .* Uranium trihydride does not deviate from stoichiometry to any measurable degree at room temperature but does so to a significant degree at high temperatures. For example, at 450°C (840°F) the existence range is  $UH_{2.98-3.00}$  and at 800°C (1500°F),  $UH_{0.9-3}$ . The hydrogen deficiency comes about from hydrogen vacancies. The interaction energy between the vacancies is rather high (4.3 kcal/mole), as there is a great tendency for the vacancies to cluster and cause nucleation of the metal phase as the temperature is lowered.

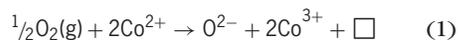
*Tungsten trioxide,  $WO_{3-\delta}$ .* The tungsten-oxygen system shows a variety of materials corresponding to the general formula  $WO_x$ , but many of these appear to be discrete compounds, such as  $WO_{2.90}$  (which is  $W_{20}O_{58}$ ) and  $WO_{2.72}$  (which is  $W_{18}O_{49}$ ), or mixtures of compounds belonging to general series such as  $W_nO_{3n-2}$ . Still, it is possible to prepare single crystals of  $WO_{3-\delta}$ , where  $\delta$  is continuously variable, provided that  $\delta$  is very small. One way in which this can be done is to anneal a stoichiometric crystal at, for example, 1100°C (2000°F) for several days in a controlled oxygen pressure. The oxygen pressure can be set by using an appropriate flow ratio of argon and oxygen, or for very low values a mixture of carbon monoxide and carbon dioxide. Although stoichiometric  $WO_3$  is pale yellow, removal of oxygen darkens the crystals first through green, then blue-green, and finally black. The conductivity is *n*-type. Removal of oxygen produces oxygen vacancies, each of which can act as an electron donor. (This can be seen by considering that removal of an oxide ion  $O^{2-}$  from  $WO_3$  would disturb electric neutrality, so the two minus charges, or two electrons, would have to be put back into the lattice for each vacancy created.) At low oxygen defect, for example,  $WO_{2.999}$ , the oxygen vacancies are independent of each other and randomly distributed in the structure. As the defect increases, the oxygen vacancies begin to line up, coalesce, and produce shear planes where the octahedral  $WO_6$  units share edges rather than corners. In  $WO_{2.994}$ , the average spacing between the crystallographic shear planes, which are along  $\{120\}$ , would be about 28 nanometers. However, they are observed to be quasi-ordered at spacings of 4–5 nm. The oxygen defect in  $WO_{3-\delta}$  has been measured with great precision by



**Fig. 5. Cobalt-deficient cobaltous oxide, Co<sub>1-δ</sub>O.** The squares represent missing cobalt atoms. Note the presence of two Co<sup>3+</sup> to compensate for each Co<sup>2+</sup> missing.

determining the reducing equivalence of a given sample, as by reaction with Ag(SCN)<sub>4</sub><sup>3-</sup> to form elemental Ag.

*Cobaltous oxide, Co<sub>1-δ</sub>O.* Stoichiometric CoO cannot be made. All preparations whether by dehydration of Co(OH)<sub>2</sub>, decomposition of CoCO<sub>3</sub>, or controlled oxidation of cobalt wire lead to a product that is deficient in cobalt. As shown in Fig. 5, the electrical neutrality is maintained by having the doubly positive charge of each missing Co<sup>2+</sup> ion taken up by the existence of two Co<sup>3+</sup> ions someplace in the structure. The resulting material is a good *p*-type conductor in which charge transport occurs by transfer of a “hole” from a Co<sup>3+</sup> to a neighboring Co<sup>2+</sup>. Since the conductivity depends on the concentration of Co<sup>3+</sup>, the electric properties of Co<sub>1-δ</sub>O are dependent on the oxygen pressure. [In fact, they have been used as a way of measuring oxygen pressures over the enormous range of 10<sup>-1</sup> to 10<sup>-13</sup> atm (or 10<sup>4</sup> to 10<sup>-8</sup> pascals).] How this comes about can be seen from the following considerations: Oxidation of CoO by oxygen can be viewed as resulting from (1) incorporation of an oxygen atom at a surface site, (2) transfer of an electron from each of two interior Co<sup>2+</sup> ions to the neutral oxygen to form a normal oxide ion, and (3) migration of a Co<sup>2+</sup> to the surface and creation of Co<sup>2+</sup> vacancy in the interior. The net reaction can be written as (1), where □ represents the



additional cation vacancy created. The condition for equilibrium is shown as Eq. (2) where  $P_{O_2}$  represents pressure of the oxygen gas.

$$\frac{[O^{2-}][Co^{3+}]^2[\square]}{P_{O_2}^{1/2}[Co^{2+}]^2} = K' \quad (2)$$

Since the concentrations of oxide ions [O<sup>2-</sup>] and of

cobaltous ions [Co<sup>2+</sup>] are practically constant, they can be absorbed in  $K$ . The result is Eq. (3). The chem-

$$\frac{[Co^{3+}]^2[\square]}{P_{O_2}^{1/2}} = K' \quad (3)$$

ical equation shows that one vacancy is created for every two cobaltic ions formed, so assuming there were negligible vacancies at the start, the concentration of vacancies in the final crystal is just half that of the cobaltic ions, or Eq. (4). Substituting and solving for [Co<sup>3+</sup>], one gets Eq. (5).

$$[\square] = \frac{1}{2}[Co^{3+}] \quad (4)$$

$$[Co^{3+}] = \left(2K'P_{O_2}^{1/2}\right)^{1/3} \text{ or } \sim P_{O_2}^{1/6} \quad (5)$$

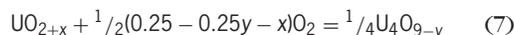
Since the conductivity is directly proportional to the concentration of Co<sup>3+</sup> ions, it follows that it will show a one-sixth power dependence on oxygen pressure.

If the concentration of vacancies at the start is not negligible but on the contrary intrinsically large, then [□] is also a constant. In such a case, the conductivity would follow a 1/4th power dependence on oxygen pressure. In practice, plots of log conductivity versus log  $P_{O_2}$  show slopes between 1/4 and 1/6 depending on the biographical defects in the specimen.

*Uranium dioxide, UO<sub>2+δ</sub>.* The fluorite structure in which UO<sub>2</sub> crystallizes is capable of great defect concentration; still the nonstoichiometry range accessible UO<sub>2.00-2.25</sub> is exceptionally broad. Several independent lines of evidence suggest there are two kinds of defects in this range—oxygen interstitials in the first part, UO<sub>2+x</sub>, and oxygen vacancies in the second part, U<sub>4</sub>O<sub>9-y</sub>. Precision lattice-constant determination gives two linear segments for the lattice parameter  $a$ , as illustrated by relationships (6).

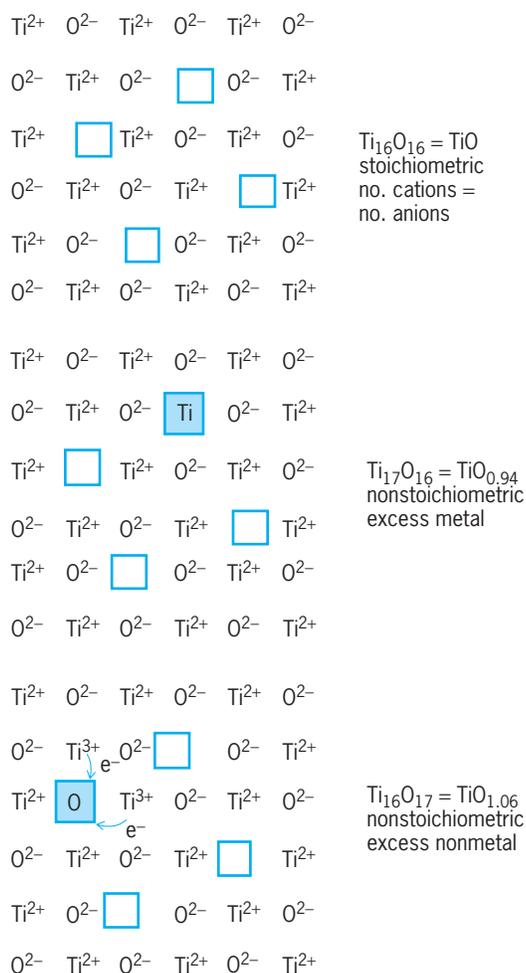
$$\begin{aligned} a &= 5.4705 - 0.094x \text{ for } 0 < x < 0.125 \\ a &= 5.4423 + 0.029y \text{ for } 0 < y < 0.31 \end{aligned} \quad (6)$$

Similarly, a plot of the partial molal free energy of oxygen in UO<sub>2+δ</sub> as a function of  $\delta$  is best fitted by two straight lines intersecting at UO<sub>2.125</sub>. Finally, a determination of the entropy change  $\Delta S$  for reaction (7) shows  $-\Delta S$  tending to very high values



[about 200 entropy units (eu) per mole of O<sub>2</sub> reacting] as the composition of the product approaches U<sub>4</sub>O<sub>9</sub>. Normally, the entropy loss accompanying the fixing of 1 mole of O<sub>2</sub> is only about 35 eu; therefore it is believed that considerable ordering of the interstitial oxygens must be occurring on conversion from UO<sub>2+x</sub> to U<sub>4</sub>O<sub>9-y</sub>. X-ray and neutron diffraction studies support such a model.

*Titanium monoxide, TiO<sub>1±δ</sub>.* This material is interesting because of the wide range of composition possible, from TiO<sub>0.85</sub> to TiO<sub>1.18</sub>, and because of the metallic character observed. The structure is cubic over the whole range and is frequently described as rock salt,



**Fig. 6.** How nonstoichiometry arises in  $TiO_{1\pm\delta}$ . The squares represent vacant lattice sites, and the tint indicates added atoms.

though ordering of the cation and anion vacancies makes it more like spinel. **Figure 6** shows schematically how the Ti/O ratio can deviate both above and below unity because of vacancy imbalance. Stoichiometric TiO has no imbalance but an unusually large number of vacant lattice sites, amounting to about 15%. This can be determined by noting that the observed density is less than that calculated on the basis of measured x-ray spacings. If TiO is heated in various oxygen pressures, either above or below that corresponding to equilibrium with the stoichiometric material, excess oxygen or excess titanium atoms can be incorporated at the vacant sites. If neutral oxygen is added, the dinegative oxide-ion charge is obtained by converting two  $Ti^{2+}$  ions to  $Ti^{3+}$  ions; if titanium is added, the  $Ti^0$  moves to a cation site and forms  $Ti^{2+}$  by donating two electrons to the metallic orbitals. The metallic behavior of  $TiO_{1\pm\delta}$  is attributable to electron delocalization resulting from overlap of the  $d$  electrons.

**Tungsten bronzes,  $M_\delta WO_3$  and analogs.** The tungsten bronzes are a curious set of compounds, in which alkali metals, alkaline earth metals, copper, silver, thallium, lead, thorium, uranium, the rare-earth elements, hydrogen, or ammonium can be inserted in

a  $WO_3$  structure. In the case of sodium the materials are semiconducting for  $\delta < 0.25$  but metallic for  $\delta > 0.25$ . Colors range from blue to violet to coppery to yellow-gold as  $\delta$  changes from 0.4 to 0.98. The materials are remarkably inert to most reagents at room temperature. They can be prepared by electrolysis of molten  $Na_2WO_4 \cdot WO_3$  mixes; heating of  $Na_2WO_4$ ,  $WO_3$ , and W; and vapor-phase reaction of Na with  $WO_3$ .

The cubic sodium tungsten bronzes,  $Na_\delta WO_3$  ( $\delta > 0.43$ ), have a simple structure in which a cube containing tungsten atoms at the corners and oxygen atoms at the edge centers is random-statistically occupied by a sodium atom in the cube center. The sodium atom transfers its electron to the conduction band of the  $WO_3$  host structure, a conduction band which may be formed by overlap of the  $5d_{t_{2g}}$  orbitals of the tungsten atoms or by interaction of these orbitals with the oxygen orbitals. Rubidium and potassium form hexagonal tungsten bronzes which become superconducting at temperatures in the range 2–6 K, depending on composition.

Vanadium bronzes,  $M_\delta V_2O_5$ , are analogous compounds based on insertion of metals in  $V_2O_5$ . However, unlike the tungsten bronzes, which are generally metallic and show small temperature-independent paramagnetism, they are semiconductors and normally paramagnetic.

Molybdenum bronzes,  $M_\delta MoO_3$ , are intermediate in properties. For example,  $K_{0.30} MoO_3$  has a low temperature-independent magnetic susceptibility, semiconductive behavior below 180 K, and metallic behavior above 180 K. Somewhat startling in this material is the finding that  $n$ -type behavior in the semiconducting range changes to  $p$ -type conductivity in the metallic range, opposite to what is generally observed for semiconductors as temperature rises. Titanium bronzes,  $Na_\delta TiO_2$ , and platinum bronzes,  $Na_\delta Pt_3O_4$ , have also been reported.

**Intercalation compounds.** The intercalation compounds include (a) the clathrates, where guest molecules occupy isolated cavities of the host (for example, gas hydrates such as  $Cl_2 \cdot xH_2O$ ,  $5.75 < x < 7.66$ ); (b) tunnel compounds, where molecules such as hydrocarbons fit into tunnels of the host structure (for example, urea); (c) layer compounds, where molecules such as ammonia or alkali metal atoms are inserted between the layers of a transition-metal dichalcogenide; and (d) the zeolites, or molecular sieves, where guest molecules move through a three-dimensional network of tunnels. In all these cases, saturation of site occupancy would lead to stoichiometric products; however, the ratio of guest to host is generally less than saturation and is variable, so the materials can be regarded as nonstoichiometric compounds. The potential number of combinations is enormous. More than 100 intercalation hosts are already recognized, and just one of these, graphite, can accommodate more than 12,000 types of guest.

Typical of the inorganic intercalates is  $Li_\delta TiS_2$  ( $0 < \delta < 1$ ). The host,  $TiS_2$ , has a layered structure in which a sheet of titanium atoms is sandwiched

between two sheets of hexagonally close-packed sulfur atoms. Adjacent sulfur sheets are weakly bonded to each other by van der Waals forces, and guest species can move into this van der Waals gap. Intercalation with lithium can be achieved by exposing  $\text{TiS}_2$  to lithium vapor, to lithium dissolved in liquid ammonia, or to *n*-butyl lithium dissolved in a nonpolar solvent. It can also be achieved by electrointercalation, using a cell in which lithium metal is the anode,  $\text{TiS}_2$  is the cathode, and the electrolyte is a lithium salt dissolved in an organic solvent such as propylene carbonate. Intercalation proceeds with an expansion of the lattice *c* parameter by about 0.07–0.085 nm. The expansion increases with lithium content. Nuclear magnetic resonance studies of  $\text{Li}_x\text{TiS}_2$  indicate that the lithium is quite mobile and is almost 100% ionic. The use of  $\text{Li}_x\text{TiS}_2$  in lithium batteries is possible. See INTERCALATION COMPOUNDS.

Michell J. Sienko

Bibliography. C. Catlow and W. Mackrodt (eds.), *Nonstoichiometric Compounds*, 1987; L. Eyring and M. O'Keefe (eds.), *The Chemistry of Extended Defects in Non-Metal-Imperfect Crystals*, vols. 1–3, 1974–1975; P. Kofstad, *Nonstoichiometry, Diffusion and Electrical Conductivity in Binary Metal Oxides*, 1972, reprint 1983.

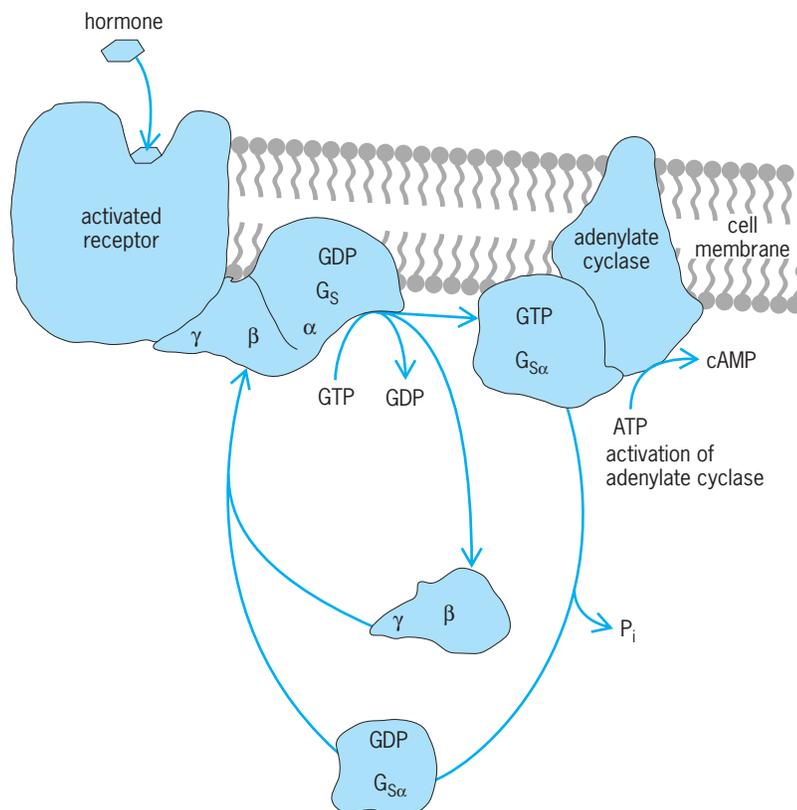
## Noradrenergic system

A neuronal system that is responsible for the synthesis, storage, and release of the neurotransmitter norepinephrine. Norepinephrine, also known as noradrenalin, consists of a single amine group and a catechol nucleus (a benzene ring with two hydroxyl groups) and is therefore referred to as a monoamine or catecholamine. It exists in both the central and peripheral nervous systems. In the autonomic system in the periphery, norepinephrine is the primary neurotransmitter released by the sympathetic nervous system. The sympathetic nervous system mediates the “fight or flight” reaction, preparing the body for action by affecting cardiovascular function, gastrointestinal motility and secretion, bronchiole dilation, glucose metabolism, and so on. Within the central nervous system, norepinephrine has been associated with several brain functions, including sleep, memory, learning, and emotions. The major clustering of norepinephrine-producing neuron cell bodies in the central nervous system is in the locus coeruleus. This center, located in the pons with extensive projections throughout the brain, produces more than 70% of all brain norepinephrine.

**Norepinephrine synthesis, storage, release.** Norepinephrine is synthesized through a succession of steps. The amino acid tyrosine is first transported from blood into noradrenergic neurons, where it is transformed to dopa. Dopa is synthesized to dopamine, which is also a catecholamine neurotransmitter. In dopaminergic cells the pathway stops at this point, whereas in noradrenergic neurons the dopamine is converted to norepinephrine. In adrenergic cells the pathway continues for one more step

to synthesize a third catecholamine, epinephrine (also known as adrenaline). Synthesis of all of the catecholamines is limited by tyrosine hydroxylase activity that converts tyrosine to dopa. Although at physiological concentrations of tyrosine the enzyme is saturated, the activity of the enzyme is limited by the availability of a cofactor, bipterin. Due to its rate-limiting nature in this pathway, tyrosine hydroxylase is a good candidate for regulation by end-product concentrations and by phosphorylation of the enzyme. After synthesis, the majority of norepinephrine is transported into synaptic vesicles in the nerve terminals, where it remains until needed. When the nerve terminal is activated by depolarization, calcium flows into the nerve terminal, leading to the docking of the vesicles onto the presynaptic membrane and release of norepinephrine into the synaptic cleft. See EPINEPHRINE.

**Receptors and regulation.** Once released into the synaptic cleft, norepinephrine is free to bind to specific receptors located on the presynaptic or postsynaptic terminal, which mediate various effects on either the presynaptic or postsynaptic cells. Binding of norepinephrine to a receptor initiates a chain of events (the effector system) in the target cell that can be mediated by a number of different second messenger systems. The exact effect is determined by the identity of the receptor activated. All noradrenergic receptors are linked to G proteins (see *illus.*).



Regulatory cycle of G-protein-mediated signal transduction. Hormone binds to the receptor, triggering the exchange of GDP for GTP. The  $\alpha$  subunit–GTP dissociates from the  $\beta\gamma$  subunit. The GTP-liganded  $\alpha$  subunit activates several processes (adenylate cyclase activation);  $\beta\gamma$  subunits regulate shared or distinct effectors. (After D. Nelson and M. Cox, *Lehninger Principles of Biochemistry*, 3d ed., 2000)

Binding of norepinephrine to a receptor leads to activation of the G protein, which in turn modulates the activity of another protein (enzyme or ion channel) in the cell. Noradrenergic receptors can be broken down into three categories:  $\alpha 1$ ,  $\alpha 2$ , and  $\beta$  receptors. These alpha and beta categories are constantly being subdivided with the explosion in molecular biology, allowing investigators to identify small differences in receptor structure that may have little impact on the physiological effect of the receptor. For example, there are currently four different  $\alpha 1$  receptors:  $\alpha 1A$ ,  $1B$ ,  $1C$ , and  $1D$ . Depending on the subtype of  $\alpha 1$  receptor activated, the ultimate effect may be a modulation of a calcium channel or activation of phospholipase C (the enzyme which converts phospholipids into inositol trisphosphate and diacyl glycerol). All  $\alpha 2$  receptors are linked to the inhibition of adenylate cyclase and thus reductions in cyclic 3',5'-adenosine monophosphate (cAMP). Included in the category of  $\alpha 2$  receptors are presynaptic autoreceptors that regulate further release of norepinephrine. This negative-feedback mechanism acts as a brake on the noradrenergic neuron. When the autoreceptor is stimulated, it decreases noradrenergic firing; when it is blocked, it increases noradrenergic firing. There are three main  $\beta$  receptors that have been identified, all linked to adenylate cyclase stimulation and thus leading to increases in cAMP.

Noradrenergic receptors are highly regulated. Both homologous desensitization due to prolonged activation of specific receptors and heterologous desensitization due to prolonged activation of other receptor subtypes (other noradrenergic receptors or other neurotransmitter receptors) occur. Phosphorylation, of the  $\beta$  receptors in particular, has been well documented as a mechanism of regulating the receptors' activity.

**Inactivation.** Termination of norepinephrine occurs by a high-affinity reuptake mechanism in the presynaptic membrane. The active transport of norepinephrine is fueled by the sodium gradient across the cell membrane. Once transported back into the presynaptic terminal, norepinephrine can be stored in vesicles for future use or enzymatically degraded by monoamine oxidase. Any stray norepinephrine that is not taken up into the terminal is degraded by catechol-O-methyl transferase.

Certain medications achieve their effect by altering various stages of synthesis, storage, release, and inactivation of norepinephrine. The behavioral manifestations of these alterations have led to a better understanding of norepinephrine's role in various psychiatric disorders. It has been suggested that altered noradrenergic function may be of etiological significance in depression, posttraumatic stress disorder, anxiety disorder, and schizophrenia. Interpretation of the pharmacological data has been complicated by the recent studies on neurogenesis. It has been shown that neurogenesis occurs in specific brain areas and that this process is modulated by environmental conditions such as stress or therapeutic

agents used in psychiatric disorders such as modulators of serotonergic activity or noradrenergic activity, and electroconvulsive shock therapy. Thus, the etiology underlying these disorders may be related to a lack or excess of neurogenesis rather than defects in a specific neurotransmitter system.

**Depression.** The monoamine deficiency hypothesis postulates that depressive illness arises from a deficiency of synaptic norepinephrine. This hypothesis is supported by the fact that catecholamine-depleting agents, such as reserpine, cause depression, while agents that increase synaptic norepinephrine, such as tricyclics and monoamine oxidase inhibitors, are effective treatments for depression. However, advances in neuropharmacology have challenged the monoamine deficiency hypothesis on grounds that a more complex model is necessary to explain depression and its treatment. First, certain drugs, including cocaine and amphetamines, that inhibit reuptake of norepinephrine are not effective in treating depression. Second, several "atypical" antidepressants are effective for treating depression but do not block uptake of norepinephrine or affect its breakdown through monoamine oxidase. Third, the inhibitory effects of tricyclic antidepressants and monoamine oxidase inhibitors occur within minutes after drug administration, but the antidepressant effects generally do not occur for 2-3 weeks. *See* MONOAMINE OXIDASE.

Because of the temporal discrepancy between initiation of antidepressants and the occurrence of therapeutic effects, research has turned to the examination of the effects of chronic antidepressant treatment on receptor function. A number of investigators have seen both in vivo and in vitro that chronic treatment with antidepressants and electroconvulsive shock therapy (an alternative treatment for depression) lead to a decrease in  $\beta$  receptors. This correlates well with an increase in  $\beta$  receptors which has been documented in suicide victims with a history of depression. The  $\alpha$  receptors are also decreased with chronic antidepressant treatment and show supersensitivity in suicide victims. These changes in receptors have not been documented by every study. The changes in the noradrenergic system with antidepressant treatment suggest that manipulations of the system can alleviate depression but do not prove that alterations in the system are responsible for the depression in the first place. It is also possible that the stimulatory effects of norepinephrine on neurogenesis may in part explain the temporal discrepancy between onset of therapy and alleviation of symptoms. During neurogenesis it would take weeks to produce a fully functional differentiated neuron. The different neurotransmitter systems in the brain interact in complex ways, making a definitive cause difficult to elucidate. *See* AFFECTIVE DISORDERS.

**Anxiety disorders.** Anxiety disorders are broken down into at least four categories: panic disorder, posttraumatic stress disorder, generalized anxiety disorder, and obsessive-compulsive disorder. The

extent of involvement of the noradrenergic system with each disorder varies. There is no indication that obsessive-compulsive disorder includes changes in the noradrenergic system. Panic disorder and generalized anxiety disorder are the most similar in their neuroendocrine involvement and response to noradrenergic manipulation. As with most psychological disturbances, the changes in the central nervous system are not restricted to changes with a single chemical system or anatomical region. The hypothalamic-pituitary-somatotropin axis appears to have altered regulation, and it responds abnormally to  $\alpha_2$  receptor activation. Patients with panic disorder more frequently show an anxiety reaction in response to  $\alpha_2$  antagonist administration than patients with other psychological abnormalities. Agents known to decrease noradrenergic function such as tricyclic antidepressants and monoamine oxidase inhibitors are often effective treatments for patients with panic disorder or generalized anxiety disorder.

Involvement of the noradrenergic system has also been implicated in the pathophysiology of posttraumatic stress disorder. The immediate response to severe stress includes the activation of the sympathetic nervous system as well as an increase in the firing rate of the locus coeruleus neurons, both of which lead to the release of norepinephrine. These responses are adaptive and protective, allowing a person to cope with the stress. These responses become maladaptive when they are chronically activated due to depletion of norepinephrine stores, changes in noradrenergic receptor numbers, and chronic exposure of brain areas to glucocorticoids. Although posttraumatic stress disorder can occur with just a single exposure to an event, it is assumed that the stress reaction is repeated due to reliving of the trauma, which leads to an exacerbation of the symptoms. Traumatized veterans, when reexposed to combat-associated stimuli, respond with abnormal elevations of autonomic parameters such as blood pressure and heart rate. Elevated levels of norepinephrine in plasma and urine have been reported in severely stressed individuals and in patients with posttraumatic stress disorder. At the same time, platelet  $\alpha_2$  adrenergic receptor numbers, lymphocyte adenylate cyclase activity, and monoamine oxidase levels have been reported decreased. Consistent with these abnormalities in norepinephrine metabolism is the suggestion that several antihypertensive medications, which decrease the effects of norepinephrine, may diminish the symptoms of arousal-related posttraumatic stress disorder. In addition to dysregulation of the noradrenergic system, patients with posttraumatic stress disorder may also have anomalies in their cortisol secretion and dopaminergic systems. See STRESS (PSYCHOLOGY).

**Schizophrenia.** The action of antipsychotic agents on the dopaminergic system has led to many investigations into the role of dopamine in schizophrenia. In 1989 a new type of antipsychotic agent, clozapine, was approved for use in schizophrenia.

Clozapine demonstrated superior efficacy as an antipsychotic agent when compared to traditional neuroleptics, without many of the extrapyramidal side effects. It was shown to increase plasma levels of norepinephrine levels fivefold, as well as increasing the number of spontaneously active cells in the locus coeruleus and increasing the firing rate of locus coeruleus neurons. The  $\alpha_1$  and  $\alpha_2$  antagonist activity of clozapine is insufficient to explain the changes in plasma norepinephrine levels. Evidence from a number of studies suggests the involvement of multiple neurotransmitters in this disease. It has been hypothesized that the effect of increasing norepinephrine alleviates schizophrenia by enhancing synaptic strength in cortical regions responsible for the psychotic symptoms. Thus, changes in the noradrenergic system may not be directly responsible for schizophrenia but may modulate other neurotransmitters that are affected in this disease. In contradiction to the studies with clozapine that demonstrate a correlation between increased plasma norepinephrine levels and alleviation of symptoms, other studies have found elevated levels of norepinephrine in cerebrospinal fluid of schizophrenics, both living and deceased. In addition, medications such as cocaine and amphetamines that are known to increase norepinephrine levels through uptake inhibition can cause psychosis. Receptor studies have shown no changes in  $\alpha$  receptors, with a possible decrease in  $\beta$  receptors, in the hippocampus. See NEUROBIOLOGY; PSYCHOPHARMACOLOGY; SCHIZOPHRENIA.

Michelle Mynlieff; Dennis S. Charney;  
Alan Breier; Steven Southwick

**Bibliography.** D. Goldstein, G. Eisenhofer, and R. McCarty (eds.), *Catecholamines Bridging Basic Science with Clinical Medicine*, vol. 42 in *Advances in Pharmacology*, Academic Press, San Diego, 1998; J. G. Hardman et al. (eds.), *Goodman and Gilman's The Pharmacological Basis of Therapeutics*, McGraw-Hill, New York, 1996; G. J. Siegel et al. (eds.), *Basic Neurochemistry Molecular, Cellular and Medical Aspects*, 6th ed., Lippincott-Raven, Philadelphia, 1999.

## Normal (mathematics)

A term generically synonymous with perpendicular, which often refers specifically to a line that goes through a point  $P$  of a curve  $C$  and is perpendicular to the tangent to  $C$  at  $P$ . If a plane curve  $C$  has equation  $y = f(x)$ , in rectangular coordinates the normal (line) to  $C$  at  $P(x_0, y_0)$  has slope  $-1/f'(x_0)$ , provided  $f'(x_0) \neq 0$ . The expression  $f'(x_0)$  denotes the derivative of  $f(x)$ , evaluated for  $x = x_0$ , and so the normal has equation  $y - y_0 = [-1/f'(x_0)](x - x_0)$ . If curve  $C$  is not a plane curve, all normal lines of  $C$  at point  $P$  on  $C$  lie in a plane, the normal plane of  $C$  at  $P$ . For other uses of normal (for example, normal form of equation of a line or a plane) see ANALYTIC GEOMETRY.

Leonard M. Blumenthal

## North America

The third largest continent, extending from the narrow isthmus of Central America to the Arctic Archipelago. The physical environments of North America, like the rest of the world, are a reflection of specific combinations of the natural factors such as climate, vegetation, soils, and landforms. While the distribution of life-giving solar heat and water seemingly should determine the character and distribution of the natural regions, in reality the landforms modify the expected environmental patterns. In addition, human activities affect the quality of some environments, and natural hazards alter some regions permanently or semipermanently. *See* CONTINENT.

In order to understand the characteristics and distributional patterns of the natural regions of North America, it is necessary to consider the continent's absolute and relative location, and the types and patterns of its climates, vegetation, soils, and landforms.

### Location

North America covers 9,400,000 mi<sup>2</sup> (24,440,000 km<sup>2</sup>) and extends north to south for 5000 mi (8000 km) from Central America to the Arctic. It is bounded by the Pacific Ocean on the west and the Atlantic Ocean on the east. The Gulf of Mexico is a source of moist tropical air, and the frozen Arctic Ocean is a source of polar air. With the major mountain ranges stretching north-south, North America is the only continent providing for direct contact of these polar and tropical air masses, leading to frequent climatically induced natural hazards such as violent spring tornadoes, extreme droughts, subcontinental floods, and winter blizzards, which are seldom found on other continents. *See* AIR MASS; ARCTIC OCEAN; ATLANTIC OCEAN; GULF OF MEXICO; PACIFIC OCEAN.

### Environment

Several environmental factors contribute to the characteristics of the generally recognized natural regions of North America. These include geologic structure, climatic types and regions, vegetation regions, and soil regions.

**Geologic structure.** The North American continent includes (1) a continuous, broad, north-south-trending western cordilleran belt stretching along the entire Pacific coast; (2) a northeast-southwest-trending belt of low Appalachian Mountains paralleling the Atlantic coast; (3) an extensive rolling region of old eroded crystalline rocks in the north-central and northeastern part of the continent called the Canadian Shield; (4) a large, level interior lowland covered by thick sedimentary rocks and extending from the Arctic Ocean to the Gulf of Mexico; and (5) a narrow coastal plain along the Atlantic Ocean and the Gulf of Mexico. These broad structural geologic regions provide the framework for the natural regions of this continent and affect the location and nature of landform, climatic, vegetation, and soil regions.

**Climatic types.** The North American continent, while dominated by broad areas of continental cli-

mates, has regions representing every climate on Earth. Thus, in combination with its location and terrain, the continent is subject to almost every climatic hazard known: intense tornadoes in the southern Great Plains and the Midwest; intense tropical hurricanes in the southeast and in Central America; extreme winter cold and blizzards in the north; droughts of the intensity of the Dust Bowl; and floods.

The North American climatic regions (**Fig. 1**) result from the combined effect of numerous controlling factors that create the specific climates characterizing different parts of the continent. The most important controls of the North American climates are (1) latitude, which controls the distribution of solar energy throughout the year; (2) location of semipermanent high-pressure cells (over the oceans around 30°N latitude); (3) the large size of the continent, which creates continentality in the interior; (4) polar and tropical air masses; (5) wind belts, especially the westerly winds with the cyclonic storms; and (6) mountain barriers. The two air masses that dominate the central part of the continent are the very cold and dry Continental Polar (cP) air mass, whose source region is the Arctic and north central Canada, and the warm and wet Maritime Tropical (mT) air mass, whose source region is the tropical and subtropical ocean, including the Caribbean region. The meeting of these two air masses creates most of the violent weather in the central part of the continent. The other air masses affecting North America are the hot and dry Continental Tropical (cT) in the southwest, and two cool and wet Maritime Polar (mP) air masses originating over the oceans northwest and northeast of the continent. Their effect is more local.

The resulting climates dominating large parts of the continent are polar and high-latitude humid continental (snow) climate, including tundra, subarctic, and humid continental; midlatitude temperate, moist; midlatitude dry-summer (Mediterranean); semiarid and arid (desert); and tropical.

*Polar and High-latitude humid continental.* Also known as snow climates, these include tundra, subarctic and humid continental climates.

The tundra climate has an average monthly mean temperature below 50°F (10°C), and trees do not grow. Precipitation is minimal.

The subarctic climate is located between 50 and 70°N latitude and extends from western Alaska across Canada to the east coast of North America. Its winter is dark, long, and bitterly cold, with averages of -35°F (-38°C). Lakes are frozen from September to May. Summer is very short, but because of long days the temperature increases rapidly, reaching the 50s or low 60s on the Fahrenheit scale (10 or 15°C). Precipitation is meager [5-20 in. (13-50 cm)].

The humid continental climate is located between 35 and 55°N latitude, from the east coast toward the interior of the continent. Dominated by the westerly wind belt and its migratory pressure systems, the weather changes frequently in this climate, both from day to day and from season to season. Daily variability is the major characteristic of this climate

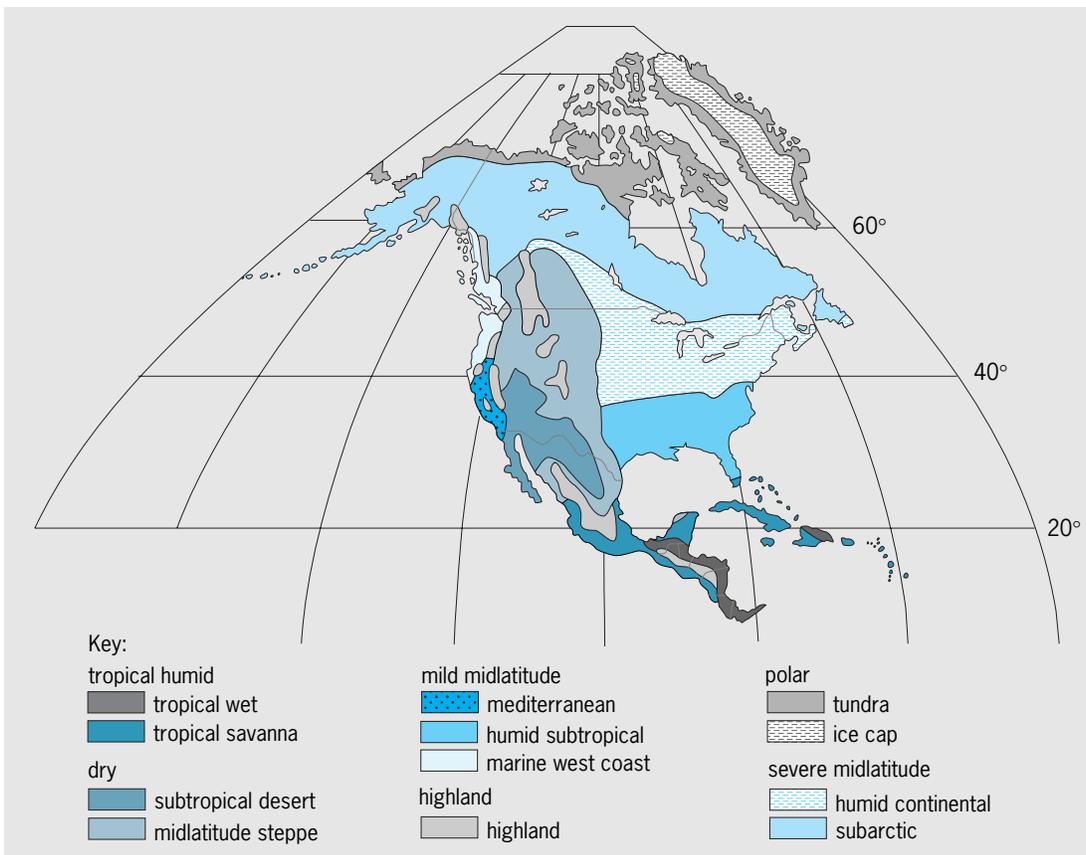


Fig. 1. Climatic regions of North America (modified Köppen system). (After T. L. McKnight, *Physical Geography: A Landscape Appreciation*, 3d ed., Prentice Hall, 1990)

in all seasons, including hazardous winter blizzards, thunderstorms, tornadoes, heat waves and prolonged drought, and great floods.

The summer temperature averages 70°F (20°C) but reaches the upper 80s on the Fahrenheit scale (upper 20s on Celsius) on many days. The average winter temperature is around 15°F (−9°C), and there are about five months with averages below freezing. This is a humid climate with up to 50 in. (125 cm) of monthly rainfall, which diminishes toward the interior. There is more precipitation in summer than in winter. Snow stays on the ground for less than a month in the south and up to five months in the north. See DROUGHT; PRECIPITATION (METEOROLOGY); THUNDERSTORM; TORNADO.

*Mid-latitude.* The temperate climate known as humid subtropical lies between 25 and 40°N latitude and occupies the southeastern part of the continent. The summers are warm to hot—monthly averages reach 85°F (27°C)—with high humidity day and night. The winters are mild, but invasion of cold polar air sometimes brings severe weather for a few days, and killing frost can occur in late spring. Annual precipitation is between 40 and 60 in. (100 and 150 cm), but occasional tropical hurricanes bring very heavy precipitation, high wind, and wave destruction to coastal areas. See FROST; HURRICANE; TROPICAL METEOROLOGY.

*Mediterranean climate.* The regional extent of this climate in North America is small and coastal. This dry

subtropical climate occurs on the western side of the continent between 30 and 40°N latitude. The climate is characterized by a modest amount of precipitation, about 15 in. (38 cm), concentrated in a mild (50°F or 10°C) winter and a very hot (75–85°F or 24–29°C) and dry summer. Abundant sunshine is typical for this climate. Winter rains come from the frontal cyclonic storms of the westerly winds, and summer drought from the effect of the Subtropical High Pressure cell located just offshore.

*Marine West Coast climate.* The marine west coast climate lies north of the Mediterranean climate between 40 and 60°N latitude. It is exceptionally mild for its latitude due to its coastal location and the effect of the oceanic westerly winds. Its average summer temperature is between 60 and 70°F (16 and 21°C) and the average cold-month temperature is between 35 and 45°F (2 and 7°C). This is one of the wettest climates in the middle latitudes, with average rainfall of 30 to 50 in. (75–125 cm) but reaching 200 in. (500 cm) on windward slopes.

*Dry climate.* The semiarid and arid climates are characterized by greater evaporation than precipitation. The semiarid steppe climates have sufficient moisture to support short-grass vegetation. The desert climates are so dry that they support only xerophytic vegetation that can withstand extreme heat. Because of extreme drought, dry climates may experience large diurnal and seasonal temperature differences.

*Tropical climates.* The tropical climates include the tropical wet climate, stretching 5 to 10° latitude on both sides of the equator. It is hot all year, with an average monthly temperature over 64°F (18°C), and wet all year, with 60 to 100 in. (150 to 250 cm) of annual rainfall. The tropical wet and dry (savanna) climate stretches up to 25° latitude and has a distinct wet summer and very dry winter pattern. This type of climate characterizes the southern tip of Florida. Tropical wet, and tropical wet and dry, climates are present in the coastal parts of Mexico and Central America.

**Climatic regions.** In the northern part of the North American continent, the major climatic control is high latitude, affecting the distribution of solar radiation throughout the year. Since in the summer the days are long and the angle of the Sun is reasonably high, solar heat is accumulated. During the rest of the year, and especially in the winter, the low angle of the Sun and short days significantly limit the amount of heat received. As a result, there is a broad northern belt that has four-season humid continental climates with cold and snowy winters and warm to cool summers; the severity of the winters increases northward (for example, north central Canada) and decreases southward. Cyclonic storms bring most of the precipitation. North of these climates stretches the tundra climate, where the average monthly temperatures never rise above 50°F (10°C), snow lies on the ground year-round, and subsoil is permanently frozen (permafrost). See INSOLATION; PERMAFROST.

In the southeastern part of the North American continent, there is no snow cover; the climate is temperate, of humid subtropical type with moist and warm to hot summers and short, cool winters. Here the winds from the Atlantic High Pressure cell sweep heat and moisture from the subtropical oceans onto the land, providing for a warm and moist temperate climate with much summer rain.

Along the west coast of the continent, between 35 and 45°N latitude the Mediterranean climate is very hot and dry in the summer because of the descending and drying air of the Pacific High Pressure cell located off the coast. Winters are much cooler and rainy because the cyclonic storms of the westerly winds move southward to replace the High Pressure cell. Northward of the Mediterranean climate, the coastal belt has the Marine West Coast climate, which has moderate temperatures and copious precipitation all year long, with total annual rainfall reaching 200 in. (500 cm). The combined effect of coastal and windward locations, and the mountain barriers along the Pacific coast, give the Marine Coast climate the highest precipitation in North America outside the tropics.

The remaining part of the continent, located mainly in the west, has a complex mosaic of semidry steppe, desert, and mountain climates. The mountain barriers, including the Coastal Ranges and Sierra Nevada, have deserts on their lee (east) sides, where the descending air becomes very drying. The orographic effect of the Rocky Mountains and the con-

tinental location create a broad north-south belt of dry steppes in the Great Plains. The descending air of the Pacific High Pressure cell creates the driest deserts of North America in the southwestern part of the continent and parts of northern Mexico. See DESERT; MOUNTAIN METEOROLOGY; PLAINS.

**Tropical climatic zones.** The rest of Mexico and Central America have tropical and mountain climates. Since this area lies in tropical latitudes, the diurnal temperature range is small throughout the year. However, since temperature decreases with altitude, there is an altitudinal climatic zonation on high mountain slopes. These zones are called tierra caliente (hot land), tierra templada (temperate land), tierra fria (cold land), and tierra helada (frozen land). Here, weather, climate, vegetation, soils, and land use vary according to elevation.

The tierra caliente extends from sea level to about 2500 ft (750 m) and includes coastal lowlands, plains like the Yucatán Peninsula, and interior basins, like Balsas River valley, covering about half of the tropics. Here, daytime temperatures are about 85–90°F (30–32°C) and nights are about 70–75°F (21–23°C).

The tierra templada lies roughly between 2500 and 5500 ft (750 and 1660 m). This zone includes the intermediate mountain slopes, the Plateau of Central Mexico, and Central America. Mild daytime temperatures of 75–80°F (23–26°C) increase to 85–90°F (30–32°C) in April and May, while nights have pleasant temperatures of about 60–70°F (15–21°C). Northward, the range between summer and winter temperatures increases, and night frost is possible. This zone has much of the population and agricultural production of the tropics.

The tierra fria lies above 5500 ft (1660 m) elevation and represents only about 10% of Middle America. It is present in high basins and on mountain slopes of Mesa Central of Mexico, and Central America. Its day temperatures are about 75–80°F (23–26°C), but night temperatures fall to 50–55°F (10–12°C). Frost is common in cooler months.

The tierra helada lies above 12,000 ft (3630 m). It occupies the highest mountain peaks, has temperatures below 50°F (10°C) all year, and has permanent ice and snow.

In terms of rainfall, northern Mexico, with its dry deserts and steppes, contrasts sharply with its southern part and Central America, which are humid and rainy. Generally, the east coasts have more rain than the west coasts because of the dominance of the easterly trade winds. The hot months (May to October) are more rainy, often receiving 75–80% of annual rainfall. These windward coasts receive between 80 and 120 in. (200 and 300 cm) of rainfall annually; the lee sides of the mountains receive 20–40 in. (50–100 cm) of precipitation. Most of the rain comes from daily thunderstorms, midlatitude cyclones, and tropical hurricanes, especially those from the Atlantic Ocean. Some of the hurricanes are true weather hazards, since they are extremely destructive. See CLIMATOLOGY.

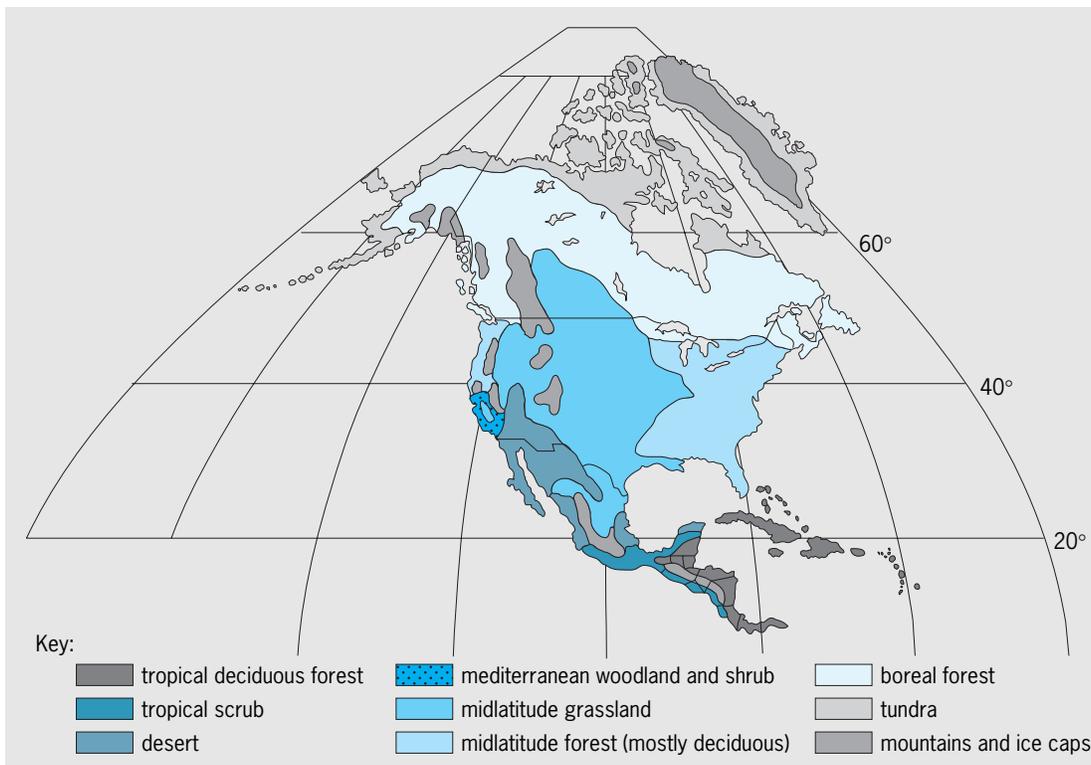


Fig. 2. Natural vegetation regions of North America. (After T. L. McKnight, *Physical Geography: A Landscape Appreciation*, 3d ed., Prentice Hall, 1990)

**Vegetation regions.** A major component of the natural environment directly related to climatic conditions, and of vital importance to humans as an indicator of the potential for food growing, is the natural vegetation. Though vegetation in midlatitudes has often been cleared to convert the land for agricultural use, its nature must be understood as a prerequisite for proper utilization of the land on which it once grew. Comparison of a climatic map (Fig. 1) with a natural vegetation map (Fig. 2) reveals a striking resemblance between the patterns of distribution of these two factors. See ALTITUDINAL VEGETATION ZONES; PLANT GEOGRAPHY.

The zone of the tundra climates is covered by ice or by tundra vegetation. Since this area is really a cold desert, with extreme shortage of water (most of it is frozen), and very short and cool summers with mean temperatures below 50°F (10°C), trees cannot grow. Instead, the tundra plant assemblages consist of dwarf shrubs, mosses, lichens, grasses, and flowering herbs, all of which go through their life cycle in the short summer. The tundra supports reindeer and birds (which migrate southward in the winter), and has enormous numbers of insects. See TUNDRA.

The northern part of the humid continental snowy climate, where mean winter temperatures rise above 50°F (10°C), supports an evergreen needleleaf forest, which is known as the boreal forest in North America. This great northern forest is one of the most extensive biomes in the world, with the simplest assemblage of plants. Almost all the trees are coniferous, needleleaf evergreens, mostly pines, firs, and spruces, although patches of deciduous trees

can be found, such as birch, aspen, and poplar. Along the northern fringe, trees are short and widely spaced, but toward the south they grow taller and are more dense. The undergrowth is generally not dense; mosses and lichens create a ground cover. Drainage is poor in the summer because of frozen subsoil, and bogs and swamps are common. Everything is frozen in the winter. Animal life consists mainly of fur-bearing mammals, and numerous birds in the summer, as well as abundant insects.

The southern zone of the humid continental climate supports the midlatitude deciduous forest, although along its northern boundary the forest is mixed with evergreen conifers. Previously, extensive areas were covered by this forest, but most of it has been cleared for agriculture. This is a dense forest of broadleaf deciduous trees that grow tall but have generally little undergrowth. The animal life consists of birds, mammals, and reptiles.

The southeastern part of the United States, dominated mainly by the humid subtropical climate, supports subtropical broadleaf deciduous as well as evergreen forests, including the southern pine forest. The trees grow tall and have luxurious crowns.

The central and western parts of the United States and southern Canada, which are dominated by the steppe and desert climates, are covered by grassland and desert vegetation. East of the Rocky Mountains, in the Great Plains, short steppe grasses dominate. The tall-grass prairies in the Midwest have virtually disappeared under the plow. The steppes, where trees grow only along the few streams that cross these areas, provide pasture to large herbivores

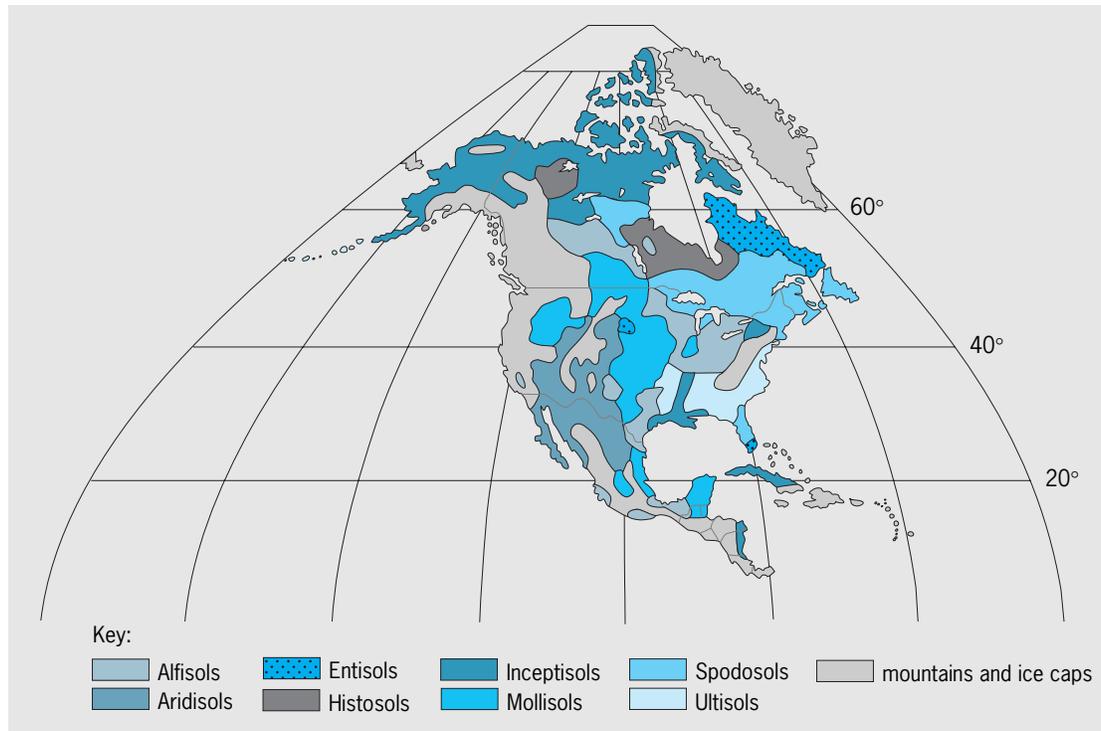


Fig. 3. Soil regions of North America. (After T. L. McKnight, *Physical Geography: A Landscape Appreciation*, 3d ed., Prentice Hall, 1990)

and smaller animals that often live underground. See GRASSLAND ECOSYSTEM.

The desert lands of the southwestern part of North America are so dry that only drought-resisting plants thrive. They either have structures, like cacti, that allow them to conserve moisture, or they are smaller plants that go through their entire life cycle during the brief, infrequent rainy periods. The plant cover in the desert is not continuous, and there are large areas of bare ground. The animal life is quite rich, though barely visible since it lives mainly underground. See PLANT-WATER RELATIONS.

The Mediterranean woodland and shrub covers a very small part of North America. It coincides with the Mediterranean climate in coastal California. It grows under the hot dry summer and rainy winter conditions, and consists of dense woody shrubs called chaparral, or grassy areas with broadleaf evergreen trees. The animal life is dominated by rodents and reptiles. See CHAPARRAL.

Mexico has desert vegetation in Baja California and in the northern part of the Mexican Plateau, with semidry steppe grassland around it, and temperate highland forest, broadleaf and evergreen, on mountain slopes, with tundra on the highest peaks. In Central America, the lowlands and lower mountain slopes have tropical rainforest, with a great number of species of broadleaf evergreens, some of which are 150 ft (45 m) tall. Multiple levels of tree crowns restrict sunlight and undergrowth. Pure stands of trees are rare. A unique type of vegetation in these moist and hot tropical areas of abundant rainfall are the expanses of savanna (tall grasses) found in the lowlands of eastern Nicaragua and Honduras. They have

developed on porous gravelly soils that will support only drought-tolerant vegetation. See SAVANNA.

**Soil regions.** The climatic and biotic regimes are so closely related that they generally coincide regionally. However, the soils, while strongly related to both regimes, also depend on the type of rock from which they develop (Fig. 3). There are specific factors (rock material, climate, and vegetation) whose interaction results in physical and biological processes that form the soils. To understand the characteristics and the regional distribution of North American soils, consideration must be given to the chemical composition of the rock material, the climatic and biotic environment, and the length of time during which these factors interact. Also, as the time of formation gets longer, the rock factor becomes less important (except in its limiting role as to the type of chemicals initially provided), and the environmental factors become more important. The length of time is especially important in North America, because more than half of the continent was overridden by glaciers, which removed whatever soils were there prior to glaciation. Since the glaciers did not totally disappear from central Canada until about 7000 years ago, most of the present soils in the northern parts of the continent are less than 7000 years old. See GLACIAL GEOLOGY.

Inasmuch as the processes that transform the rock materials into soils are determined by the climate (temperature and moisture) and vegetation, they differ from one climatic region to another. These processes are laterization (weathering that produces a high concentration of laterites), podzolization (increase in acid content), gleization (conversion to

clay), calcification (saturation of soil colloids with calcium), and salinization (accumulation of soluble salts). Laterization, podzolization, and gleization occur in moist climates and are based on downward percolation of rainwater. Calcification and salinization are based mainly on upward migration of capillary water due to intense evaporation. During the movement of water downward or upward, soil minerals and organic particles are redistributed throughout the earth material and form horizontal layers called horizons. Soil profiles composed of different horizons identify mature soils. The soil profiles differ from climate to climate since the soil-forming processes differ.

In North America the process of podzolization occurs in high and middle latitudes in humid, cool climates and in regions of coniferous boreal forests, which provide limited acidic needle-type litter. In these cool climates, chemical weathering is slow, but the abundance of acids and adequate precipitation make leaching a very effective process. This process produces podzolic soil profiles that have a leached-out, gray, sandy upper horizon; a brown middle horizon of accumulated minerals, like clay; and a lower horizon consisting of regolith, or broken rock material. These soils are identified mainly as spodosols (Fig. 3). *See* REGOLITH.

In the southeastern part of North America, the warmer humid climate affects the soil-forming processes by including some components of the laterization process, which dominates the hot and moist tropical regions with broadleaf tree and shrub vegetation. This process is characterized by rapid decompositions of organic matter and intense weathering of parent material, leading to dissolution of most minerals and leaving behind mainly iron and aluminum oxides which give it a red color. The soils of part of the Midwest are alfisols, and the southeastern soils are ultisols (Fig. 3), developed under warm to hot, and moist but not tropical conditions. *See* WEATHERING PROCESSES.

A large part of central North America, characterized by lower moisture due to its continentality, has mollisol soils. These soils form in subhumid conditions under dense grass vegetation, which is present in this area. Mollisols form through the process of calcification and have a dark, humus-rich top horizon. In this process, both leaching and removal of minerals are limited by the lack of precipitation. Calcium carbonate is carried downward by limited percolation and is concentrated in a horizon below the surface forming a dense layer; some of it is then carried upward by capillary water during intense evaporation, and by grass roots. In areas of undisturbed grasslands, mollisols are very fertile, but they lack moisture. *See* HUMUS.

The western belt of North America, dominated by mountain and desert climates, has a mosaic of mountain and desert soils. The desert soils form by the salinization process under drought conditions. Since there is less precipitation than evaporation, moisture drawn upward by intense evaporation brings various salts to the surface of the desert, which often be-

comes toxic. These desert soils continue into northern Mexico.

The mountain soils are highly complex, very thin, and generally immature because steep slopes and low temperatures prevent them from developing profiles and maturing.

The soils of Mexico and Central America are strongly related to the dominant climates and terrain configuration. The North Mexican deserts have thin desert soils formed by the processes of salinization; they are fringed by better steppe soils formed by calcification. The tropical coastal lowlands have red latosol soils formed by the process of laterization. They are greatly leached by excessive precipitation and inherently infertile. In some basins, there are heavy, clay, leached soils, called rendzina, which develop on soft limestones, such as those of northern Guatemala. On the very steep mountain slopes of tropical Central America, soils are very thin and poorly developed. Alluvial soils along the rivers, especially in the plains, are the most fertile. *See* PEDOLOGY; SOIL; SOIL CHEMISTRY.

### Landform Regions

In North America the climatic, vegetation, and soil regions exhibit approximate regional coincidence forming natural regions clearly dominated by climatic factors. However, the natural regions are almost equally determined by the landforms of North America (Fig. 4).

The major landform subdivisions of the North American continent are the Canadian Shield, Appalachian Highlands, Southeastern Coastal Plain, Interior Provinces including the Great Plains, North American Cordillera, Arctic Margin Plains, and Central America. These regions consist of subregions that differ in rock type or in elevation, slope, relief, pattern, or texture of their terrain. The result is an extremely varied and beautiful natural landscape.

**Canadian Shield.** Properly referred to as the geological core of the continent, the exposed Canadian Shield extends about 2500 mi (4000 km) from north to south and almost as much from east to west. The rest of it dips under sedimentary rocks that overlap it on the south and west. The Canadian Shield consists of ancient Precambrian rocks, over 500 million years old, predominantly granite and gneiss, with very complex structures indicating several mountain-building episodes. It has been eroded into a rolling surface of low to moderate relief with elevations generally below 2000 ft (600 m). Its surface has been warped into low domes and basins, such as the Hudson Basin, in which lower Paleozoic rocks, including Ordovician limestones, have been preserved. Since the end of the Paleozoic Era, the Shield has been dominated by erosion. Parts of the higher surface remain at about 1500–2000 ft (450–600 m) above sea level, particularly in the Labrador area. The Shield remained as land throughout the Mesozoic Era, but its western margins were covered by a Cretaceous sea and by Tertiary terrestrial sediments derived from the Western Cordillera. *See*



Key:

- |                                    |                                     |
|------------------------------------|-------------------------------------|
| 1 = New England-Maritime Provinces | 13 = Northern Rockies               |
| 2 = Piedmont                       | 14 = Brooks Range                   |
| 3 = Blue Ridge Mountains           | 15 = Central Alaska                 |
| 4 = Ridge and Valley Province      | 16 = Sierra-Cascade-Coast Mountains |
| 5 = Appalachian Plateau            | 17 = Pacific Troughs                |
| 6 = Interior Low Plateaus          | 18 = Pacific Coast Ranges           |
| 7 = Ozark Province                 | 19 = Interior Plateaus of Canada    |
| 8 = Ouachita Province              | 20 = Columbia Plateau               |
| 9 = Central Lowland Province       | 21 = Basin and Range Province       |
| 10 = Great Plains Province         | 22 = Colorado Plateaus              |
| 11 = Southern Rocky Mountains      | 23 = Mexican Highland               |
| 12 = Middle Rocky Mountains        |                                     |

Fig. 4. Physical divisions of North America.

CRETACEOUS; ORDOVICIAN; MESOZOIC; PALEOZOIC; PRECAMBRIAN; TERTIARY.

The entire exposed Shield was glaciated during the Pleistocene Epoch, and its surface was intensely eroded by ice and its meltwaters, erasing major surface irregularities and eastward-trending rivers that were there before. The surface is now covered by glacial till, outwash, moraines, eskers, and lake sediments, as well as drumlins formed by advancing ice. A deranged drainage pattern is evolving on this surface with thousands of lakes of various sizes. *See* DRUMLIN; ESKER; GLACIAL EPOCH; GLACIATED TERRAIN; MORAINES; PLEISTOCENE; TILL.

At the contact of the old, hard Shield rocks and the overlapping younger sedimentary rocks around

it, large erosional basins have formed. They now contain all the Great Lakes, and lakes Winnipeg, Athabasca, Great Slave, and Great Bear. A line drawn through these lakes, together with the St. Lawrence River, forms a circular boundary of the exposed Shield. Within this boundary lies a desolate flat to hilly area, with thousands of lakes between rocky knobs, frozen much of the year, covered by coniferous forest, and dominated by extremely cold subarctic and humid continental climate. Toward the north, the trees become smaller and fewer, and eventually tundra vegetation of mosses and lichen takes over, as the climate becomes even colder and there are no months with average temperatures above 50°F (10°C).

The Hudson Bay basin extends for 1000 mi (1600 km) from north to south. On its southern shore, the Ordovician limestones outcrop in a lowland nearly 180 mi (300 km) wide but lying less than 500 ft (150 m) above sea level. These limestones dip into the bay and are nearly 6000 ft (1800 m) thick in its center. During the glacial period, the ice was centered upon Hudson Bay. During deglaciation, till and outwash were widely spread around the bay; and isostatic rebound of about 500–800 ft (150–250 m) took place, and still continues, at the rate of about 1.6 ft (0.5 m) per 100 years. Former shorelines of Hudson Bay run nearly parallel to the present shoreline. Permafrost exists around Hudson Bay, and much of its vegetation is muskeg, or bogs with small black spruce. *See* BASIN; MUSKEG.

In great contrast to Hudson Bay Lowland is the Baffin Island/Labrador rim, which is the uplifted edge of the Shield rising 5000–6000 ft (1510–1820 m) and was highly eroded by preglacial rivers. During glaciation these river valleys were occupied by glaciers that eroded them into fiords. *See* FIORD.

The southern edge of the Shield is lower than the eastern rim, but it still forms a topographic barrier, particularly along the north shore of the St. Lawrence River and Gulf, where it rises in a 1500–3000 ft (450–900 m) escarpment. In contrast, the west and southwest sides of the Shield are relatively low, and they either disappear under the overlying sediments or face a small sedimentary escarpment. *See* ESCARPMENT.

The Canadian Shield extends into the United States as Adirondack Mountains in New York State, and Superior Upland west of Lake Superior.

**Southeastern Coastal Plain.** The Southeastern Coastal Plain is geologically the youngest part of the continent, and it is covered by the youngest marine sedimentary rocks. This flat plain, which parallels the Atlantic and Gulf coastline, extends for over 3000 mi (4800 km) from Cape Cod, Massachusetts, to the Yucatán Peninsula in Mexico. It is very narrow in the north but increases in width southward along the Atlantic coast and includes the entire peninsula of Florida. As it continues westward along the Gulf, it widens significantly and includes the lower Mississippi River valley. It is very wide in Texas, narrows again southward in coastal Mexico, and then widens in the Yucatán Peninsula and continues as a wide submerged plain, or a continental shelf, into the sea. *See* COASTAL PLAIN.

The surface of the emerged coastal plain is very low, less than 100 ft (60 m) above sea level, and very flat near the coast; it slowly rises inland up to several hundred feet and becomes rolling, with belts of hills, especially in the Gulf portion. The key to these characteristics is its geologic structure. *See* CONTINENTAL MARGIN.

The entire coastal plain is underlain by gently seaward dipping sedimentary rocks of varying resistance, dating back to Cretaceous, lower and upper Tertiary, and Quaternary. The rocks are often weakly consolidated, and Quaternary deposits that fringe the shoreline and fill the Mississippi Embayment are unconsolidated. *See* QUATERNARY.

The contact between these young rocks and the older rocks of the continent, especially along the Atlantic plain, is called the Fall Line; it is marked by falls and rapids. As the layers of sedimentary rocks are slowly eroding seaward, land-facing cuesta escarpments are formed, and they eventually become belts of hills, such as the Red Hills of Alabama and Mississippi. In addition, the coastal plain rocks are gently arched, or downwarped, affecting the plain's surface and coastal characteristics, and resulting in different coastal subregions. Along the Atlantic coast, the Delaware downwarp creates the Embayed Section with a large Delaware Bay, Chesapeake Bay, and many estuaries; the downwarp in South Carolina and Georgia creates the Sea Island section with several belts of offshore islands. In contrast, the Cape Fear Arch between these two subsections leads to the exposure of the underlying Cretaceous rocks. The broad Peninsular Arch uplifts Florida, exposing upper Tertiary limestone at about 50 ft (15 m) above sea level. Solutional karst features, such as sinkholes, lakes, and springs, dominate this area. *See* KARST TOPOGRAPHY.

On the Gulf Plain, wide belts of Cretaceous rocks are exposed along its inner boundary, while the downwarped Mississippi embayment, reaching deep into the continent, is deeply filled with the youngest Quaternary sediments. The Gulf Coastal Plain is wide and is referred to as a belted plain, because the landward-facing cuesta escarpments, developed on more resistant rocks and forming belts of hills, alternate with parallel lowlands developed on weaker rocks. *See* COASTAL LANDFORMS.

Another characteristic of the Coastal Plain is its complex shoreline, partly related to the depression of the northern part of the continent during glaciations. The Embayed section consists of endless estuaries, including Delaware Bay and Chesapeake Bay, many fringed by miles-long offshore sandbars. Further south, these bars create big lagoons, such as Pamlico Sound. In this area, Cape Lookout and Cape Hatteras are examples of wave-formed capes on the offshore bars enclosing the lagoons. Behind the sandbars are either big lagoons or extensive swamplands. South of Cape Fear, offshore sandbars are replaced by Sea Islands just offshore, and "Carolina Bays" on the coastal plain; the latter are oval-shaped depressions, often holding lakes. The Atlantic coast of Florida is fringed by very long sandbars and narrow lagoons. Sandbars and coral islands form the Florida Keys.

The Everglades are a fresh-water marsh with sawgrass growing up to 12 ft (3.5 m) high, among which are tree-covered islands. This is the largest remaining subtropical wilderness in North America; it includes rich animal and bird life, and is a National Park covering 2200 mi<sup>2</sup> (5700 km<sup>2</sup>). *See* WETLANDS.

The shoreline characteristics of the Gulf include sandbars, small estuaries, lagoons, swamp lands, and the Mississippi River floodplain and delta. The floodplain of the Lower Mississippi River is about 600 mi (960 km) long and over 50 mi (80 km) wide, and most of it is subject to flooding. The Mississippi drains over 1 million square miles (2.6 million square kilometers) and delivers over 2 million tons of rock material to the Gulf daily. Its floodplain includes a multitude of active river meanders, oxbow lakes, and abandoned meanders, as well as parallel tributaries such as the Yazoo River, and islands such as the Crowley's Ridge. The delta itself is a unique part of the shoreline where land and water intermingle and numerous distributary channels carry water and silt to the Gulf. *See* DELTA; ESTUARINE OCEANOGRAPHY; FLOODPLAIN; FLUVIAL SEDIMENTS.

A unique inner boundary of the Coastal Plain in Texas is the Balcones Escarpment, a fault scarp that rises sharply nearly 1500 ft (450 m) above the Black Prairie Lowland west and north of San Antonio. The Rio Grande, forming the boundary with Mexico, is the largest river crossing the flat western part of the Coastal Plain.

Extending from Cape Cod, Massachusetts, to Mexico and Central America, the Coastal Plain is affected by a variety of climates and associated vegetation. While a humid, cool climate with four seasons affects its northernmost part, subtropical air masses affect the southeastern part, including Florida, and hot and arid climate dominates Texas and northern Mexico; Central America has hot, tropical climates.

Varied soils characterize the Coastal Plain (Fig. 4), including the fertile alluvial soils of the Mississippi Valley. Broadleaf forests are present in the northeast, citrus fruits grow in Florida, grasslands dominate the dry southwest, and tropical vegetation is present on Central American coastal plains.

**Eastern Seaboard Highlands.** Between the Southeastern Coastal Plain and the extensive interior provinces lies a belt of mountains that, by their height and pattern, create a significant barrier between the eastern seaboard and the interior of North America. These mountains consist of the Adirondack Mountains and the New England Highlands.

The Adirondack Mountains are a domal extension of the Canadian Shield, about 100 mi (160 km) in diameter, composed of complex Precambrian rocks. They rise to 5600 ft (1700 m), with the highest peak being Mount Marcy, and include more than a dozen peaks whose elevations exceed 4000 ft (1210 m). They are bounded on the south and west by early Paleozoic sedimentary rocks forming inward-facing cuestas, on the north by the St. Lawrence River Lowland, and on the east by the Lake Champlain Lowland. They have been glaciated and are rugged, with hundreds of picturesque lakes, like Lake Placid. *See* LAKE.

The New England Highlands consist of a north-south belt of mountains east of the Hudson Valley, including the Taconic mountains in the south (Mount Equinox is 3800 ft or 1150 m), and the Green mountains in the north (Mount Mansfield is 4400 ft or 1330 m), and continuing as the Notre Dame Mountains along the St. Lawrence Valley and the Chic-Choc Mountains of the Gaspé Peninsula. The large area of New England east of these mountains is an eroded surface of old crystalline rocks culminating in the center as the White Mountains, with their highest peak of the Presidential Range, Mount Washington, reaching over 6200 ft (1880 m). This area has been intensely glaciated, and it meets the sea in a rugged shoreline. Nova Scotia and Newfoundland have a similar terrain.

The rocks underlying New England are Precambrian and early Paleozoic igneous or metamorphic granites, gneisses, schists, quartzites, slates, and marbles. At least one downfaulted valley is filled with Triassic rocks—the 100-mi-long (160-km) Triassic Lowland of Connecticut and Massachusetts with volcanic trap ridges, followed by the Connecticut River, which breaks out of the lowland across the old surface to empty into the Long Island Sound.

New England is a hilly to mountainous region carved out of ancient rocks, eroded by glaciers, and covered by glacial moraines, eskers, kames, erratics, and drumlins, with hundreds of lakes scattered everywhere. It has a cool and moist climate with four seasons, thin and acid soils, and mixed coniferous and broadleaf forests.

**Appalachian Highlands.** The Appalachian Highlands are traditionally considered to consist of four parts: the Piedmont, the Blue Ridge Mountains, the Ridge and Valley Section, and the Appalachian Plateau (Fig. 5). These subregions are all characterized by different geologic structures and rock types, as well as different geomorphologies, but their origin and relationships can best be understood when discussed together.

From southeast to northwest is the Fall Line, which is the contact zone between the young, sedimentary rocks of the Coastal Plain and the old crystalline rocks of the Piedmont. At the western edge of the rolling Piedmont, the old rocks rise rapidly and form the Blue Ridge Mountains. West of the Blue Ridge Mountains lies the Great Valley, filled with Paleozoic sedimentary rocks, which is part of the extensive Ridge and Valley Section composed of parallel ero-

sional ridges and valleys carved out of folded Paleozoic rocks. West of this section rises the Allegheny Front, an escarpment which forms the eastern edge of the Appalachian Plateau, a hill land carved out of nearly horizontal Paleozoic sedimentary rocks.

The northern boundary of the entire Appalachian System is an escarpment of Paleozoic rocks trending eastward along Lake Erie, Lake Ontario, and the Mohawk Valley. The boundary then swings south along Hudson River Valley and continues southwestward along the Fall Line to Montgomery, Alabama. The western boundary trends northeastward through Cumberland Plateau in Tennessee, and up to Cleveland, Ohio, where it joins the northern boundary. Together with New England, this region forms the largest mountainous province in eastern United States.

*The Piedmont.* The Piedmont extends for almost 1000 mi (1600 km) in a northeast-southwest direction from southern New York to central Alabama. Its width varies from less than 20 mi (32 km) in New Jersey to over 120 mi (192 km) in the south. The Piedmont is underlain by complex Precambrian and early Paleozoic crystalline rocks. The youngest rocks are Triassic sedimentaries that have been preserved in downfaulted grabens called Triassic Lowlands. The soils developed on these Triassic sediments are often the best in the Piedmont. The three Triassic basins often include igneous rock layers that form ridges such as the Watchung Mountains in New Jersey, and the columnar basaltic Palisades along Hudson River (50 mi or 80 km long and 1000 ft or 300-m high).

The surface of the Piedmont is rolling to hilly, and its relief increases toward the Blue Ridge Mountains. Several monadnocks such as the granite Stone Mountain (1.5 mi or 2.5 km wide, 200 ft or 60 m high) near Atlanta, Georgia, are spectacular reminders that this is an old erosional plain. Rivers starting in the Blue Ridge Mountains cross the Piedmont and the Coastal Plain to reach the Atlantic. However, beginning with the Roanoke River, the rivers reaching the Piedmont first cross the Blue Ridge where they cut water gaps; others, like the Potomac with the Shenandoah River, cross the Ridge and Valley region; further north, the Susquehanna and the Delaware rivers head in the Appalachian Plateau. Thus, the drainage divide in the Appalachian System shifts from the Blue Ridge Mountains in the south to the Appalachian Plateau in the north.

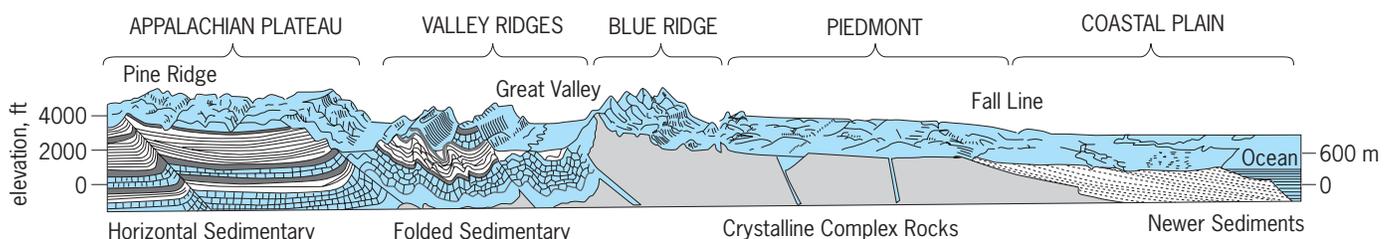


Fig. 5. Profile of the Coastal Plain and the Appalachian Highlands. (After W. H. Yoho and E. C. Pirkle, *Natural Landscapes of the United States*, 4th ed., Kendall/Hunt, 1985)

**Blue Ridge Mountains.** These mountains lie west of the Piedmont and stretch from Pennsylvania to Georgia for over 540 mi (860 km). They begin as a single ridge in Pennsylvania, and become a high and wide mountain range south of Roanoke, Virginia, where they reach elevations of over 6000 ft (1820 m) for about 45 peaks, and over 5000 ft (1510 m) for nearly 300 peaks. Mount Mitchell is the highest peak at 6684 ft (2025 m) in the most rugged part called the Great Smoky Mountains. Geologically, the Blue Ridge Mountains are a continuation of the Piedmont (Fig. 5). Their national parks include Shenandoah National Park and Great Smoky Mountains National Park.

**Ridge and Valley Province.** Paralleling the Blue Ridge Mountains, this is the most unusual mountain province in eastern North America. It consists of nearly parallel, alternating ridges and valleys hundreds of miles long, extending 2880 mi (4600 km) from southeastern New York to Alabama. These are a result of intense folding of lower Paleozoic sedimentary rocks of alternating resistant and nonresistant layers, and subsequent erosion of the resulting anticlines and synclines into parallel ridges and valleys (Fig. 5). The width, length, and spacing of the ridges and valleys vary from north to south. In the north, the high, parallel ridges and the valleys are very long, narrow, and closely spaced. The ridges in the central part are densely spaced and are the highest. Approximately south of Roanoke, Virginia, the ridges become lower and the valleys widen. This linear terrain continues for hundreds of miles but eventually becomes less distinct and ends south of Birmingham, Alabama.

Among the unique features of the Ridge and Valley Province are superposed rivers which had cut across ridges forming water gaps, such as the Delaware Gap across the Kittatinny Mountains with 1000-ft-high (300-m) cliffs rising above the river; or the Susquehanna River Gap north of Harrisburg, Pennsylvania; and the Potomac River Gap at Harpers Ferry, West Virginia.

The thick Ordovician Paleozoic limestones of the Ridge and Valley section give rise to karstic features, such as the extensive and beautiful caverns with striking stalactites and stalagmites in Luray and Endless caverns near Luray, Virginia. See CAVE; STALACTITES AND STALAGMITES.

**Appalachian Plateau.** Covering large parts of seven states, the Appalachian Plateau is the widest part of the Appalachian System (Fig. 5). It is bound on the north by the escarpment of the Devonian rocks rising over the Mohawk Valley, and by Lake Ontario and Lake Erie lowlands. Its characteristic rock structure is nearly horizontal layers of Paleozoic rocks, mostly sandstones, conglomerates, and shales, from Devonian to Permian.

The terrain of the Appalachian Plateau is mainly hilly. The Paleozoic rocks that rise from 1000 ft (300 m) in the west to 5000 ft (1510 m) in West Virginia have been eroded by dendritically branching streams into hills and low mountains with deep valleys. The northern section is drained predominantly by eastward-flowing rivers, like the Delaware and

Susquehanna, while the central and southern parts are drained by southward-flowing rivers, mainly the Ohio and Tennessee.

In the northeast corner of the Plateau the rocks rise to nearly 4000 ft (1200 m) and form the Catskill Mountains. Other mountainous sections are the Pocono Mountains in eastern Pennsylvania, the Allegheny Mountains in Pennsylvania and West Virginia, and the Cumberland Mountains in Kentucky and Tennessee. These mountains are formed from resistant Pottsville and Pocono sandstones.

The northern edge of the Appalachian Plateau has been glaciated, leaving a large drumlin field between Syracuse and Rochester, New York, and the Finger Lakes, located south of the drumlin field. This is a belt of parallel, north-south-trending, narrow lakes, up to 40 mi (64 km) long, formed when advancing glaciers deepened the valleys of preglacial northward-flowing rivers, and subsequently blocked them with moraines to form the lakes.

The southern part of the Appalachian Plateau becomes narrow and forms the Cumberland Plateau, with a pronounced series of 700 mi (1120 km) long, east-facing escarpments up to 1000 ft (300 m) high.

The Appalachian Plateau is one of the largest continuous bituminous coal areas in the world, and its northern part is the major coal mining region in North America.

**Interior Domes and Basins Province.** The southwestern part of the Appalachian Plateau, overlain mainly by the Mississippian and Pennsylvanian sedimentary rocks, has been warped into two low structural domes called the Blue Grass and Nashville Basins, and a structural basin, drained by the Green River; its southern fringe is called the Pennyroyal Region. The Interior Dome and Basin Province is contained roughly between the Tennessee River in the south and west, and the Ohio River in the north.

There is no boundary on the east, because the domes are part of the same surface as the Appalachian Plateau. However, erosional escarpments, forming a belt of hills called knobs, clearly mark the topographic domes and basins. The northern dome, called the Blue Grass Basin or Lexington Plain, has been eroded to form a basin surrounded by a series of inward-facing cuesta escarpments. The westernmost cuesta reaches about 600 ft (180 m) elevation while the central part of the basin lies about 1000 ft (300 m) above sea level, which is higher than the surrounding hills. This gently rolling surface with deep and fertile soils exhibits some solutional karst topography. See FLUVIAL EROSION LANDFORMS.

The southern dome, the Nashville Basin, is truly a topographic basin with average elevations in the center of about 600 ft (180 m) being lower than the surrounding Eastern Highland Rim, which reaches 1300 ft (400 m), and the Western Highland Rim, with elevations of 900 ft (270 m). The basin is mainly a rolling and fertile plain surrounded by the cuesta escarpments.

The third part of this province is both a topographic and a structural basin, drained northward by the Green River. Its south side is the

Pennyroyal limestone plain, above which rise two escarpments. The Green River is the only surface stream that crosses the karst plain. Thousands of sinkholes, lost creeks, underground rivers, and caves make it a classic karst topography. Within it lies the Mammoth Cave, an enormous solutional cave system developed in thick limestones.

**Ozark and Ouachita Highlands.** The Paleozoic rocks of the Pennyroyal Region continue westward across southern Illinois to form another dome of predominantly Ordovician rocks, called the Ozark Plateau. This dome, located mainly in Missouri and Arkansas, has an abrupt east side, and a gently sloping west side, called the Springfield Plateau. Its surface is stream eroded into hilly and often rugged topography that is developed mainly on limestones, although shales, sandstone, and chert are present. Much residual chert, eroded out of limestone, is present on the surface. There are some karst features, such as caverns and springs. In the northeast, Precambrian igneous rocks protrude to form the St. Francois Mountains, which reach an elevation of 1700 ft (515 m).

In Arkansas, the Plateau surface rises to 2500 ft (757 m) and forms the 200-mi-long (320-km) and 35-mi-wide (56-km) Boston Mountains, which have stream-eroded tops, and a 300–600-ft-high (90–180-m) escarpment along its southern side. South of the Arkansas River Valley lie the Ouachita Mountains, over 200 mi (320 km) long and 100 mi (160 km) wide, and best characterized by their east-west-trending ridges and valleys. These ridges and valleys have been eroded from folded and thrust-faulted sedimentary rocks, such as shales and sandstones mainly of Pennsylvanian age. The mountains reach elevations of over 2600 ft (780 m) in the center, with ridges rising 1500 ft (450 m) above the valleys. In the Ouachita Mountains Hot Springs National Park, there are 47 hot springs with water temperature of 143°F (61°C) and a combined flow of over a million gallons (4000 m<sup>3</sup>) a day. *See* PENNSYLVANIAN.

**Central Lowlands.** One of the largest subdivisions of North America is the Central Lowlands province which is located between the Appalachian Plateau on the east, the Interior Domes and Basins Province and the Ozark Plateau on the south, and the Great Plains on the west. It includes the Great Lakes section and the Manitoba Lowland in Canada. This huge lowland in the heart of the continent (whose elevations vary from about 900 ft or 270 m above sea level in the east and nearly 2000 ft or 600 m in the west) is underlain by Paleozoic rocks that continue from the Appalachian Plateau and dip south under the recent coastal plain sediments; meet the Cretaceous rocks on the west; and overlap the crystalline rocks of the Canadian Shield on the northeast.

The present surface of nearly the entire Central Lowlands, roughly north of the Ohio River, and east of the Missouri River, is the creation of the Pleistocene ice sheets. When the ice formed and spread over Canada, and southward to the Ohio and Missouri rivers, it eroded much of the preexisting surface. During deglaciation, it left its deposits over the Canadian Shield and the Central Lowlands.

During its four (Nebraskan, Kansan, Illinoian, and Wisconsin) or more glacial periods, the ice eroded the contact zone between the Canadian Shield and the Central Lowlands into great lowlands, some of which now hold big lakes. Preglacial river valleys, as well as structurally controlled erosional lowlands between cuestas around the Michigan Basin, directed the ice movement southward and helped form the Great Lakes basins. As the ice gradually melted, the Great Lakes basins filled with water along the edge of the ice, forming the huge Lake Agassiz and the predecessors of the Great Lakes; they eventually established their drainage to the east, and acquired their present shapes and sizes.

Toward the end of the evolution of the Great Lakes, the Niagara Falls were cut. Initially, when Lake Erie drained through the Niagara River into Lake Ontario, the latter's elevation was about 150 ft (45 m) higher than at present and there were not falls on the Niagara River. When ice melted and the Mohawk outlet opened to the east, Lake Ontario began to drain rapidly, and the Niagara River began to cut its valley to keep up with the lake's falling level. The falls developed on the Niagara escarpment, and have since retreated about 7 mi (11 km), leaving a deep gorge. They continue to retreat at a rate of about 4 ft (1.2 m) per year, mostly by rock falls, as the underlying soft shale and sandstone are undermined, and the dolomite edge breaks off and falls down. Eventually the falls will become lower, undermining of the underlying rocks will cease, and the falls may become just rapids.

The overall surface characteristics of the Central Lowlands were created by the ice sheets. During ice advance, debris deposition was often greater than erosion. In areas where surface resistance equalled ice stress, the ice just moved over its own debris and streamlined it into oval-shaped hills, called drumlins. Among the world's largest fields of drumlins are those in eastern Wisconsin, and north and south of Lake Ontario.

During ice melting, long belts of hilly moraines formed at the end of ice lobes. Between the moraines are stretches of rolling till plains composed of heterogeneous glacial debris, including silt, sand, gravel, boulders, and large erratics from distant areas. Flat sandy surfaces, called outwash plains, were formed by meltwaters; clay-covered lacustrine flats are formed lake bottoms. Smaller features include kames and eskers. Scattered between these features are numerous lakes and swamps, and deranged river systems. *See* KAME.

All these features are most striking in areas covered by the most recent ice advance, the Wisconsin, especially in Minnesota, Wisconsin, and Michigan.

A unique terrain is the Driftless Area of southwestern Wisconsin, which is surrounded by glacial debris of different ice lobes but shows no evidence of having been covered by ice. Here, the Paleozoic rock surface, which has been buried by glacial material everywhere else, is being eroded by streams into deep valleys and flat uplands.

The Central Lowlands are drained by the third

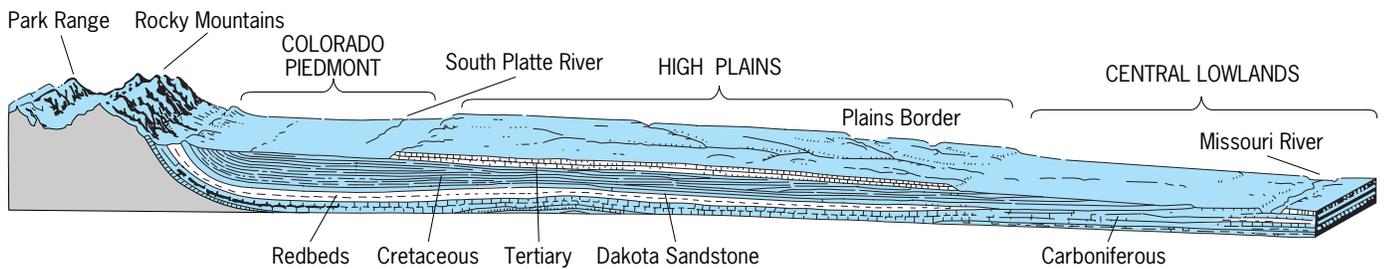


Fig. 6. Rock structure of the Great Plains. (After W. H. Yoho and E. C. Pirkle, *Natural Landscapes of the United States*, 4th ed., Kendall/Hunt, 1985).

longest river system in the world, the Missouri-Mississippi, which is 3740 mi (6000 km) long. This mighty river system, together with the Ohio and the Tennessee, drains not only the Central Lowlands but also parts of the Appalachian Plateau and the Great Plains, before it crosses the Coastal Plain and ends in the huge delta of the Mississippi. The river carries an enormous amount of water and alluvium and continues to extend its delta into the Gulf. In 1993 it reached a catastrophic level of a hundred-year flood, claimed an enormous extent of land and many lives, and created an unprecedented destruction of property. This flood again alerted the population to the extreme risk of occupying a river floodplain. See FLOODPLAIN; RIVER.

**Great Plains.** The Great Plains, which lie west of the Central Lowlands, extend from the Rio Grande and the Balcones Escarpment in Texas to central Alberta in Canada. On the east, they are bounded by a series of escarpments, such as the Côteau du Missouri in the Dakotas. The dry climate with less than 20 in. (50 cm) of precipitation, and steppe grass vegetation growing on calcareous soils, help to determine the eastern boundary of the Great Plains. On the west, the Great Plains meet the abrupt front of the Rocky Mountains, except where the Colorado Piedmont and the lower Pecos River Valley separate them from the mountains (Fig. 6).

These extensive Great Plains are high and slope eastward from 5500 ft (1660 m) elevation at the foothills of the Rocky Mountains to 1500 ft (450 m) along their eastern boundary. They are overlain by terrestrial Tertiary deposits, resting on marine Cretaceous formations. The latter were laid down in a Mesozoic Sea and are exposed in the Piedmont along the eastern edge of the Plains, and in the northwest and north. In Colorado, the Cretaceous rocks rise against the mountains, where their resistant Dakota sandstone layers form long, striking hogback ridges that parallel the mountain front.

The Tertiary rocks forming the surface of the Great Plains were deposited during the prolonged erosion of the Rocky Mountains during which the sediments were carried by streams eastward and spread over the Cretaceous surface. The most important top formation is the Ogallala, which consists of sands, gravels, silts, and an irregular massive calcrete formed under dry conditions. The Ogallala is the most valuable aquifer, or water-bearing formation, of the Great Plains. However, much of it has been eroded, form-

ing pediments in Colorado, and Canadian badlands along Deer River near Calgary. Erosion of the Tertiary deposits gave rise to the Sand Hills of Nebraska, and extensive layers of loess, or windblown silt, farther east. Tertiary igneous features such as lava mesas are present in the Great Plains in the Raton section in northeastern New Mexico. Volcanic ash deposits are present farther east. See AQUIFER; LOESS.

The Great Plains region shows distinct differences between its subsections from south to north. The southernmost part, called the High Plains or Llano Estacado, and Edwards Plateaus are the flattest. While Edwards Plateau, underlain by limestones of the Cretaceous age, reveals solutional karst features, the High Plains have the typical Tertiary bare cap rock surface, devoid of relief and streams.

The central part of the Great Plains has a recent depositional surface of loess and sand. The Sand Hills of Nebraska form the most extensive sand dunes area in North America, covering about 24,000 mi<sup>2</sup> (62,400 km<sup>2</sup>). They are overgrown by grass and have numerous small lakes. The loess region to the south provides spectacular small canyon topography. See DUNE.

The northern Great Plains, stretching north of Pine Ridge and called the Missouri Plateau, have been intensely eroded by the western tributaries of the Missouri River into river breaks and interfluvies. In extreme cases, badlands were formed, such as those of the White River and the Little Missouri.

Another unusual feature of northwestern Great Plains are the small mountain groups of intrusive, extrusive, and even diastrophic origin, some reaching over 11,000 ft (3330 m) elevation (Crazy Mountains). The largest are the Black Hills of South Dakota, a classical partially eroded dome 60 mi (96 km) wide and 120 mi (190 km) long. The highest elevation is in Harney Peak (7240 ft or 2190 m); and the best-known peak is Mount Rushmore, with heads of four presidents carved in its granite face.

The terrain of the Canadian Great Plains consists of three surfaces rising from east to west: the Manitoba, Saskatchewan, and Alberta Prairies developed on level Cretaceous and Tertiary rocks. Climatic differences between the arid and warm southern part and the cold and moist northern part have resulted in regional differences. The eastern boundary of the Saskatchewan Plain is the segmented Manitoba Escarpment, which extends for 500 mi (800 km) northwestward, and in places rises 1500 ft (455 m)

above the Manitoba Lowland. Côteau du Missouri marks the eastern edge of the higher Alberta Plain.

**Western Cordillera.** The mighty and rugged Western Cordilleras stretch along the Pacific coast from Alaska to Mexico. There are three north-south-trending belts: (1) Brooks Range, Mackenzie Mountains, and the Rocky Mountains to the north and Sierra Madre Oriental in Mexico; (2) Interior Plateaus, including the Yukon Plains, Canadian Central Plateaus and Ranges, Columbia Plateau, Colorado Plateau, and Basin and Range Province stretching into central Mexico; and (3) Coastal Mountains from Alaska Range to California, Baja California, and Sierra Madre Occidental in Mexico.

This subcontinental-size mountain belt has the highest mountains, greatest relief, roughest terrain, and most beautiful scenery of the entire continent. It has been formed by earth movements resulting from the westward shift of the North American lithospheric plate. The present movements, and the resulting devastating earthquakes along the San Andreas fault system paralleling the Pacific Ocean, are part of this process. See CORDILLERAN BELT; PLATE TECTONICS.

**Rocky Mountains.** This very high, deeply eroded and rugged mountainous region comprises several distinct parts: Southern, Middle, and Northern Rockies, plus the Wyoming Basin in the United States, and the Canadian Rockies. The Southern Rockies, extending from Wyoming to New Mexico, include the Laramie Range, the Front Range, and Spanish Peaks with radiating dikes on the east; Medicine Bow, Park, and Sangre de Cristo ranges in the center; and complex granite Sawatch Mountains and volcanic San Juan Mountains of Tertiary age on the west. Most of the ranges are elongated anticlines with exposed Precambrian granite core, and overlapping Paleozoic and younger sedimentary rocks which form spectacular hogbacks along the eastern front. There are about 50 peaks over 14,000 ft (4200 m) high, while the Front Range alone has about 300 peaks over 13,000 ft (3940 m) high. The southern Rocky Mountains, heavily glaciated into a beautiful and rugged scenery with permanent snow and small glaciers, form a major part of the Continental Divide.

The Middle Rocky Mountains consist of the Wyoming and Big Horn basins, and separate elongated anticlinal ranges with granite cores and flanking hogbacks, including the Big Horn, Owl Creek, Wind River, and Uinta. The highly eroded volcanic Absaroka Mountains, the crystalline Beartooth-

Snowy Mountains, the volcanic Yellowstone Plateau, the faulted Teton Range, and the Wasatch Range are also part of the Middle Rockies. The maximum peak elevations are over 13,000 ft (3940 m), with the Grand Teton reaching over 13,700 ft (4150 m). All the Middle Rocky ranges have been intensely glaciated, and have deep U-shaped valleys, cirques, horns, moraines, and thousands of lakes.

The Yellowstone Plateau is a high basin with elevations from about 7000 to 8500 ft (2120 to 2575 m) surrounded by mountains whose elevations exceed 10,000 ft (3030 m). It is filled with basaltic and colorful rhyolitic lavas of Tertiary age, has over 4000 hot springs and brightly colored mineral springs, and numerous geysers that spout steam intermittently, including The Old Faithful which erupts regularly. The large Yellowstone Lake, 300 ft (90 m) deep, at 7750 ft (2350 m) elevation, is drained northward by Yellowstone River which cuts a 1200-ft-deep (360-m) canyon. See GEYSER; SPRING (HYDROLOGY).

The large rivers of the Middle Rockies and the Wyoming Basin, including the Yellowstone, Big Horn, Green, and Snake, leave this province through deep canyons, which they cut across the high ranges on which they were superposed from a higher surface.

The Northern Rocky Mts. are the widest, although not the highest, of the United States Rockies. They are located mainly in Idaho and western Montana. Their southern part, 6000–8000 ft (1820–2420 m) high, is carved into huge ranges from the nearly homogeneous granite mass of the Idaho Batholith of Jurassic and Cretaceous age. The highest peaks rise to 12,000 ft (3630 m). The northern part is underlain by Precambrian metamorphic and sedimentary rocks, which have been folded, faulted, and thrust eastward. The most spectacular example is the Lewis overthrust (Chief Mountain) on the United States–Canadian boundary, which moved over 120 mi (200 km) eastward (Fig. 7). The southeastern part of the Northern Rockies consists of short fault block ranges of Paleozoic and Mesozoic sedimentary rocks.

The Northern Rockies are drained by the Solomon, Clearwater, and Clark rivers. They were glaciated but, being lower, did not develop a glacial terrain as spectacular as the higher Middle and Southern Rockies, except in Glacier National Park on the Canadian border. Here, more than 50 small glaciers are still present, and majestic horns, aretes, cirques, U-shaped valleys, and lakes dominate the scenery.

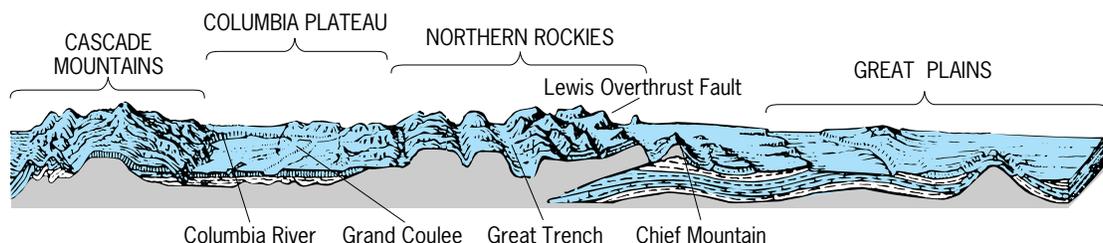


Fig. 7. Profile of the Great Plains, Northern Rockies, Columbia Plateau, and Cascade Mountains. (After W. H. Yoho and E. C. Pirkle, *Natural Landscapes of the United States*, 4th ed., Kendall/Hunt, 1985).

The Canadian Rocky Mountains begin as Purcell, Selkirk, Monashee and Cariboo ranges, together known as Columbia Mountains, merging northward into the Eastern Front Ranges. They are composed of Precambrian cores with overlapping Paleozoic and Mesozoic sediments; the Paleozoic limestones are thrust eastward over the Mesozoic rocks. The glaciated Front Range reaches almost 13,000 ft (3940 m) in elevation (Mount Robson). Its Columbian ice field (300 ft or 90 m thick) lies at 10,000–12,000 ft (3030–3640 m) elevation on the Continental Divide. West of the Canadian Rockies and parallel to them trends the nearly 1000-long (1600-km) Rocky Mountain Trench, a straight, flat-bottom, downfaulted valley, several miles wide and up to 3000 ft (910 m) deep, followed by such rivers as the Columbia, Kootenay, and Frazer. North of Liard River begins the Tintina Trench that continues for 450 mi (720 km). East of this trench lie the Mackenzie Mountains, which continue as Richardson Mountains, and the treeless 600-mi-long (960-km) 9000-ft-high (2730-m) Brooks Range. North of Brooks Range and its foothills lies the extensive frozen Arctic Coastal Plain, covered by gravels, sands, and silts of Pleistocene age, and exhibiting frozen ground features such as patterned ground.

*Interior Plateaus and Ranges Province.* This portion of the Western Cordillera lies between the Rocky Mountains and the Coastal Mountains. It is an extensive and complex region. It begins in the north with the wide Yukon Plains and Uplands; narrows into the Canadian Central Plateaus and Ranges; widens again into the Columbia Plateau, Basin and Range Province, and Colorado Plateau; and finally narrows into the Mexican Plateau and the Central American isthmus.

1. The Yukon Plains and Uplands is a region of low valleys, dissected plateaus, and low mountains. The lowlands include the Yukon Flats, underlain by scattered permafrost and covered by loess; and the lower Yukon plain and delta. The dissected upland between Yukon and Tanaka rivers continues east as the Klondike Region.

Continuing southward, the Canadian Central Plateaus and Ranges region, lower than the mountains to the east and west, remains complex and mountainous, and reveals widespread glaciation, including numerous lakes. The Stikine Mountains in the north and the Columbia Mountains in the south, with peaks over 11,000 ft (3330 m), are the roughest and highest ranges. The rocks of this area are mainly Paleozoic and Precambrian sedimentaries and

metamorphics. The Frazer Plateau west of Columbia Mountains is a rolling upland developed mainly on intrusive and extrusive igneous rocks of Tertiary age.

2. The Columbia Plateau (Fig. 7), which lies south of the United States–Canadian border, and between the Northern Rockies and the Cascade Ranges, is one of the largest lava-covered surfaces in the Western Hemisphere. It covers over 100,000 mi<sup>2</sup> (260,000 km<sup>2</sup>), and consists of many flows of Tertiary lavas with individual layers reaching up to 200 ft (60 m) thick. The total thickness of the lavas probably reaches to 8000 ft (2420 m). *See* LAVA.

In the center of Columbia Plateau, the Blue Mountains, reaching over 9000 ft (2730 m), project through the lavas. They consist of Paleozoic and Mesozoic sedimentary rocks that were part of the pre-lava surface. The northwestern part of Columbia Plateau, called the Walla Walla, is its lowest part; at the junction of the Columbia and Snake rivers, the elevations are only about 300–400 ft (90–120 m). In the south, the elevations are about 5000 ft (1510 m); in the northeast part of the Snake River Plain, about 6000 ft (1810 m); and the mountains in the center reach, over 9000 ft (2730 m).

In the northern section, near Yakima, Washington, there are broad anticlinal and synclinal ridges and valleys. In the northeast are the unique Channeled Scablands, a 2000-mi<sup>2</sup> (5200-km<sup>2</sup>) area of dry channels separated by eroded and nearly bare lava surfaces. It appears that the channels were cut by ice meltwaters when a dam holding glacial Lake Missoula broke and created a sudden flood of enormous proportions that spilled across the area. The largest channel, the Grand Coulee, is about 50 mi (80 km) long, several miles wide, and nearly 1000 ft (300 m) deep. Its Dry Falls have a 450-ft (135-m) drop.

The Snake River Plain is drained by Snake River, which begins at about 6000 ft (1810 m) elevation in the Yellowstone Plateau and flows westward over the American Falls (50 ft or 15 m), Twin Falls (180 ft or 55 m), and Shoshone Falls (212 feet or 64 m). Here, it creates a deep gorge with the spectacular Thousand Springs, which emerge 185 ft (56 m) above the valley floor from porous layers of basalt, yielding a flow of over 2 million gallons (7.5 million liters) per minute. The Plain is quite level, except for the striking Craters of the Moon, with lunarlike landscape, where 60 volcanic cones rise over 300 ft (90 m).

3. The Great Basin, or the Basin and Range Province (Fig. 8), an enormous desert region of over 300,000 mi<sup>2</sup> (780,000 km<sup>2</sup>), lies south of the

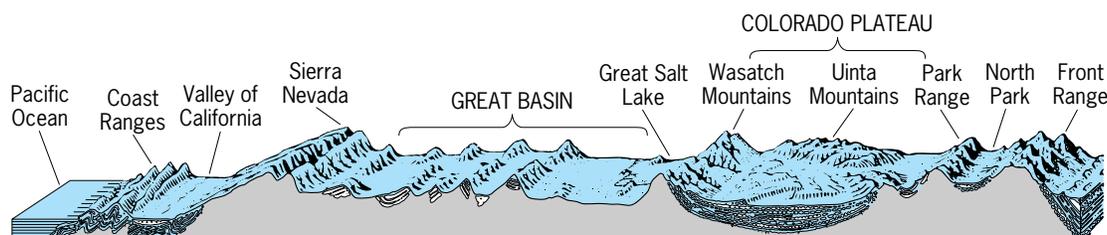


Fig. 8. Geologic cross section from the Rocky Mountains to the Coastal Ranges. (After W. H. Yoho and E. C. Pirkle, *Natural Landscapes of the United States*, 4th ed., Kendall/Hunt, 1985).

Columbia Plateau; it occupies Nevada, western Utah, southern California, southern Arizona, and parts of New Mexico. It has a strikingly different landscape from that of any surrounding regions. It is composed of nearly straight, parallel, and north-south-trending high ranges, with steep sides separated by flat desert basins. The ranges are a result of block faulting, which began in Tertiary time and continues. The basins often contain intermittent lakes, or playas. See PLAYA.

The ranges are 50–100 mi (80–160 km) long and 5–15 mi (8–25 km) wide, rising above the basins up to 6000 ft (1810 m), and reaching elevations of 10,000–13,000 ft (3030–3940 m). The ranges are generally asymmetrical, with one steeper face.

In addition to being unique in terrain characteristics, the region is a complex mosaic of different rock types and ages found adjacent to each other. Another unique characteristic of this area is the internal drainage of the northern part (the Great Basin) and its hydrologic history dating back to the cooler and more rainy climates of the Pleistocene Epoch (pluvial periods). The major river of this internal drainage area is the Humboldt, which flows across Nevada and disappears in Carson Lake. The largest lake is the Great Salt Lake in northwestern Utah, covering about 2000 mi<sup>2</sup> (5200 km<sup>2</sup>), with a depth of about 15–20 ft (4.5–6 m). This is a remnant of the former Lake Bonneville, which was 10 times as large and probably 1000 ft (300 m) deep.

In southern California lies the Death Valley, a downfaulted basin, 282 ft (85 m) below sea level, surrounded by some of the highest ranges and creating the greatest amount of relief in the Basin and Range region.

The Basin and Range area has several distinct subsections: the Great Basin, the Mojave and Sonoran desert, the Mexican Highlands, and the Sacramento Mountains; it continues as an extensive intermountain basin and range region into the Central Plateau of Mexico.

The Great Basin in Nevada has the highest, longest, and most closely spaced north-south-trending ranges. The Sonoran Desert in southern

California has lower, irregular, and less oriented and less parallel ranges, which show a great amount of erosion. The Salton Sea, a downfaulted basin at 235 ft (71 m) below sea level, and Imperial Valley together cover 2500 mi<sup>2</sup> (6500 km<sup>2</sup>).

The Mexican Highlands to the southeast have larger and higher ranges reaching elevations up to 10,000 ft (3030 m; Mount Graham), but the orientation of the ranges remains complex. The Sacramento Mountains include the Guadalupe Mountains with the well-known peak, El Capitan, over 8000 ft (2420 m) high. This is a sheer cliff of fossil limestone reef, the world's largest such cliff known. Nearby are the Carlsbad Caverns, one of the largest limestone cavern systems in the world, reaching over 1000 ft (300 m) deep.

The entire southern part of the Basin and Range Province, in contrast to the Great Basin, has an open drainage mainly by the Colorado and Gila rivers, and the Rio Grande, accounting for greater erosion and removal of debris from this area.

4. The Colorado Plateau (Fig. 9), occupying large parts of Utah, Arizona, Colorado, and New Mexico, covers over 150,000 mi<sup>2</sup> (390,000 km<sup>2</sup>) and is surrounded by the Rocky Mountains on the north and east, and the Basin and Range Province on the southeast, south, and west. Its major characteristics are nearly horizontal sedimentary rocks of Paleozoic, Mesozoic, and Cenozoic age, hundreds of canyons reaching the scope of the Grand Canyon of the Colorado River, and endless erosional escarpments that form as the upper rock layers gradually erode away. Large areas of lava flows and hundreds of volcanic cones are also present, as are several igneous intrusive domes. Precambrian metamorphic rocks are present in the lowest part of the Grand Canyon, which reveals the entire sedimentary sequence of the overlying rocks. The north-south profile of the Colorado Plateau (Fig. 9) shows nearly level Paleozoic sedimentary rocks of the southern part. The surface rises northward toward the Grand Canyon, where in the Kaibab Plateau it reaches 8000–9000 ft (2420–2730 m) elevation. From there, the enormous sequence of sedimentary rocks gently dips

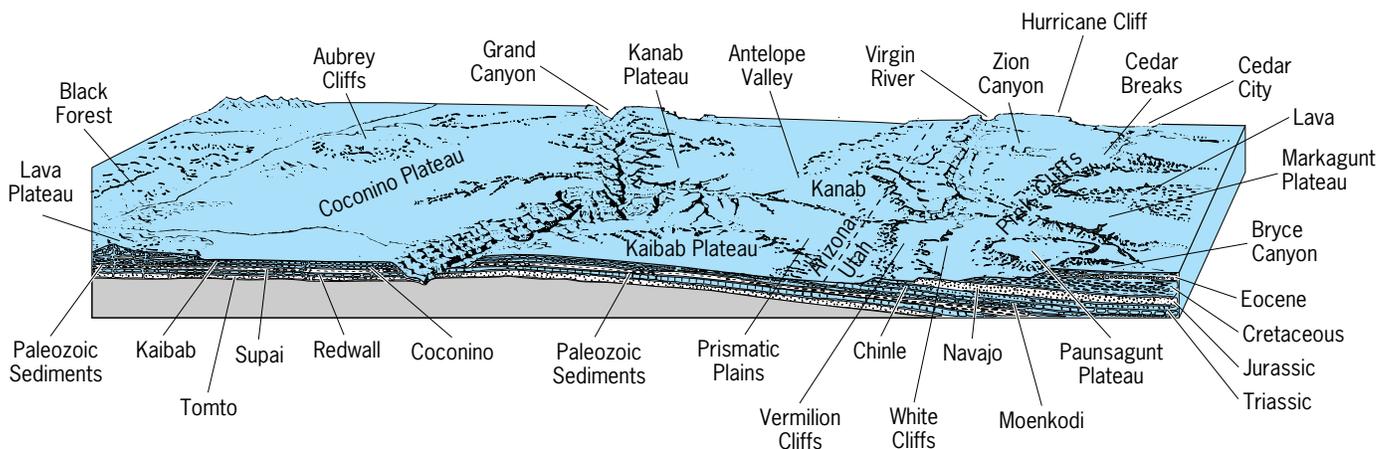


Fig. 9. Profile of the Colorado Plateaus (north-south). (After W. H. Yoho and E. C. Pirkle, *Natural Landscapes of the United States*, 4th ed., Kendall/Hunt, 1985).

northward, and the top layers are eroded into enormous escarpments. These escarpments rise steplike from the lowest Triassic Chocolate Cliffs, to Jurassic Vermillion and White Cliffs, to Cretaceous Gray Cliffs, and to Tertiary Pink Cliffs, each over 1000–2000 ft (300–600 m) high. In the north, the Uinta Basin, containing the same sequence of rocks, is bounded on the south by high escarpments reaching over 4000 ft (1210 m).

The east-west Colorado Plateau profile shows a series of upwarps and enormous faults that have cut the surface into north-south Utah plateaus, including Kaibab (9000 ft or 2730 m), Kanab, Uinkaret, and Shivwits (5000 ft or 1510 m), separated by escarpments, and all bounded on the west by the Grand Cliffs, 3000–4000 ft (900–1210 m) high.

The southwestern part of the Plateau, around San Francisco Mountains (12,800 ft or 3880 m), and the southeastern part, around Mount Taylor (11,400 ft or 3450 m), are extensive Tertiary lava surfaces with hundreds of volcanic cones.

The Central Canyon Lands section is best known for its spectacular 217-mi-long (347-km) Grand Canyon of the Colorado River (Fig. 9), which cuts a 6000-ft-deep (1820-m) canyon into the Paleozoic rocks, with the lowest 1000 ft (300 m) cut into Precambrian metamorphics. The colorful walls of the canyon, which is over 5 mi (8 km) wide, include cliff makers like the White Permian Kaibab limestone (the cap rock), the Mississippian Red Wall limestone, and the Permian Supai sandstone, as well as shales that form the platforms. These horizontal beds of Paleozoic and later sedimentary rocks have been elevated from the bottom of the ocean up to 9000 ft (2730 m), and now extend over large surfaces, like the Kaibab Plateau, or the lower Coconino Plateau. *See PALEOGEOGRAPHY.*

Over the length of the Grand Canyon, the Colorado River flows rapidly over many rapids, dropping from over 2600 ft (780 m) to 1000 ft (300 m) in elevation. Other big canyons are those of the Green, Gunnison, San Juan, and Little Colorado rivers.

Surface features of intrusive origin are the Abaja, La Sal, and Henry mountains, as well as eroding anticlinal folds like San Rafael Swell and Zuni Uplift in the southeast.

The Navajo section south of the Grand Canyon consists of numerous mesas, Painted Desert, and Petrified Forest with great fossilized logs of conifers of Trassic age. *See PETRIFIED FORESTS.*

The Colorado Plateau, with its enormous variety of unique terrain features, has an exceptional number of national parks, including the Grand, Zion, Bryce Canyons, Canyonland, Arches, Petrified Forest, Capital Reef, and Mesa Verde, and Sunset Crater National Monument.

*Coastal Lowlands and Ranges.* These extend along the entire length of North America and include Alaskan Coast Ranges, Aleutian Islands, Alaska Range, Canadian Coast Ranges, and a double chain of the Cascade Mountains and Sierra Nevada on the east, and Coast Ranges on the west, separated by the Puget Sound, Willamette Valley, and the Great Val-

ley of California. These ranges continue southward as Lower California Peninsula, Baja California, and Sierra Madre Occidental in Mexico.

The Alaskan Coast Ranges include the Chugach-Kenai Mountains, Wrangel Mountains, and glacier-covered St. Elias Range. They are extremely rugged due to deep ice erosion. This entire windward coast receives heavy precipitation which, combined with high latitude, forms snow and ice and results in intense erosion by glaciers. The spectacular St. Elias Range of folded and faulted Paleozoic and Mesozoic sedimentary rocks, with Mount Logan at 19,850 ft (6000 m), and many peaks higher than 15,000 ft (4550 m), are covered by numerous glaciers. The Seward Valley glacier spreads out to form the extensive piedmont Malaspina Glacier.

North of these coastal mountains lies the huge arc of the majestic Alaska Range, which continues westward as Alaska Peninsula and Aleutian Islands. The Alaska Range of folded and faulted sedimentary Paleozoic and Mesozoic rocks, and extensive granite intrusions forming the lofty peaks, is 600 mi (960 km) long, and is the highest range in North America, with permanently snow-covered Mount McKinley reaching 20,320 ft (6160 m).

The Alaska Peninsula and the Aleutian Range form a volcanic arc with dozens of volcanoes, many active in recent centuries. Though the Aleutian volcanoes are generally less than 4000 ft (1210 m) high, the northern ones rise to 10,000 ft (3030 m). One of the most violent recent eruptions in this area was that of Mount Katmai, at the beginning of the twentieth century, which spread an enormous amount of ash, gas, and steam at extremely high temperatures.

South of the Aleutian Island chain lies the deep oceanic Aleutian Trench, an area of frequent earthquakes, with submerged peaks exceeding 25,000 ft (7575 m) in height. *See SEAMOUNT AND GUYOT.*

The Canadian Coastal Mountains continue southward from St. Elias Mountains along the Alaska Panhandle and, beginning with Glacier Bay near Juneau, become the beautiful Alexander Archipelago of hundreds of islands separated by deep, narrow, and steep-sided fiords carved out by glaciers. The two largest islands are the Queen Charlotte Island and Vancouver Island, reaching over 7000 ft (2120 m) elevation. These islands represent the outer coastal ranges, while the straits represent the submerged lowland, separating the outer from the inner coastal mountains. The magnificent fiorded coast is backed inland by the high granite Canadian Coast Ranges, reaching up to 13,260 ft (4020 m) in Mount Waddington. These mountains are capped by big glaciers and snow fields and possess striking glacial features such as aretes, cirques, and horns. *See CIRQUE.*

South of the United States boundary begins the double chain of the coastal mountains: the Cascade-Sierra Nevada and the Coastal Ranges. The Cascade Mountains, over 600 mi (960 km) long, consists of two parts differing in geology and terrain. The Northern Cascades (Fig. 7) begin in Canada and continue to Mount Rainier. They are composed mainly of Paleozoic and Mesozoic sedimentary and metamorphic

rocks, eroded into broad ranges, rugged peaks, and very deep valleys carved out by glaciers. Their crest elevations are about 8500 ft (2575 m), but the largest volcanic cones reach over 10,000 ft (3030 m). They are covered by glaciers.

The Middle and Southern Cascades are predominantly a constructed range that was built by lava accumulations from volcanic vent extrusions. The range has long western slopes and a shorter steep slope on the east, along which is a line of volcanic cones including glacier-capped Mount Rainier, the highest (14,410 ft or 4370 m); glacier-capped Mount Adams (12,300 ft or 3730 m); Mount St. Helen, which erupted three times in 1980, reducing its height by 1300 ft (400 m); and Mount Hood, Mount Jefferson, and Three Sisters. The biggest volcanic cone in the south is the beautiful double cone of Mount Shasta, 14,160 ft (4290 m) high and covered by several glaciers. One of the most interesting features is Crater Lake (5 mi or 8 km wide and 2000 ft or 600 m deep), with Little Wizard Island volcano inside. It is located in a caldera of former Mount Mazama, which erupted about 7000 years ago and collapsed, leaving the caldera with rims standing over 2000 ft (600 m) high above the lake. *See CALDERA; VOLCANO.*

Sierra Nevada (Fig. 8), while topographically connected to the Cascades, is a different range geologically. It is a great granite fault block uplifted on its east side, which forms a complete drainage divide with no mountain passes for 150 mi (240 km). The granite intruded the area in Jurassic and Cretaceous time, but remnants of Paleozoic sedimentary rocks exist in the north and in the western foothills. The range is 400 mi (640 m) long and up to 75 mi (120 km) wide, with peaks over 14,000 ft (4250 m), including the highest peak in the 48 contiguous states, Mount Whitney (14,495 ft or 4392 m). This enormous asymmetrical block has its highest crest near the east scarp, which rises at least 9000 ft (2730 m) above the desert basins to the east. Along this east face lies Lake Tahoe (1650 ft or 500 m deep) at an elevation of 6625 ft (2000 m). The western slopes descend gently to the Great Valley of California, but are cut by deep river canyons. The highest parts of the Sierras have been intensely glaciated into horns, cirques, aretes, and spectacular U-shaped valleys, such as the Yosemite Valley. This valley, drained by Merced River, is eroded out of solid granite, forming vertical walls, down which descend waterfalls, including the Upper Yosemite Falls, 1430 ft (430 m) high, and Ribbon Falls, 1600 ft (485 m) high. The granite forms include domes, like the Half Dome, and spires, like the Cathedral Spires. The beautiful trees of this area, the giant sequoia, can be seen in Mariposa Grove, but the most famous are in the Sequoia National Park, near Kings Canyon. These trees are among the largest and oldest living things on Earth. *See SEQUOIA.*

The Coastal Ranges of the United States (Fig. 8) are separated from the Cascade-Sierra Nevada by the Puget Sound Lowland in the north, glaciated by ice lobes descending from the mountains; the Willamette Valley crossed by the lower Columbia River; and the Great Valley of California (Fig. 8)

[400 mi or 640 km long and 50 mi or 80 km wide], the biggest downwarped trough in North America completely surrounded by mountains, and deeply filled with coarse alluvium eroded from the Sierras. The Great Valley is drained by the Sacramento River in the north and the San Joaquin River in the south, but the two rivers join and form an inland delta where they enter the San Francisco Bay.

The Coastal Ranges begin in the north with the Olympic Mountains. This domal mountain group, averaging 5000 ft (1510 m) but with peaks nearly 8000 ft (2420 m) in Mount Olympus, and deep valleys, are very rugged because of intense glacial erosion; over 50 glaciers are still present. The oldest rocks are Cretaceous volcanics and metamorphics, but Cenozoic rocks are found on the margins. The Olympics support great forests of Douglas-fir and red cedar.

South of the Olympic Mountains and parallel to the coast lie the Oregon Coast Ranges, a uniform and low belt of mountains, generally below 2000 ft (600 m) elevation, with a gentle western slope and a short, steep east slope. These mountains are composed of Tertiary sedimentary rocks, mostly sandstones, shales, and limestones, with some igneous rocks forming the higher peaks. Their coastline is an alternation of erosional headlands and narrow sandy beaches.

The Oregon Coast Ranges merge into the high Klamath Mountains, which are composed of hard metamorphic Paleozoic and Mesozoic rocks, with old intrusive rocks forming the highest peaks (up to 9000 ft or 2730 m). They are very rugged and in accessible, with steep slopes and narrow valleys cut by the Klamath river system. The highest central part, the Trinity Alps, shows signs of glaciation. Douglas-fir, redwoods, and sugar pine grow in these wild mountains.

South of the Klamath Mountains stretch the 400-mi-long (640-km) California Coast Ranges (Fig. 8). In the north, they are composed mainly of sedimentary and metamorphic Mesozoic rocks; they have low (2000–4000 ft or 600–1210 m), narrow, even-crested ranges, which are closely spaced, and have a distinct northwest-southeast orientation. The highest peaks reach 8000 ft (2420 m). In northwestern California in the Redwood National Park are giant redwood trees, including the tallest trees known, with heights of 367 ft (111 m).

South of San Francisco, the Santa Cruz and Santa Lucia ranges and Santa Clara and Salinas valleys retain the northwest-southeast orientation, but become much larger. The elevations of the ranges increase to 5000–6000 ft (1520–1830 m). The ranges have complex rock composition of Mesozoic and Cenozoic sedimentary and metamorphic rocks, with some Paleozoics.

At the southern end of the Coast Ranges, the mountain orientation changes to east-west, and elevations increase to a maximum of 8826 ft (2675 m) in Mount Pinos. Here begin the Transverse Ranges, which include the huge fault blocks of Precambrian metamorphics, like San Gabriel, San Jacinto, and

Santa Rosa, rising over 10,000 ft (3030 m), and San Bernardino, nearly 11,500 ft (3480 m). The largest basins, filled with Quaternary sediments, include the Los Angeles, San Bernardino, Ventura, and San Fernando. These ranges merge into the Peninsular Range of California, which is composed of Mesozoic igneous rocks, and has a short, steep eastern face, and long gradual western slopes, with distinct marine terraces north of San Diego. The Los Angeles Basin, extending over 1000 mi<sup>2</sup> (2600 km<sup>2</sup>), is open to the ocean, but is surrounded by high mountains on all other sides. While beautifully located, it is a perfect trap for air pollutants forming smog, held down by both surface and upper-level temperature inversions.

To the south trends Baja California, a nearly 800-mile long (1280-km) belt of mountains, with the highest peak over 10,000 ft (3000 m). It is a fault block with a steep east face overlooking the downfaulted Gulf of California. Its west side is covered by lava flows eroded into plateaulike surfaces. A mass of granite forms its southern tip.

*Mexico and Central America.* The basin-and-range type of terrain of the southwest United States continues into northern Mexico and forms its largest physiographic region, the Mexican Plateau. This huge tilted block stands more than a mile above sea level—from about 4000 ft (1200 m) in the north, it rises to about 8000 ft (2400 m) in the south. The arid northern part, called Mesa del Norte, extends southward to about 22°N latitude, where it merges into higher and more moist southern part, called Mesa Central. The plateau surface consists of north-south-trending ranges rising 500–2500 ft (150–750 m) above adjacent flat desert basins. Limestone and shale rocks dominate the eastern part of the plateau, and volcanic materials the western. The southern half is dominated by recent volcanic activity, and it ends in a volcanic rim, the Transverse Volcanic Axis, which includes snow-capped peaks such as the Popocatepetl (17,900 ft or 5425 m) and Ixtaccihuatl (17,300 ft or 5240 m). These volcanoes formed about 10 million years ago, but some are still active. Immediately north of these volcanoes are the former lake basins now forming fertile plains. One of the largest is the Valley of Mexico, in which Mexico City is located.

The Plateau is surrounded by striking escarpments. On the east is the Sierra Madre Oriental of elongated limestone ranges trending north-south and rising to 8000 ft (2420 m), with peaks reaching 13,000 ft (3940 m). It merges in the south with volcanic snow-capped peaks such as Orizaba (18,700 ft or 5660 m), the highest peak in Middle America. The western rim, the Sierra Madre Occidental, consists mainly of volcanic rocks, cut by deep and beautiful canyons.

The Mexican Plateau is separated from the Southern Mexican Highlands (Sierra Madre del Sur) by a low, hot and dry Balsas Lowland drained by the Balsas River. The average elevators of the Southern Highlands are about 8000 ft (2420 m), with peaks reaching over 10,000 ft (3030 m), and V-shaped valleys cut 4000 ft (1200 m) down, leaving narrow crests between.

To the east of the Southern Highlands lies a lowland, the Isthmus of Tehuantepec, which is considered the divide between North and Central America. Here the Pacific and Gulf coasts are only 125 mi (200 km) apart.

The lowlands of Mexico are the coastal plains. The Gulf Coastal Plain trends southward for 850 mi from the Rio Grande to the Yucatan Peninsula. It is about 100 mi (160 km) wide in the north, just a few miles wide in the center, and very wide in the Yucatan Peninsula. Barrier beaches, lagoons, and swamps occur along this coast.

The Pacific Coastal Plains are much narrower and more hilly. North-south-trending ridges of granite characterize the northern part, and islands are present offshore. Toward the south, sandbars, lagoons, and deltaic deposits are common.

East of the Isthmus of Tehuantepec begins Central America with its complex physiographic and tectonic regions. This narrow, mountainous isthmus is geologically connected with the large, mountainous islands of the Greater Antilles in the Caribbean. They are all characterized by east-west-trending rugged mountain ranges, with deep depressions between them. One such mountain system begins in Mexico and continues in southern Cuba, Puerto Rico, and the Virgin Islands. North of this system, called the Old Antillia, lies the Antillian Foreland, consisting of the Yucatán Peninsula and the Bahama Islands, which are level, low-lying plains underlain by limestone. Northern Yucatán is a limestone surface highly affected by solution, so that there are no surface streams. Rainwater sinks down through sinkholes and flows underground.

Central American mountains are bordered on both sides by active volcanic belts. Along the Pacific, a belt of young volcanoes extends for 800 mi (1280 km) from Mexico to Costa Rica. The youngest volcanoes include over 40 large peaks, which culminate in the Fuego (12,600 ft or 3800 m) in Guatemala.

Costa Rica and Panama are mainly a volcanic chain of mountains extending to South America. Nicaragua is dominated by a major crustal fracture trending northwest-southeast. In the center of this lowland are the two largest lakes of Middle America: Lake Managua and Lake Nicaragua. At the northwestern end of this lowland lies the Gulf of Fonseca, the largest gulf along the Pacific shore of Middle America.

### Natural Disasters

North America is subjected to more natural disasters than most continents. For example, along the entire length of the California coast runs a northwest-southeast zone of active crustal fractures, the San Andreas Fault Zone. The continent on the western side of this fault is moving northwestward at an estimated rate of about 2 in. (5 cm) per year, making this region the greatest earthquake hazard zone in North America. *See* EARTHQUAKE; FAULT AND FAULT STRUCTURES.

In addition to the earthquakes, steep-sloped mountainous California is subject to great mudflows when

it rains, as well as enormous forest fires during the drought season.

All coastal western cordillera are subject to volcanic eruptions, like those of Mount Katmai in Alaska and Mount St. Helens in Washington State. *See* VOLCANO.

The interior deserts are subject to dust storms, and flash-floods and mud flows.

The Great Plains are subject to great summer droughts and dust storms, as well as blizzards in the winter, and are annually devastated by violent tornadoes (average of 50 per year in Oklahoma). *See* DUST STORM.

The Central Lowlands have suffered from both prolonged great droughts and excessive precipitation, often leading to enormous floods, such as the extremely destructive hundred-year floods of the Mississippi River system in 1993.

The southeastern part of the United States is subject to frequent tropical hurricanes in which violent winds, enormous rainfall, and towering oceanic waves destroy the coastal areas. *See* HURRICANE.

The northern and northeastern parts of the continent suffer nearly every winter from extremely low temperatures, and hazardous ice and snow storms.

Mexico and Central America are intensely affected by frequent and devastating volcanic eruptions, earthquakes, and tropical hurricanes.

Barbara Zakrzewska Borowiecka

*Bibliography.* W. W. Atwood, *The Physiographic Provinces of North America*, 1940; J. B. Bird, *The Natural Landscapes of Canada*, 1972; E. M. Bridges, *World Geomorphology*, 1990; N. M. Fenneman, *Physiography of Eastern United States*, 1938; N. M. Fenneman, *Physiography of Western United States*, 1931; W. L. Graf, *Geomorphic Systems of North America*, Geological Society of America, Centennial Special Volume 2, 1987; D. V. Harris, *The Geologic Story of the National Parks and Monuments*, 1980; J. A. Henry and J. Mossa, *Natural Landscapes of the United States*, 5th ed., 1995; C. B. Hunt, *Natural Regions of the United States and Canada*, 1974; T. L. McKnight, *Physical Geography: A Landscape Appreciation*, 1992; A. N. Strahler and A. H. Strahler, *Elements of Physical Geography*, 4th ed., 1990; W. D. Thornbury, *Regional Geomorphology of the United States*, 1965; R. C. West and J. P. Augelli, *Middle America, Its Lands and People*, 3d ed., 1989.

## North Pole

That end of the Earth's axis which points toward the North Star, Polaris (Alpha Ursae Minoris). It is the geographical pole located at 90° N latitude where all meridians converge, and should not be confused with the north magnetic pole, which is in the Canadian Archipelago. The North Pole's location falls near the center of the Arctic Sea. Being at the end of the Earth's axis, which is inclined 23½° (23°.45) from a line perpendicular to the plane of the ecliptic, the North Pole has phenomena unlike any other place

except the South Pole. For 6 months the Sun does not appear above the horizon, and for 6 months it does not go below the horizon. During this latter period, March 21–September 23, the Sun makes a very gradual spiral around the horizon, gaining altitude until June 21; then it starts to lose altitude until it disappears below the horizon after September 23. The Sun's highest altitude is 23 1/2°. As there is a long period (about 7 weeks) of continuous twilight before March 21 and after September 23, the period of light is considerably longer than the period of darkness.

There is no natural way to determine local Sun time because there is no noon position of the Sun, and all shadows, at all times, fall to the south, the only direction possible from the North Pole. *See* MATHEMATICAL GEOGRAPHY; SOUTH POLE; TERRESTRIAL COORDINATE SYSTEM.

Van H. English

*Bibliography.* A. H. Strahler and A. Strahler, *Introducing Physical Geography*, 4th ed., 2005.

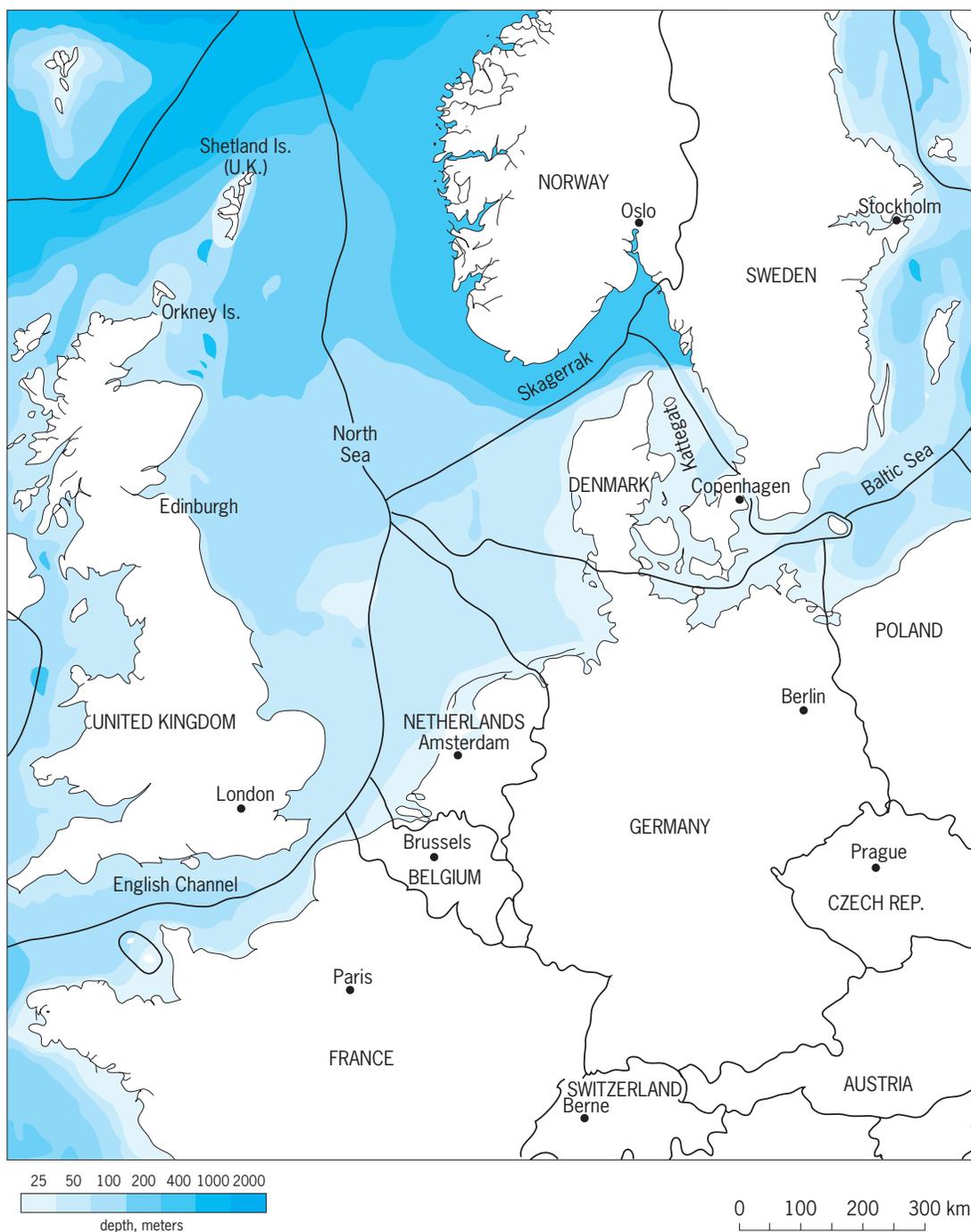
## North Sea

A flooded portion of the northwest continental margin of Europe occupying an area of over 200,000 mi<sup>2</sup> (500,000 km<sup>2</sup>). The North Sea has extensive marine fisheries and important offshore oil and gas reserves. In the south, its depth is less than 150 ft (50 m), but north of 58° it deepens gradually to 600 ft (200 m) at the top of the continental slope. A band of deep water down to 1200 ft (400 m) extends around the south and west coast of Norway and is known as the Norwegian Trench (see *illus.*).

**Geologic history.** The area of the North Sea overlies a large sedimentary basin that has gradually been accumulating sediments since Permian time (about 2.5 × 10<sup>8</sup> years ago). Since then, gradual subsidence has been occurring between the stable massifs of Scotland, Scandinavia, and Central Europe where the old basement rocks lie at or near the surface. *See* MASSIF; PERMIAN.

The most recent geological event (about 2 million years ago) to affect the area was the Pleistocene glaciation when ice from the Scandinavian ice caps extended across the North Sea to the British coast. Many of the present topographical features of the North Sea date from this period. *See* PLEISTOCENE.

**Currents and water level.** Fluctuations in currents and water level are predominantly due to the semidiurnal tide which propagates into the North Sea from the adjacent Atlantic Ocean. The largest tidal amplitudes (6–18 ft or 2–5 m) and largest currents (1.5–6 ft/s or 0.5–2 m/s) are in the west and south. Adjacent to the Norwegian coast, tidal amplitudes and currents are low (<3 ft or 1 m; 0.3 ft/s or 0.1 m/s). Numerical modeling of the North Sea tide has demonstrated the existence of amphidromic points (no change in tidal amplitude) off southeastern England, east of Denmark, and close to the southern coast of Norway. Storms can considerably modify mean levels and currents, especially in the southern North Sea where storm surges can raise the level by more than 6 ft (2 m), and real-time storm warning



North Sea, showing depth. (Adapted from *The North Sea Secretariat, Oslo, Norway*)

schemes to predict their occurrence are in operation. See STORM SURGE; TIDE.

The nontidal residual current circulation of the southern North Sea is mainly determined by wind velocity, but in the north, well-defined non-wind-driven currents have been identified, especially in the summer. Two of these currents bring in water from outside the North Sea; one flows through the channel between Orkney and Shetland (the Fair Isle current), and the other follows the continental slope north of Shetland and merges with the Fair Isle current southwest of Norway before entering the Skagerrak. The north-flowing Norwegian coastal cur-

rent provides the exit route for North Sea waters, and is formed from the waters of these two major inflows and from other much smaller inputs such as river runoff, the English Channel, and the Baltic Sea.

The transport of the Norwegian coastal current is typically  $5.3\text{--}8.8 \times 10^7 \text{ ft}^3/\text{s}$  ( $1.5\text{--}2.5 \times 10^6 \text{ m}^3/\text{s}$ ), the smaller figure being associated with winter. The Fair Isle current has a transport of about  $1.1 \times 10^7 \text{ ft}^3/\text{s}$  ( $300,000 \text{ m}^3/\text{s}$ ), the Norwegian Trench inflow of Atlantic water making up the balance. In late winter the Fair Isle current is a weakly defined feature, especially in the west. See ATLANTIC OCEAN; OCEAN CIRCULATION.

**Water properties and biology.** Average salinities have very small seasonal variations and are typically 35, increasing to 35.3 in the north, adjacent to the Atlantic Ocean, and also occasionally off the east end of the English Channel. In coastal regions, and the southern North Sea generally, salinities are less than 35 but are mostly greater than 30. Vertical variations in salinity are minor except adjacent to the Norwegian coast, within the coastal current, where salinities in the upper 120 ft (40 m) are some 2–3 less than midwater and near-bottom salinities.

Minimum surface temperatures occur in February and range from 45°F (7°C) in the northwest to 36°F (2°C) adjacent to the Danish and German coasts. In the summer, this gradient is reversed, with August temperatures of about 63°F (17°C) in the south ranging to only 52.7°F (11.5°C) in the northwest. The near-bottom waters of the central and northern parts of the North Sea maintain winter temperatures until late October, when convective overturn associated with autumnal cooling and strong winds mixes the water column. Before this overturn occurs, however, the bottom waters become slightly depleted in oxygen; otherwise, oxygen levels are near saturation level throughout the North Sea. Levels of inorganic nutrients such as phosphate, silicate, and nitrate reflect the high fractional content of nutrient-rich Atlantic waters, but summer surface values are usually depleted because of the spring outburst of phytoplankton. This outburst occurs in March in the permanently stable Norwegian coastal waters, but not until late May and June in the tidally mixed coastal waters of the British Isles.

During late autumn and winter, there is deep mixing within the water column and nutrient concentrations are high. Nevertheless, phytoplankton growth is extremely low, mainly due to limited daylight. During spring and summer, the water column stratifies and the phytoplankton are retained within a shallow wind-mixed surface layer. When this stratification first develops, conditions within this shallow surface layer are ideal for phytoplankton growth as nutrients are abundant and there is plenty of light. This can result in the development of algal blooms. These blooms are usually short-lived, however, as the nutrients in the mixed surface layer are soon used up. Blooms generally occur in the spring and autumn, that is, at the beginning and end of stratification. See PHYTOPLANKTON.

There is a rich diversity of zooplankton within the North Sea. Copepods are of particular importance in the food web, with *Calanus finmarchicus* being the characteristic species of the northern North Sea, while smaller copepods (for example, *Temora longicornis*) are most important elsewhere. Some species of zooplankton enter the North Sea from the Atlantic Ocean during the summer and autumn. In general, they arrive in small numbers, but *Salpa fusiformis* often forms dense swarms and may reduce the amount of food available for the indigenous plankton. See COPEPODA; MARINE ECOLOGY; SEAWATER; ZOOPLANKTON.

**Marine contamination.** Problems resulting from the input of anthropogenically derived nutrients are widespread in coastal areas throughout the North Sea, in various estuaries and fiords, in the Wadden Sea, and in the German Bight. They contribute to the development of exceptional algal blooms and exceptional densities of macroalgae, to changes in the composition of the phytoplankton and zoobenthos communities, and to intermittent oxygen depletion in the bottom water and the subsequent death of benthic fauna. Ninety percent or more of the freshwater-borne nutrients to the North Sea are from a few main rivers (over half from the Rhine).

Many harmful algal blooms have been observed in the North Sea, ranging from small, localized blooms to blooms covering vast areas, such as the *Phaeocystis* bloom that has been observed intermittently in the Wadden Sea since the early 1970s. *Phaeocystis* and *Coscinodiscus* blooms recur on the southeastern and eastern North Sea coasts; *Coscinodiscus* is harmful to fish because it causes increased mucous production, which occasionally restricts gill function. *Gyrodinium aureolum*, first noticed in northern European waters in 1966, is now one of the most common dinoflagellates in the autumn. There have been frequent blooms since 1981, and these have often caused significant mortalities of farmed fish. A number of these blooms appear to have started in the open waters of the Skagerrak and Kattegat and have then spread northward with the coastal current. In 1988 an extremely large bloom of the small flagellate *Chrysochromulina polylepis* affected the Kattegat, Skagerrak, and Sound, covering an area of 75000 km<sup>2</sup>, and produced a toxin which killed or affected various marine organisms, including mussels and wild and cultured fish.

There are many studies concerning the levels of inorganic and organic contaminants in fish from the North Sea. Concentrations in fish are often elevated in areas with fine-grained sediments, particularly in bays and estuaries close to industrialized areas, and depositional areas such as the Dogger Bank, the Oyster Ground, the Wadden Sea, the German Bight, and the Norwegian Trench.

In recent years, general decreases in metal concentrations have been reported. Examples include cadmium, lead, and copper in the region to the north of the Rhine-Meuse and in the Wadden Sea. Metal concentrations in sediments from estuaries tend to be higher than those from coastal areas; in some estuaries, however, there is evidence of decreasing concentrations, such as cadmium and lead in the Scheldt.

**Fisheries.** There is a wide range of fish stocks in the North Sea and adjacent waters and, in terms of species exploited by commercial fisheries, they constitute the richest area in the northeast Atlantic. The commercially important stocks exploited for human consumption include cod, haddock, whiting, pollock, plaice, sole, herring, mackerel, lobster, prawn, and brown shrimp (*Crangon crangon*). A number of stocks are used for fishmeal and oil; these stocks include sand eel, Norway pout, blue whiting, and sprat. Beyond these stocks, there are landings from a

variety of demersal (found near the sea floor) species such as turbot, anglerfish, gurnards, lemon sole, rays, and sharks.

A basic problem is that the North Sea has been exposed to intensive fishing for more than a century. This has implications for both the state of the fish stocks and the environment of the North Sea. As a result, the Intermediate Ministerial Meeting of the 5th International Conference on the Protection of the North Sea called for "further integration of fisheries and environmental protection, conservation and management measures, drawing upon the development and application of an ecosystem approach" and "integration of environmental objectives into fisheries policy." See MARINE FISHERIES.

**Petroleum exploration.** The first major discovery of oil in the North Sea was made in 1969 at Ekofisk. The major developments of the offshore oil industry have been in the northern North Sea, in the United Kingdom and Norwegian sectors. Gas deposits have been exploited mainly in the shallower southern regions in the United Kingdom, Dutch, and Danish sectors, as well as in Norwegian waters. From 1990 to 1996 the number of offshore platforms and oil production almost doubled, primarily reflecting increased activity in the Norwegian and United Kingdom sectors.

Offshore installations are significant sources of oil contamination in the North Sea. Overall, inputs of oil have reduced from a maximum of some 28,000 tons in 1985 to about 10,000 tons in 1997, but the contribution from produced water has increased pro-

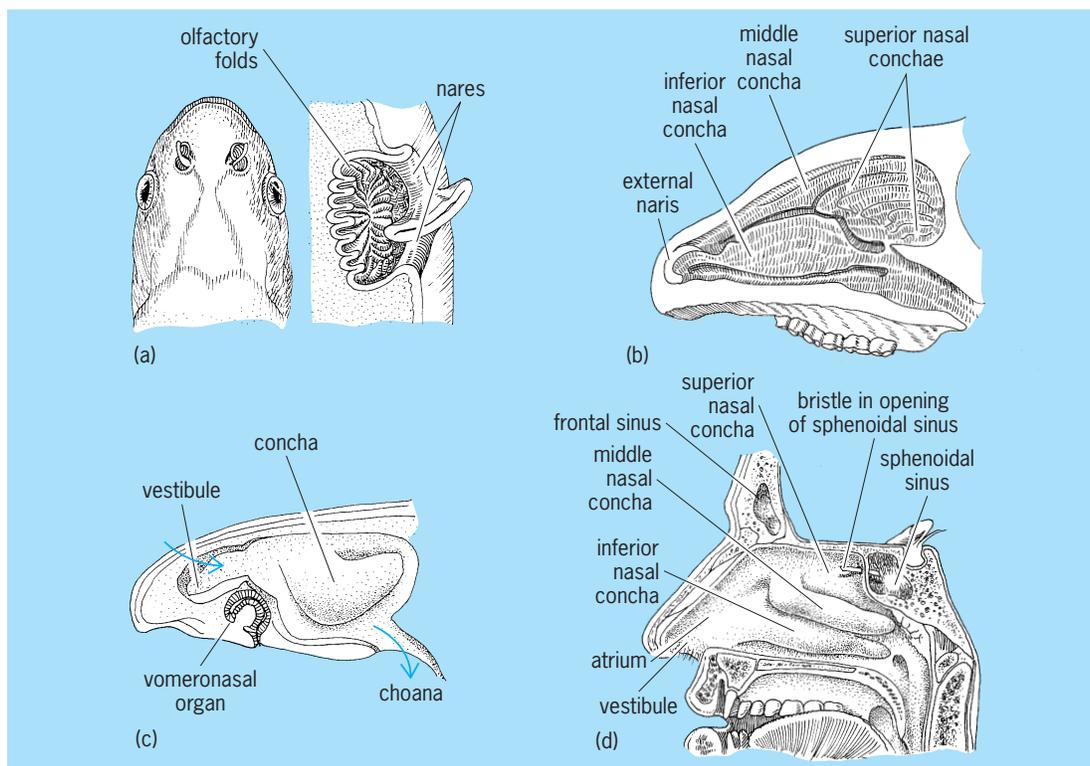
gressively as oil fields have matured and the number of installations has increased. See OIL AND GAS, OFFSHORE.

H. D. Dooley

**Bibliography.** J. Andersen and T. Niilonen (eds.), *Progress Report. 4th International Conference on the Protection of the North Sea*, Esbjerg, Denmark, June 8-9, 1995; 5th North Sea Conference Secretariat, *Intermediate Ministerial Meeting on the Integration of Fisheries and Environmental Issues*, Bergen, Norway, March 13-14, 1997; J. S. Gray et al., Managing the environmental effects of the Norwegian oil and gas industry: From conflict to consensus, *Mar. Pollut. Bull.*, 38:525-530, 1999; NSTF Oslo and Paris Commissions, *North Sea Quality Status Report*, Olsen and Olsen, Fredensborg, Denmark, 1993; G. Radach and H. J. Lenhart, Nutrient dynamics in the North Sea: Fluxes and budgets in the water column derived from ERSEM, *Neth. J. Sea Res.*, 33:301-335, 1995; Report of the ICES Advisory Committee on Fishery Management, *ICES Coop. Res. Rep.*, no. 229, 1999; E. Svendsen and A. K. Magnusson, Climatic variability in the North Sea., *ICES Mar. Sci. Symp.*, 195:144-158, 1992.

## Nose

The nasal cavities and the structures surrounding and associated with them. The nose functions primarily as the organ of smell and in most tetrapods also assumes a respiratory function, forming the anterior end of the air passage through which air is drawn in and in which it is warmed and moistened.



Nose in various vertebrates. (a) Minnow. (b) Sheep (after H. Giersberg and P. Rietschel, *Vergleichende Anatomie der Wirbeltiere, Erster Band, Fischer, 1967*). (c) Lizard (after W. W. Ballard, *Comparative Anatomy and Embryology, Ronald, 1964*). (d) Human (after W. J. Hamilton et al., *Textbook of Human Anatomy, Macmillan, 1956*).

In most fishes the nasal cavities are a pair of pits, one on each side of the snout. Although the cavities frequently have two external openings, they are rarely connected with the mouth. In lampreys and hagfishes there is only a single median nasal cavity, and in a few fishes there are choanae, or openings between the nasal and oral cavities. The nose contains folds which increase the surface area covered by sensory epithelium (*illus. a*).

In tetrapods each nasal cavity has two openings, one to the outside (external naris) and one to the oral cavity (internal naris or choana). In some forms the cavity is a simple sac, but commonly projections develop on the lateral wall (*illus. b, c, and d*); these are termed conchae or, in mammals, turbinals. Reptiles have one to three conchae, and most birds have three. However, in mammals the conchae may be more numerous. They are relatively few and small in humans, whose sense of smell is not very well developed.

In humans the nasal cavities are triangular openings that pass from the external nares back to the dorsal part of the pharynx (*illus. d*). The lateral walls are composed principally of portions of the ethmoid and sphenoid bones and projections of three turbinate bones, or conchae, on each side. The floor of the nose is formed by the palate, which is also the roof of the mouth. The nasal cavities are lined with respiratory epithelium, which also lines the paranasal sinuses. The latter are cavities in the frontal, ethmoid, sphenoid, and maxillary bones which communicate with the nasal passages. The external nose consists of the two nasal bones that form the bony bridge and two pairs of lower nasal cartilages. These together with the tightly adherent skin determine the individual shape and size of the human nose.

Numerous blood vessels, lymphatics, and nerves supply and drain both the external and the internal portions of the nose. *See* OLFACTION.

Thomas S. Parsons

## Nose cone

The forward portion of a spacecraft that is designed for atmospheric entry. Nose cones are utilized for intercontinental ballistic missiles and crewed spacecraft such as Apollo and space shuttles. The nose cone is required to withstand heating encountered during atmospheric entry, maintain the structural integrity of the spacecraft, prevent overheating of the payload, and usually maintain the aerodynamic characteristics of the spacecraft. The requirements for a safe entry pose many challenging problems to the designer. Nature provides a dramatic demonstration of the scope of the problem in the form of so-called shooting stars. These bodies are meteoroids which are heated to incandescence as they enter the Earth's atmosphere and which, with few exceptions, destroy themselves prior to Earth impact. *See* ATMOSPHERIC ENTRY; METEOR.

**Amount of heating.** An Apollo spacecraft returning from the Moon enters the Earth's atmosphere at

about 7 mi/s (11 km/s). Braking action must reduce this velocity  $V$  to practically zero at Earth impact; that is, most of the initial kinetic energy of the spacecraft must be absorbed by the atmosphere prior to impact. The spacecraft's initial kinetic energy, per unit of spacecraft mass, is  $V^2/2$ , which is equivalent to a significant thermal energy. A calculation shows each unit of mass has the energy capacity to melt 250 times as much steel. Hypothetically, then, if the spacecraft is made of steel, it possesses enough energy to destroy itself 250 times over. As the spacecraft decelerates in the Earth's atmosphere, it experiences frictional heating in the boundary layer at its surface, and the nose cone is also subjected to heat from gases that are at elevated temperatures as a result of being decelerated by the bow shock wave. The designer must prevent all but a small fraction of the energy in the gas surrounding the nose cone from reaching it in the form of heat; that is, the designer must prevent the spacecraft from absorbing all its kinetic energy. *See* AEROTHERMODYNAMICS; BOUNDARY-LAYER FLOW; SHOCK WAVE.

**Heating rate.** The amount of heat transferred into the nose cone depends on its shape and the materials from which it is made. The bow shock wave heats the gases behind it. The heat reaches the nose cone in the form of convection and radiation through the boundary layer adjacent to the surface. *See* HEAT TRANSFER.

For a slender body with a sharp nose (**Fig. 1**), the bow shock is relatively weak and lies fairly close to the body, resulting in small wave drag. A proportionately small amount of air is heated, and the friction drag is high. For a blunt body (**Fig. 2**) the bow shock is much stronger, extends farther away from the vehicle sides, creates larger wave drag, and heats a considerably greater gas volume than its sharp, thin counterpart. The heat applied to the spacecraft in the latter case is markedly less than the former case, because a greater fraction is absorbed in heating the atmosphere. Many spacecraft, including the Apollo, have employed the blunt-body concept to survive the entry maneuver when convection is the primary mode of heat transfer. However, the blunt body is not a panacea for all designs, since radiation from the incandescent gas behind the stronger shock wave of the blunt body becomes a significant mode of heat transfer as the entry velocity increases.

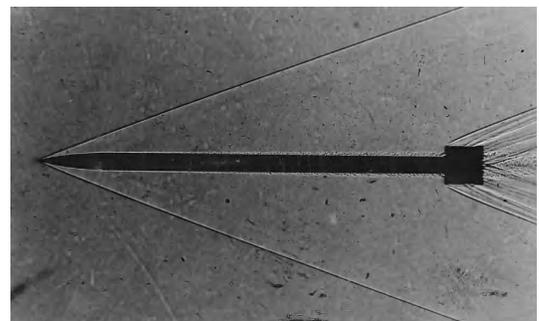


Fig. 1. Flow about slender body.

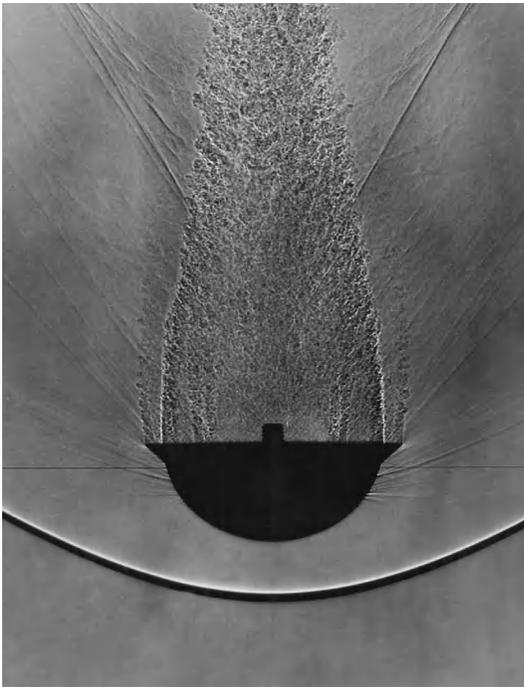


Fig. 2. Flow about blunt body.

**Real-gas effects.** At high entry velocities the translational energy of the gas behind the bow shock as measured by its temperature becomes comparable with the energies associated with various molecular and atomic processes, such as dissociation and ionization. Under these conditions the energy that is involved in these processes must be taken into account when calculating the flow field. The thermodynamic and transport properties of the gas differ from those of a perfect gas, and these changes are significant. For example, the distance of the bow shock from the body depends on the specific-heat ratio, which can change significantly as a result of real-gas effects. At high enough velocities the number of free electrons produced becomes appreciable and the gas becomes a conductor of electricity. The presence of free electrons in the shock layer alters the dielectric properties of the medium and can have large effects on radio communication to the vehicle. The design of the entry body requires the solution of serious problems associated with real-gas effects, such as the prediction of shock shapes, flow fields, and pressure distributions for high velocities and Mach numbers. *See* GAS DYNAMICS.

**Surface thermal protection.** Even for a properly designed shape, it is inevitable that some fraction of the spacecraft's initial kinetic energy will finally reach the nose cone in the form of heat. The design of the heat shield for the nose cone is a complex procedure, which is highly dependent on the heating level. There are a variety of surface-protection or cooling systems which have been used. Generally these systems consist of heat sinks of various types: the absorption of heat by virtue of a material's sensible heat capacity, latent heat capacity, or chemical heat capacity. That is, heat absorption is accomplished

by a temperature rise, a phase change, or a chemical reaction. Aerodynamic lift is employed by reusable space shuttles to lower heating rates so that the nose cone material can radiate away much of the incident heating.

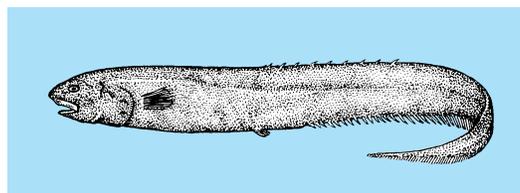
Ablation is used to provide surface protection. The designer can divert heat from the spacecraft by allowing the nose cone's outer layer of material to melt, vaporize, or sublime. Materials which vaporize or sublime have the advantage of a blocking action which, by thickening and cooling the boundary layer, prevents a sizable amount of heat from being convected from the hot gas stream to the spacecraft. Additionally, some materials form a char layer of porous graphite refractory residue as the outer layers decompose. The char's outer surface can reach high temperatures, so that it radiates sizable amounts of energy away from the spacecraft and thereby decreases the heat transfer. While large ablation rates provide excellent thermal protection, the resulting change in profile due to surface recession can adversely change the aerodynamic characteristics of the spacecraft. The designer must account for this change. In addition, the designer must provide adequate strength to prevent mechanical erosion of the char by aerodynamic shear stresses. *See* SPACE FLIGHT; SPACE SHUTTLE; SPACECRAFT STRUCTURE.

Philip R. Nachtsheim

## Notacanthoidei

A suborder of Albuliformes consisting of two families, Halosuridae (halosaurs) and Notacanthidae (spiny eels). They are also known as Lyopomi and Heteromi, respectively. The body of these fishes is elongate and tapers posteriorly. The caudal fin is absent or essentially absent; pectoral fins are high on the body; pelvic fins are abdominal; and the anal fin is long (see **illustration**). There is no duct to the swim bladder (physoclistous); the orbitosphenoïd, pterosphenoid, intercalary, and basisphenoid bones are absent; the posttemporal is simple or ligamentous; the transverse processes are not suturally joined to vertebral centra; and there is no mesocoracoid arch. Some species have photophores (light emitting organs), a characteristic of many deep-sea fishes.

The halosaurs differ from spiny eels in lacking spines in the dorsal fin and having gill membranes that are completely separate rather than joined or partly joined.



Spiny eel (*Notacanthus nasus*). (After D. S. Jordan and B. W. Evermann, *The Fishes of North and Middle America*, U.S. Nat. Mus. Bull. 47, 1900)

The Notacanthoidei is a small group with a history extending back to the Upper Cretaceous; it includes 2 families, 6 Recent genera, and about 25 species. The species inhabit deep seas worldwide, in waters between 125 and 4900 m (413 and 16,1700 ft). They are like true eels (Anguilliformes) in lacking a firm suspension of the pectoral girdle from the skull, but some have fin spines similar to those of perciform fishes. Of much interest was the discovery of a pelagic leptocephalous elongate and flattened from side to side larval stage among the halosaurs. The larvae are strikingly similar to those of true eels, testifying to a close relationship, a conclusion that is strengthened by the similarities in the swim bladders of these groups. The leptocephalous larva is characteristic of teleost fishes in the subdivision Elopomorpha. See ALBULIFORMES; ANGUILLIFORMES; ELOPOMORPHA; PHOTOPHORE GLAND. Reeve M. Bailey; Herbert Boschung

**Bibliography.** P. H. Greenwood, Notes on the anatomy and classification of elopomorph fishes, *Bull. Brit. Mus. Nat. Hist. (Zool.)*, 32(4): 65-102, 1977; N. B. Marshall, Observations on the Heteromi, an order of teleost fishes, *Bull. Brit. Mus. (Nat. Hist.)*, vol. 9, no. 6, 1962; J. S. Nelson, *Fishes of the World*, 3d ed., 1994.

## Nothosauria

An extinct clade (evolutionary lineage) of eosauroptrygian reptiles (that is, a suborder of Eosauroptrygia) that also includes pachypleurosaur (suborder Pachypleurosauria) and pistosaurs (suborder Pistosauroidea). Representatives of these groups are known from uppermost Lower Triassic to Upper Triassic marine deposits in Europe, the Near and Middle East, Northern Africa, China, and North America. Triassic nothosaurs and their relatives are typically restricted to near-shore habitats or shallow epicontinental seas. Some derived members of the pistosaurs constitute the sister group (closest relatives) to plesiosaurs, a clade that greatly diversified during the Jurassic and Cretaceous. See PLACODONTIA; PLESIOSAURIA; SAUROPTERYGIA.

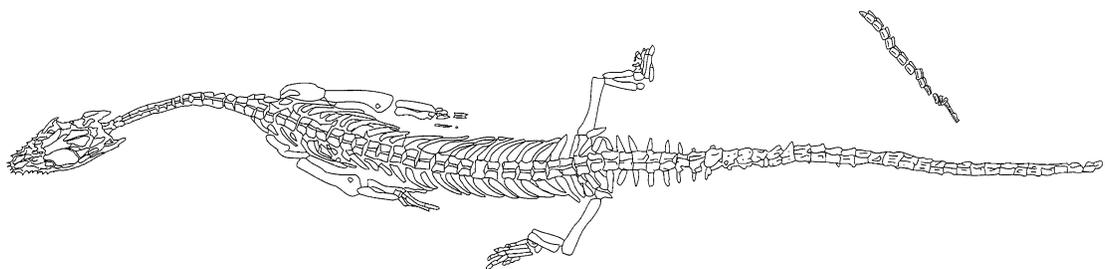
**Morphology.** Nothosaurs in the broad sense (that is, including pachypleurosaur and pistosaurs) range approximately 3-13 ft (1-4 m) in length (see **illustration**). Although their limbs are reduced in size relative to those of their terrestrial predecessors, they are not highly modified for aquatic propulsion. The neck is long, with at least 15 cervical vertebrae.

Like aquatic lizards and crocodiles, early nothosaurs probably swam primarily by lateral undulation of the trunk and tail, with the limbs held to the side to reduce drag. Advanced nothosaurs show modifications of the front limbs to act as paddles; the rear limbs are somewhat reduced. In many genera, the ribs are expanded and heavily ossified (a condition termed pachyostosis) to reduce buoyancy. Some genera have as many as six pairs of sacral ribs.

Nothosaurs (and placodonts) are distinguished from other aquatic reptiles by the closure of the openings in the palate that occur in more primitive reptiles, and by the covering of the base of the braincase. Secondarily developed openings in the palate indicate close relationships of advanced pistosaurs with plesiosaurs. The underside of the shoulder girdle is characterized by a wide gap between the base of the scapulae and the coracoids, behind a stout transverse bar formed by the interclavicle and the blades of the clavicles. As in plesiosaurs (and placodonts), part of the scapula is located on the outside of the clavicle, a reversal of the relationship of these bones in other reptiles.

**Evolutionary trends.** Interesting evolutionary trends are observed in the skull structure of nothosaurs. The relatively small pachypleurosaur had a short and broad skull, and the jaws were furnished with small, vertically oriented peglike teeth. Rapid opening of the mouth and simultaneous depression of the throat skeleton would have created a suction effect to capture prey, which is similar to the method used by aquatic turtles today. More advanced members of the nothosaurs evolved a narrow and strongly depressed skull with a much-elongated rostrum (elongated snout) that was furnished with strongly procumbent, fanglike teeth in its anterior part. These pincer-jaws allowed the capture of fish and cephalopod prey with rapid lateral strikes of the head. The very low profile of the skull reduced drag as the head struck sideways, but necessitated important changes in the arrangement of the jaw-closing musculature. *Nothosaurus giganteus* was one of the top carnivores of Triassic near-shore habitats, with a documented skull length of nearly 3 ft (up to 85 cm).

**Phylogeny.** Primitive nothosaurs resemble primitive lepidosauromorphs in the pattern of the skull roof but have lost the lower temporal bar. However, the precise relationships of Sauroptrygia (including nothosaurs and placodonts) within evolutionarily advanced diapsid reptiles (those branches that include the tuatara and squamates on one side



*Neusticosaurus*, a small, primitive nothosaur (pachypleurosaur) from the Middle Triassic of the Alps. The specimen is about 25 cm long, excluding detached portion.

and crocodiles, dinosaurs, and birds on the other) remain incompletely understood. See DIAPSIDA; REPTILIA.

Robert L. Carroll; Olivier C. Rieppel

Bibliography. R. L. Carroll, *Vertebrate Paleontology and Evolution*, 1988; O. Rieppel, *Encyclopedia of Paleoberpetology*, Pt. 12A: *Sauropterygia I. Placodontia, Pachypleurosauria, Notbosauroida, Pistosauroida*, F. Pfeil, Munich, 2000.

## Notomyotida

A small and distinctive order of valvatecan Asteroidea with long, straight-sided arms and relatively small disks. The skeleton is differentiated into marginals, actinals, and abactinals, and there is a clear calycinal ring. Both supra- and inframarginals are present and are offset or alternate rather than vertically aligned as they are in other asteroids. Intermarginal ossicles are not present. On the oral surface the longest actinal row lies adjacent to the ambulacral groove. Digestive and reproductive organs are confined to the disk and most proximal parts of the arms and do not extend well into the arms as in other valvatecanes. Dorsal longitudinal muscle bands lie free within the arms and are anchored into the body wall only at their distal and proximal ends.

Only a single family, Benthoplectinidae, is included.

Extant benthoplectinids are specialist deep-water suspension feeders that live at depths usually between 1000 and 10,000 ft (300 and 3000 m). There are eight genera and 80 species, distributed in all the world's major oceans. Four fossil species can be attributed to this group, the oldest being *Plesiastropecten ballovensis* from near the base of the Lower Jurassic in Switzerland. See ASTEROIDEA; ECHINODERMATA.

Andrew B. Smith

## Notostraca

An order of branchiopod crustaceans, sometimes called tadpole shrimps. Generally they range from about 20 mm (0.4 in.) to (exceptionally) about 90 mm (3.5 in.) in length. The multisegmented trunk, up to 44 segments in some species, is elongate and cylindrical. The anterior part of the body is covered dorsally by a domed carapace beneath which lie the paired sessile eyes. Each of the first 11 trunk segments bears a pair of limbs, while a varying number of more posterior segments bear up to six pairs of smaller limbs per segment—a very unusual situation. At least four, often more, posterior segments are limbless. The trunk terminates in a telson that bears a pair of slender, segmented, caudal filaments. Each trunk limb bears several endites. These are mostly used for handling the food and become progressively reduced on the smaller posterior limbs, but are drawn out into long, filiform sensory structures on the first pair. In the female, the eleventh pair bears a bowl-shaped pouch with a hinged lid that acts as a temporary receptacle for the eggs that are then shed, or are stuck to surfaces.

The antennules and antennae are minute. The mandibles are massive biting structures that, unlike the rolling crushing mandibles of most branchiopods, are capable of both abduction and adduction and therefore of true biting. They are assisted by robust maxillules. Notostracans feed on a variety of small organisms that are seized by the trunk limbs and passed forward to the mouthparts, but they also collect and eat detritus with the same limbs.

Some species are bisexual, but in some parts of their range some are self-fertilizing hermaphrodites. The highly resistant eggs, which can withstand desiccation if necessary, hatch as nauplii, but in some cases this stage is transient, molting almost at once to a more advanced stage. The two extant genera, *Triops* and *Lepidurus*, are essentially worldwide in distribution and occur mostly in temporary waters. See BRANCHIOPODA.

Geoffrey Fryer

## Notoungulata

An order of dominant, hoofed herbivores of the Cenozoic of South America that are abundantly represented in Paleocene through Pleistocene nonmarine sedimentary rocks of that continent. There are also isolated occurrences in the Paleocene of Central Asia and early Eocene of North America. Diverging from a primitive condylarth ancestry at an early date, they radiated into a wide diversity of forms, some of which were convergent with Northern Hemisphere ungulates.

Notoungulates were characterized by a skull with an expanded temporal region due to the presence of a large sinus in the squamosal (**Fig. 1**) and no postorbital bar. The dentition is primitive with full eutherian formula and a tendency to retain a closed tooth row, although in some groups the median incisors are enlarged, and a gap between the incisor row and cheek teeth is developed by reduction of the posterior incisors, canines, and anterior premolars. There was always a complete molar row and incomplete molarization of the posterior premolars. The teeth were low- to high-crowned, some groups

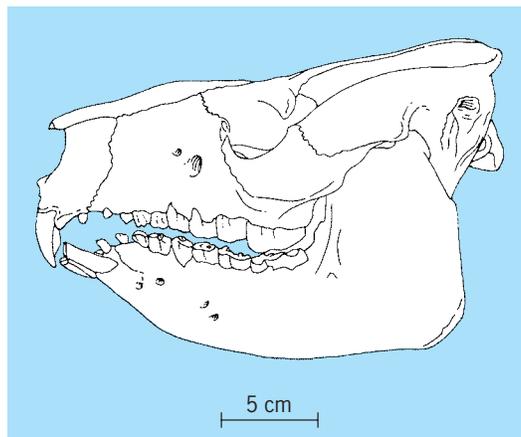


Fig. 1. Skull and jaw of *Adinotherium ovinum*, an early Miocene toxodontid notoungulate from the Santa Cruz formation of Patagonia, Argentina.

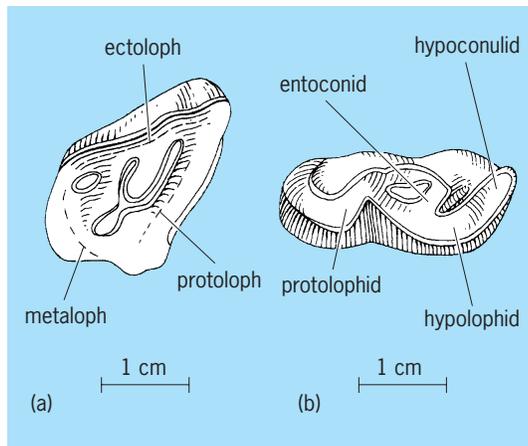


Fig. 2. Examples of molars of *Adinotherium ovinum* from Santa Cruz formation of Patagonia, Argentina. (a) Upper molar. (b) Lower molar.

developing high-crowned ever-growing teeth. The cusps were joined by ridges in the upper and lower molars (lophodonty) even in the earliest forms. Characteristically, the upper molars had a straight ectoloph (Fig. 2a), oblique protoloph, and transverse metaloph; the median valley enclosed by the three lophes usually carried accessory cusps. The lower molars typically were crescentic, the protolophid and hypolophid resembling the Perissodactyla. The entoconid was isolated (Fig. 2b) or connected transversely to the hypoconid rather than to the hypoconulid. The feet were primitive, with five toes (three or two in some advanced forms), and the weight was borne mainly by the third digit. See DENTITION.

Notoungulates are represented in the earliest known mammalian faunas of South America (late Paleocene) by a diversified archaic stock (suborder Notioprogonia), at that time already a specialized sideline rather than a truly ancestral group, and by the earliest members of a central stock, suborder Toxodontia, which gave rise to most of the middle and late Tertiary forms. In addition, a third suborder, the rodentlike Typotheria, appeared in the late Paleocene. By early Eocene time a fourth suborder, the rodentlike Hegetotheria, had appeared. Surprisingly, at about the same time in Central Asia the most abundant mammalian herbivores were members of the Notioprogonia. Two specimens record the presence of this group in the early Eocene of North America. A late Mesozoic origin, either in Asia or South America, seems possible; however, knowledge of late Mesozoic and Paleocene faunas of the world is too incomplete at present to rule out other possibilities.

In late Paleocene time South America was isolated from North America, and from that time until the Pliocene notoungulate evolution proceeded unharassed by competition from the herbivores of the Northern Hemisphere. Toward the close of the Eocene the Notoungulata displayed their greatest diversity, with members of all four suborders being present. Notioprogonians did not survive the Eocene. In the middle and late Tertiary the toxo-

donts, typotheres, and hegetotheres tended to specialize along rather restricted lines; a decrease in diversity within the order resulted. The late Pliocene and Pleistocene emigrant ungulates from North America eventually replaced the large, hippopotamuslike toxodontids, bear-sized typotheres, and rabbitlike hegetotheres in South America. Toxodontids spread north into Central America in the Pleistocene. See EUTHERIA.

Richard H. Tedford

Bibliography. M. J. Benton, *Vertebrate Paleontology*, 1991; R. L. Carroll, *Vertebrate Paleontology and Evolution*, 1988.

## Nova

The sudden brightening of a previously inconspicuous star. The name, short for nova stella (new star), formerly included objects now classified as supernovae and as other kinds of cataclysmic variables. Classical novae now include only those events where the energy source is hydrogen fusion (burning) on the surface of a white dwarf in a close binary system and the white dwarf is not destroyed in the process. See SUPERNOVA.

A handful of novae are discovered each year in the Milky Way Galaxy, and the total rate is probably 20–50 per year. A comparable number are found in other, nearby galaxies. The system consists of a normal, hydrogen-burning star in a close orbit (periods of a few days or less) around a white dwarf or degenerate star. A stream of gas flows from the normal star into a disk around the white dwarf and then accretes onto its surface. Hydrogen gradually builds up there until it is hot and dense enough for nuclear burning, normally with carbon, oxygen, neon, or magnesium from the white dwarf itself acting as a catalyst. Any nuclear fuel ignited under degenerate conditions explodes, because energy released does not cause the gas to expand, so temperature rises rapidly. See BINARY STAR; WHITE DWARF STAR.

The flow occurs at a rate of about  $10^{-9}$  solar mass per year, and  $10^{-5}$  to  $10^{-4}$  solar masses must accumulate to trigger an explosion. Thus novae will recur every  $10^4$ – $10^5$  years. Only part of the accumulated hydrogen burns, but it is all expelled, at speeds of several thousand kilometers per second, together with products of the nuclear reactions. These include rare isotopes such as carbon-13, nitrogen-15, and aluminum-26 (which decays to magnesium-26), and novae are believed to be significant contributors to the galactic supply of these isotopes. See ISOTOPE; NUCLEOSYNTHESIS.

Novae brighten in a few days and fade in months to years. The peak brightness is more than 100 times the solar luminosity, and the total energy release more than  $10^{45}$  ergs ( $10^{38}$  joules). See CATAclysmic VARIABLE; LIGHT CURVES; VARIABLE STAR.

Bibliography. M. F. Bode and A. Evans (eds.), *Classical Novae*, 1989; C. H. Payne-Gaposchkin, *The Galactic Novae*, 1957, reprint 1964; N. Vogt (ed.), *Cataclysmic Variable Stars*, 1992.

## Nozzle

A conduit with a variable cross-sectional area in which a fluid accelerates into a high-velocity stream. The effect of the changing cross-sectional area on the fluid velocity can be explained by the principle of mass conservation applied to successive cross-sectional planes of the nozzle. Equation (1) must be

$$\dot{m} = \rho VA \quad (1)$$

satisfied, where  $\rho$  is the mass density of the fluid [in kilograms per cubic meter ( $\text{kg}/\text{m}^3$ )],  $V$  is the average velocity in the cross section [in meters per second (m/s)],  $A$  is the cross-sectional area [in square meters ( $\text{m}^2$ )], and  $\dot{m}$  is the rate of mass flow through the nozzle [in kilograms per second (kg/s)]. Decreasing  $A$  along the length of the nozzle must result in an increase in  $\rho V$  since  $\dot{m}$  is the same at every cross section. See FLUID FLOW.

**Design and operation.** An important parameter of nozzle operation is the difference in pressure [in pascals (Pa)] between the inlet and outlet. Higher pressure differences push the fluid to higher velocities and achieve higher mass flow rates for a given nozzle size.

*Liquid nozzles.* For liquid nozzles it can be assumed that  $\rho$  is constant and therefore  $V$  increases as  $A$  decreases and vice versa. Liquid nozzles, such as those on fire hoses, are called converging because the area decreases along the length of the nozzle to increase the speed. Typical liquid nozzles have a simple conical shape and are designed to a specific ratio of inlet to outlet areas. To produce good uniformity of the velocity profiles in the issuing jet, the total cone angle should be less than  $30^\circ$ . In cases where the pressure difference across the nozzle is fixed, the mass flow rate from the nozzle can be regulated by pushing a conical spear into the exit plane of the nozzle from the inside. This has the effect of reducing the outlet flow area and hence the mass flow rate.

The difference between the reservoir pressure ( $P_o$ ) and the nozzle exit pressure ( $P_b$ ) required for a given exit velocity  $V_e$  can be determined approximately, using Bernoulli's theorem, as given by Eq. (2). In cases where the pressure difference is less

$$P_o - P_b \cong \frac{\rho V_e^2}{2} \quad (2)$$

than 10 kPa, air may be treated as if its density were constant. See BERNOULLI'S THEOREM.

*High-speed gas nozzles.* In the case of gas nozzles the gas density can change dramatically as a result of the pressure reduction between the inlet and outlet of the nozzle. At very high gas speeds this effect is so significant that the basic shape of the nozzle must change to a converging-diverging form (Fig. 1). The diverging portion is necessary to accommodate the expansion of the gas as it accelerates to lower pressure.

An understanding of the operation of the converging-diverging nozzle requires knowledge of the Mach number, which is the ratio of the gas speed

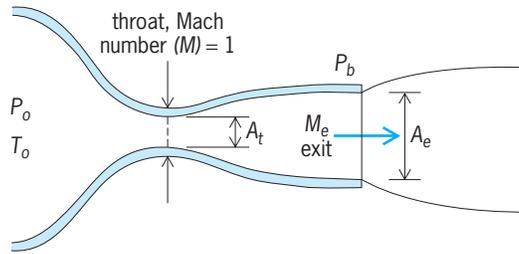


Fig. 1. Typical convergent-divergent nozzle with a jet plume.  $P_o$ ,  $T_o$  = pressure and temperature upstream of the nozzle;  $A_t$  = area at the throat;  $P_b$  = back pressure;  $M_e$ ,  $A_e$  = Mach number and area at exit.

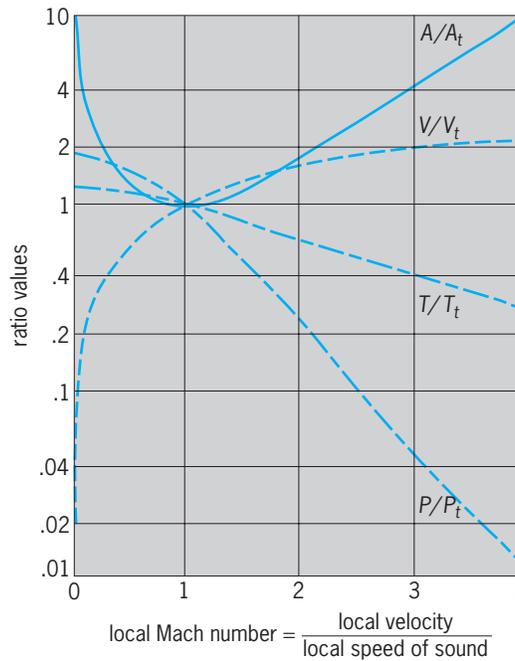


Fig. 2. Conditions for isentropic flow of air through a nozzle when the Mach number at the throat is unity. Ratios of area  $A$ , pressure  $P$ , temperature  $T$ , and velocity  $V$  to their values at the throat ( $A_t$ ,  $P_t$ ,  $T_t$ , and  $V_t$ ) are plotted as functions of Mach number.

to the speed of sound in the gas  $c$ ;  $Ma = V/c$ . A subsonic flow has  $Ma < 1$ , a sonic flow has  $Ma = 1$ , and a supersonic flow has  $Ma > 1$ . Figure 2 shows how the pressure, temperature, and speed of air flowing in a nozzle are related to Mach number and profile of nozzle cross-sectional area. From this figure, it can be seen that the flow in the converging portion of the nozzle must be subsonic, the flow at the throat can at most be sonic, and supersonic flow can occur only in the diverging portion. To achieve supersonic flow in the diverging section, the reservoir pressure must be sufficient to achieve sonic flow at the throat. The fall of air temperature with increasing  $Ma$  indicated in Fig. 2 is a direct result of the gas expansion. See MACH NUMBER; SOUND.

To achieve a particular Mach number at the nozzle exit ( $Ma_e$ ), the ratio of the pressures at the outlet ( $P_b$ ) and inlet ( $P_o$ ) of the nozzle must follow Eq. (3).

$$\frac{P_b}{P_o} \cong \left(1 + 0.2Ma_e^2\right)^{-3.5} \quad (3)$$

If  $P_b/P_o$  is somewhat higher than this value and the same supersonic flow condition prevails inside the divergent part of the nozzle, then an oblique shock wave will originate at the edge of the nozzle. For even higher back-pressure ratios, supersonic flow cannot be realized at the exit, although the throat can still be sonic. Under this condition, a normal shock wave occurs in the divergent part of the nozzle. This operating condition is not desirable, because the flow is subsonic at the exit of the nozzle and the useful mechanical energy is degraded into thermal energy whenever a shock wave occurs. When the back-pressure ratio is lower than the value given by Eq. (3), the jet stream from the nozzle will expand to form a plume (Fig. 1). See JET FLOW; SHOCK WAVE.

There is a very interesting phenomenon observed in the operation of gas nozzles known as choking. Once the velocity at the throat reaches sonic speed, the back pressure has no further effect on throat conditions and the mass flow is entirely determined by the reservoir conditions and throat area. For air flow, this relationship is Eq. (4), where  $T_o$  is the ab-

$$\dot{m}_{\max} = 0.0404 \frac{P_o}{T_o^{0.5}} A_t \quad (4)$$

solute reservoir temperature [in kelvins (K)] at the nozzle inlet. This equation specifies the maximum mass flow for a given throat area, or alternatively, it specifies the minimum throat area for a given mass flow. See CHOKED FLOW.

**Viscous effects.** The above discussion is based upon the assumption that the fluid is frictionless (inviscid). Although all real fluids are viscous, the influence due to viscosity is generally very minor. Only when using a very viscous liquid, such as heavy oil, or when a gas accelerates to a very high Mach number inside the nozzle ( $Ma > 5$ ) do viscous effects become important. See GAS DYNAMICS; VISCOSITY.

**Applications.** A nozzle can be used for a variety of purposes. It is an indispensable piece of equipment in many devices employing fluid as a working medium. The reaction force that results from the fluid acceleration may be employed to propel a jet aircraft or a rocket. In fact, most military jet aircraft employ the simple convergent conical nozzle, with adjustable conical angle, as their propulsive device. If the high-velocity fluid stream is directed to turn a turbine blade, it may drive an electric generator or an automotive vehicle. High-velocity streams may also be produced inside a wind tunnel so that the conditions of flight of a missile or an aircraft may be simulated for research purposes. The nozzle must be carefully designed in this case to provide uniformly flowing fluid with the desired velocity, pressure, and temperature at the test section of the wind tunnel. Nozzles may also be used to disperse fuel into an atomized mist, such as that in diesel engines, for combustion purposes. See ATOMIZATION; IMPULSE TURBINE; INTERNAL COMBUSTION ENGINE; JET PROPULSION; ROCKET PROPULSION; WIND TUNNEL.

Nozzles may also be used as metering devices for gas or liquid. A convergent nozzle inserted into a pipe

produces a pressure difference between its inlet and outlet that can be measured and related directly to the mass flow. For this purpose, the shape of the nozzle must be designed according to standard specifications or carefully calibrated against a mass flow rate standard. A unique feature of gas metering nozzles is that the mass flow rate through the nozzle depends only on the inlet pressure if the pressure ratio across it is sufficient to produce sonic velocity at the throat. See FLOW MEASUREMENT.

Wen L. Chow; A. Gordon L. Holloway

Bibliography. J. D. Anderson, *Modern Compressible Flow*, 3d ed., McGraw-Hill, 2003; R. D. Blevins, *Applied Fluid Dynamics*, Krieger Publishing, 1992; F. M. White, *Fluid Mechanics*, 5th ed., McGraw-Hill, 2003.

## Nuclear battery

A battery that converts the energy of particles emitted from atomic nuclei into electric energy. Two basic types have been developed: (1) a high-voltage type, in which a beta-emitting isotope is separated from a collecting electrode by a vacuum or a solid dielectric, provides thousands of volts but the current is measured in picoamperes; (2) a low-voltage type gives about 1 V with current in microamperes.

**High-voltage nuclear battery.** In the high-voltage type, a radioactive source is attached to one electrode, emitting charged particles. The source might be strontium-90, krypton-85, or hydrogen-3 (tritium), all of which are pure beta emitters. An adjacent electrode collects the emitted particles. A vacuum or solid dielectric separates the source and the collector electrodes.

One high-voltage model, shown in Fig. 1, employs tritium gas sorbed in a thin layer of zirconium metal as the radioactive source. This source is looped around and spot-welded to the center tube of a glass-insulated terminal. A thin coating of carbon applied to the inside of a nickel enclosure acts as an efficient collector having low secondary emission. The glass-insulated terminal is sealed to the nickel enclosure. The enclosure is evacuated through the center tube, which is then pinched off and sealed.

The Radiation Research Corporation model R-1A is  $\frac{3}{8}$  in. (0.95 cm) in diameter and 0.531 in. (1.35 cm) in height. It weighs 0.2 oz (5.7 g) and occupies 0.05 in.<sup>3</sup> (0.8 cm<sup>3</sup>). It delivers about 500 V at 160 pA. Future batteries are expected to deliver 1  $\mu$ A at 2000 V, with a volume of 64 in.<sup>3</sup> (1048 cm<sup>3</sup>).

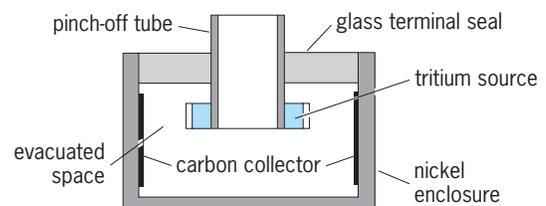


Fig. 1. Tritium battery in cross section.

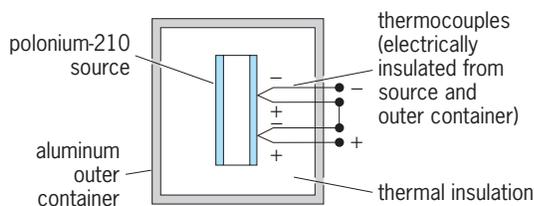


Fig. 2. Thermoelectric nuclear battery.

Earlier models employed strontium-90. This isotope has the highest toxicity in the human body of the three mentioned. Tritium has only one one-thousandth the toxicity of strontium-90. Both strontium-90 and krypton-85 require shielding to reduce external radiation to safe levels. Tritium produces no external radiation through a wall that is thick enough for any structural purpose. Tritium was selected on the basis of these advantages.

The principal use of the high-voltage battery is to maintain the voltage of a charged capacitor. The current output of the radioactive source is sufficient for this purpose.

This type of battery may be considered as a constant-current generator. The voltage is proportional to the load resistance. The current is determined by the number of emissions per second captured by the collector and does not depend on ambient conditions or the load. As the isotope ages, the current declines. For tritium, the intensity drops 50% in a 12-year interval. For strontium-90, the intensity drops 50% in a 25-year interval.

**Low-voltage nuclear battery.** Three different concepts have been employed in the low-voltage type of nuclear batteries: (1) a thermopile, (2) the use of an ionized gas between two dissimilar metals, and (3) the two-step conversion of beta energy into light by a phosphor and the conversion of light into electric energy by a photocell.

*Thermoelectric-type nuclear battery.* This low-voltage type, employing a thermopile, depends on the heat produced by radioactivity (Fig. 2). It has been calculated that a sphere of polonium-210 of 0.1 in. (0.25 cm) diameter, which would contain about 350 curies ( $1.3 \times 10^{13}$  becquerels), if suspended in a vacuum, would have an equilibrium surface temperature of  $4000^\circ\text{F}$  ( $2200^\circ\text{C}$ ), assuming an emissivity of 0.25. For use as a heat source, it would have to be hermetically sealed in a strong, dense capsule. Its surface temperature, therefore, would be lower than  $4000^\circ\text{F}$  ( $2200^\circ\text{C}$ ).

To complete the thermoelectric battery, the heat source must be thermally connected to a series of thermocouples which are alternately connected thermally, but not electrically, to the heat source and to the outer surface of the battery. After a short time, a steady-state temperature differential will be set up between the junctions at the heat source and the junctions at the outer surface. This creates a voltage proportional to the temperature drop across the thermocouples. The battery voltage decreases as the age of the heat source increases. With polonium-210

(half-life, 138 days) the voltage drops about 0.5%/day. The drop for strontium-90 is about 0.01%/day (20-year half-life).

A battery containing 57 curies ( $2.1 \times 10^{12}$  becquerels) of polonium-210 sealed in a sphere 0.4 in. (1 cm) in diameter and seven chromel-constantan thermocouples delivered a maximum power of 1.8 mW. It had an open-circuit voltage of 42 mV with a  $140^\circ\text{F}$  ( $78^\circ\text{C}$ ) temperature differential. Over a 138-day period, the total electrical output would be about  $1.5 \times 10^4$  joules (watt-seconds).

Total weight of the battery was 1.20 oz (34 g). This makes the energy output per pound equal to

$$\frac{1.5 \times 10^4}{3600} \text{ watt-hours} \times \frac{1}{1.2} \times 16 = 55.6$$

This is the same magnitude as with conventional electric cells using chemical energy. This nuclear energy, however, is being dissipated whether or not the electric energy is being used.

The choice of isotope for a thermoelectric nuclear battery is somewhat restricted. Those with a half-life of less than 100 days would have a short useful life, and those with a half-life of over 100 years would give too little heat to be useful. This leaves 137 possible isotopes. This number is further reduced by the consideration of shielding.

The trend is to use plutonium-238 (unable to support a chain reaction) and strontium-90. The most frequently used thermocouple material is a doped lead-telluride alloy that performs between  $400$  and  $900^\circ\text{F}$  ( $200$  and  $480^\circ\text{C}$ ). Another material is silicon-germanium alloy, which can operate at  $1500^\circ\text{F}$  ( $800^\circ\text{C}$ ) (hot junction).

The thermoelectric systems developed so far are in the small power range (5–60 W), and work with a maximum efficiency of 7%.

When shielding is not a critical problem, as in unmanned satellites, they are extremely convenient, being reliable and light. SNAP-3 and SNAP-9A (Systems for Nuclear Auxiliary Power) have weights in the 0.5–1.0 kg/W range.

The above discussion applies only to a portable power source. Drastic revision would be required if a thermoelectric-type nuclear battery were to be designed for central-station power. See THERMOELECTRICITY.

*Gas-ionization nuclear battery.* In this battery a beta-emitting isotope ionizes a gas situated in an electric field (Fig. 3). Each beta particle produces about 200 current carriers (ions), so that a considerable current multiplication occurs compared with the

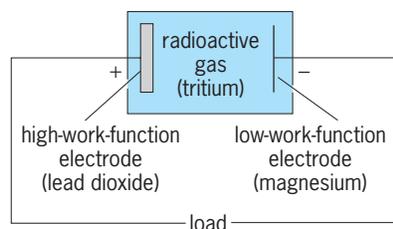


Fig. 3. Gas-ionization nuclear battery.

rate of emission of the source. The electric field is obtained by the contact potential difference of a pair of electrodes, such as lead dioxide (high work function) and magnesium (low work function). The ions produced in the gas move under the influence of the electric field to produce a current.

A cell containing argon gas at 2 atmospheres, electrodes of lead dioxide and magnesium, and a radioactive source consisting of 1.5 millicuries ( $5.6 \times 10^7$  becquerels) of tritium has a volume of 0.01 in.<sup>3</sup> (0.16 cm<sup>3</sup>) and an active plate area of 0.2 in.<sup>2</sup> (1.29 cm<sup>2</sup>), and gives a maximum current of  $1.6 \times 10^{-9}$  A. The open-circuit voltage per cell depends on the contact potential of the electrode couple. A practical value appears to be about 1.5 V. Voltage of any value may be achieved by a series assembly of cells.

The ion generation is exploited also in a purely thermal device, the thermionic generator. It consists of an emitting electrode heated to 2200–3300°F (1200–1800°C) and a collecting electrode kept at 930–1650°F (500–900°C). The hot electrode is made of tungsten or rhenium, metals which melt at more than 5400°F (3000°C) and have low vapor pressures at the working temperature. The collecting electrode is made out of a metal with a low work function, such as molybdenum, nickel, or niobium. Electrons from the hot electrode traverse a gap of 0.04–0.08 in. (1–2 mm) to the collector and recover the emitter by an external circuit.

The ionization space is filled with cesium vapor which acts in two ways. First, it covers the surface of the two electrodes with adsorbed cesium atoms and thus reduces the work function to the desired level. Second, it creates an ionized atmosphere, thus controlling the electron space charge.

The heating of the emitting electrode can be obtained by any of the known means: concentrated solar energy, radioisotopes, or conventional nuclear reactors. While reaching the high temperature was a problem easy to solve, the cooling of the collecting electrode appeared for a long time to be a critical technical problem. This problem was solved by the development of the "heat pipe."

In principle the heat pipe is an empty cylinder absorbing heat at one end by vaporization of a liquid and releasing heat at the other end by condensation of the vapor; the liquid returns to the heat-absorbing end by capillarity through a capillary structure covering the internal face of the cylinder. The heat transfer of a heat pipe may be 10,000 times or more higher than that of copper or silver.

The prototype developed under the sponsorship of NASA (NASA/RCA Type A1279) represents an advanced design capable of supplying 185 W at 0.73 V with an efficiency of 16.5%.

Since no major limitations are foreseen in the development of thermionic fuel cells, it is expected that they will provide the most convenient technique for using nuclear energy in the large-scale production of electricity.

*Scintillator-photocell nuclear battery.* This type of cell is based on a two-step conversion process (Fig. 4).

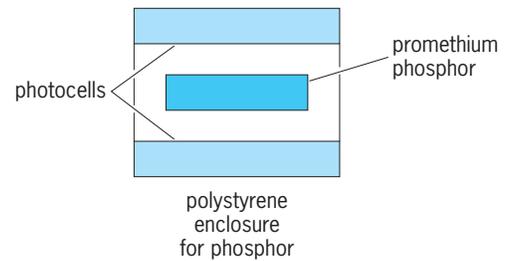


Fig. 4. Scintillator-photocell battery.

Beta-particle energy is converted into light energy; then the light energy is converted into electric energy. In order to accomplish these conversions, the battery has two basic components, a light source and photocells.

The light source consists of a mixture of finely divided phosphor and promethium oxide ( $\text{Pm}_2\text{O}_3$ ) sealed in a transparent container of radiation-resistant plastic. The light source is in the form of a thin disk. The photocells are placed on both faces of the light source. These cells are modified solar cells of the diffused-silicon type.

Since the photocells are damaged by beta radiation, the transparent container of the light source must be designed to absorb any beta radiation not captured in the phosphor. Polystyrene makes an excellent light-source container because of its resistance to radiation.

The light source must emit light in the range at which the photocell is most efficient. A suitable phosphor for the silicon photocell is cadmium sulfide or a mixture of cadmium and zinc sulfide.

In a prototype battery, the light source consisted of 0.0018 oz (50 milligrams) of phosphor and about 0.00018 oz (5 mg) of the isotope promethium-147. This isotope is a pure beta-emitter with a half-life of 2.6 years. It is deposited as a coating of hydroxide on the phosphor particles, which are then dried to give the oxide. For use with the low light level (about 0.001 times sunlight) of the light source, special treatment is necessary to make the equivalent shunt resistance of the cell not less than 100,000 ohms. For a description of the photocell see SOLAR CELL.

The prototype battery, when new, delivers  $20 \times 10^{-6}$  A at 1 V. In 2.6 years (half-life) the current drops about 50% but the voltage drops only about 5%.

The power output improves with decreasing temperature, as a result of improved photocell diode characteristics which more than compensate for a decrease in short-circuit current. At  $-100^\circ\text{F}$  ( $-73^\circ\text{C}$ ), the power output is 1.7 times as great as at room temperature. At  $144^\circ\text{F}$  ( $62^\circ\text{C}$ ), the power output is only 0.6 times as great as at room temperature.

The battery requires shielding to reduce the weak gamma radiation to less than 9 milliroentgens per hour (mR/h), which is the tolerance for continuous exposure of human extremities. The unshielded battery has a radiation level of 90 mR/h. By enclosing the cell in a case of tungsten alloy, density 16.5, the external radiation becomes less than 9 mR/h.

The unshielded battery has a volume of  $0.014 \text{ in.}^3$  ( $0.23 \text{ cm}^3$ ) and a weight of  $0.016 \text{ oz}$  ( $0.45 \text{ g}$ ). Over a 2.5 year period, the total output would be  $0.32 \text{ Wh}$  (whether or not used). This gives a unit output of  $320 \text{ Wh/lb}$  ( $704 \text{ Wh/kg}$ ), about six times as great as chemical-battery output. But the shielded battery has a volume of  $0.07 \text{ in.}^3$  ( $1.15 \text{ cm}^3$ ) and a weight of  $0.6 \text{ oz}$  ( $17 \text{ g}$ ). This reduced the unit output to  $8.5 \text{ Wh/lb}$  ( $18.7 \text{ Wh/kg}$ ). The cell can undergo prolonged storage at  $200^\circ\text{F}$  ( $93^\circ\text{C}$ ). See BATTERY.

Jack Davis; L. Rozeanu; Kenneth Franzese

## Nuclear binding energy

The amount by which the mass of an atom is less than the sum of the masses of its constituent protons, neutrons, and electrons expressed in units of energy. This energy difference accounts for the stability of the atom. In principle, the binding energy is the amount of energy which was released when the several atomic constituents came together to form the atom. Most of the binding energy is associated with the nuclear constituents (protons and neutrons), or nucleons, and it is customary to regard this quantity as a measure of the stability of the nucleus alone. See NUCLEAR STRUCTURE.

A widely used term, the binding energy (BE) per nucleon, is defined by the equation below, where

$$\text{BE/nucleon} = \frac{[ZH + (A - Z)n - zM^A]c^2}{A}$$

$zM^A$  represents the mass of an atom of mass number  $A$  and atomic number  $Z$ ,  $H$  and  $n$  are the masses of the hydrogen atom and neutron, respectively, and  $c$

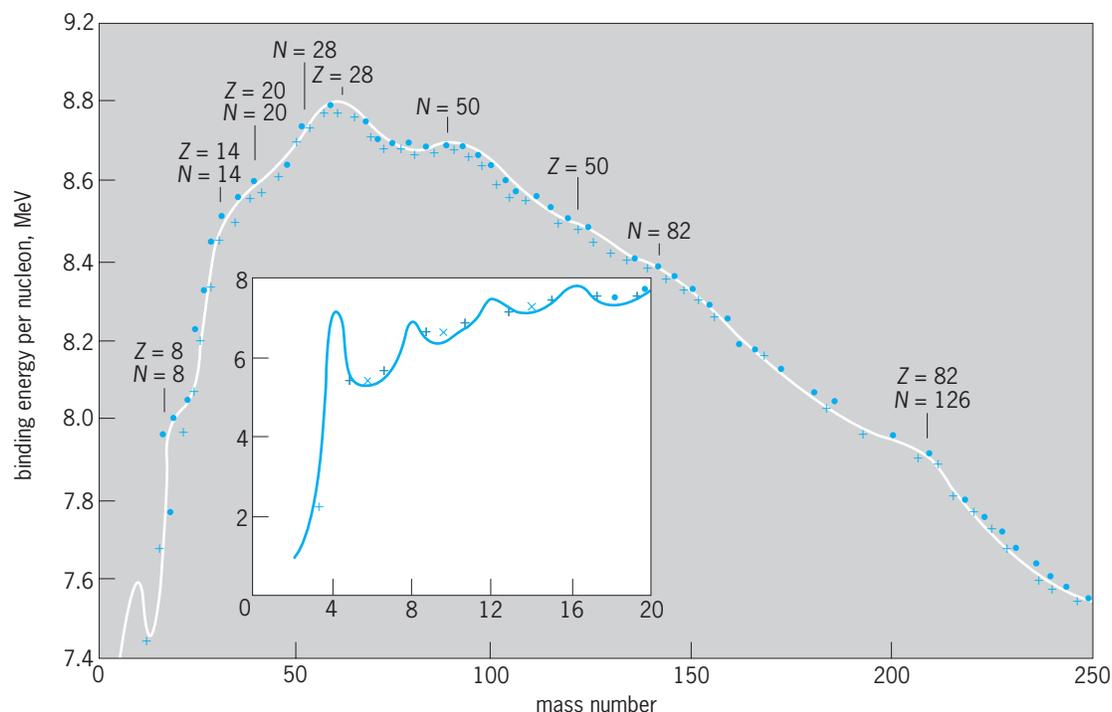
is the velocity of light. The binding energies of the orbital electrons, here practically neglected, are not only small, but increase with  $Z$  in a gradual manner; thus the BE/nucleon gives an accurate picture of the variations and trends in nuclear stability. The **illustration** shows the BE/nucleon (in megaelectronvolts) plotted against mass number for  $A > 40$ .

The BE/nucleon curve at certain values of  $A$  suddenly changes slope in such a direction as to indicate that the nuclear stability has abruptly deteriorated. These turning points coincide with particularly stable configurations, or nuclear shells, to which additional nucleons are rather loosely bound. Thus there is a sudden turning of the curve over  $A = 52$  (28 neutrons); the maximum occurs in the nickel region (28 protons,  $\sim A = 60$ ); the stability rapidly deteriorates beyond  $A = 90$  (50 neutrons); there is a slightly greater than normal stability in the tin region (50 protons,  $\sim A = 118$ ); the stability deteriorates beyond  $A = 140$  (82 neutrons) and beyond  $A = 208$  (82 protons plus 126 neutrons).

The BE/nucleon is remarkably uniform, lying for most atoms in the range 5–9 MeV. This near constancy is evidence that nucleons interact only with near neighbors; that is, nuclear forces are saturated.

The binding energy, when expressed in mass units, is known as the mass defect, a term sometimes incorrectly applied to quantity  $M - A$ , where  $M$  is the mass of the atom. See MASS DEFECT. Henry E. Duckworth

The term binding energy is sometimes also used to describe the energy which must be supplied to a nucleus in order to remove a specified particle to infinity, for example, a neutron, proton, or alpha particle. A more appropriate term for this energy is the separation energy. This quantity varies greatly



Graph of binding energy per nucleon (in megaelectronvolts) plotted against mass number.  $N$  = number of neutrons. (After A. H. Wapstra, *Isotopic measure, part 1, where  $A$  is less than 34, Physica, 21:367–384, 1955*)

from nucleus to nucleus and from particle to particle. For example, the binding energies for a neutron, a proton, and a deuteron in  $^{16}\text{O}$  are 15.67, 12.13, and 20.74 MeV, respectively, while the corresponding energies in  $^{17}\text{O}$  are 4.14, 13.78, and 14.04 MeV, respectively. The usual order of neutron or proton separation energy is 7–9 MeV for most of the periodic table.

D. H. Wilkinson

### Nuclear chemical engineering

The branch of chemical engineering that deals with the production and use of radioisotopes, nuclear power generation, and the nuclear fuel cycle.

A nuclear chemical engineer requires training in both nuclear and chemical engineering. As a nuclear engineer, he or she should be familiar with the nuclear reactions that take place in nuclear fission reactors and radioisotope production, with the properties of nuclear species important in nuclear fuels, with the properties of neutrons, gamma rays, and beta rays produced in nuclear reactors, and with the reaction, absorption, and attenuation of these radiations in the materials of reactors. *See* BETA PARTICLES; GAMMA RAYS; NEUTRON.

As a chemical engineer, he or she should know the properties of materials important in nuclear reactors and the processes used to extract and purify these materials and convert them into the chemical compounds and physical forms used in nuclear systems. *See* CHEMICAL ENGINEERING.

**Radioisotopes.** The principal nuclear species important in nuclear fuels are the fissionable isotopes uranium-235, which makes up 0.7% of natural uranium; plutonium-239, produced by neutron irradiation of uranium-238, the other 99.3% of natural uranium; and uranium-233, produced by neutron irradiation of natural thorium. Other nuclear species of interest to nuclear chemical engineers are the fission products, which make irradiated nuclear fuel intensely radioactive; radioactive isotopes of heavy elements such as protactinium, neptunium, plutonium, americium, and curium, produced by neutron absorption in nuclear fuels; and isotopes of light elements, such as deuterium used in the form of heavy water in reactors fueled with natural uranium, tritium (hydrogen-3) used with deuterium as fuel for fusion reactors, and lithium-6, the starting material for tritium production. *See* AMERICIUM; CURIUM; DEUTERIUM; LITHIUM; NEPTUNIUM; PLUTONIUM; PROTACTINIUM; THORIUM; TRANSURANIUM ELEMENTS; TRITIUM; URANIUM.

**Power generation.** The principal types of nuclear power reactors are light-water reactors, of either the pressurized or boiling-water type; pressurized heavy-water reactors; and sodium-cooled, liquid-metal fast breeder reactors. Aspects of light- and heavy-water reactors of concern to nuclear chemical engineers include production and purification of the uranium dioxide fuel, production of the hafnium-free zirconium tubing used for fuel cladding, and control of corrosion and radioactive corrosion products by

chemical treatment of coolant. Another chemical engineering aspect of heavy-water reactor operation is control of the radioactive tritium produced by neutron activation of deuterium. Aspects of liquid-metal fast-breeder reactors of concern to nuclear chemical engineers include fabrication of the mixed uranium dioxide–plutonium dioxide fuel, purity control of sodium coolant to prevent fouling and corrosion, and reprocessing of irradiated fuel to recover plutonium and uranium for recycle. *See* NUCLEAR POWER; NUCLEAR REACTOR.

**Nuclear fuel cycle.** The nuclear fuel cycle comprises the industrial operations involved in producing fuel for nuclear reactors, and processing and storing safely the intensely radioactive fuel discharged from them after irradiation. For light-water reactors, the principal steps of the fuel cycle are: mining of uranium ores; concentration of the contained uranium; purification of uranium; conversion to uranium hexafluoride; enrichment of uranium-235 from 0.7% in natural uranium to about 3% by gaseous diffusion or the gas centrifuge; fabrication of enriched uranium into fuel elements; irradiation of fuel elements in a nuclear reactor to produce energy and neutrons; storage of irradiated fuel to permit some of its intense radioactivity to decay; packaging irradiated fuel for long-term, safe storage; and emplacement of fuel in terminal storage facilities. Alternatively, because irradiated fuel contains as much uranium-235 as natural uranium and about 0.5% of fissionable plutonium, irradiated fuel may be reprocessed to recover these valuable materials for reuse as reactor fuel, and to separate and convert the radioactive fission products into an insoluble glass. Such glass is more inert, more compact, and more suitable for long-term storage than unprocessed irradiated fuel. *See* HAZARDOUS WASTE; NUCLEAR FUEL CYCLE; NUCLEAR FUELS; NUCLEAR FUELS REPROCESSING; RADIOACTIVE WASTE MANAGEMENT; URANIUM METALLURGY.

Manson Benedict

**Bibliography.** M. Benedict, T. H. Pigford, and H. W. Levi, *Nuclear Chemical Engineering*, 2d ed., 1981; R. A. Knief, *Nuclear Engineering: Theory and Technology of Commercial Nuclear Power (SCPP)*, 2d ed., 1992; J. R. Lamarsh and A. J. Baratta, *Introduction to Nuclear Engineering*, 3d ed., 2001; R. H. Perry and D. Green (eds.), *Perry's Chemical Engineers' Handbook*, 7th ed., 1997; Jan Rydberg et al., *Radiochemistry and Nuclear Chemistry*, 3d ed., 2001; P. D. Wilson (ed.), *The Nuclear Fuel Cycle: From Ore to Wastes*, 1996.

### Nuclear chemistry

An interdisciplinary field that, in general, encompasses the application of chemical techniques to the solution of problems in nuclear physics. The discovery of the naturally occurring radioactive elements and of nuclear fission are classical examples of the work of nuclear chemists.

Although chemical techniques that are employed in nuclear chemistry are essentially the same as those

in radiochemistry, these fields may be distinguished on the basis of the aims of the investigation. Thus, a nuclear chemist utilizes chemical techniques as a tool for the study of nuclear reactions and properties, whereas a radiochemist utilizes the radioactive properties of certain substances as a tool for the study of chemical reactions and properties. There is considerable overlap between the two fields and in some cases (for example, the preparation and study of synthetic elements) a distinction may be somewhat arbitrary. Nuclear chemistry and nuclear physics are also closely related, and nuclear chemists may employ purely physical techniques in their studies. The term hot-atom chemistry, originally used to describe the chemical reactions of atoms produced in nuclear transformations, is sometimes extended to include the chemical reactions of any excited atoms, such as those produced in accelerated beams or by absorption of radiation. For the application of radioactive tracers to chemical problems. For the chemical effects of radiation on various systems *see* RADIATION CHEMISTRY; RADIOCHEMISTRY.

**Scope.** The chemical identification of radioactive nuclides and the determination of their nuclear properties (half-life, mode and energy of decay, mass number, nucleon binding energies, and so on) has been one of the major activities of nuclear chemists. Such studies have produced an extensive array of radioisotopes, and ongoing studies are concerned mainly with the more difficult identification of nuclides of very short half-life at the boundaries of the "valley of beta stability" and in the regions of neutron and proton instability, as well as short-lived isomeric states. Nuclear chemical investigations led to the discovery of the synthetic radioactive elements which do not have any stable isotopes and are not formed in the natural radioactive series (technetium, promethium, astatine, and the transuranium elements). The synthesis of element 107 has been announced and claims made for the observation of elements 108 and 109. Attempts to extend the periodic table even further and to prepare larger quantities of the new elements for intensive study are being made by using various nuclear reactions, particularly accelerated heavy-ion bombardments of targets of the heaviest elements available. Accelerators capable of accelerating even uranium ions are available, and investigations of the possible existence of superheavy nuclei in a new region of relative stability are being carried out. *See* TRANSURANIUM ELEMENTS.

Other major areas of nuclear chemistry include studies of nuclear structure and spectroscopy (involving the determination of the energy spectra of emitted particles and the construction of energy-level diagrams, or decay schemes) and of the probability and mechanisms of various nuclear reactions (involving the determination of excitation functions, that is, the variation of cross section—or probability of reaction—with the energy of the bombarding particle, and the angular and energy distributions and correlations of the reaction products).

Many of these investigations involve the chemical isolation and identification of the products of nu-

clear reactions or chemical techniques for the preparation of thin, uniform deposits of pure materials for use as bombardment targets and as sources for nuclear spectroscopy. However, the nuclear chemist also carries out many experiments of a purely physical nature, observing the products of nuclear reactions with counters, ionization chambers, nuclear emulsions, and other particle track detectors. Moreover, advances in the technology of nuclear radiation detectors and associated electronic equipment have led to instruments of such sophistication that classical methods of chemical analyses are often not necessary. The borderline between nuclear chemistry and nuclear physics has, in fact, practically disappeared in many areas of investigation (for example, studies of heavy-ion reactions and many aspects of nuclear structure and spectroscopy), and if a distinction is to be made at all, it may be only that the nuclear physicist tends to concentrate on interactions between the fundamental particles and observes the nucleons and light nuclei (neutrons, protons, deuterons, tritons, and alpha particles) emitted in nuclear reactions, while the nuclear chemist examines interactions with complex nuclei and observes the heavier fragments and residual nuclei produced. This distinction is perhaps most meaningful in the area of high-energy nuclear phenomena where the high-energy, or particle, physicist is concerned mainly with the production and properties of and the interactions between the elementary (or strange) particles and their role in the structure of matter. The nature of the interaction of high-energy projectiles with complex nuclei and studies of the products formed is mainly the concern of nuclear chemists. Studies involving the use of highly radioactive isotopes of the actinide elements and investigations of many aspects of the nuclear fission process have been carried out largely by nuclear chemists. *See* NUCLEAR REACTION.

The contributions of nuclear chemists have not been limited to empirical studies of nuclear reactions and properties. They are also active in theoretical investigations and have made significant contributions to theories of nuclear structure, radioactive decay (particularly alpha emission), and nuclear reactions of many types, including fission. *See* NUCLEAR STRUCTURE.

**Historical background.** Nuclear chemistry began with the work of Pierre and Marie Curie on the isolation and identification of polonium and radium in 1898. The discovery of actinium, thoron (radon-222), and radon in the next few years, the identification of the alpha particle as the helium ion, and the characterization of the relationships in the natural radioactive decay series played an essential role in the development of atomic and nuclear science. The interpretation of the results of these early experiments led to the establishment of the laws of radioactive transformation and the concept of isotopes (E. Rutherford and F. Soddy). The phenomenon of nuclear isomerism was discovered by O. Hahn (1921), who showed that the radioactive species UZ and UX<sub>2</sub> represented different excitation levels of <sup>234</sup>Pa.

See ACTINIUM; ISOTOPE; POLONIUM; PROTACTINIUM; RADIUM; RADON.

Artificial transmutation reactions were first carried out utilizing the alpha particles emitted in the decay of the natural radioactive isotopes as bombarding particles (Rutherford and coworkers, 1919). Such bombardments led to the discovery of artificial radioactivity by Irene Curie and F. Joliot in 1934. The discovery of the neutron (J. Chadwick, 1932) and deuterium (H. Urey and coworkers, 1932) and the development of devices to accelerate the ions of hydrogen (proton), deuterium (deuteron), and helium (alpha particle) gave tremendous impetus to the production and study of radioactive nuclides. By 1937 about 200 were known, and the number has grown steadily to nearly 2100 in 1984.

The discovery of nuclear fission by O. Hahn and F. Strassmann in 1939 culminated a dramatic episode in the history of nuclear science and opened the doors to the atomic age. Nuclear chemistry became a mature science during the period 1940–1945, largely as a result of the concentrated effort to produce the atomic bomb. It has continued to grow and expand in scope with the advancing technology in particle accelerators, nuclear reactors, radiation detection instrumentation, mass spectrometers, electromagnetic isotope separators, and computers. See NUCLEAR FISSION.

**Nuclear spectroscopy.** The energy spectra of the radiations emitted by radioactive nuclides are determined by the use of a number of specialized instruments. Alpha-particle spectra may be determined quite well with solid-state detectors coupled to multichannel pulse-height analyzers. The pulse output of the detector is proportional to the energy deposited and the individual events are stored in the analyzer memory. Magnetic spectrometers are used for higher resolution but require sources of higher intensity. See NUCLEAR SPECTRA.

Beta-particle spectra may be examined with proportional or scintillation counters, but precise energy determinations and studies of the shapes of the spectra require magnetic spectrometers.

Gamma-ray and x-ray spectra are very conveniently measured using thallium-activated sodium iodide (NaI) crystal scintillation counters which give a fluorescent light output proportional to the energy deposited in the crystal. The light is amplified with a photomultiplier tube and converted to a pulse-height that is stored in a multichannel analyzer. Solid-state detectors, particularly intrinsic germanium and lithium-drifted germanium and silicon crystals, possess higher resolution and are generally used for studies requiring high accuracy. Bent-crystal spectrometers, which depend on Bragg scattering, are instruments of high resolution but require intense sources of the radioactive nuclide. They are often used in measurements of gamma-ray spectra accompanying nuclear reactions—capture gamma rays in the  $(n,\gamma)$  reaction, for example.

Nuclear structure determinations may also involve nuclear reaction spectroscopy, in which the energy levels of the nuclide of interest are derived from the

projectile properties (type and energy) and the measured (discrete) energies of the ejected nucleon or fragment [for example, in a  $(d,p)$ ,  $(d,t)$ , or more complex reaction]. Combinations of several techniques may be required to establish the energy-level structure of a nuclide and characterize the energy, spin, and parity of each level.

**Studies of nuclear reactions.** In a nuclear reaction, a bombarding particle interacts with a target nucleus to produce one or more product nuclei and, perhaps, other particles. (Spontaneous transformations of unstable nuclei, such as alpha and beta decay or spontaneous fission, may be considered a special type of nuclear reaction in which no external excitation by bombarding particles is involved.) If the target is thin with respect to the ranges of the product nuclei, or if it is dispersed in a medium such as a nuclear emulsion, the reaction may be studied event by event by recording the emitted nuclei and particles with counters or observing the tracks they leave in nuclear emulsions or other particle track detectors, such as mica or polycarbonate films. Such studies are especially useful for determinations of the angular distributions of the recoil fragments and for information on the multiplicity of fragments emitted in a single event. More sophisticated techniques of track detection employing bubble, spark, and streamer chambers can provide more information on the particle identity and energy from the track density and its curvature in magnetic fields.

By suitable in-line arrangements of thin detectors (through which the particles pass, depositing only part of their kinetic energy) and thick detectors (which record the total kinetic energy), the mass and charge of many light nuclei can be uniquely identified. Time-of-flight (velocity) and kinetic energy measurements may also be combined to establish the masses of recoiling nuclei from nuclear reactions, and measurements of the characteristic x-rays may be made to establish the nuclear charge. These measurements require rather sophisticated instrumentation, and they cannot be applied in all cases. Where applicable, however, they can provide information on the correlations between particles or fragments emitted in the same event, information that is essential for an understanding of the reaction mechanism. See PARTICLE DETECTOR.

When a thick target is bombarded, the reaction products cannot escape, and an accumulation of many individual events is retained in the target matrix. (Catcher foils are placed adjacent to the target to retain the small fraction of the products escaping from the target surfaces.) The recoil products from thin targets may also be accumulated on catcher foils at various angles to the beam direction for angular distribution studies and recoil range measurements. Chemical separations are employed to identify the elements formed in the reactions and measurements of the radiations made to characterize the isotopes. When a number of isotopes of the same element are produced in the target and cannot be conveniently differentiated on the basis of half-life or radiation characteristics, chemical separations alone

may not suffice. Separation of the isotopes can be achieved by the acceleration and deflection of ions in electromagnetic isotope separators. The separated isotopes may be collected and their radiations measured free from any interference. The abundances of stable as well as radioactive species produced in nuclear reactions may also be determined by mass spectrometry. *See* MASS SPECTROMETRY.

The radiochemical separation and identification of the accumulated reaction products from the matrix of a thick target following bombardment is a relatively simple means of determining the identity and yields (or formation cross sections) of the products. The high sensitivity and specificity of the technique make it particularly useful for measurements of rare events. Its relative simplicity makes it especially valuable for general surveys and systematics of nuclear reactions to establish trends and identify regions of particular interest that may warrant more detailed investigation with other, more sophisticated techniques (such as event-by-event particle detection with counter telescopes as described above). This approach has been particularly important for initial studies when a new energy regime or new projectile becomes available. For example, it has led to the identification of new phenomena and stimulated further investigation of the reactions induced by heavy ions and relativistic projectiles.

Nuclear chemists have played a major role in the determination of the yields, as well as the energy and angular distributions, of various reaction products resulting from the bombardment of a broad range of targets with projectiles varying in type from neutrons, protons, and pions to uranium ions and in energy from thermal (0.025 eV) to highly relativistic energies (500 GeV). These studies are well suited to the investigation of the modes of target breakup through such processes as nuclear fission, spallation, and others involving more extensive and violent target explosions. *See* NUCLEAR FISSION; SPALLATION REACTION.

**Search for new nuclides.** Although nearly 2100 radioactive nuclides have been identified, there are still gaps where nuclides are known to exist but have yet to be identified and unknown regions yet to be probed. In particular, the limits of neutron and proton stability in nuclei are actively being pursued. As these limits are approached, the half-lives of the nuclear species become very short (milliseconds to seconds) and specialized techniques have been developed for such investigations. These include (1) rapid, automated radiochemical separations; (2) gas transport techniques, in which the nuclide is entrained in a flowing gas stream and transported rapidly from the target to a detector; (3) target wheels which turn in a bombarding beam and carry the target rapidly and repeatedly from the activating beam position to detector positions; (4) moving tapes that collect the products recoiling from a thin target and rapidly transport them to detectors for measurement; (5) on-line isotope separators associated with heated targets (which vaporize the products and serve as the ion source for the separator) and

various detectors at the collector position; and (6) other innovative methods.

One of the most exciting and active areas in the search for new nuclides is the attempt to extend the periodic table to elements beyond 107. In particular, theoretical prediction of a new "island of stability" near atomic number 114 and neutron number 184 has stimulated attempts to reach it. An intensive experimental program has been under way to form such a product at low enough excitation energy that it will survive long enough to be identified. Heavy-ion, neutron-rich projectiles (such as  $^{48}\text{Ca}$ ) and neutron-rich targets (such as  $^{247}\text{Cm}$ ) have been prepared in the hope that a fusion product in the expected region may be identified. *See* PERIODIC TABLE.

**Technique of radiochemical analysis.** The chemical manipulations and separations of radioactive materials, generally referred to as techniques of radiochemical analysis, differ from the ordinary analytical techniques in a number of respects. Procedures in radiochemical analysis are not required to provide quantitative recovery but are selected for specificity and speed, with reasonably good yields (usually the order of 50% or better) generally sufficing. The criteria of radiochemical purity in a radioactive preparation are somewhat more stringent than those of ordinary chemical purity. Thus, trace quantities of impurities are rarely of importance in ordinary quantitative analyses, but in a radioactive preparation, contamination by trace quantities of radioactive impurities may completely negate the results of an experiment.

The handling of highly radioactive materials presents a health hazard, and special techniques for the manipulation of samples behind shielding walls must be utilized. Some effects of high levels of radioactivity on the solutions, such as heating and the decomposition of solvents which produces bubbling, also may affect normal procedures.

The mass of radioactive material produced in nuclear reactions is usually very small. The concentrations of nuclear reaction products in the solutions of target materials are generally of the order of  $10^{-10}$  M or less. Many normal chemical operations, such as precipitation, are not feasible with such small concentrations. Although separations can be carried out with these tracer quantities using such techniques as solvent extraction and ion exchange, it then is difficult to determine the efficiency for the recovery of the product. Moreover, the chemical behavior of such dilute solutions may differ considerably from that normally encountered. For example, radiocolloid formation, adsorption on the walls of vessels and on the surfaces of dust particles and precipitates, and the concentration dependence of some equilibrium constants become prominent at such extremely high dilution. To avoid these difficulties, an isotope dilution technique may be employed in which macroscopic quantities of stable isotopes of the element are added to serve as a carrier for the radioactive species. *See* ISOTOPE DILUTION TECHNIQUES.

The amount of carrier used represents a compromise between considerations of convenience in

chemical manipulations and yield determination and the preparation of high specific activity sources in which counting corrections for absorption and scattering of radiations in the sample itself are minimized. Quantities of 10–20 mg are used most often. Chemical procedures are simplified if macroscopic quantities of only a few elements are present. When many elements are produced in a nuclear reaction (in nuclear fission, for example), aliquots of the solution usually are taken for the analysis of each element or small group of elements. It is then necessary to add carriers for only relatively few products of interest. Trace quantities of the other elements present are removed in the chemical procedures by the use of scavenging precipitations of a compound of high surface area, such as iron(III) hydroxide or manganese dioxide, which tend to occlude traces of foreign substances or of a representative precipitate for an insoluble group of elements, such as bismuth sulfide to carry trace quantities of other insoluble sulfides, lanthanum fluoride for insoluble fluorides, or iron(III) hydroxide for insoluble hydroxides. If the element of interest itself forms a precipitate which may occlude traces of other elements, it may be necessary to add holdback carriers for the latter to dilute the effect of radioactive contamination of the product precipitate.

For the isolation of products of high specific activity without regard to yield, the carrier may be an element with chemical properties similar to those of the desired product, but which can be separated from it in the last stages of the procedure, leaving the product essentially carrier-free. Such carriers are referred to as nonisotopic carriers. When it is necessary to determine the yield of a nuclear reaction product, a known quantity of an isotopic carrier must be used. It is also imperative that complete interchange between the valence states of the carrier and the active species be achieved before any chemical separation is begun. In the case of elements which do not have any stable isotopes, or when a carrier-free procedure is desired, a known quantity of an isotopic radioactive tracer may be used. Radiations of the tracer should be easily distinguishable from those of the product. The fractional recovery of the added carrier or tracer then will represent the yield of the product of interest.

Classical precipitation methods of chemical separation are time-consuming, and in order to study short-lived radioactive species, rapid procedures are essential and other techniques are generally employed. Ion-exchange, solvent-extraction, and volatilization techniques have proved most useful, and their development is closely associated with radiochemical investigations. In cases where volatile products (rare gases, halogens, and so on) are formed in the nuclear reaction, a solution of the target may be bombarded in a closed system and the products swept out continuously. Some solid matrices that emanate gases readily may also be used. The products may be chemically separated by passing them through a suitable series of rapid steps in the flowing system before collection in a form suitable for direct measurement of the radiations. Nuclides with half-

lives of the order of seconds may be investigated with such on-line techniques. Products recoiling out of thin targets may also be collected in gas streams or on moving tapes for rapid removal from the target and on-line or subsequent analysis. An isotope separator may form part of the on-line system, and in favorable cases the target may serve as the ion source of the separator. *See* ION EXCHANGE; RADIOCHEMICAL LABORATORY; SOLVENT EXTRACTION.

**Sample preparation and counting techniques.** For studies in nuclear chemistry, the object of the radiochemical separations is the preparation of a pure sample in a form suitable for the radioactive assay of the nuclide of interest or for the determination of its nuclear properties. The detector used will, of course, depend on the type of radiation involved and the kind of information desired.

Alpha particles and fission fragments have short ranges in matter, and to prevent absorption losses samples of less than 100 micrograms/cm<sup>2</sup> surface density are generally required. A uniform sample deposit is necessary for accurate alpha-particle and fission fragment measurements. This is best accomplished by volatilizing, electroplating, or spraying on metal foils. Samples collected with an isotope separator are well suited for such measurements. Evaporation from solution may also be used if the amount of solid residue is small, but uniformity of the deposit is difficult to achieve. The samples are counted internally in ionization chambers or gas proportional counters or with solid-state detectors.

Beta particles may cover a wide range of energies, and the techniques of sample preparation and counting will vary accordingly. The most commonly used detectors are Geiger, flow-type proportional, and scintillation counters. Samples may be prepared as indicated for alpha emitters, in the form of precipitates or filter-paper disks or sample cups, as gases for internal counting, and as liquids. External sample counting usually is employed for convenience whenever feasible. *See* GEIGER-MÜLLER COUNTER; SCINTILLATION COUNTER.

Gamma radiation is highly penetrating, and the size or form of the sample is generally not very critical. Because of much higher efficiency and resolution, scintillation counters and solid-state detectors have displaced all other types of detectors for gamma radiation.

Whenever possible it is advisable to design experiments so that relative counting of samples will suffice. It is then necessary only to reproduce the counting conditions for each sample. The determination of absolute disintegration rates is a more difficult task, and special techniques are required. Counters which detect all the radiations emanating from the source ( $4\pi$  counters) are used, and the samples are either dispersed in the counting medium of a proportional or Geiger counter as a gas, dissolved in the medium of a liquid scintillation counter, or counted as a very thin deposit on a very thin film which is placed between two identical counters. Beta-gamma coincidence counting may be used for determining absolute disintegration rates when the decay scheme of the nuclide is not too complicated.

**Applied nuclear chemistry.** The radiochemical and counting techniques outlined above are powerful tools for the study of nuclear reactions and the properties of nuclides. New techniques and instruments are constantly being adapted to the needs of nuclear chemistry and, conversely, investigations in nuclear chemistry have indicated the need and provided stimulation for the development of many new instruments. The techniques of nuclear chemistry have been applied to studies in a number of related fields, and nuclear chemists have contributed to studies in reactor chemistry and physics, isotope production, nuclear engineering, the reactor fuel/reprocessing cycle, nuclear weapons development, geo- and cosmochemistry, environmental chemistry, and accelerator beam studies and monitoring as well as to basic studies in chemistry and the life sciences and the industrial and agricultural application of radioisotopes. The field of analytical chemistry has been especially influenced by the technique of activation analysis, which utilizes many of the results and methods of nuclear chemistry. *See* ACTIVATION ANALYSIS.

The preparation of radioactively tagged compounds and radioactive tracers has been of particular importance in the field of nuclear medicine. The radioactive properties (half-life and type of energy of decay) and chemical or biochemical properties of tracers and tagged compounds can be designed and prepared to fit the needs of specific diagnostic or therapeutic requirements. *See* RADIOLOGY.

Ellis P. Steinberg

**Bibliography.** W. T. Carnall and G. R. Choppin (eds.), *Plutonium Chemistry*, 1980; G. R. Choppin and J. Rydberg (eds.), *Radiochemistry and Nuclear Chemistry: Theory and Applications*, 1994; G. Friedlander et al., *Nuclear and Radiochemistry*, 4th ed., 1981; B. G. Harvey, *Introduction to Nuclear Physics and Chemistry*, 2d ed., 1969; C. Hevesy and F. A. Paneth, *A Manual of Radioactivity*, 1938; E. K. Hyde, I. Perlman, and G. T. Seaborg, *The Nuclear Properties of the Heavy Elements*, 1964; M. Lefort, *Nuclear Chemistry*, 1968; J. D. Navratil, *Nuclear Chemistry*, 1993; G. T. Seaborg and W. Loveland (eds.), *Nuclear Chemistry*, 1982.

## Nuclear engineering

The branch of engineering that deals with the production and use of nuclear energy and nuclear radiation. The multidisciplinary field of nuclear engineering is studied in many universities. In some it is offered in a special nuclear engineering department; in others it is offered in other departments, such as mechanical or chemical engineering. Primarily, nuclear engineering involves the conception, development, design, construction, operation, and decommissioning of facilities in which nuclear energy or nuclear radiation is generated or used.

Examples of facilities include nuclear power plants, which convert nuclear energy to electricity; nuclear propulsion reactors, which convert nuclear energy to mechanical force used for the propul-

sion of ships and submarines; space nuclear reactors, which convert nuclear energy to heat and electricity to power satellites, probes, and vehicles; nuclear production reactors, which produce fissile or fusible materials used in nuclear weapons; nuclear research reactors, which generate nuclear radiation (neutrons and gamma rays) for scientific research and for a variety of medical and industrial applications; gamma cells, which contain intense sources of gamma rays that are used for sterilizing medical equipment and food and for manufacturing polymers; particle accelerators, which produce nuclear radiation for use in medical and industrial applications; and repositories for safe storage of nuclear waste. *See* PARTICLE ACCELERATOR; SHIP NUCLEAR PROPULSION; SPACE POWER SYSTEMS; SUBMARINE.

**Nuclear power plants.** All operating nuclear power plants use the fission process. A nuclear power plant is a large, complex system consisting of a nuclear reactor, steam generators, heat exchangers, a turbogenerator, condensers, and many auxiliary systems. The central part of the nuclear reactor is the core, which contains the nuclear fuel and in which the nuclear chain reaction is established. The core is where fission energy is generated and converted to heat. This heat is removed from the reactor by a liquid or gaseous coolant and converted into electricity in the turbogenerator. *See* ELECTRIC POWER GENERATION; HEAT TRANSFER; NUCLEAR POWER; NUCLEAR REACTOR.

A typical reactor core is made up of fuel assemblies, which contain the fuel rods. A fuel rod consists of many fuel pellets enclosed in a sealed tube called clad. Uranium and other heavy nuclei in the fuel interact with passing neutrons and undergo fission reactions that split them into lighter nuclei and emit energy and neutrons. The succession of fissions due to absorption of neutrons emitted in previous fissions is the chain reaction. Spaces are provided between fuel assemblies or between fuel rods for control rods, which absorb neutrons. The control rods are inserted into or withdrawn from the core to control the chain reaction. *See* CHAIN REACTION (PHYSICS); NEUTRON; NUCLEAR FISSION; NUCLEAR FUELS; URANIUM.

Significant quantities of radioactive materials are produced in the reactor, most from the fission process itself and the remainder from the absorption, by various materials, of neutrons emitted in the fission process. Some of these radioactive materials have useful applications, and the rest are nuclear waste. *See* NUCLEAR REACTION; RADIOACTIVITY.

**Nuclear facility design and operation.** Nuclear engineering encompasses a wide variety of disciplines that need to collaborate for the conception, development, design, construction, operation, and decommissioning of nuclear facilities. These disciplines include reactor physics, thermal hydraulics, thermodynamics, mechanics, metallurgy, chemistry, control and instrumentation, radiation control and shielding, health physics, and economics. *See* NUCLEAR CHEMISTRY.

Reactor physicists determine the dimensions and composition of the core required for establishing a

chain reaction and for maintaining it at full power for a desirable duration—typically 18–24 months (after which the reactor is shut down for 3–6 weeks for refueling)—and determine the composition, number, and location of the control rods necessary for safe operation of the reactor. They also calculate the fission rate at any given point in the core; the rate of change of the chain reaction due to variation in operating conditions, such as in the coolant inlet temperature and flow rate; the change in the fuel composition as a function of time; and the rate of interaction of neutrons with different materials. Finally, reactor physicists design shields against the leakage of nuclear radiation. *See* NUCLEAR PHYSICS; RADIATION SHIELDING; REACTOR PHYSICS.

Thermal hydraulics engineers determine how much coolant is needed in the core volume; the coolant flow rate; the size of the pumps required for circulating the coolant; and the temperature anywhere in the core and in sensitive components, such as the pressure vessel. Thermal hydraulics engineers and reactor physicists closely collaborate in analyzing reactor safety, using computer simulation to predict changes in the core power level and temperature distribution due to a variety of accident scenarios. The basic safety requirement is that no conceivable accident should lead to a breach in the fuel rod clad in the reactor vessel and in the reactor containment structures that could result in a release of radioactive materials to the environment.

Thermodynamic engineers design the systems needed to convert nuclear-generated heat to electricity at maximum efficiency.

Structural engineers perform mechanical analyses of many important components in nuclear power plants to determine the thickness of components so that the stresses they are subjected to will not exceed permissible limits. Among the most sensitive components are the fuel rod clad and the reactor pressure vessel.

Metallurgical engineers develop nuclear fuel and clad that can reliably operate in the reactor for long periods of time; predict the change in the mechanical properties of the nuclear fuel due to the accumulation of fission products and the relatively high temperature at the center of the fuel pellets in the operating reactor; predict the change in the mechanical properties of structural materials, as a result of being continuously bombarded by the neutrons emitted in the fission process; and develop structural materials that will maintain good mechanical strength at high temperatures and resist the corrosive effects of the coolant. *See* RADIATION DAMAGE TO MATERIALS.

Chemical engineers develop processes for converting uranium ores into the uranium dioxide used to manufacture the fuel pellets; for uranium isotopic enrichment; for processing spent fuel from the reactor; and for safe storage of nuclear waste. *See* DECONTAMINATION OF RADIOACTIVE MATERIALS; ISOTOPE SEPARATION; NUCLEAR CHEMICAL ENGINEERING; NUCLEAR FUEL CYCLE; NUCLEAR FUELS REPROCESSING.

Instrumentation and control engineers design detectors for monitoring the reactor's operating con-

ditions; design systems to process the signals from these detectors and transmit them to the central control room; and design systems to move the reactor control rods, change coolant flow rates, and open or close valves in response to the detectors' signals or to signals provided by the power-plant operators. They also design the control room where operators obtain detailed information on the power plant's condition and can modify this condition.

Health physicists and radiation control specialists monitor the radiation level of neutrons and gamma rays anywhere in the power plant accessible to the staff. They also measure the radiation doses to which staff members are exposed, to ensure that they will not exceed the permitted dose rate. *See* GAMMA RAYS; HEALTH PHYSICS; RADIATION INJURY (BIOLOGY).

Economists estimate the cost of generating electricity in the nuclear power plant, taking into account a host of factors such as the investment in the design and construction of the power plant; the interest on that investment; the time it takes to construct the plant; the cost of the nuclear fuel; the amount of energy generated by the nuclear fuel; the duration of plant shutdown for refueling and for maintenance; the cost of the operation and maintenance of the power plant.

**Non-power plant applications.** Starting in the 1990s, a significant number of nuclear engineers were engaged in activities associated with the design of facilities for the safe storage of spent fuel (the fuel that has been taken out of nuclear reactors) and of other nuclear waste; the design of underground repositories for safely storing the long-lived nuclear waste; and the design of special facilities for reducing the volume of and hazard from the nuclear waste. *See* RADIOACTIVE WASTE MANAGEMENT.

Nuclear engineers are also involved in the production of radioactive isotopes for medical and industrial applications. Additional nuclear engineering activities which are not associated with nuclear power plants include the design of facilities which use nuclear radiation for the diagnosis and treatment of cancers and other illnesses; for the detection of land mines, other explosives, and drugs; for the diagnosis of imperfections in sensitive components, such as aircraft components; for food sterilization; for the measurement of underground moisture level and automatic irrigation control; for the search for oil and other minerals; as well as for many other industrial and agricultural applications. *See* NUCLEAR MEDICINE; RADIOACTIVITY AND RADIATION APPLICATIONS; RADIOISOTOPE.

**Fusion research and development.** Many nuclear engineers are also involved in the research and development of future fusion power plants—plants that will be based on the fusion reaction for generating nuclear energy. Many challenging engineering problems are involved, including the development of large powerful magnets and the development of technologies for heating the fusion fuel to hundreds of millions of degrees; confining this ultrahot fuel;

and compressing submillimeter spherical capsules containing fusion fuel to many thousand times their natural solid density. New structural and coolant materials and new chemical processes are also being developed. See NUCLEAR FUSION. Ehud Greenspan

Bibliography. K. Almeras and R. Lee, *Nuclear Engineering: An Introduction*, 1992; S. Glasstone and A. Sesonske, *Nuclear Reactor Engineering*, 2 vols., 4th ed., 1994; R. A. Knief, *Nuclear Engineering: Theory and Technology of Commercial Nuclear Power*, 2d ed., 1992; J. R. Lamarsh and A. Baratta, *Introduction to Nuclear Engineering*, 3d ed., 2001; R. L. Murray, *Nuclear Energy*, 5th ed., 2000.

## Nuclear explosion

An explosion whose energy is produced by a nuclear transformation, either fission (**Fig. 1**) or fusion (**Fig. 2**). See NUCLEAR FISSION; NUCLEAR FUSION.

This article focuses on the effects of nuclear explosions. For discussions of how these explosions are produced See ATOMIC BOMB; HYDROGEN BOMB.

The energy of a nuclear explosion is usually stated in terms of the mass of trinitrotoluene (TNT) which would provide the same energy. The complete fissioning of 1 kg of uranium or plutonium would be equivalent to 17,000 metric tons of TNT (17 kilotons); 2 lb would be equivalent to 17,000 short tons. The indicated yield-to-mass ratio of  $1.7 \times 10^7$  cannot be realized, largely because of the ancillary equip-



**Fig. 1.** Nuclear explosion "Priscilla," with a yield of about 37 kilotons, detonated from a balloon on June 24, 1957, at the Nevada test site. (Los Alamos National Laboratory)



**Fig. 2.** Nuclear explosion "Mike," the first test of a thermonuclear device with a yield of about 10 megatons, detonated on the surface at Eniwetok Atoll on October 31, 1952. (Los Alamos National Laboratory)

ment necessary to assemble the nuclear components into an explosive configuration in the very short time required. The first nuclear weapons (1945) weighed about 5 tons, but with yields of 15 to 20 kT their yield-to-weight ratio was, nevertheless, close to 4000 times larger than that of previous weapons. By 1960 the United States had developed a weapon with a yield of about 1 megaton (MT) in a weight of about 1 ton for use in an intercontinental missile. With this, the yield-to-weight ratio was raised to  $10^6$ . Though improvements may be available, and may have been realized, there is no room for further changes of the qualitatively significant sort already achieved.

Although weapons with yields of up to 15 and 60 MT were fired by the United States and Soviet Union respectively, about 1960 the main interest of the major nuclear powers focused on adapting nuclear devices for delivery by missile carriers, and this called for smaller weapons with so-called moderate yields of tens or hundreds of kilotons, up to a few megatons. See MISSILE.

It is sometimes supposed that nuclear weapons are distinguished by the large amounts of energy released. In fact, there have been a number of accidental chemical explosions, particularly of ships loaded with ammunition or of large accumulations of ammonium nitrate, with energies in the range of 1–3 kT. In addition, since the great explosion of Krakatoa (1883), there have been a dozen or so volcanic explosions with energies in the tens of megatons. Though the size of a typical nuclear explosion is appalling, the most significant feature of a nuclear device derives from its yield-to-weight ratio. See EXPLOSIVE.

The damage mechanism of a conventional explosion is blast—the pressure, or shock, wave transmitted in the surrounding medium. The blast wave from a nuclear explosion is similar, except for the great difference in scale. A nuclear explosion also produces several kinds of effects not experienced with ordinary explosives.

**Blast.** The destructive action of blast correlates with the peak overpressure in the wave. The pressure near the surface is of primary interest. This is affected by the reflection that occurs at the surface as well as by the height of the explosion. From 1 kT exploded on the surface, the distance to which an overpressure of 5 lb/in.<sup>2</sup> (34.5 kilopascals) extends is about 1500 ft (450 m). This pressure extends to 2300 ft (700 m) from ground zero for an explosion at a height of approximately 800 ft (250 m). For heights greater than 1800 ft (550 m), 1 kT does not result in 5 lb/in.<sup>2</sup> on the surface. For a given yield and specified overpressure, there is a height of burst which maximizes the area exposed to the chosen overpressure. For very high pressures (hundreds of pounds per square inch), a low height is necessary, and a surface burst may do about as well.

The distance at which some stated overpressure will be realized scales with the cube root of the yield, providing the height of burst is similarly scaled. From the above example for 1 kT, the 5-lb/in.<sup>2</sup> (34.5-kPa) radius for a 1-megaton explosion at 8000 ft (2500 m) above the surface would be 23,000 ft (7000 m), and this radius would be 230 ft (70 m) for a 1-ton bomb detonated at a height of 80 ft (25 m).

Some heavy structures can withstand an overpressure of 5 lb/in.<sup>2</sup> (34.5 kPa), but such a blast would destroy most commercial buildings, apartments, and residences. People in residences exposed to 5 lb/in.<sup>2</sup> would have about a 50% chance of being killed. In many studies it is assumed that the number of persons surviving within the 5-lb/in.<sup>2</sup> ring is equal to the number killed in the area beyond, and the number of fatalities is taken to be equal to the number present inside the ring.

This is consistent with the experience at Hiroshima where, with a yield of 15 kT fired at a height of 1850 ft (560 m), the 5-lb/in.<sup>2</sup> (34.5-kPa) radius was about 5700 ft (1730 m). The number of early fatalities was about equal to the number of people estimated to have been present within about 1 mi (1.6 km) of ground zero.

Damage at decreasing levels extends to larger distances, and fatal injuries may still occur. Out to 2.5 times the 5-lb/in.<sup>2</sup> (34.5-kPa) range, where pressures would be only 1 lb/in.<sup>2</sup> (7 kPa), panes of window glass will be shattered into flying splinters.

**Thermal radiation.** Very high energy densities are realized in nuclear explosions. Consequently, when the energy generation is complete, the materials of the device constitute a ball of very hot gas with temperatures in excess of 10<sup>7</sup> K (2 × 10<sup>7</sup> °F). This material expands violently, engulfing the air it encounters, increasing the mass and volume of the hot zone while the temperature drops accordingly. By the time the expanding front has reached a radius of 33 ft (10 m), the mass of air entrained will be more than 5 tons, so that beyond this stage the product of the explosion behaves as a ball of strongly heated air. The hot globe radiates energy as a blackbody, the temperature determining the spectrum of quanta. Initially these quanta, with energies in the kilovolt range, are absorbed in an inch or two of air at sea-level den-

sity. Most of the radiation at this stage merely heats up the air being engulfed by the hydrodynamic process. As the falling temperature approaches 10<sup>4</sup> K (2 × 10<sup>4</sup> °F), the spectrum peaks in the visible range, the range of wavelengths for which air is transparent. The energy radiated in this band travels through the air until it impinges on some absorbing surface. As the temperature falls toward 3000°F (2000 K), the emission of thermal radiation (which varies as the fourth power of the temperature) ceases to be significant. About one-third of the energy of the explosion is distributed as thermal radiation on line-of-sight trajectories. *See* HEAT RADIATION.

Thermal radiation is not important in connection with conventional explosives since the temperature of freshly burned chemical explosives (about 5000°F or 3000 K) is too low for an appreciable fraction of the energy to be radiated.

Exposure to 5–10 cal/cm<sup>2</sup> (2–4 × 10<sup>5</sup> J/m<sup>2</sup>) of thermal radiation energy in the short time (a second, or so) during which the thermal pulse is delivered will ignite many combustible materials (fabrics, paper, dry leaves, and so forth). It will cause serious flash burns on exposed skin. Such energy levels will be delivered in clear air to 0.3–0.4 mi (0.5–0.6 km) by an explosion of 1 kT. The thermal energy is proportional to the yield at a given distance, and falls off (nearly) as the square of the distance. The ranges will be reduced by factors which reduce “visibility” in the atmosphere. At Hiroshima, burn injuries alone would have been fatal to almost all persons in the open without protection out to a little over 1 mi (1.6 km), and burns serious enough to require treatment were experienced at distances greater than 2 mi (3.2 km).

Another consequence of the thermal radiation is flash blindness. The eye, after being exposed to an intense burst of light, may not be able to receive new images for some seconds, or even for hours. A person need not be facing the burst point to be dazzled. The range at which this effect could occur is much greater than that at which burns would be experienced. It is also much greater at night, when the eye is dark-adapted, than in daylight. Intense light from the explosion can cause permanent burns on the retina in those who view the explosion directly.

**Prompt ionizing radiation.** Neutrons and gamma rays are released from a nuclear explosion. Since the fissions occur before the weapon materials expand, almost all the gamma rays accompanying fission are absorbed in these materials. But as soon as the hot ball of vaporized weapon components expands, the gamma rays being emitted by the radioactive fission fragments can escape to the air. During the few tens of seconds before the fireball rises away from the point at which the explosion occurred, it provides an intense source of gamma rays. The radiation emitted during this interval is referred to as prompt radiation. The remaining radioactivity, which is swept upward from the scene of the explosion to the altitude at which the fireball stops rising, and some of which ultimately returns to the surface,

constitutes the residual radiation. See ALPHA PARTICLES; NEUTRON; RADIOACTIVITY.

Neutrons and gamma rays are attenuated by passage through the air, the gamma rays by a factor of  $e$  (2.718) in a distance of about 1080 ft (330 m), the neutrons in about 650 ft (200 m). At distances greater than 3300 ft (1000 m) the gamma rays dominate, and the prompt radiation exposure varies as  $(1/R^2)e^{-R/a}$ , where  $R$  is the distance from the explosion and  $a$  is 1080 ft or 330 m. The biological damage caused by radiation results from the deposition of energy in tissue. See RADIATION INJURY (BIOLOGY).

At Hiroshima a dose equal to or greater than 450 rads (4.5 grays) extended to almost 1 mi (1.6 km). About half of the persons exposed to this dose in a short time will die within a few weeks, so that persons in this area who were not protected by heavy building walls experienced severe hazard from radiation. This is about the same distance for severe hazards from blast and thermal effects. From the scaling laws discussed above, it follows that the prompt radiation exposure, with its exponentially decreasing term, falls off more rapidly than the blast effect which, in turn, falls more rapidly than the intensity of thermal radiation. For an explosion much larger than 15 or 20 kT, the hazard range from thermal radiation or blast will be larger than that from prompt radiation, and prompt radiation will be a relatively unimportant effect. For much smaller yields, this order of importance will be reversed. To illustrate: from a 1-MT explosion the ranges of 500 rads (5 grays), 5 lb/in.<sup>2</sup> (34.5 kPa), and 10 cal/(cm<sup>2</sup> · s) [4.2 × 10<sup>5</sup> W/m<sup>2</sup>] are, respectively, about 1.6, 4.4, and 7.5 mi (2.5, 7, and 12 km); from 1 kT explosion they are 0.5, 0.4, and 0.3 mi (0.8, 0.7, and 0.5 km).

**Residual ionizing radiation.** For an explosion high enough above the surface that the fireball does not reach the ground, the only condensable material carried aloft is the small mass of the weapon itself. This is mixed with an enormous quantity of heated air. When the material cools to the point where the normally solid materials recondense, the particles formed are very small. The fission fragments, with the main exception of the gaseous elements xenon and krypton, will be incorporated in these particles. They will descend very slowly, reaching the surface only in days or months after the explosion. Most of these particles are distributed over the hemisphere in which the explosion occurred. The residual radioactivity distributed in this way would add a small component to the natural background radioactivity. Occasional exceptions to this would occur should rain, falling through or from the debris cloud before it had dispersed widely, bring an appreciable fraction of the radioactivity to the surface in a small area.

The situation is different for an explosion on or very close to the surface. In this case a large mass of surface material will be vaporized or swept up into the rising fireball. Much larger particles will be formed, which will fall back relatively rapidly, with the result that about half the residual radioactivity may be deposited on the surface in the first 24 h

and within a few hundred miles in the downwind direction. Assuming a steady 15 mi/h (7 m/s) wind this local fallout from a 1-MT surface burst could have the effect that an unprotected person who remained in place on the edge of a cigar-shaped contour about 90 mi (145 km) long and up to 13 mi (21 km) wide would receive a radiation dose in excess of 450 rads (4.5 grays) by 24 h after the explosion, and another 450 rads by the end of a week. At points inside the contour higher exposures would apply, while in a large area outside the contour exposures would still be at serious levels. To avoid death or injury, persons in an area several times larger than the contour would have to evacuate quickly or find adequate shelter for an extended period. Some areas (extending up to a few tens of miles in the downwind direction) could not safely be restored to use for several years. In actual situations, where variability in the wind is likely, the contour of the area involved will be quite irregular. See RADIOACTIVE FALLOUT.

**Electromagnetic pulse (EMP).** The gamma rays emitted from an explosion will collide with electrons in the air and create a sheath of electric current. This will never be fully symmetric, so that at least some electromagnetic radiation will be generated. At very high altitudes the variation in the density of the atmosphere provides a large-scale asymmetry which is important in this respect. A 1-MT explosion at an altitude 200–300 mi (300–500 km) over a point in Nebraska would result in an electromagnetic pulse in excess of 25,000 V/m over the entire area of the contiguous United States. Such a pulse would induce currents in power lines, antennas, and exposed conductors generally. Unless these, or the circuits connected to them, were appropriately shielded, the effects could be similar to those sometimes caused by lightning: burnt-out components, short circuits, opened breakers, and so forth. Nationwide interruptions or breakdowns could occur on communication networks, power systems, and control equipment. See ELECTROMAGNETIC PULSE (EMP).

**Radio and radar.** As a result of the radioactivity of the fission fragments, the cloud containing the weapon debris will be in an ionized state for a considerable period. An explosion can also cause temporary regional changes in the ionosphere. The operation of long-range radio communication or radar observation in channels affected by such perturbations could be degraded or blocked.

**Ozone depletion.** At the very high temperatures in the fireball of a nuclear explosion, chemical reactions occur in normal air. By the time the fireball cools, nitric oxide composes about 1% of its molecules. Nitric oxide can catalyze the destruction of ozone. It is only from rather large explosions (greater than 1 MT) that the fireball rises to the altitude (greater than 12 mi or 20 km) of the Earth's ozone blanket in the stratosphere. Hundreds of large explosions would be required to introduce enough nitric oxide into the ozone to deplete it appreciably. However, nuclear weapon stockpiles favor weapons having relatively small yields. An ozone catastrophe in the event of a nuclear war may no longer

belong on a list of realistic threats. See STRATOSPHERIC OZONE.

**Worldwide fallout.** A large-scale nuclear war was usually pictured as involving several thousand megatons of nuclear explosions in a short period, mainly on, or over, the territory of the United States and the Soviet Union. Following such a war, residual radiation would be deposited over the remaining area of the Northern Hemisphere, particularly in the North Temperate Zone. Radiation exposure at levels lower than those causing any immediate perceptible ill effects can still result in cancer after several years, and in genetic effects. See MUTATION.

Estimates of the magnitude of such effects have ranged from the possible extinction of biological life down to quite mild levels. Though such calculations can be made, the results depend strongly on many assumptions. The best estimates at present suggest that over the three or four decades following a nuclear war some millions of deaths would result from delayed radiation effects outside the target countries. Though this would appear to be a very large number, it would still represent only a small fractional increase in the normal incidence of fatal cancer. See CANCER (MEDICINE).

**Nuclear winter.** Many large urban fires and massive forest or brush fires could be started by a nuclear war. Large amounts of soot and smoke would be injected into the atmosphere by these near-simultaneous fires. Such smoke clouds would merge and be distributed around the hemisphere, and they could persist for weeks. They would reduce the amount of sunlight reaching the surface and could cause a precipitous drop in temperature in the midcontinental regions of the whole temperate zone. By some estimates, temperatures would average below 0°F (−20°C) for months or more, the amount of cooling and the duration depending on the magnitude of the war, but also on many of the assumptions necessary for the estimates. Such an outcome of a possible nuclear war has come to be referred to as nuclear winter. Even some fraction of such an effect would be serious.

J. Carson Mark

**Bibliography.** S. Glasstone and P. J. Dolan (eds.), *The Effects of Nuclear Weapons*, 1977, reprint 1983; C. C. Grace, *Nuclear Weapons: Principles, Effects, and Survivability*, 1993; M. A. Harwell and T. C. Hutchinson, *Environmental Consequences of Nuclear War*, vol. 2, 2d ed., 1989; Office of Technology Assessment, Congress of the United States, *The Effects of Nuclear War*, 1980; A. B. Pittock et al., *Environmental Consequences of Nuclear War*, vol. 1, 2d ed., 1989.

## Nuclear fission

An extremely complex nuclear reaction representing a cataclysmic division of an atomic nucleus into two nuclei of comparable mass. This rearrangement or division of a heavy nucleus may take place naturally (spontaneous fission) or under bombardment with neutrons, charged particles, gamma rays, or other

carriers of energy (induced fission). Although nuclei with mass number  $A$  of approximately 100 or greater are energetically unstable against division into two lighter nuclei, the fission process has a small probability of occurring, except with the very heavy elements. Even for these elements, in which the energy release is of the order of  $2 \times 10^8$  eV, the lifetimes against spontaneous fission are reasonably long. See NUCLEAR REACTION.

**Liquid-drop model.** The stability of a nucleus against fission is most readily interpreted when the nucleus is viewed as being analogous to an incompressible and charged liquid drop with a surface tension. Such a droplet is stable against small deformations when the dimensionless fissility parameter  $X$  in Eq. (1) is less than unity, where the charge is in

$$X = \frac{(\text{charge})^2}{10 \times \text{volume} \times \text{surface tension}} \quad (1)$$

esu, the volume is in  $\text{cm}^3$ , and the surface tension is in  $\text{ergs/cm}^2$ . The fissility parameter is given approximately, in terms of the charge number  $Z$  and mass number  $A$ , by the relation  $X = Z^2/47A$ .

Long-range Coulomb forces between the protons act to disrupt the nucleus, whereas short-range nuclear forces, idealized as a surface tension, act to stabilize it. The degree of stability is then the result of a delicate balance between the relatively weak electromagnetic forces and the strong nuclear forces. Although each of these forces results in potentials of several hundred megaelectronvolts, the height of a typical barrier against fission for a heavy nucleus, because they are of opposite sign but do not quite cancel, is only 5 or 6 MeV. Investigators have used this charged liquid-drop model with great success in describing the general features of nuclear fission and also in reproducing the total nuclear binding energies. See NUCLEAR BINDING ENERGY; NUCLEAR STRUCTURE; SURFACE TENSION.

**Shell corrections.** The general dependence of the potential energy on the fission coordinate representing nuclear elongation or deformation for a heavy nucleus such as  $^{240}\text{Pu}$  is shown in **Fig. 1**. The expanded scale used in this figure shows the large decrease in energy of about 200 MeV as the fragments separate to infinity. It is known that  $^{240}\text{Pu}$  is deformed in its ground state, which is represented by the lowest minimum of  $-1813$  MeV near zero deformation. This energy represents the total nuclear binding energy when the zero of potential energy is the energy of the individual nucleons at a separation of infinity. The second minimum to the right of zero deformation illustrates structure introduced in the fission barrier by shell corrections, that is, corrections dependent upon microscopic behavior of the individual nucleons, to the liquid-drop mass. Although shell corrections introduce small wiggles in the potential-energy surface as a function of deformation, the gross features of the surface are reproduced by the liquid-drop model. Since the typical fission barrier is only a few megaelectronvolts, the magnitude of the shell correction need only be small for irregularities to be

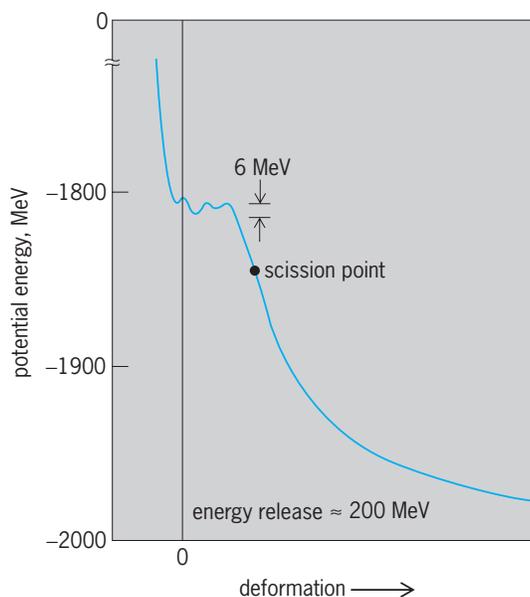


Fig. 1. Plot of the potential energy in MeV as a function of deformation for the nucleus  $^{240}\text{Pu}$ . (After M. Bolsteli et al., *New calculations of fission barriers for heavy and superheavy nuclei*, *Phys. Rev.*, 5C:1050–1077, 1972)

introduced into the barrier. This structure is schematically illustrated for a heavy nucleus by the double-humped fission barrier in Fig. 2, which represents the region to the right of zero deformation in Fig. 1 on an expanded scale. The fission barrier has two maxima and a rather deep minimum in between. For comparison, the single-humped liquid-drop barrier is also schematically illustrated. The transition in the shape of the nucleus as a function of deformation is represented in the upper part of the figure.

**Double-humped barrier.** The developments which led to the proposal of a double-humped fission barrier were triggered by the experimental discovery of

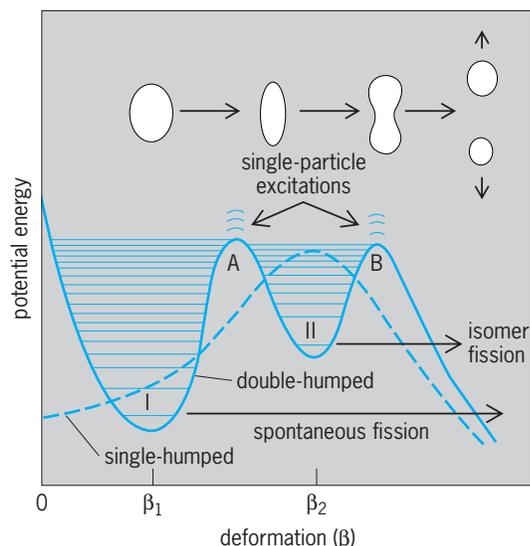


Fig. 2. Schematic plots of single-humped fission barrier of liquid-drop model and double-humped barrier introduced by shell corrections. (After J. R. Huizenga, *Nuclear fission revisited*, *Science*, 168:1405–1413, 1970)

spontaneously fissionable isomers by S. M. Polikanov and colleagues in the Soviet Union and by V. M. Strutinsky's pioneering theoretical work on the binding energy of nuclei as a function of both nucleon number and nuclear shape. The double-humped character of the nuclear potential energy as a function of deformation arises, within the framework of the Strutinsky shell-correction method, from the superposition of a macroscopic smooth liquid-drop energy and a shell-correction energy obtained from a microscopic single-particle model. Oscillations occurring in this shell correction as a function of deformation lead to two minima in the potential energy, shown in Fig. 2, the normal ground-state minimum at a deformation of  $\beta_1$  and a second minimum at a deformation of  $\beta_2$ . States in these wells are designated class I and class II states, respectively. Spontaneous fission of the ground state and isomeric state arises from the lowest-energy class I and class II states, respectively. See NUCLEAR ISOMERISM.

The calculation of the potential-energy curve illustrated in Fig. 1 may be summarized as follows. The smooth potential energy obtained from a macroscopic (liquid-drop) model is added to a fluctuating potential energy representing the shell corrections, and to the energy associated with the pairing of like nucleons (pairing energy), derived from a non-self-consistent microscopic model. The calculation of these corrections requires several steps: (1) specification of the geometrical shape of the nucleus, (2) generation of a single-particle potential related to its shape, (3) solution of the Schrödinger equation, and (4) calculation from these single-particle energies of the shell and pairing energies.

The oscillatory character of the shell corrections as a function of deformation is caused by variations in the single-particle level density in the vicinity of the Fermi energy. For example, the single-particle levels of a pure harmonic oscillator potential arrange themselves in bunches of highly degenerate shells at any deformation for which the ratio of the major and minor axes of the spheroidal equipotential surfaces is equal to the ratio of two small integers. Nuclei with a filled shell, that is, with a level density at the Fermi energy that is smaller than the average, will then have an increased binding energy compared to the average, because the nucleons occupy deeper and more bound states; conversely, a large level density is associated with a decreased binding energy. It is precisely this oscillatory behavior in the shell correction that is responsible for spherical or deformed ground states and for the secondary minima in fission barriers, as illustrated in Fig. 2. See NONRELATIVISTIC QUANTUM THEORY.

More detailed theoretical calculations based on this macroscopic-microscopic method have revealed additional features of the fission barrier. In these calculations the potential energy is regarded as a function of several different modes of deformation. The outer barrier B (Fig. 2) is reduced in energy for shapes with pronounced left-right asymmetry (pear shapes), whereas the inner barrier A and deformations in the vicinity of the second minimum are stable

against such mass asymmetric degrees of freedom. Similar calculations of potential-energy landscapes reveal the stability of the second minimum against gamma deformations, in which the two small axes of the spheroidal nucleus become unequal, that is, the spheroid becomes an ellipsoid.

**Experimental consequences.** The observable consequences of the double-humped barrier have been reported in numerous experimental studies. In the actinide region, more than 30 spontaneously fissionable isomers have been discovered between uranium and berkelium, with half-lives ranging from  $10^{-11}$  to  $10^{-2}$  s. These decay rates are faster by 20 to 30 orders of magnitude than the fission half-lives of the ground states, because of the increased barrier tunneling probability (Fig. 2). Several cases in which excited states in the second minimum decay by fission are also known. Normally these states decay within the well by gamma decay; however, if there is a hindrance in gamma decay due to spin, the state (known as a spin isomer) may undergo fission instead.

Qualitatively, the fission isomers are most stable in the vicinity of neutron numbers 146 to 148, a value in good agreement with macroscopic-microscopic theory. For elements above berkelium, the half-lives become too short to be observable with available techniques; and for elements below uranium, the prominent decay is through barrier A into the first well, followed by gamma decay. It is difficult to detect this competing gamma decay of the ground state

in the second well (called a shape isomeric state), but identification of the gamma branch of the 200-ns  $^{238}\text{U}$  shape isomer has been reported. See RADIOACTIVITY.

Direct evidence of the second minimum in the potential-energy surface of the even-even nucleus  $^{240}\text{Pu}$  has been obtained through observations of the E2 transitions within the rotational band built on the isomeric  $0^+$  level. The rotational constant (which characterizes the spacing of the levels and is expected to be inversely proportional to the effective moment of inertia of the nucleus) found for this band is less than one-half that for the ground state and confirms that the shape isomers have a deformation  $\beta_2$  much larger than the equilibrium ground-state deformation  $\beta_1$ . From yields and angular distributions of fission fragments from the isomeric ground state and low-lying excited states, some information has been derived on the quantum numbers of specific single-particle states of the deformed nucleus (Nilsson single-particle states) in the region of the second minimum.

At excitation energies in the vicinity of the two barrier tops, measurements of the subthreshold neutron fission cross sections of several nuclei have revealed groups of fissioning resonance states with wide energy intervals between each group where no fission occurs. Such a spectrum is illustrated in Fig. 3a, where the subthreshold fission cross section of  $^{240}\text{Pu}$  is shown for neutron energies between

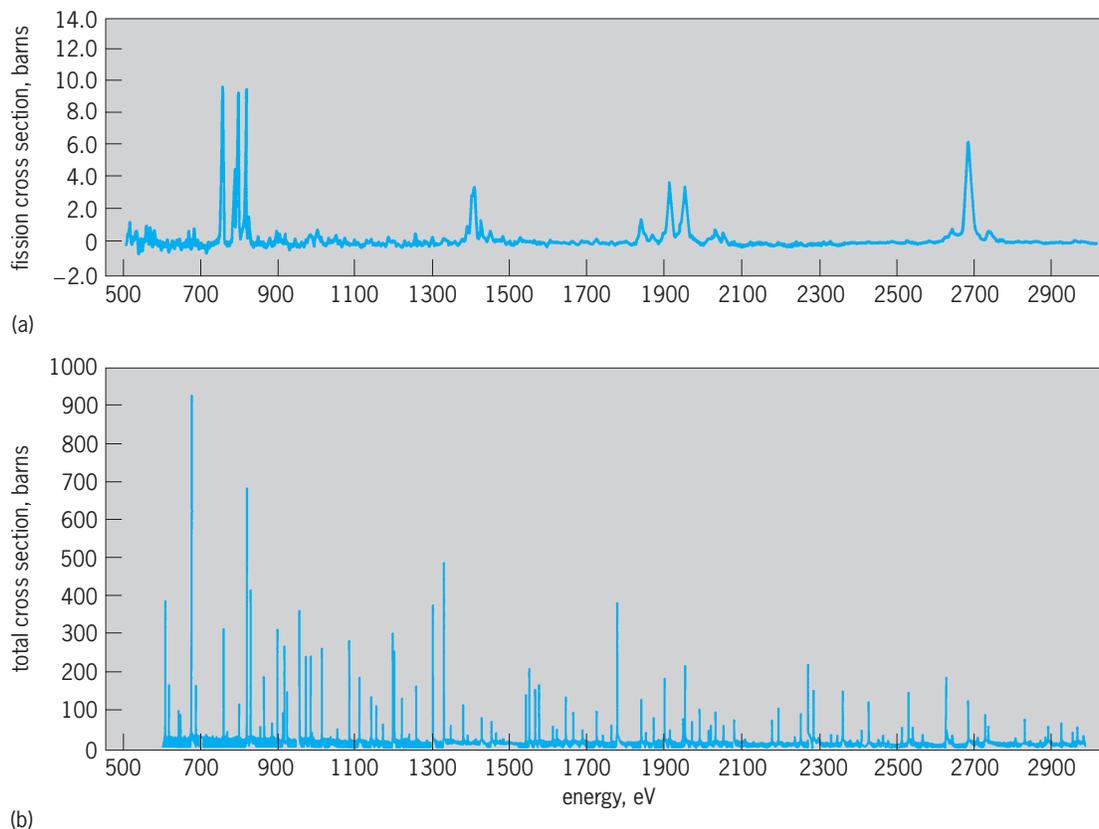


Fig. 3. Grouping of fission resonances demonstrated by (a) the neutron fission cross section of  $^{240}\text{Pu}$  and (b) the total neutron cross section. (After V. M. Strutinsky and H. C. Pauli, *Shell-structure effects in the fissioning nucleus*, *Proceedings of the 2d IAEA Symposium on Physics and Chemistry of Fission*, Vienna, pp. 155–177, 1969)

500 and 3000 eV. As shown in Fig. 3*b*, between the fissioning resonance states there are many other resonance states, known from data on the total neutron cross sections, which have negligible fission cross sections. Such structure is explainable in terms of the double-humped fission barrier and is ascribed to the coupling between the compound states of normal density in the first well to the much less dense states in the second well. This picture requires resonances of only one spin to appear within each intermediate structure group illustrated in Fig. 3*a*. In an experiment using polarized neutrons on a polarized  $^{237}\text{Np}$  target, it was found that all nine fine-structure resonances of the 40-eV group have the same spin and parity:  $I = 3+$ . Evidence has also been obtained for vibrational states in the second well from neutron ( $n, f$ ) and deuteron stripping ( $d, pf$ ) reactions at energies below the barrier tops ( $f$  indicates fission of the nucleus). See NEUTRON SPECTROMETRY.

A. Bohr suggested that the angular distributions of the fission fragments are explainable in terms of the transition-state theory, which describes a process in terms of the states present at the barrier deformation. The theory predicts that the cross section will have a steplike behavior for energies near the fission barrier, and that the angular distribution will be determined by the quantum numbers associated with each of the specific fission channels. The theoretical angular distribution of fission fragments is based on two assumptions. First, the two fission fragments are assumed to separate along the direction of the nuclear symmetry axis so that the angle  $\theta$  between the direction of motion of the fission fragments and the direction of motion of the incident bombarding particle represents the angle between the body-fixed axis (the long axis of the spheroidal nucleus) and the space-fixed axis (some specified direction in the laboratory, in this case the direction of motion of the incident particle). Second, it is assumed that during the transition from the saddle point (corresponding to the top of the barrier) to scission (the division of the nucleus into two fragments) the Coriolis forces do not change the value of  $K$  (where  $K$  is the projection of the total angular momentum  $I$  on the nuclear symmetry axis) established at the saddle point.

In several cases, low-energy photofission and neutron fission experiments have shown evidence of a

double-humped barrier. In the case of two barriers, the question arises as to which of the two barriers A or B is responsible for the structure in the angular distributions. For light actinide nuclei like thorium, the indication is that barrier B is the higher one, whereas for the heavier actinide nuclei, the inner barrier A is the higher one. The heights of the two barriers themselves are most reliably determined by investigating the probability of induced fission over a range of several mega-electronvolts in the threshold region. Many direct reactions have been used for this purpose, for example, ( $d, pf$ ), ( $t, pf$ ), and ( $^3\text{He}, df$ ). There is reasonably good agreement between the experimental and theoretical barriers. The theoretical barriers are calculated with realistic single-particle potentials and include the shell corrections.

**Fission probability.** The cross section for particle-induced fission  $\sigma(y, f)$  represents the cross section for a projectile  $y$  to react with a nucleus and produce fission, as shown by Eq. (2). The quantities  $\sigma_R(y)$ ,  $\Gamma_f$ ,

$$\sigma(y, f) = \sigma_R(y) \frac{\Gamma_f}{\Gamma_t} \quad (2)$$

and  $\Gamma_t$  are the total reaction cross sections for the incident particle  $y$ , the fission width, and the total level width, respectively, where  $\Gamma_t = \Gamma_f + \Gamma_n + \Gamma_\gamma + \dots$  is the sum of all partial-level widths. All the quantities in Eq. (2) are energy-dependent. Each of the partial widths for fission, neutron emission, radiation, and so on, is defined in terms of a mean lifetime  $\tau$  for that particular process, for example,  $\Gamma_f = \hbar/\tau_f$ . Here  $\hbar$ , the action quantum, is Planck's constant divided by  $2\pi$  and is numerically equal to  $1.0546 \times 10^{-34}$  J s =  $0.66 \times 10^{-15}$  eV s. The fission width can also be defined in terms of the energy separation  $D$  of successive levels in the compound nucleus and the number of open channels in the fission transition nucleus (paths whereby the nucleus can cross the barrier on the way to fission), as given by expression (3), where  $I$  is the angular momentum and  $i$  is an index

$$\Gamma_f(I) = \frac{D(I)}{2\pi} \sum_i N_{fi} \quad (3)$$

labeling the open channels  $N_{fi}$ . The contribution of each fission channel to the fission width depends upon the barrier transmission coefficient, which, for

**Cross sections for neutrons of thermal energy to produce fission or undergo capture in the principal nuclear species, and neutron yields from these nuclei\***

Nucleus	Cross section for fission, $\sigma_f$ , $10^{-24}$ cm <sup>2</sup>	$\sigma_f$ plus cross section for radiative capture, $\sigma_r$	Ratio, $1 + \alpha$	Number of neutrons released per fission, $\nu$	Number of neutrons released per slow neutron captured, $\eta = \nu/(1 + \alpha)$
$^{233}\text{U}$	$525 \pm 2$	$573 \pm 2$	$1.093 \pm 0.003$	$2.50 \pm 0.01$	$2.29 \pm 0.01$
$^{235}\text{U}$	$577 \pm 1$	$678 \pm 2$	$1.175 \pm 0.002$	$2.43 \pm 0.01$	$2.08 \pm 0.01$
$^{239}\text{Pu}$	$741 \pm 4$	$1015 \pm 4$	$1.370 \pm 0.006$	$2.89 \pm 0.01$	$2.12 \pm 0.01$
$^{238}\text{U}$	0	$2.73 \pm 0.04$			0
Natural uranium	4.2	7.6	1.83	$2.43 \pm 0.01$	1.33

\*Data from Brookhaven National Laboratory 325, 2d ed., suppl. no. 2, vol. 3, 1965. The data presented are the recommended or least-square values published in this reference for 0.0253-eV neutrons.

a two-humped barrier (see Fig. 2), is strongly energy-dependent. This results in an energy-dependent fission cross section which is very different from the total cross section shown in Fig. 3 for  $^{240}\text{Pu}$ .

When the incoming neutron has low energy, the likelihood of reaction is substantial only when the energy of the neutron is such as to form the compound nucleus in one or another of its resonance levels (Fig. 3*b*). The requisite sharpness of the “tuning” of the energy is specified by the total level width  $\Gamma$ . The nuclei  $^{233}\text{U}$ ,  $^{235}\text{U}$ , and  $^{239}\text{Pu}$  have a very large cross section to take up a slow neutron and undergo fission (see **table**) because both their absorption cross section and their probability for decay by fission are large. The probability for fission decay is high because the binding energy of the incident neutron is sufficient to raise the energy of the compound nucleus above the fission barrier. The very large, slow neutron fission cross sections of these isotopes make them important fissile materials in a chain reactor. See CHAIN REACTION (PHYSICS); REACTOR PHYSICS.

**Scission.** The scission configuration is defined in terms of the properties of the intermediate nucleus just prior to division into two fragments. In heavy nuclei the scission deformation is much larger than the saddle deformation at the barrier, and it is important to consider the dynamics of the descent from saddle to scission. One of the important questions in the passage from saddle to scission is the extent to which this process is adiabatic with respect to the particle degrees of freedom. As the nuclear shape changes, it is of interest to investigators to know the probability for the nucleons to remain in the lowest-energy orbitals. If the collective motion toward scission is very slow, the single-particle degrees of freedom continually readjust to each new deformation as the distortion proceeds. In this case, the adiabatic model is a good approximation, and the decrease in potential energy from saddle to scission appears in collective degrees of freedom at scission, primarily as kinetic energy associated with the relative motion of the nascent fragments.

On the other hand, if the collective motion between saddle and scission is so rapid that equilibrium is not attained, there will be a transfer of collective energy into nucleonic excitation energy. Such a nonadiabatic model, in which collective energy is transferred to single-particle degrees of freedom during the descent from saddle to scission, is usually referred to as the statistical theory of fission. See PERTURBATION (QUANTUM MECHANICS).

The experimental evidence indicates that the saddle to scission time is somewhat intermediate between these two extreme models. The dynamic descent of a heavy nucleus from saddle to scission depends upon the nuclear viscosity. A viscous nucleus is expected to have a smaller translational kinetic energy at scission and a more elongated scission configuration. Experimentally, the final translational kinetic energy of the fragments at infinity, which is related to the scission shape, is measured. Hence, in principle, it is possible to estimate the nuclear viscosity coefficient by comparing the calculated dependence

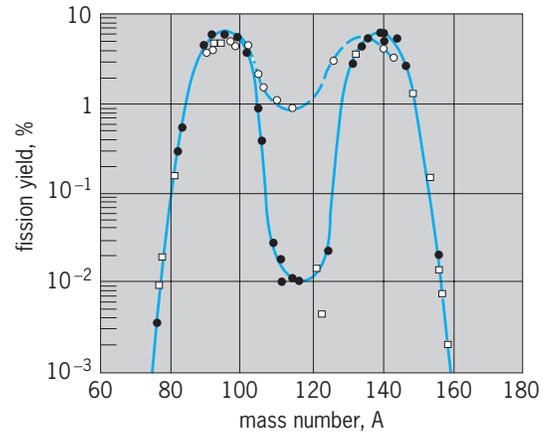


Fig. 4. Mass distribution of fission fragments formed by neutron-induced fission of  $^{235}\text{U} + n = ^{236}\text{U}$  when neutrons have thermal energy, solid curve (*Plutonium Project Report, Rev. Mod. Phys.*, 13:539, 1964), and 14-MeV energy, broken curve (based on R. W. Spence, *Brookhaven National Laboratory, AEC-BNL (C-9)*, 1949). Quantity plotted is  $100 \times (\text{number of fission decay chains formed with given mass})/(\text{number of fissions})$ .

upon viscosity of fission-fragment kinetic energies with experimental values. The viscosity of nuclei is an important nuclear parameter which also plays an important role in collisions of very heavy ions.

The mass distribution from the fission of heavy nuclei is predominantly asymmetric. For example, division into two fragments of equal mass is about 600 times less probable than division into the most probable choice of fragments when  $^{235}\text{U}$  is irradiated with thermal neutrons. When the energy of the neutrons is increased, symmetric fission (Fig. 4) becomes more probable. In general, heavy nuclei fission asymmetrically to give a heavy fragment of approximately constant mean mass number 139 and a corresponding variable-mass light fragment (Fig. 5).

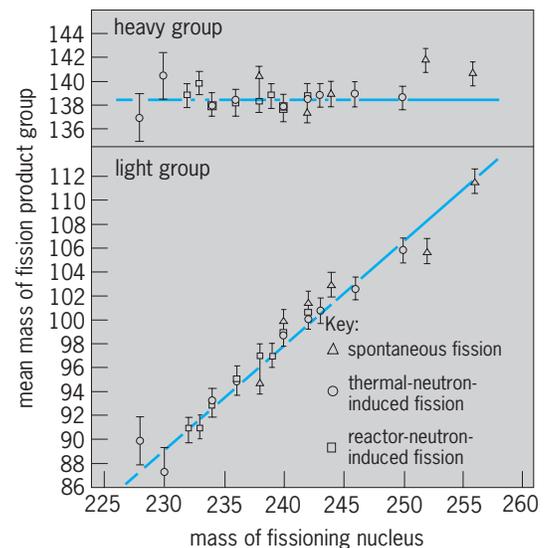


Fig. 5. Average masses of the light- and heavy-fission product groups as a function of the masses of the fissioning nucleus. Energy spectrum of reactor neutrons is that associated with fission. (After K. F. Flynn et al., *Distribution of mass in the spontaneous fission of  $^{256}\text{Fm}$* , *Phys. Rev.*, 5C:1725-1729, 1972)

These experimental results have been difficult to explain theoretically. Calculations of potential-energy surfaces show that the second barrier (B in Fig. 2) is reduced in energy by up to 2 or 3 MeV, if octuple deformations (pear shapes) are included. Hence, the theoretical calculations show that mass asymmetry is favored at the outer barrier, although direct experimental evidence supporting the asymmetric shape of the second barrier is very limited. It is not known whether the mass asymmetric energy valley extends from the saddle to scission; and the effect of dynamics on mass asymmetry in the descent from saddle to scission has not been determined. Experimentally, as the mass of the fissioning nucleus approaches  $A \approx 260$ , the mass distribution approaches symmetry. This result is qualitatively in agreement with theory.

A nucleus at the scission configuration is highly elongated and has considerable deformation energy. The influence of nuclear shells on the scission shape introduces structure into the kinetic energy and neutron-emission yield as a function of fragment mass. The experimental kinetic energies for the neutron-induced fission of  $^{233}\text{U}$ ,  $^{235}\text{U}$ , and  $^{239}\text{Pu}$  have a pronounced dip as symmetry is approached, as shown in Fig. 6. (This dip is slightly exaggerated in the figure because the data have not been corrected for fission fragment scattering.) The variation in the neutron yield as a function of fragment mass in the neutron yield as a function of fragment mass for these same nuclei (Fig. 7) has a “saw-toothed” shape which is asymmetric about the mass of the symmetric fission fragment. Both these phenomena are reasonably well accounted for by the inclusion of closed-shell structure into the scission configuration.

A number of light charged particles (for example, isotopes of hydrogen, helium, and lithium) have been observed to occur, with low probability, in fission. These particles are believed to be emitted very near the time of scission. Available evidence also indicates

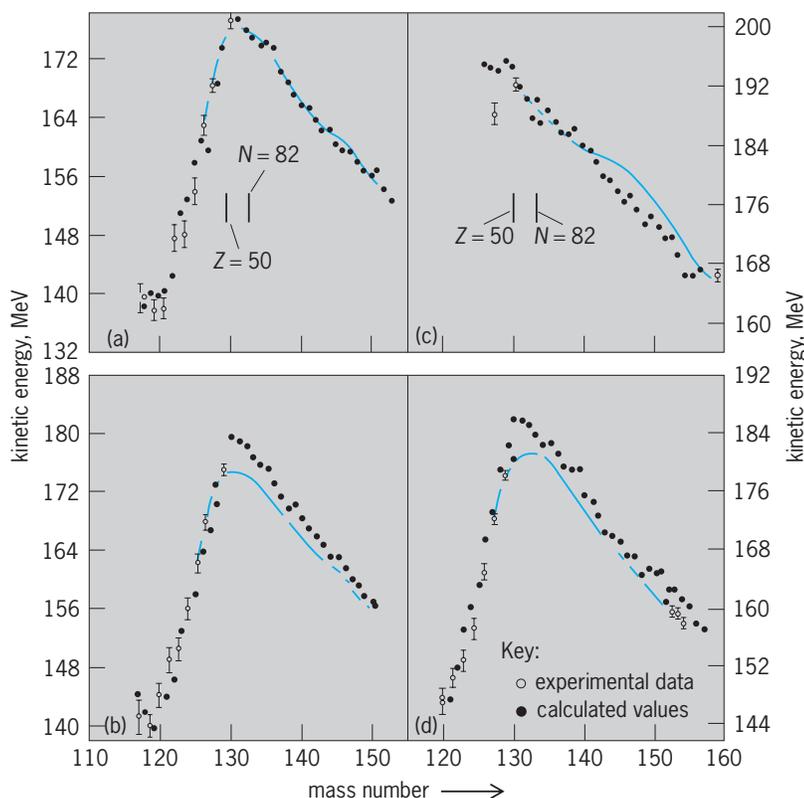


Fig. 6. Average total kinetic energy of fission fragments as a function of heavy fragment mass for fission of (a)  $^{233}\text{U}$ , (b)  $^{235}\text{U}$ , (c)  $^{252}\text{Cf}$ , and (d)  $^{239}\text{Pu}$ . Curves indicate experimental data. (After J. C. D. Milton and J. S. Fraser, *Time-of-flight fission studies on  $^{233}\text{U}$ ,  $^{235}\text{U}$  and  $^{239}\text{Pu}$* , *Can. J. Phys.*, 40:1626–1663, 1962)

that neutrons are emitted at or near scission with considerable frequency.

**Postscission phenomena.** After the fragments are separated at scission, they are further accelerated as the result of the large Coulomb repulsion. The initially deformed fragments collapse to their

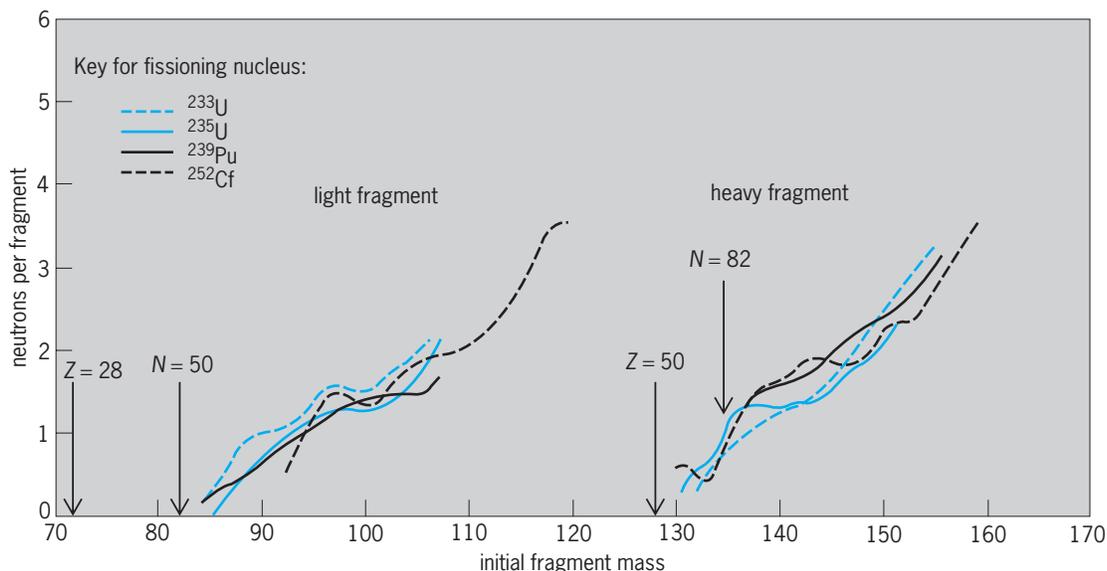


Fig. 7. Neutron yields as a function of fragment mass for four types of fission as determined from mass-yield data. Approximate initial fragment masses corresponding to various neutron and proton “magic numbers”  $N$  and  $Z$  are indicated. (After J. Terrell, *Neutron yields from individual fission fragments*, *Phys. Rev.*, 127:880–904, 1962)

equilibrium shapes, and the excited primary fragments lose energy by evaporating neutrons. After neutron emission, the fragments lose the remainder of their energy by gamma radiation, with a life-time of about  $10^{-11}$  s. The kinetic energy and neutron yield as a function of mass are shown in Figs. 6 and 7. The variation of neutron yield with fragment mass is directly related to the fragment excitation energy. Minimum neutron yields are observed for nuclei near closed shells because of the resistance to deformation of nuclei with closed shells. Maximum neutron yields occur for fragments that are “soft” toward nuclear deformation. Hence, at the scission configuration, the fraction of the deformation energy stored in each fragment depends on the shell structure of the individual fragments. After scission, this deformation energy is converted to excitation energy, and hence, the neutron yield is directly correlated with the fragment shell structure. This conclusion is further supported by the correlation between the neutron yield and the final kinetic energy. Closed shells result in a larger Coulomb energy at scission for fragments that have a smaller deformation energy and a smaller number of evaporated neutrons.

After the emission of the prompt neutrons and gamma rays, the resulting fission products are unstable against beta decay. For example, in the case of thermal neutron fission of  $^{235}\text{U}$ , each fragment undergoes on the average about three beta decays before it settles down to a stable nucleus. For selected fission products (for example,  $^{87}\text{Br}$  and  $^{137}\text{I}$ ) beta decay leaves the daughter nucleus with excitation energy exceeding its neutron binding energy. The resulting delayed neutrons amount, for thermal neutron fission of  $^{235}\text{U}$ , to about 0.7% of all the neutrons that are given off in fission. Although small in number, these neutrons are quite important in stabilizing nuclear chain reactions against sudden minor fluctuations in reactivity. See DELAYED NEUTRON; NEUTRON; THERMAL NEUTRONS.

John R. Huizenga

**Bibliography.** P. David (ed.), *Dynamics of Nuclear Fission and Related Collective Phenomena, Proceedings*, Bad Honnef, Germany, 1981; J. Kristiak and B. I. Pustylnik (eds.), *Dynamical Aspects of Nuclear Fission: Proceedings of the 2d International Conference*, 1994; H. Marten and D. Seelinger (eds.), *Physics and Chemistry of Fission*, 1991; R. Vandenbosch and J. R. Huizenga, *Nuclear Fission*, 1973.

## Nuclear fuel cycle

The nuclear fuel cycle typically involves the following steps: (1) finding and mining the uranium ore; (2) refining the uranium from other elements; (3) enriching the uranium-235 content to 3–5%; (4) fabricating fuel elements; (5) interim storage and cooling of spent fuel; (6) reprocessing of spent fuel to recover uranium and plutonium (optional); (7) fabricating recycle fuel for added energy production (optional); (8) cooling of spent fuel or reprocessing waste, and its eventual transport to a repos-

itory for disposal in secure long-term storage. See NUCLEAR FUELS.

Steps 6 and 7 are used in Britain, France, India, Japan, and Russia. They are no longer used in the United States, which by federal policy has been restricted to a “once through” fuel cycle.

**Raw materials.** Nuclear reactors produce energy using fuel made of uranium slightly enriched in the isotope  $^{235}\text{U}$ . The basic raw material is natural uranium that contains 0.71%  $^{235}\text{U}$  (the only naturally occurring isotope that can sustain a chain reaction). The other isotope of natural uranium consists of  $^{238}\text{U}$ , part of which converts to plutonium-239, during reactor operation. The isotope  $^{239}\text{Pu}$  also sustains fission, typically contributing about one-third of the energy produced per fuel cycle. See NUCLEAR FISSION; NUCLEAR REACTOR.

Similarly, thorium-232, itself nonfissionable, when exposed in a reactor partly converts to fissionable  $^{233}\text{U}$ , but this cycle has not been used commercially.

**Uranium resources and mining.** Large deposits of commercially extractable uranium ores have been found in the United States (Colorado plateau), eastern Europe, Russia, Canada, Africa, and Australia. Commercial production occurs in 23 countries. World reserves are estimated at over  $4 \times 10^6$  tons ( $3.6 \times 10^6$  metric tons) of recoverable uranium ore. Resource estimates including low-grade ores range up to  $20 \times 10^7$  tons ( $1.8 \times 10^7$  metric tons). See URANIUM.

Deposits with no surface outcrops can be found by detecting the tiny amounts of radon gas that escape from radioactive decay of uranium. Most production is now in near-surface mining. Tunnel mining requires high ventilation rates due to the presence of radon gas.

**Refining.** Depending on the type of ore, uranium may be purified by precipitation from aqueous solutions with ammonia or by solvent extraction from nitrate solutions. Uranium oxide is treated with hydrogen fluoride to make  $\text{UF}_4$ , and then with fluorine to make  $\text{UF}_6$  gas. This step also completes the purification of uranium since most impurities do not have volatile fluorides.

**Enrichment methods.** The initial enrichment of gram quantities of  $^{235}\text{U}$  was done in a large mass spectrograph adapted from a cyclotron magnet in Berkeley, California. The first tonnage production was done in the United States using  $\text{UF}_6$  in a gaseous diffusion cascade, a method that was also used by Argentina, China, and France. Production plants in the United States and eight other countries currently use gas centrifuges. Other methods of separation applicable to uranium isotopes have been demonstrated on a kilogram scale, including AVLIS (atomic vapor laser isotope separation), MLIS (molecular laser isotope separation), and several processes based on ion-exchange columns. See ISOTOPE SEPARATION.

**Fuel fabrication.** Uranium in the form of  $\text{UF}_6$  is enriched to 3–5%  $^{235}\text{U}$ , then converted to  $\text{U}_3\text{O}_8$  and  $\text{UO}_2$ . These oxides are compacted and furnace-fired to make high-density ceramic pellets. The pellets are put into zirconium tubes that are arranged in

precisely spaced bundles for loading into a reactor. When recycle is used, the fuel pellets are made of a solid solution of uranium and plutonium oxides called MOX (mixed oxides) fuel.

Water-cooled reactors (over 92% of the world's nuclear capacity) use fuels consisting of uranium oxide pellets in zirconium tubes. Four percent of these water-cooled reactors use blocks of graphite as a moderator. There are nearly a dozen of these Chernobyl-type reactors remaining in service, principally in Russia, Ukraine, and Lithuania. The remaining types of reactors—gas-cooled reactors and fast neutron reactors—use different fuel types; the fuel consists of bundles of small-diameter stainless steel tubes filled with pellets of highly enriched uranium (HEU) in the form of  $UO_2$  or MOX. Fast reactor prototypes are in operation in Russia, India, and Japan, and were formerly used in the United States, France, and England.

**Interim storage and cooling.** Spent fuel is removed from the reactor after 3–5 years of energy output. Initially, it is stored in water-filled pools adjacent to the reactor. After about 10 years, the radioactivity and decay heat output become small enough that air-cooled storage in concrete-shielded cylinders is used, usually at or near the reactor site.

**Reprocessing of spent fuel.** Reprocessing of spent fuel from commercial reactors is done in large radiation-shielded facilities. Operation is entirely by remote control, using robotic devices for maintenance. Large-scale reprocessing facilities for commercial fuels have been built and operated in at least eight countries (Belgium, China, England, France, Germany, India, Japan, and Russia), and were formerly operated in the United States.

Spent fuel is chopped up and then dissolved in nitric acid. The uranium and plutonium are separated from each other and then from the radioactive fission products by using a liquid-liquid solvent extraction process called PUREX. The residues from reprocessing constitute medium- and high-level wastes. Another method, called pyroprocessing, has also been developed. See NUCLEAR FUELS REPROCESSING; SOLVENT EXTRACTION.

The United States has reprocessed large tonnages of uranium from production reactors to make its military stockpile of over 50 tons of plutonium. However, the United States policy since the Carter administration has been to discourage reprocessing internationally (and to set a good example domestically) in the hope that this will reduce the likelihood of proliferation of nuclear weapons. Reprocessing to recycle plutonium and uranium for power production also became uneconomic in the 1980s due to the unexpectedly high abundance and low costs of uranium ores. This development has also reduced the interest in fast breeder reactors that can produce additional fuel material. The sharp increases that have occurred since then in the costs of energy produced from fossil fuels (coal, oil, and gas), and the high costs of most alternative energy sources, now make nuclear power highly competitive economically and environmentally more attractive. (It is the only large-scale

energy source that produces no emission of carbon dioxide or the other greenhouse gases that cause global warming.) Reprocessing and recycle have become correspondingly more practical economically for some countries, but they are unlikely to occur in the United States due to public concerns and continued adherence to fossil fuels for most energy production.

**Recycle.** After reprocessing to recover the plutonium, the latter is mixed with uranium and converted to mixed oxides. The MOX, processed to the form of high-density ceramic pellets, is loaded into zirconium tubes and fuel bundles. Recycle in thermal reactors is used to increase the energy extractable from a ton of ore by over 50%. Recycle in fast breeder reactor can increase the extractable energy by more than 500%.

Belgium, China, France, Germany, Japan, and Russia, with large and growing nuclear power capacities, use recycled plutonium. Disposal of highly enriched uranium from nuclear weapons is beginning to be undertaken by blending with natural or depleted uranium to make the 3–5% low-enrichment fuel. Similarly, MOX fuel capability can be used to dispose of plutonium stockpiled for nuclear weapons. This option is being planned in Europe and Russia, and is beginning to be considered in the United States. See PLUTONIUM.

**Once-through fuel cycle.** All United States commercial reactors use a once-through cycle, meaning without recycle. The spent fuel is stored on-site at each reactor prior to planned eventual disposal in a government-owned repository. Several long-term research efforts have been under way to evaluate the geology and hydrology of Yucca Mountain, Nevada, and other candidate sites for a repository. All sites are subject to public opposition due to the various environmental factors that must be taken into account.

**Fuel performance improvements.** Improvements in the fuel cycle have centered on increasing the amount of energy produced from about 4–7 megawatt-days of thermal energy (MWDt) per pound of uranium (10–15 MWDt/kg) to over 18 MWDt/lb (40 MWDt/kg). Development work is reaching for 27–45 MWDt/lb (60–100 MWDt/kg). This corresponds to an increase from about 40 megawatt-hours to over 300 megawatt-hours of electricity produced per pound of fuel (or from 88 MWh/kg to 660 MWh/kg). This has made it possible to produce power for 18–24 months between extended shutdowns for refueling. The longer periods between reloadings increase capacity factors and reduce production costs per unit energy generated.

**Nuclear waste disposal options.** The principal options for the disposal of radioactive wastes from civilian nuclear energy are the following:

1. Storing the fission product waste as a vitrified glass or ceramic after reprocessing, and removing the uranium and plutonium for recycle. The vitrified cylinder is encapsulated in a corrosion-resistant metal jacket (often double-walled) and emplaced in a

deep underground repository. The repository is chosen for its geology that demonstrates it has been stable for tens of millions of years and appears likely to remain stable for at least 10,000 years more.

2. Storing vitrified waste as in option 1 above but with added processing to sequester certain long-lived fission products and actinide elements for separate disposal or for transmutation. Research studies are exploring the feasibility of separating and transmuted several of the long-lived radioelements to more innocuous forms by exposure in a reactor or a large accelerator. The resulting products are either stable or have short half-lives, simplifying the disposal requirements for the remaining materials.

3. Vitrifying and emplacing the spent fuel as above but without recycle or removing the uranium, plutonium, fission products, or actinides.

4. Mechanically compacting bundles of spent fuel rods, encapsulating them in specially designed jackets, and emplacing them as above.

Many countries are implementing option 1 and are researching option 2. The United States is following option 4 for commercial spent fuel. The United States is using option 1 above for the waste remaining from the former production reactors and Navy ship propulsion reactors. There has been a small-scale use of option 4 by Sweden. See RADIOACTIVE WASTE MANAGEMENT.

**Costs.** The sum of the costs of the eight steps in the complete nuclear fuel cycle can be expressed as X dollars (U.S.) per pound of fuel (or in comparable units of other nations). Where recycle is used, there is a credit for the fuel value of the recovered uranium and plutonium. A pound of fuel produces heat energy, which is converted to electricity, producing Y kilowatt-hours per pound. The nuclear fuel cycle cost is then X/Y in dollars per kilowatt-hour. For most countries, including the United States, the sum of these costs is lower than the fuel costs for burning oil or gas, and about half that of burning coal. However, the plant costs and labor costs for nuclear generating units are generally higher than for fossil fuel generating units. See COAL.

**Issues.** Various issues revolve around the type of nuclear fuel cycle chosen. For instance, the question is still being argued whether "burning" weapons materials in recycle reactors is more or less subject to diversion (that is, falling into unauthorized hands) than storing and burying these materials. In either case, the degree of security is highly dependent on the effectiveness of national and international regimes of inspection, accountancy, and control. The International Atomic Energy Agency in Vienna is one such agency for control of recycled material from civilian reactors.

Another issue involves the composition of radioactive wastes and its impact on repository design. The nuclear fuel cycles that include reprocessing make it possible to separate out the most troublesome long-lived radioactive fission products and the minor actinide elements that continue to produce heat for centuries. The remaining waste decays to radiation levels comparable to natural ore bodies in about 1000

years. The separated actinides are a small volume fraction of the total waste and can be sequestered for transmutation to more stable products by irradiation in reactors or by using particle accelerators. The shorter time for the resulting wastes to decay away simplifies the design, management, and costs of the repository. See ACTINIDE ELEMENTS.

Edwin L. Zebroski

**Bibliography.** L. C. Hebel et al., Report of the Study Group on Nuclear Fuel Cycles and Waste Management, *Rev. Mod. Phys.*, 50(1)Part II:S114-S117, 1978; R. Lee et al., *External Costs and Benefits of Fuel Cycles*, ORNL and Resources for the Future, McGraw-Hill Utility Data Institute, Washington, DC, 1995; C. McCombie, The hazards presented by radioactive wastes and the safety levels to be provided by repositories, *Nuc. Eng. Design*, 176(1-2):43-50, 1997; T. H. Pigford and P. L. Chambré, *Scientific Basis for Nuclear Waste Management: Proceedings of the Material Research Society*, Elsevier, 1983.

## Nuclear fuels

Materials whose ability to release energy derives from specific properties of the atom's nucleus. In general, energy can be released by combining two light nuclei to form a heavier one, a process called nuclear fusion; by splitting a heavy nucleus into two fragments of intermediate mass, a process called nuclear fission; or by spontaneous nuclear decay processes, which are generically referred to as radioactivity. Although the fusion process may significantly contribute to the world's energy production in future centuries and although the production of limited amounts of energy by radioactive decay is a well-established technology for specific applications, the only significant industrial use of nuclear fuel so far utilizes fission. Therefore, the term nuclear fuels generally designates nuclear fission fuels only. See NUCLEAR BATTERY; NUCLEAR FISSION; NUCLEAR FUSION; RADIOACTIVITY AND RADIATION APPLICATIONS.

**Nuclear characteristics.** Large releases of energy through a fission or a fusion reaction are possible because the stability of the nucleus is a function of its size. The binding energy per nucleon provides a measure of the nucleus stability. By selectively combining light nuclei together by a fusion reaction or by fragmenting heavy nuclei by a fission reaction, nuclei with higher binding energies per nucleon can be formed. The result of these two processes is a release of energy, the magnitude of which is several orders greater than that obtained by the combustion of carbon compounds (such as wood, coal, or oil). For example, the fissioning of one nucleus of uranium releases as much energy as the oxidation of approximately  $5 \times 10^7$  atoms of carbon. See NUCLEAR BINDING ENERGY.

Many heavy elements can be made to fission by bombardment with high-energy particles. However, only neutrons can provide a self-sustaining nuclear fission reaction. Upon capture of a neutron by a heavy nucleus, the latter may become unstable and

split into two fragments of intermediate mass. This fragmentation is generally accompanied by the emission of one or several neutrons, which can then induce new fissions. Only a few long-lived nuclides have been found to have a high probability of fission:  $^{233}\text{U}$ ,  $^{235}\text{U}$ , and  $^{239}\text{Pu}$ . Of these nuclides, only  $^{235}\text{U}$  occurs in nature as 1 part in 140 of natural uranium, the remainder being mostly  $^{238}\text{U}$ . The other nuclides must be produced artificially:  $^{233}\text{U}$  from  $^{232}\text{Th}$ , and  $^{239}\text{Pu}$  from  $^{238}\text{U}$ . The nuclides  $^{233}\text{U}$ ,  $^{235}\text{U}$ , and  $^{239}\text{Pu}$  are called fissile materials since they undergo fission with either slow or fast neutrons, while  $^{232}\text{Th}$  and  $^{238}\text{U}$  are called fertile materials. The latter, however, can also undergo the fission process at low yields with energetic neutrons; therefore, they are also referred to as being fissionable.

The term nuclear fuel applies not only to the fissile materials, but often to the mixtures of fissile and fertile materials as well. Using a mixture of fissile and fertile materials in a reactor allows capture of excess neutrons by the fertile nuclides to form fissile nuclides. Depending on the efficiency of production of fissile elements, the process is called conversion or breeding. Breeding is an extreme case of conversion corresponding to a production of fissile material at least equal to its consumption.

**Physical characteristics.** The type and design of nuclear fuels vary with the reactor design. For a discussion of the steps by which the nuclear fuel is prepared for use in a reactor and eventually disposed of, see NUCLEAR FUEL CYCLE

Most modern reactors that are used to produce electricity are light-water-moderated and -cooled (light-water reactors or LWRs); the fuel is uranium isotope 235 ( $^{235}\text{U}$ ) diluted in  $^{238}\text{U}$ , and it is in the form of uranium dioxide ( $\text{UO}_2$ ). Several prototype power reactors use liquid sodium as a coolant (for example, liquid-metal fast breeder reactors or LMFBRs), and the fuel is usually  $^{235}\text{U}$  and  $^{239}\text{Pu}$  in the oxide form. Gas-cooled reactors usually use oxides or carbides of  $^{232}\text{Th}$  and  $^{235}\text{U}$  or  $^{233}\text{U}$  in the shape of small spheres.

The designs of nuclear fuels have covered a wide range of chemical forms, sizes, shapes, containment types, and detailed characteristics. The reactor design, neutron-physics properties, economics, safety and licensing requirements, and proven reliability affect fuel design.

The chemical form is determined by the effects of the nonfissile atoms on the neutron flux and on the physical characteristics, which in turn influence the reliability and licensability of the fuel design. The only influence that the chemical form has on the energy production is secondary, that is, by the nonfissile atoms absorbing or moderating neutrons. This insensitivity of the nuclear fission process to chemical form allows flexibility in fuel types. The primary forms used have been metals, oxides, carbides, and nitrides. Almost all commercial reactors use oxides ( $\text{UO}_2$ ). Fissile material also has been alloyed with or dispersed in solid materials. Other forms that have been tried include fissile material in a molten salt, water, acid, or liquid metal. In general, these solutions are too corrosive to be practical.

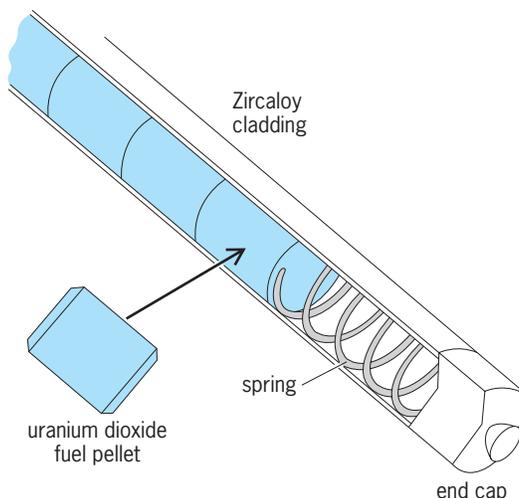


Fig. 1. Fuel rod design for a light-water reactor.

The solid fuels have been designed with many shapes, including rods, spheres, plates, and curved plates. The fissile material is contained in a structural material that is intended to contain fission products and provide structural stability. The most common fuel design uses small cylindrical pellets of uranium dioxide ( $\text{UO}_2$ ) sealed inside a long tube (Figs. 1 and 2). These designs and their properties are discussed in more detail below.

The design of nuclear fuels must provide physical properties that ensure safe operation, reliable

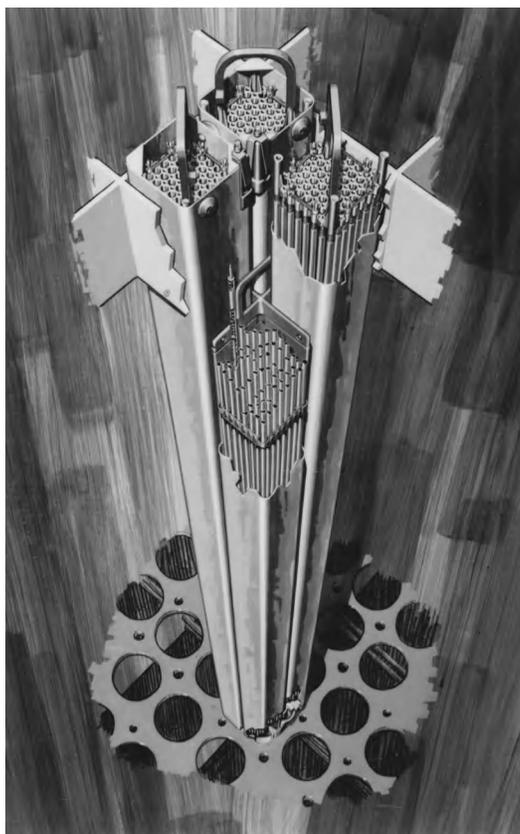


Fig. 2. Boiling-water reactor fuel assembly. (General Electric Co.)

performance, and economical power production. For light-water reactors, safe operation requires that the sealed tube that contains the fissile material remain intact during normal operation and most abnormal events, which are usually associated with a power increase or a loss of cooling. Reliable performance requires that the fuel remain dimensionally stable and that the fuel container not leak during normal operation. Economical power production, that is, low fuel-cycle cost, requires that neutrons be used efficiently to produce fission and to convert fertile nuclides to fissile nuclides; that a high fraction of fissile isotopes be burned; and that waste disposal costs be minimized.

**Reliability.** During normal operation, fuel performance is characterized by the integrity of the fuel container and the stability of component dimensions. Fuel containers that leak will release radioactive fission products to the coolant. Strict limits on coolant radioactivity, which protect plant personnel and the public, prevent operation with more than approximately one leaking rod in 2000 for light-water reactors (plants normally operate with less than one leaking rod in 10,000). Excessive leakage results in costly reactor shut-downs for locating and replacing defective fuel elements.

Dimensional stability is important to ensure proper coolant flow and power distribution throughout the core. In light-water reactors the primary concerns, which have been eliminated with design improvements, have been with fuel-pellet swelling and resultant interaction with the cladding, fuel-rod elongation and resultant interaction with the top and bottom plates, and fuel-assembly distortion. As fuel performance has improved, fuel burnup has been increased to reduce costs. This increase in residence results in corrosion of the cladding tube being the limiting performance attribute. In the 1990s many nuclear fuel vendors developed zirconium alloys with improved corrosion resistance to support increased fuel burnup.

**Economic power production.** Fuel-cycle costs, which are a significant but not the greatest contributor to electrical-energy production costs, are mostly determined by raw material (yellowcake) purchase costs, enrichment costs, interest rates, energy extracted, and waste storage and disposal costs. Fuel design and enrichment requirements are influenced by the desired energy extraction and neutron economics (efficient use of the fissile atoms). In light-water reactors, fuel designers try to maximize the amount of water and minimize the amount of structural material (principally a zirconium alloy called Zircaloy) in the core. Increased water provides better moderation of neutrons, and reduction of structural material decreases the number of neutrons absorbed by parasitic atoms (neither fissile nor fertile). Most of the cost of electricity production by nuclear power is due to capital and operating expenditures, not fuel-cycle costs.

**Properties.** The important properties of nuclear fuel are thermal, mechanical, chemical, and nuclear.

Thermal properties influence component stored energy and temperatures, which affect phenomena that are temperature sensitive. Mechanical properties, such as the creep rate and yield strength of structural components, determine dimensional stability and influence reliability. Chemical properties determine reaction rates among different types of atoms and influence reliability, for example, by assisting in stress-corrosion cracking. Irradiation changes fuel properties by changing the chemical composition and the atomic structure, and these changes must be reflected in performance and licensing calculations.

The basic thermal properties of the fuels, such as thermal conductivity, are only modestly changed by the irradiation and the resultant damage. The conversion of fertile atoms to fissile atoms and of fissile atoms to the fission-product atoms changes the composition of the materials, but the changes in thermal properties are modest. *See* THERMAL CONDUCTION IN SOLIDS.

The situation is quite different for gases. For light-water reactor fuel designs, a gas, such as helium, is used to transfer heat from the fuel to the structural container. Fission gases, such as xenon and krypton, are poor conductors of heat and degrade the thermal conductivity of the heat-transfer gas if they escape from the fuel. Heat transfer also is affected by cracking and slight movement of fuel, by microstructural changes in the fuel, and by bonding of the fuel and container, which can occur by chemical diffusion. *See* CONDUCTION (HEAT).

The mechanical properties of the core structural components and the fuel change with irradiation. High-energy particles, for example, neutrons, interact with atoms in solids and knock significant numbers of atoms out of their normal positions. This is called radiation damage. The effects of this damage are different during and after the radiation and different for time-dependent than for time-independent properties. *See* RADIATION DAMAGE TO MATERIALS.

The extra free atoms that are knocked out of their normal position often make time-dependent processes, such as creep, occur at faster rates. Time-independent phenomena do not have time during irradiation to notice the softening effect of the knocked-out atoms; they just see the hardening effect of the damage. However, after irradiation, the damage left behind makes it more difficult for deformation to occur, and the material properties reflect those of a harder material for both time-dependent and time-independent properties.

The fission process also causes changes in the chemical composition of materials. For example, fuel that started out as essentially all uranium dioxide slowly accumulates plutonium and many types of fission atoms. Some of these such as iodine are corrosive, some are radioactive with long half-lives, and some are gases and are not stable in solid fuel. Fuels are usually designed to retain the products in the fuel so that they do not corrode or stress the fuel container. Structural materials are selected to have few and benign interactions during radiation, thus

producing few undesirable isotopes. However, some radioactive atoms are produced, for example,  $^{60}\text{Co}$ , in stellite and steel components. In general, the change in composition is not sufficient to significantly change the material properties. Most changes in properties are due to irradiation hardening.

The fission products have some influence on neutron economics. Some of the products capture a significant number of neutrons. As their concentration increases, they will eventually steal a significant amount of reactivity. Reprocessing of spent fuels has to remove isotopes with high neutron cross section as well as recover fissile material, thus obtaining high-reactivity fuel from spent fuel. *See* NUCLEAR FUELS REPROCESSING.

**Breeder reactor fuels.** The capture of neutrons by abundant, fertile nuclides, such as  $^{238}\text{U}$ , leads to conversion of those atoms to fissile nuclides, such as  $^{239}\text{Pu}$ . If nuclear power eventually becomes the major source of electric power for many industrialized countries, there may be a need to build reactors that breed fuel by converting more fertile material to fissile material than is consumed during operation, thus providing fuel for themselves and for nonbreeding light water reactors.

High conversion ratios are obtained by designing the reactor to have a flux spectrum that has a high density of high-energy (fast) neutrons (that is, a hard flux spectrum). This increases the number of neutrons released by the fission process; a greater amount of excess neutrons is then available for breeding purposes. The predominant current design uses a liquid metal, usually sodium, instead of water to cool the fuel. This removes the primary material (hydrogen in water) that slows down (thermalizes) the neutrons. Since fast neutrons are not as efficient as thermal neutrons at inducing fission, the enrichment (concentration) of atoms (usually  $^{235}\text{U}$  and  $^{239}\text{Pu}$ ) in the fuel is increased to maintain the fissile chain reaction.

The main differences in the design of the fuel elements from those of light-water reactors are that the cladding is usually stainless steel rather than an alloy of zirconium, and fuel elements are grouped together into assemblies that are placed inside hexagonal stainless steel cans. The fuel pellets are uranium dioxide (approximately 80%) and plutonium dioxide (approximately 20%), although the nitrides and carbides of uranium have been tested and the Experimental Breeder Reactor II in Idaho operated with highly enriched metal fuel. The primary advantage of these alternative designs is their higher fuel thermal conductivity compared to that of uranium and plutonium dioxides. However, other characteristics of these fuels, particularly changes in crystallographic structure at low temperature, low melting temperature, fission-gas retention, compatibility with cladding, or swelling under irradiation, are either inferior to those of the oxide or are not reliably established.

The stainless steel cladding, usually alloy 316 or 304, is stronger than zirconium alloys at operating temperatures, although stainless steels have a lower

melting temperature which is a disadvantage for transients. *See* ALLOY; STAINLESS STEEL.

**Gas-cooled reactor fuels.** The first gas-cooled reactor (GCR) power station was built at Calder Hall in the United Kingdom and began operation in 1956. Since then, a number of gas-cooled reactor plants have been constructed and operated. The gas-cooled reactors have graphite cores comprising multisided prisms stacked into a horizontal cylinder, about 50 ft (15 m) in diameter and 33 ft (10 m) long. The coolant is carbon dioxide ( $\text{CO}_2$ ). The fuel is natural uranium that is metal clad with a thin cylindrical tube of a magnesium alloy called Magnox. In order to achieve good heat transfer, the Magnox cans are provided with cooling fins.

The Magnox-type gas-cooled reactors are still in operation, but have been superseded by the advanced gas-cooled reactor (AGR). These reactors are similar to the Magnox type in general design. However, the Magnox fuel elements have been replaced by slightly enriched uranium dioxide clad in stainless steel, permitting higher surface temperatures and greater thermal-cycle efficiency.

To increase efficiency further, coolant temperatures in the  $1560\text{--}1830^\circ\text{F}$  ( $850\text{--}1000^\circ\text{C}$ ) range are implemented in the high-temperature gas-cooled reactor (HTGR). Helium is used in place of carbon dioxide as the coolant due to the instability of carbon dioxide at high temperatures. The development of the high-temperature gas-cooled reactor has proceeded in two directions: the pebble bed concept and the prismatic core design. In the German pebble bed design (thorium high-temperature reactor or THTR), the fuel elements consist of 2.4-in.-diameter (60-mm) spheres made of a carbon outer zone and an inner 2-in.-diameter (50-mm) region containing a mixture of fertile and fissile particles uniformly dispersed in a graphite matrix. Each ball contains 0.04 oz (1.1 g) of  $^{235}\text{UO}_2$  and 0.4 oz (11 g) of  $^{232}\text{ThO}_2$ . The core consists of about 700,000 such balls in a random arrangement. In the prismatic design, developed in the United States, the core consists of stacks of hexagonal prisms of graphite. The prismatic fuel elements are pierced with channels for cooling and for fuel rods that are made by pressing graphite powder and fuel particles together (**Fig. 3**). Each rod contains a mixture of fissile and fertile particles. The particles consist of microspheres made of fully enriched uranium carbide ( $\text{UC}_2$ ), a mixture of uranium carbide and thorium carbide ( $\text{ThC}$ ), and thorium carbide or thorium oxide ( $\text{ThO}_2$ ), coated with several concentric layers of pyrolytic carbon. The two coated-particle types in most common use are the two-layer Biso (for buffer isotropic) coating and the four-layer Triso coating with its interlayer of silicon carbide ( $\text{SiC}$ ) between two layers of high-density isotropic pyrolytic carbon. Both Biso and Triso particles are capable of essentially complete retention of gaseous fission products.

Improved coatings are needed to increase the strength and integrity of the fuel particle, reduce the diffusional release of fission products, and improve

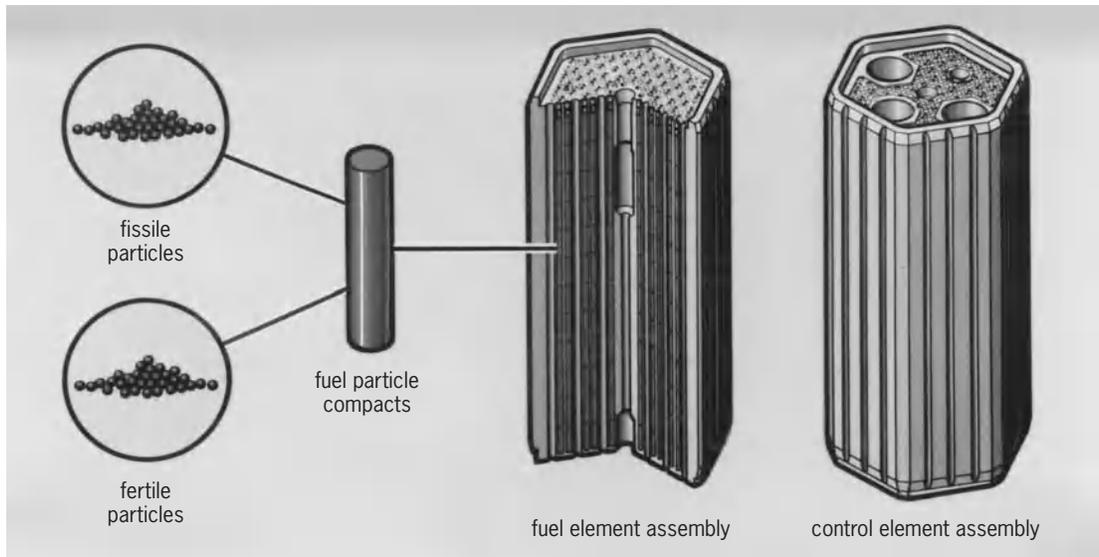


Fig. 3. Fuel element components in a high-temperature gas-cooled reactor with the prismatic core design used in the United States. (T. D. Gulden, G. A. Technologies, Inc.)

the overall chemical stability and temperature capability of coated particles.

**Fuel designs for less common reactors.** A large number of different types of reactors have been designed, constructed, and operated. However, their total capacity, including high-temperature gas-cooled reactors and liquid-metal fast-breeder reactors, has been less than 1% of the capacity of light-water reactors in the United States. Some of them were used for purposes other than the large-scale production of electricity or heat. Others are considered obsolete or have little chance for large-scale power production. Practically all concepts rely on a heterogeneous aggregation of fuel and coolant similar to those described above. A notable exception is provided by the homogeneous reactor experiments (HRE). Instead of using a solid matrix, the core consisted of a  $^{235}\text{U}$  sulfate solution in deuterium oxide ( $\text{D}_2\text{O}$ ). The molten salt reactor (MSR) was a similar project in which uranium tetrafluoride ( $\text{UF}_4$ ) and thorium tetrafluoride ( $\text{ThF}_4$ ) were kept in a eutectic melt of beryllium fluoride ( $\text{BeF}_2$ ), zirconium tetrafluoride ( $\text{ZrF}_4$ ), and sodium fluoride ( $\text{NaF}$ ) or lithium-7 fluoride ( $^7\text{LiF}$ ).

**Safety and licensing.** Nuclear fuel rods generate large amounts of energy and produce highly radioactive nuclides inside the fuel. The cladding material that contains the fuel is the first containment barrier of these radionuclides, and its integrity must be ensured. The design of nuclear fuel and reactor systems, and the operational procedures must provide this assurance. The U.S. Nuclear Regulatory Commission (NRC) represents the public to ensure that its safety is protected by these designs and procedures.

The objective of safety analyses of nuclear fuel designs is to ensure that the cladding material will retain its integrity during normal operation and during unusual events such as temperature or power increases

above normal levels. Since essentially all of the power reactors in the United States are light-water reactors with Zircaloy-clad or ZIRLO<sup>3</sup>-clad uranium dioxide fuel, the following discussion is based on safety consideration of these designs.

High temperatures can be caused by either a prolonged loss of the reactor cooling water (loss-of-cooling accident or LOCA), an increase in power, or a reduction in flow of cooling water. High temperatures (above  $1200^\circ\text{F}$  or  $650^\circ\text{C}$ ) can result in rapid cladding oxidation and hydriding (from the reaction with the cooling water), chemical interactions between the fuel and the cladding, and, at very high temperatures, melting of the fuel or cladding. Each one of these reactions would compromise the cladding integrity immediately or during further service, and must be avoided.

High power results from uncontrolled increases in the reactivity of the reactor core, for example, from ejection of a neutron-absorbing control material from the core or from the sudden injection of cold water into the core. Safety analyses are performed to ensure that, during these events, the surface of the fuel rods will continue to be cooled by a layer of cooling water and that the uranium dioxide fuel temperatures will not become excessive. Nuclear reactors are designed to have large margins of safety against all mechanically plausible means of reactivity increase, and maintenance of these margins of safety is enforced by regulations.

A loss of cooling water can be caused by pipe ruptures or stuck valves. For extreme loss of water by a pipe rupture or loss-of-coolant accident (LOCA), all light-water reactors in the United States are designed to maintain the cladding temperature below  $2200^\circ\text{F}$  ( $1200^\circ\text{C}$ ) and the cladding total oxidation at less than 17% of the metal.

The only major loss of cooling in a United States civilian light-water reactor occurred at the Three

Mile Island (Pennsylvania) Unit 2 in 1979. The loss of coolant occurred through a stuck open valve. The reactor operator, failing to recognize this situation, shut down the automatically activated emergency core cooling system (ECCS) and maintained let-down flow. For several hours the upper portion of the core was uncovered before the real situation eventually was recognized and the emergency core cooling system was activated again, resulting in recirculation of cooling water and termination of the undercooling event. During the uncovering of the upper portion of the core, the fuel rods exceeded the above limits for peak temperature and oxidation. The fuel distorted; significant portions of the structural materials melted and solidified into different shapes and formed a rubble bed on top of the lower portion of the core, and some of the material was deposited in the lower part of the reactor vessel. Despite the severity of the accident and the breakdown of the cladding as a barrier, other reactor containment systems functioned properly. The increase in radiation exposure to persons in the vicinity of the reactor at the time of the accident was within the range of local variation of annual exposure from nature. *See* NUCLEAR POWER; NUCLEAR REACTOR; REACTOR PHYSICS.

David Franklin; Albert Machiels

**Bibliography.** E. R. Bradley and G. D. Sabol (eds.), *Zirconium in the Nuclear Industry*, Eleventh International Symposium, ASTM STP 1295, 1996; C. Cangruly, *Nuclear Fuel Fabrication*, 2d ed., 1989; B. A. Ma, *Nuclear Reactor Materials and Application*, 1983; J. T. A. Roberts, *Structural Materials in Nuclear Power Systems*, 1981; F. H. Wittmann (ed.), *Fuel Elements and Assemblies*, 1987.

## Nuclear fuels reprocessing

Nuclear fuels are reprocessed for military or civilian purposes. In military applications, reprocessing is applied to extract fissile plutonium from fuels that are designed and operated to optimize production of this element. In civilian applications, reprocessing is used to recover valuable uranium and transuranic elements that remain in fuels discharged from electricity-generating nuclear power plants, for subsequent recycle in freshly constituted nuclear fuel. This military-civilian duality has made the development and application of reprocessing technology a sensitive issue worldwide and necessitates stringent international controls on reprocessing operations. It has also stimulated development of alternative processes that do not produce a separated stream of pure plutonium so that the proliferation of nuclear weapons is held in check. *See* NUCLEAR POWER; PLUTONIUM; TRANSURANIUM ELEMENTS; URANIUM.

Nuclear fuel is removed from civilian power reactors due to chemical, physical, and nuclear changes that make it increasingly less efficient for heat generation as its cumulative residence time in the reactor core increases. The fissionable material in the fuel is not totally consumed; however, the buildup of fission product isotopes (with strong neutron-absorbing

properties) tends to decrease the nuclear reactivity of the fuel. *See* NUCLEAR FISSION; NUCLEAR FUELS; NUCLEAR REACTOR.

**Factors.** A typical composition of civilian reactor spent fuel at discharge is 96% uranium, 3% fission products, and 1% transuranic elements (generally as oxides, because most commercial nuclear fuel is in the form of uranium oxide). The annual spent fuel output from a 1.2-gigawatt electric power station totals approximately 33 tons (30 metric tons) of heavy-metal content. This spent fuel can be discarded as waste or reprocessed to recover the uranium and plutonium that it contains (for recycle in fresh fuel elements). The governments of France, the United Kingdom, Russia, and China actively support reprocessing as a means for the management of highly radioactive spent fuel and as a source of fissile material for future nuclear fuel supply. The United States has maintained a national policy forbidding the reprocessing of civilian reactor fuel for plutonium recovery since the mid-1970s. Therefore, the United States is the only one of the five declared nuclear weapons states with complete fuel recycling capabilities that actively opposes commercial fuel reprocessing.

Decisions to reprocess are not made on economic grounds only, making it difficult to evaluate the economic viability of reprocessing in various scenarios. In the ideal case, a number of factors must be considered, including: (1) Cost of uranium/U<sub>3</sub>O<sub>8</sub>; prices are directly related to supply and demand. (2) Cost of enrichment; advanced technologies may reduce costs significantly. (3) Cost of fuel fabrication; plutonium-bearing fuels are more expensive to fabricate due to the toxicity of plutonium and the need for remote manufacturing techniques. (4) Cost of reprocessing; costs vary widely from country to country. (5) Waste disposal cost; high-level waste volumes from modern reprocessing should be significantly less than for the case of direct disposal, and the extraction of actinide elements may simplify the licensing of disposal systems. (6) Fissile content of spent fuel; high-value spent fuels (for example, those that contain large fractions of fissile material) make reprocessing more profitable.

The once-through fuel cycle (that is, direct disposal/no reprocessing) is favored when fuel costs and waste disposal costs are low and reprocessing costs are high. It is not necessarily true that reprocessing will inevitably be more costly than direct disposal, because technology advancements and escalating waste disposal costs can swing the balance in favor of reprocessing. *See* ACTINIDE ELEMENTS; NUCLEAR FUEL CYCLE; RADIOACTIVE WASTE MANAGEMENT.

It is generally accepted that breeder reactors, those that are designed to generate more fuel than is consumed (by absorption of excess neutrons in peripheral blankets composed of fertile isotopes such as <sup>238</sup>U or <sup>232</sup>Th), will not be deployed until the latter part of the twenty-first century. With these reactors, reprocessing will be mandatory to recover the fissile material generated in the blankets. Because

breeder reactors tend to operate with fuel of considerably higher fissile enrichment level than in the case of light-water reactors, the avoided cost of external fuel enrichment also favors reprocessing. In this fuel cycle it is assumed that all fissile material requirements of the fresh fuel are obtained from reprocessing the spent fuel and recycling its fissile content. It is reasonable to expect that the economics of reprocessing will be advantageous to the deployment of breeder reactors when the time comes for their introduction into the energy generation mix. *See* RADIOISOTOPE.

**Methods.** The technology of reprocessing nuclear fuel was created as a result of the Manhattan Project during World War II, with the purpose of plutonium production. Early reprocessing methods were refined over the years, leading to a solvent extraction process known as PUREX (plutonium uranium extraction) that was implemented on a large scale in the United States in late 1954 at the Savannah River site in South Carolina and in early 1956 at the Hanford site in Washington. The PUREX process is an aqueous method that has been implemented by several countries and remains in operation on a commercial basis. The PUREX plant at Hanford was shut in 1992, while the Savannah River plant is still operating for the treatment of residual research reactor fuel inventories. A nonaqueous reprocessing method known as pyroprocessing was developed in the 1990s as an alternative to PUREX. It has not been deployed commercially, but promises greatly decreased costs and reduced waste volumes, with practically no secondary wastes or low-level wastes being generated. It also has the important attribute of an inability to separate pure plutonium from irradiated nuclear fuel. *See* SOLVENT EXTRACTION.

*PUREX reprocessing.* The PUREX process is carried out in large shielded facilities with remote operation capabilities. The PUREX plant at the Hanford site operated with a throughput capacity of 3000 tons (2700 metric tons) of heavy metal per year. Modern plants with somewhat smaller capacity are operating in Sellafield in England and LaHague in France. In the PUREX process, spent fuel is initially dissolved with nitric acid in a large vessel. In the case of aluminum-clad fuel, the dissolution is catalyzed by the addition of mercury, whereas zirconium-alloy fuel is dissolved in a mixture of nitric and hydrofluoric acids; the hydrofluoric acid is added to complex the zirconium and prevent the occurrence of highly energetic reactions. Zirconium alloy-clad fuel can be deacid chemically with ammonium fluoride (ZRFLEX process); alternatively, the spent fuel can be chopped into short segments and the fuel material leached out with nitric acid, leaving the cladding undissolved (chop-leach process). Thorium dissolution requires the use of a catalyst such as fluoride ion. Except for a few fission products that are volatilized during dissolution, all the constituents initially in the fuel are retained in the dissolver solution as soluble nitrate salts or, in the case of minor constituents, as an insoluble sludge. Generally, the solution must be clarified to accommodate its use as feed to the solvent extraction process. *See* THORIUM.

In the solvent extraction steps, purification of the desired products (generally uranium and plutonium, possibly thorium) is achieved by the selective transfer of these constituents into an organic solution that is immiscible with the original aqueous solution. The most commonly used organic solution consists of the extractant tri-*n*-butyl phosphate (TBP) dissolved in purified kerosine or similar materials such as long-chain hydrocarbons (for example, *n*-dodecane). Uranium in aqueous nitrate solutions normally exists only in the +4 (IV) and +6 (VI) oxidation states. Nitric acid dissolution of spent fuel yields U(VI) and nitrogen oxides (NO, NO<sub>2</sub>) as products. The U(VI) is strongly hydrolyzed, and exists as the uranyl (UO<sub>2</sub><sup>2+</sup>) ion. By the same token, plutonium may exist in the Pu(III) or Pu(IV) states, with the (IV) state resulting from nitric acid dissolution. When in the proper oxidation state [that is, (UO<sub>2</sub><sup>2+</sup>), Pu(IV), and Th(IV)], the desired products are readily extractable into tributyl phosphate. Plutonium in the Pu(III) state has a very low extractability into TBP.

The solvent extraction separation is based on this ready extractability of uranyl nitrate and plutonium(IV) nitrate and the relative inextractability of fission products and plutonium(III) nitrate. The first step of the PUREX solvent extraction process thus involves the coextraction of uranium and plutonium into the organic (TBP + kerosine) phase, leaving most of the fission products in the aqueous phase. The organic solvent is then contacted with a clean aqueous solution, and the uranium and plutonium transfer into this clean aqueous phase. The solvent can now be recycled. After reducing the plutonium to the plutonium(III) oxidation state, it is then a matter of contacting the aqueous phase containing plutonium(III) nitrate and uranyl nitrate with clean organic phase; the uranium transfers to the organic phase and the plutonium remains in the aqueous phase. The uranium-organic phase can then be contacted with another clean aqueous phase to transfer the uranyl nitrate into the aqueous phase.

The products of the PUREX process described in rudimentary form here are separated nitrate solutions of uranium and plutonium. These solutions are treated separately to produce uranium and plutonium oxides. The solvent extraction process yields virtually complete recovery of the uranium and plutonium present in the spent fuel, and the uranium and plutonium are of high purity. Decontamination factors (defined as the ratio of the concentration of an impurity element to that of the valued constituent in the feed, divided by the same ratio for these elements in the product) of 10<sup>7</sup> to 10<sup>8</sup> are possible.

Because TBP is susceptible to decomposition to dibutyl phosphate and monobutyl phosphate, both of which have significant deleterious effects on plutonium stripping, care must be taken to maintain the purity of the solvent. Otherwise, additional process steps are necessary to recover plutonium, at the cost of increased volume of liquid waste. Limiting the time of exposure of sensitive solvents to high levels of nitric acid and beta/gamma radiation greatly increases their useful lifetime.

*Pyroprocessing.* Rising public concerns over costs, reactor safety, disposal of nuclear wastes, and the threat of nuclear weapons proliferation have impacted government and electric utility decisions on power plant deployment. Of particular concern is the situation that would prevail should nuclear electric power generation accelerate greatly in the lesser-developed nations. It is generally agreed that significant growth in nuclear capacity would mandate reprocessing of spent nuclear fuel for more efficient utilization of fuel resources and minimization of high-level waste quantities. However, the reprocessing technologies now in place are mid-twentieth-century technology and may prove to be impediments to such growth.

In response to these concerns, an advanced processing method is being developed to address both system economics and public acceptance issues. This method, initially developed to support the recycle of fuel materials in an advanced liquid-metal reactor system, is based on the separation of actinide elements from fission products in a molten salt electrorefining process. In this process, spent fuel in the metallic state (either as metal fuel or as the product of a head-end reduction step) is anodically dissolved in an electrorefining cell operating at a temperature of 932°F (500°C). The actinide elements present in the spent fuel are electrotransported to suitable cathodes, and the cathode deposits are treated to remove entrained electrolyte salt. A simple steel cathode is used to collect a pure uranium deposit; the transuranic elements, having chloride salts that are more stable than uranium chloride, must be collected in a liquid cadmium cathode where they deposit as solid intermetallic compounds (for example, PuCd<sub>6</sub>). The transuranic elements cannot be separated in this process, and the liquid cathode deposit typically contains about 70% (by weight) transuranics and 30% uranium, together with a minor amount (approximately 0.25%) of the rare-earth fission products. The transition-metal and noble-metal fission products remain in the anode baskets together with the cladding hulls, while the halide, pnictide, alkali metal, and alkaline-earth fission products remain in the electrolyte salt. *See* ALKALINE-EARTH METALS; ELECTROLYTE; HALIDE; RARE-EARTH ELEMENTS.

The process thus yields four distinct product streams: (1) a pure uranium product that is free of fission products and free of transuranic contamination; (2) an actinide product that contains all of the transuranic elements (such as Np, Pu, Am, and Cm) and some uranium with a trace amount of rare-earths; (3) a metal waste stream emanating from the anode basket residue; and (4) a ceramic waste stream containing fission products extracted from the electrolyte salt. The uranium product can be stored, recycled, or converted to the oxide for disposal as a low-level waste. The transuranic product is suitable for recycle as a fast reactor fuel material; independent studies have shown that this product does not represent a proliferation threat, because the plutonium therein is sufficiently contaminated that the material is not usable to create weapons. In a nuclear growth scenario, the transuranic product could be stored until required as fast reactors are introduced

into the mix of new reactor construction. In the absence of such growth, reprocessing for recovery of transuranic elements is inadvisable, and this product would more appropriately be converted to a disposable waste form by immobilizing the transuranics in glass or zeolite. Alternatively, the concentrated transuranic elements could be transmuted in an accelerator-driven system. *See* AMERICIUM; CURIUM; NEPTUNIUM.

One of the principal advantages of the pyroprocess is that it is expected to have a very low cost. This is a result of the simplicity of the process and the very compact nature of the equipment required. For example, a prototype electrorefiner with a footprint of little more than 11 ft<sup>2</sup> (1 m<sup>2</sup>) affords a throughput of 772 lb (350 kg) of heavy metal per day. The savings in the cost of the shielded facility to house the pyroprocess, compared to conventional processes, can be significant. The batchwise nature of the pyroprocess makes it economically attractive at small plant sizes. In fact, the pyroprocess is relatively insensitive to plant scale, and large, centralized plants are not necessary for economic viability, contrary to the situation that prevails with the PUREX process. The ability to operate small-scale plants at limited costs allows reactor plants and fuel reprocessing facilities to be located within proximity of each other (reducing transportation and risk of proliferation).

*Other applications.* Both the PUREX process and the pyroprocess can be used in a waste management role in support of a once-through nuclear fuel cycle if the economics of this application are favorable. The PUREX process can be operated with a low decontamination factor for plutonium by eliminating the second extraction step. The pyroprocess can be operated without the liquid cadmium cathode, placing the transuranic elements in the salt waste stream that leads to a glass-ceramic waste form. Both systems are effective in placing the fission products and actinide elements present in spent nuclear fuel into more durable waste forms that can be safely disposed in a high-level waste repository. James J. Laidler

*Bibliography.* M. Benedict, T. H. Pigford, and W. H. Levi, *Nuclear Chemical Engineering*, 2d ed., 1981; W. H. Hannum (ed.), *The Technology of the Integral Fast Reactor and Its Associated Fuel Cycle*, *Prog. Nucl. Energy*, vol. 31:1-217, 1997; J. T. Long, *Engineering for Nuclear Fuel Reprocessing*, 1978; W. W. Schulz, J. D. Navratil, and L. L. Burger, *Science and Technology of Tributyl Phosphate: Applications of Tributyl Phosphate in Nuclear Fuel Reprocessing*, 1989.

## Nuclear fusion

One of the primary nuclear reactions, the name usually designating an energy-releasing rearrangement collision which can occur between various isotopes of low atomic number. *See* NUCLEAR REACTION.

Interest in the nuclear fusion reaction arises from the expectation that it may someday be used to produce useful power, from its role in energy generation in stars, and from its use in the fusion bomb. Since a

primary fusion fuel, deuterium, occurs naturally and is therefore obtainable in virtually inexhaustible supply (by separation of heavy hydrogen from water, 1 atom of deuterium occurring per 6500 atoms of hydrogen), solution of the fusion power problem would permanently solve the problem of the present rapid depletion of chemically valuable fossil fuels. In power production the lack of radioactive waste products from the fusion reaction is another argument in its favor as opposed to the fission of uranium. See HYDROGEN BOMB; STELLAR EVOLUTION.

In a nuclear fusion reaction the close collision of two energy-rich nuclei results in a mutual rearrangement of their nucleons (protons and neutrons) to produce two or more reaction products, together with a release of energy. The energy usually appears in the form of kinetic energy of the reaction products, although when energetically allowed, part may be taken up as energy of an excited state of a product nucleus. In contrast to neutron-produced nuclear reactions, colliding nuclei, because they are positively charged, require a substantial initial relative kinetic energy to overcome their mutual electrostatic repulsion so that reaction can occur. This required relative energy increases with the nuclear charge  $Z$ , so that reactions between low- $Z$  nuclei are the easiest to produce. The best known of these are the reactions between the heavy isotopes of hydrogen, deuterium (D) and tritium (T). See DEUTERIUM; TRITIUM.

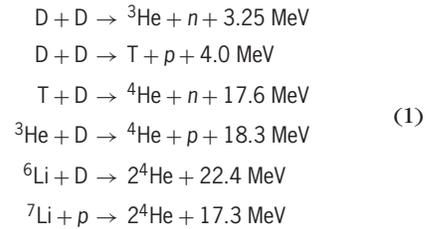
Fusion reactions were discovered in the 1920s when low- $Z$  elements were used as targets and bombarded by beams of energetic protons or deuterons. But the nuclear energy released in such bombardments is always microscopic compared with the energy of the impinging beam. This is because most of the energy of the beam particle is dissipated uselessly by ionization and interparticle collisions in the target; only a small fraction of the impinging particles actually produce reactions.

Nuclear fusion reactions can be self-sustaining, however, if they are carried out at a very high temperature. That is to say, if the fusion fuel exists in the form of a very hot ionized gas of stripped nuclei and free electrons, termed a plasma, the agitation energy of the nuclei can overcome their mutual repulsion, causing reactions to occur. This is the mechanism of energy generation in the stars and in the fusion bomb. It is also the method envisaged for the controlled generation of fusion energy. See PLASMA (PHYSICS).

### Properties of Fusion Reactions

The cross sections (effective collisional areas) for many of the simple nuclear fusion reactions have been measured with high precision. It is found that the cross sections generally show broad maxima as a function of energy and have peak values in the general range of 0.01 barn (1 barn =  $10^{-28}$  m<sup>2</sup>) to a maximum value of 5 barns, for the deuterium-tritium reaction. The energy releases of these reactions can be readily calculated from the mass differences between the initial and final nuclei or determined by direct measurement.

**Simple reactions.** Some of the important simple fusion reactions, their reaction products, and their energy releases are given by reactions (1).



If it is remembered that the energy release in the chemical reaction in which hydrogen and oxygen combine to produce a water molecule is about 1 eV per reaction, it will be seen that, gram for gram, fusion fuel releases more than  $10^6$  times as much energy as typical chemical fuels.

The two alternative D-D reactions listed occur with about equal probability for the same relative particle energies. The heavy reaction products, tritium and helium-3, may also react, with the release of a large amount of energy. Thus it is possible to visualize a reaction chain in which six deuterons are converted to two helium-4 nuclei, two protons, and two neutrons, with an overall energy release of 43 MeV—about  $10^5$  kWh of energy per gram of deuterium. This energy release is several times that released per gram in the fission of uranium, and several million times that released per gram by the combustion of gasoline.

**Cross sections.** Figure 1 shows the measured values of cross sections as a function of bombarding

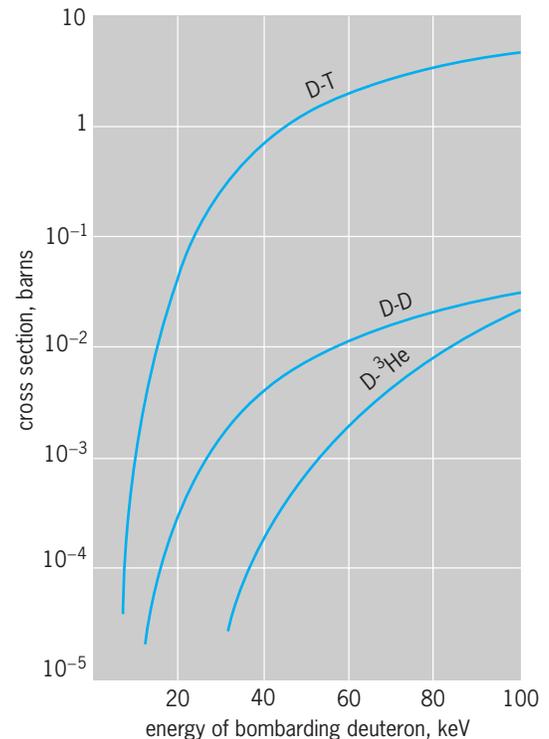


Fig. 1. Cross sections versus bombarding energy for three simple fusion reactions. (After R. F. Post, *Fusion power*, *Sci. Amer.*, 197(6):73-84, December 1957)

energy up to 100 keV for the total D-D reaction (both D-D, $n$  and D-D, $p$ ), the D-T reaction, and the D- $^3\text{He}$  reaction. The most striking characteristic of these curves is their extremely rapid falloff with energy as bombarding energies drop to a few kilovolts. This effect arises from the mutual electrostatic repulsion of the nuclei, which prevents them from approaching closely if their relative energy is small. See NUCLEAR STRUCTURE.

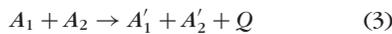
The fact that reactions can occur at all at these energies is attributable to the finite range of nuclear interaction forces. In effect, the boundary of the nucleus is not precisely defined by its classical diameter. The role of quantum-mechanical effects in nuclear fusion reactions has been treated by G. Gamow and others. It is predicted that the cross sections should obey an exponential law at low energies. This is well borne out in energy regions reasonably far removed from resonances (for example, below about 30 keV for the D-T reaction). Over a wide energy range at low energies, the data for the D-D reaction can be accurately fitted by a Gamow curve, the result for the cross section being given by Eq. (2), where the

$$\sigma_{\text{D-D}} = \frac{288}{W} \exp(-45.8/\sqrt{W}) \times 10^{-28} \text{ m}^2 \quad (2)$$

bombarding energy  $W$  is in kiloelectronvolts.

The extreme energy dependence of this expression can be appreciated by the fact that between 1 and 10 keV the predicted cross section varies by about 13 powers of 10, that is, from  $2 \times 10^{-46}$  to  $1.5 \times 10^{-33} \text{ m}^2$ .

**Energy division.** The kinematics of the fusion reaction requires that two or more reaction products result. This is because both mass energy and momentum balance must be preserved. When there are only two reaction products (which is the case in all of the important reactions), the division of energy between the reaction products is uniquely determined, the lion's share always going to the lighter particle. The energy division (disregarding the initial bombarding energy) is as in reaction (3). If reaction (3) holds,



with the  $A$ 's representing the atomic masses of the particles and  $Q$  the total energy released, then Eqs. (4) are valid, where  $W(A'_1)$  and  $W(A'_2)$  are the kinetic

$$\begin{aligned} W(A'_1) + W(A'_2) &= Q \\ W(A'_1) &= Q \left( \frac{A'_2}{A'_1 + A'_2} \right) \\ W(A'_2) &= Q \left( \frac{A'_1}{A'_1 + A'_2} \right) \end{aligned} \quad (4)$$

energies of the reaction products.

**Reaction rates.** When nuclear fusion reactions occur in a high-temperature plasma, the reaction rate per unit volume depends on the particle density  $n$  of the reacting fuel particles and on an average of their mutual reaction cross sections  $\sigma$  and relative velocity  $v$  over the particle velocity distributions. See THERMONUCLEAR REACTION.

For dissimilar reacting nuclei (such as D and T), the reaction rate is given by Eq. (5).

$$R_{12} = n_1 n_2 \langle \sigma v \rangle_{12} \text{ reactions}/(\text{m}^3 \cdot \text{s}) \quad (5)$$

For similar reacting nuclei (for example, D and D), the reaction rate is given by Eq. (6).

$$R_{11} = \frac{1}{2} n^2 \langle \sigma v \rangle \quad (6)$$

Both expressions vary as the square of the total particle density (for a given fuel composition).

If the particle velocity distributions are known,  $\langle \sigma v \rangle$  can be determined as a function of energy by numerical integration, using the known reaction cross sections. It is customary to assume a Maxwellian particle velocity distribution, toward which all others tend in equilibrium. The values of  $\langle \sigma v \rangle$  for the D-D and D-T reactions are shown in Fig. 2. In this plot the kinetic temperature is given in kiloelectronvolts; 1 keV kinetic temperature =  $1.16 \times 10^7 \text{ K}$ . Just as in the case of the cross sections themselves, the most striking feature of these curves is their extremely rapid falloff with temperature at low temperatures. For example, although at 100 keV for all reactions  $\langle \sigma v \rangle$  is only weakly dependent on temperature, at 1 keV it varies as  $T^{6.3}$  and at 0.1 keV as  $T^{13.3}$ . Also, at the lowest temperatures it can be shown that only the particles in the "tail" of the distribution, which have energies large compared with the average, will make appreciable contributions to the reaction rate, the energy dependence of  $\sigma$  being so extreme.

**Critical temperatures.** The nuclear fusion reaction can obviously be self-sustaining only if the rate of loss of energy from the reacting fuel is not greater than the rate of energy generation by fusion reactions. The simplest consequence of this fact is that there will exist critical or ideal ignition temperatures below which a reaction could not sustain itself, even under idealized conditions. In a fusion reactor, ideal

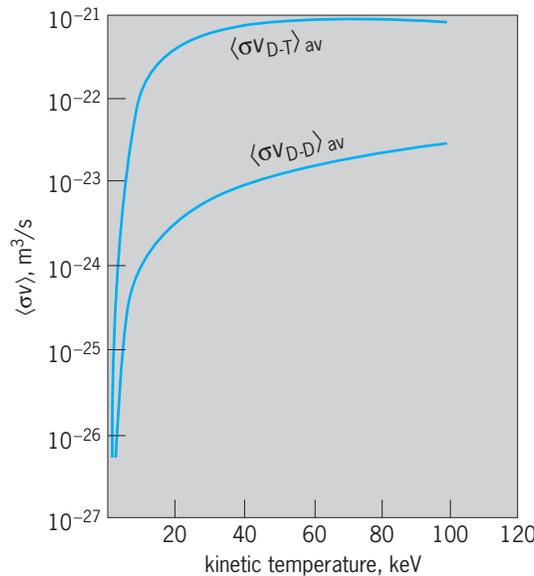


Fig. 2. Plot of the values of  $\langle \sigma v \rangle$  versus kinetic temperature for the D-D and D-T reactions.

or minimum critical temperatures are determined by the unavoidable escape of radiation from the plasma. A minimum value for the radiation emitted from any plasma is that emitted by a pure hydrogenic plasma in the form of x-rays or bremsstrahlung. Thus plasmas composed only of isotopes of hydrogen and their one-for-one accompanying electrons might be expected to possess the lowest ideal ignition temperatures. This is indeed the case: It can be shown by comparison of the nuclear energy release rates with the radiation losses that the critical temperature for the D-T reaction is about  $4 \times 10^7$  K. For the D-D reaction it is about 10 times higher. Since both radiation rate and nuclear power vary with the square of the particle density, these critical temperatures are independent of density over the density ranges of interest. The concept of the critical temperature is a highly idealized one, however, since in any real cases additional losses must be expected to occur which will modify the situation, increasing the required temperature. Richard F. Post

### Magnetic Confinement Fusion

The application of nuclear fusion to energy production requires that the nuclear fuel have sufficient temperature, density, and confinement. If the fuel is an equal mixture of deuterium and tritium, the optimum temperature is approximately 20 keV, or  $2 \times 10^8$  °C; for other fuels the optimum temperature is even higher. At such high temperatures, atoms and molecules dissociate into ions and electrons to form a state of matter called plasma. Plasmas exhibit the same relationship between pressure  $P$ , temperature  $T$ , and the number of particles per cubic meter,  $n$ , as ordinary gases,  $P = nk_B T$ , where  $k_B$  is the Boltzmann

constant. Unlike ordinary gases, such as air, plasmas are good conductors of electricity. Indeed, a fusion plasma has a conductivity more than ten times higher than copper. See BOLTZMANN CONSTANT; PLASMA (PHYSICS).

The high electrical conductivity of plasmas allows the fusion fuel to be confined by a magnetic field. The force per cubic meter exerted on a conductor carrying a current of density  $\mathbf{j}$  through a magnetic field  $\mathbf{B}$  is  $\mathbf{j} \times \mathbf{B}$ , which means the force is perpendicular to the current and the field. A confined plasma must have a gradient in its pressure,  $\nabla P$ , which is a force per cubic meter. Force balance, or equilibrium, between the magnetic field and plasma is achieved if Eq. (7) is satisfied.

$$\nabla P = \mathbf{j} \times \mathbf{B} \quad (7)$$

See MAGNETISM.

Equation (7), which is the fundamental equation for magnetic confinement, constrains the shape of magnetically confined fusion plasmas to being toroidal, or doughnut-shaped. The magnetic field is perpendicular to the direction in which the pressure varies; thus, a magnetic field line can never leave a constant-pressure surface. A mathematical theorem states that only one surface shape exists in three dimensions that can everywhere be parallel to a vector, such as a magnetic field; that shape is a torus. If Eq. (7) is violated, the plasma will expand at a speed set by its sound velocity. Such plasmas are inertially confined. The efficiency of magnetic confinement is measured by the ratio of the plasma pressure  $P$  to the magnetic pressure  $B^2/2\mu_0$ , where  $\mu_0$  is the permeability of vacuum. This ratio is called beta, given by Eq. (8).

$$\beta \equiv \frac{2\mu_0 P}{B^2} \quad (8)$$

Studies of possible designs for fusion power plants have shown that a beta of approximately 5% is required for a commercially competitive plant. In these designs, a typical plasma pressure is a few times atmospheric pressure, and a typical magnetic field strength is 5 tesla, about 100,000 times greater than the magnetic field on the Earth. See MAGNETOHYDRODYNAMICS.

**Plasma configurations.** Although the plasma configurations used for magnetic confinement fusion are toroidal, they come in a variety of forms primarily distinguished by the strength of the net plasma current. The net plasma current flows along the magnetic field lines, and does not directly change the force balance. The net current does, however, change the magnetic field following Ampère's law,  $\nabla \times \mathbf{B} = \mu_0 \mathbf{j}$ .

The two most prominent types of magnetic confinement devices are the tokamak and the stellarator (Table 1). The tokamak (Fig. 3) is toroidally symmetric and uses currents in coils to produce a strong magnetic field that encircles the torus in the toroidal direction, the long way around. A net plasma current produces a poloidal magnetic field, that is, field lines that encircle the torus the short way

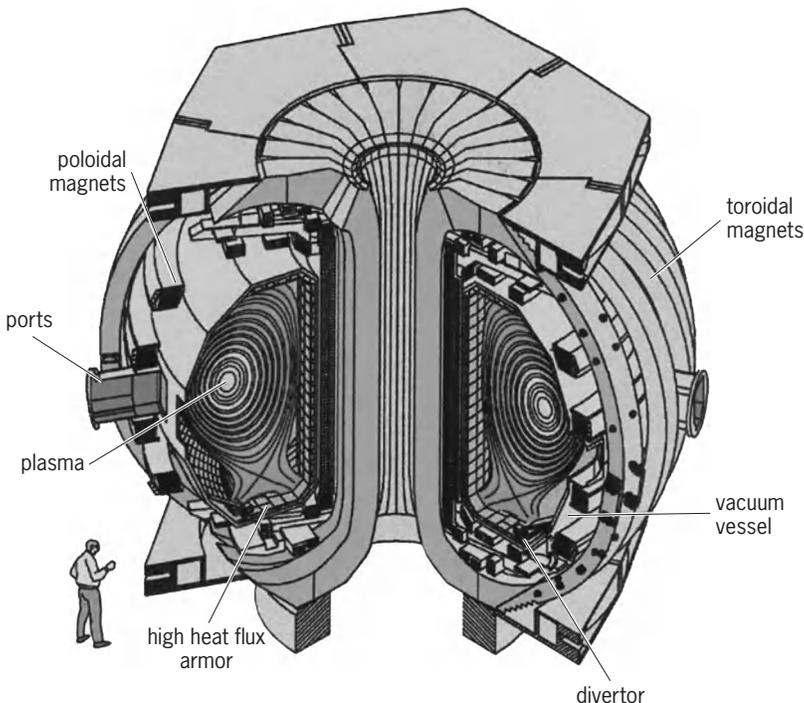


Fig. 3. Layout of the DIII-D tokamak.

TABLE 1. Some major facilities for magnetic fusion energy

Machine	Type	Location	Major radius, m (ft)	Minor radius, m (ft)	Field strength, tesla	Deuterium/tritium power, MW
DIII-D	Tokamak	United States	1.67 (5.5)	0.67 (2.2)	2.2	
JET	Tokamak	European Union	3.1 (10.2)	1.25 (4.1)	3.5	16
JT60U	Tokamak	Japan	3.4 (11.2)	1.1 (3.6)	4.2	
LHD	Stellarator	Japan	3.9 (12.8)	0.65 (2.1)	3.0	
NCSX*	Stellarator	United States	1.42 (4.7)	0.33 (1.1)	2.0	
NSTX	Spherical torus	United States	0.85 (2.8)	0.65 (2.1)	0.6	
MAST	Spherical torus	European Union	0.85 (2.8)	0.65 (2.1)	0.52	
TFTR†	Tokamak	United States	2.45 (8.0)	0.8 (2.6)	5.2	11
W7-X*	Stellarator	European Union	5.5 (18.0)	0.52 (1.7)	3.0	

\*Decommissioned. †Under construction.

around. The largest tokamaks are the Joint European Torus (JET) in Britain, the JT60U in Japan, and the decommissioned Tokamak Fusion Test Reactor (TFTR) in the United States. A tokamak, ITER, which would be capable of producing fusion power at a level comparable to a power plant, has been designed under an international agreement. The design parameters are a 6.2-m (20-ft) major radius, a 2.0-m (6.6-ft) minor radius, and a 5.3-tesla magnetic field, with a planned fusion output of 500 MW. The spherical torus is a variant of the tokamak that is more compact, which means a larger ratio of the minor to the major radius of the torus.

The stellarator (Fig. 4) requires no net plasma current to confine a toroidal plasma configuration. External coils produce field lines that lie in nested toroidal surfaces, even in the absence of a plasma. In other words, toroidal surfaces exist in the plasma region even when  $\nabla \times \mathbf{B} = 0$ . The properties of magnetic fields in a current-free region imply that the surfaces of constant pressure in a stellarator cannot be toroidally symmetric; they must have helical shaping, as must the coils that produce the field. The largest stellarator is the Large Helical Device (LHD)

in Japan, which has a steady-state magnetic field produced by superconducting coils. The W7-X is a stellarator experiment of similar scale being constructed in Germany.

**Status.** Using deuterium and tritium, the JET tokamak has produced 16 MW of fusion power, which is comparable to the power required to heat the plasma (Fig. 5). The next generation of tokamaks should achieve the basic plasma properties required for fusion power plants. Nevertheless, the current perception of plentiful and cheap energy, especially in the United States, has led to a critical examination of the attractiveness of energy production by tokamaks and by fusion in general. The most reliable estimates indicate that electrical power produced by fusion might cost twice as much as currently available sources. In addition, a large investment in research and development is required before the technology can be shown to be practical.

The importance of fusion energy is to ensure that an unlimited source of energy can be made widely available if fossil fuels become depleted or environmentally unacceptable. The effects of depletion and environmental damage are made more severe by the

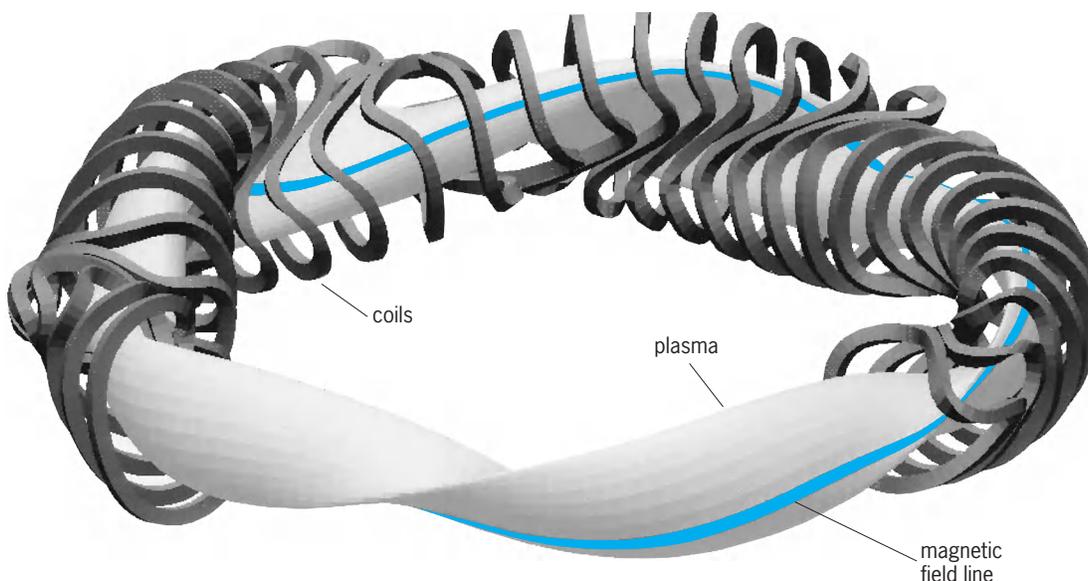
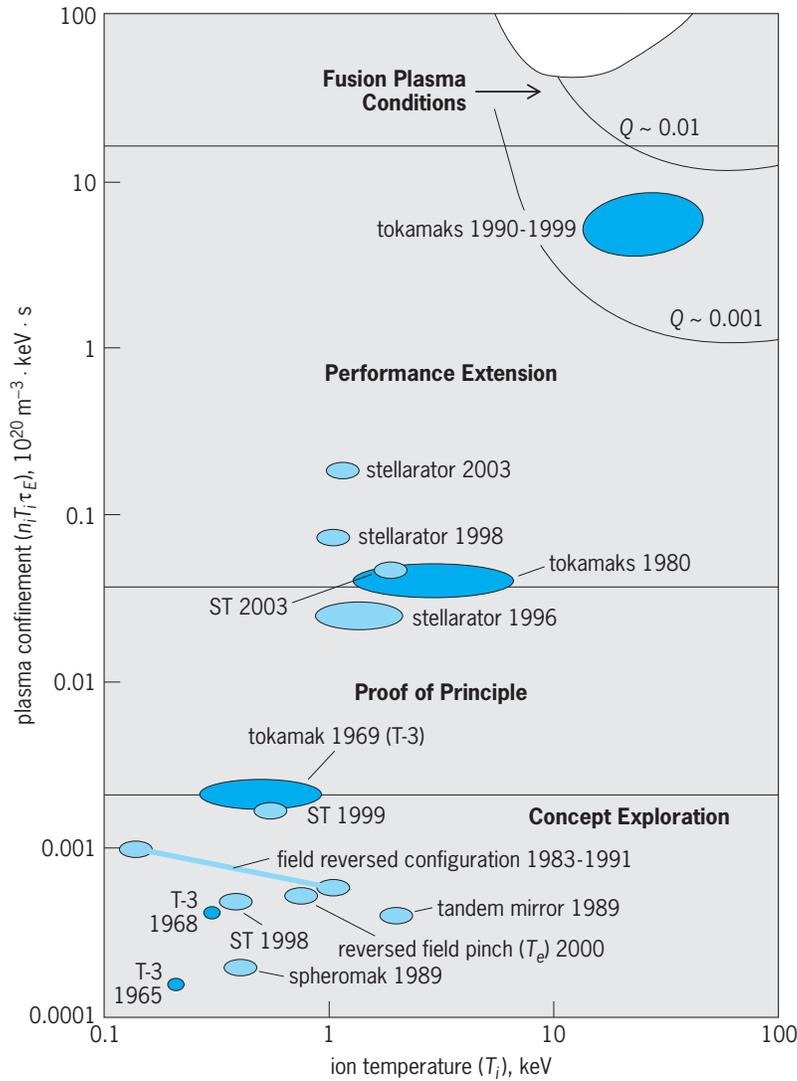


Fig. 4. Layout of the W7-X stellarator.



**Fig. 5. Progress in magnetic fusion.** The horizontal axis gives the plasma thermal energy, the temperature of the ions at the center of the plasma in kilovolts,  $T_i$ . One kilovolt is approximately  $12 \times 10^6 \text{ }^\circ\text{C}$ . The vertical axis gives the plasma confinement, which is the product of the number of ions per cubic meter,  $n_i$ , the ion temperature in kilovolts, and the energy confinement time in seconds,  $\tau_E$ , which is the thermal energy in the plasma divided by the power required to maintain the plasma temperature. The  $Q$  factors (fusion power divided by input power) are for D-D plasmas, which is the basis of most experiments. A fusion power plant would use a D-T plasma, which increases  $Q$  by an approximate factor of 1000. (Dr. Dale Meade, Princeton Plasma Physics Laboratory)

need for energy use to rise dramatically if a large fraction of the population of the world is not to remain in poverty.

**Physical considerations.** Force balance [Eq. (7)] is a necessary condition for magnetic confinement, but it is not a sufficient condition. Additional issues requiring consideration are (1) the stability of the plasma equilibrium; (2) the diffusive transport of the fuel or, more importantly, its thermal energy; and (3) the loss of thermal energy through the radiation of light. For fusion using a mixture of deuterium and tritium, the third issue is primarily one of plasma purity. Heavier elements, such as carbon or iron, can radiate at a level of power far above that produced by fusion reactions. The high temperature required for fusion can be maintained only if the concentrations of impurities are carefully controlled.

*Plasma stability.* The first issue, the stability of the plasma, is equivalent to asking whether the equilibrium is analogous to that of a ball at the top or the bottom of a hill. Instabilities in the force balance are of two types: those driven by the pressure gradient and those driven by the net plasma current.

Instabilities due to the pressure gradient arise from the tendency of a plasma to move to a region of lower magnetic field strength. The high conductivity of the plasma makes changing the magnetic flux embedded in an element of plasma difficult. (Magnetic flux is the strength of the magnetic field multiplied by an area transverse to the field.) By moving into a region of lower magnetic field strength, a plasma can release the free energy of expansion while conserving the embedded magnetic flux. The properties of a current-free magnetic field imply that in a torus a region of lower magnetic field strength exists in at least part of the periphery of the plasma. However, the plasma is not necessarily unstable toward moving into this region of reduced field strength (that is, the possibility of such movement is not necessarily analogous to the possibility of a ball rolling down from the top of a hill) because a magnetic field deformation is generally required in order for such movement to take place. The requirement for a deformation of the magnetic field stabilizes the plasma up to a critical value of the plasma pressure, or more precisely  $\beta$ . Values of  $\beta$  well above the minimal value for a power plant,  $\beta \approx 5\%$ , have been achieved in tokamak experiments.

The second type of instability is driven by the net current,  $j_n$ . Such instabilities are known as kink instabilities from the form of the plasma deformation that is produced. Kink instabilities limit the magnitude of the net plasma current. This limit is usually expressed using the safety factor,  $q$ , which is the number of times a magnetic field line must encircle the torus toroidally, the long way, before encircling the torus poloidally, the short way. The magnitude of the safety factor varies across the plasma. In a stable tokamak equilibrium,  $q$  is typically about 1 near the plasma center and 3 near the edge.

*Energy transport.* As mentioned above, the second issue concerns the transport properties of the plasma. Although magnetically confined plasmas are in force equilibrium [Eq. (7)], they are not in thermodynamic equilibrium. Temperature gradients, which represent a deviation from thermodynamic equilibrium, lead to a transport of energy just as a temperature gradient in air does. As in air, the transport of energy in a plasma due to a temperature gradient can be much larger than that predicted by the standard transport coefficients. In air the enhancement arises from convection. In fusion plasmas the transport is due to small-scale fluctuations in the electric and magnetic fields. The enhanced transport arising from these fluctuations is called microturbulent or anomalous transport. See CONVECTION (HEAT); HEAT TRANSFER.

The size of a fusion plasma, about 2 m (6 ft) for the smaller radius of the torus, is comparable by the scale of the blankets and shields that surround the plasma.

The radial transport coefficients need to have a particular magnitude to make that radius consistent with the removal of energy from the plasma at the same rate as the plasma is heated by fusion reactions. Remarkably, the level of transport observed in experiments is close to that magnitude. Nevertheless, the uncertainties associated with microturbulent transport are among the largest uncertainties in setting the minimum size, and therefore minimum cost, for an experiment to achieve the plasma confinement required for a power plant. Much progress has been made in the development of methods for computing microturbulent transport. Advances are paced by the development of faster computers and more sophisticated algorithms. In addition, experiments have shown that it is possible to manipulate the plasma in ways that give large reductions in microturbulent transport.

*Collisionality paradox.* Fusion plasmas are in a paradoxical collisionality regime. The ions and the electrons that form a fusion plasma are said to be collisionless, since they move about 10 km (6 mi) between collisions and the plasma has a scale of only a few meters. On the other hand, fusion plasmas are highly collisional, for the energy of the ions and electrons must be confined for more than a hundred collision times. This implies that the velocity of ions and electrons obeys the same probability distribution as air molecules, the Maxwellian distribution. See KINETIC THEORY OF MATTER.

In a tokamak the collisionality paradox presents no fundamental problem. The toroidal symmetry of the tokamak plus the theory of Hamiltonian mechanics ensure that a quantity called the toroidal canonical momentum is invariant along the trajectory of the particle. This invariance prevents the trajectories of the ions and electrons from straying far from the constant-pressure surfaces.

A stellarator cannot be toroidally symmetric, and careful design is required to ensure the confinement of particle trajectories. Surprisingly, the confinement of trajectories is controlled by the variation of the magnetic field strength in a constant-pressure surface and not by the shape of the surfaces. The surfaces of constant pressure in a stellarator cannot be toroidally symmetric, but the field strength can be almost independent of the toroidal angle in a pressure surface, and this quasi-symmetry confines the trajectories. Trajectories in a stellarator can also be confined using a subtle type of invariant of Hamiltonian mechanics, the longitudinal adiabatic invariant. It is this invariant that is used to confine the particle trajectories in the W7-X stellarator. See HAMILTON'S EQUATIONS OF MOTION.

*Stellarator advantages.* The reasons that stellarators are considered attractive for fusion are the robust stability of stellarator plasmas and the absence of a requirement for power to drive a net current. The robust stability of stellarator plasmas means they do not experience sudden terminations, which would exert a large stress on the surrounding structures. Sudden terminations, called disruptions, are observed in various operating conditions of tokamak

plasmas and would not be acceptable in a power plant.

**Technical considerations.** In addition to the physical constraints on the design of fusion power plants, there are technical constraints: the first wall, the blanket and shields, and the coils. The solid surface immediately surrounding the plasma is called the first wall. Suitable materials are under study. Two issues concerning the first wall have to do with the neutrons that are emitted by a fusing plasma: (1) The attractiveness of fusion depends on the first wall not becoming highly radioactive. (2) The first wall must maintain its structural integrity even though neutron bombardment tends to cause materials to swell and become brittle. A third issue is economic. Fusion power becomes cheaper the higher the energy flux that can pass through the first wall. A power flux of 5 MW/m<sup>2</sup> is a typical design value, with 80% of that power in the neutrons that are produced by deuterium/tritium fusion reactions. See RADIATION DAMAGE TO MATERIALS.

Tritium is not naturally available. In a deuterium/tritium fusion power plant, a blanket is required to produce tritium by the bombardment of lithium with neutrons. Behind the blanket are shields to prevent neutron bombardment of the current-carrying coils. The required thickness of the blanket and shield is approximately 1.5 m (5 ft) and sets a natural size scale for a fusion power plant, which is approximately 1000 MW of electric power.

The coils that produce the magnetic field in a fusion power plant must be superconducting if the plasma  $\beta$  value is to be approximately 5%. Ordinary copper coils would dissipate far too much energy. The magnetic field produced by the coils is limited both by the properties of superconductors and by the mechanical stress,  $B^2/2\mu_0$ . A commonly used limit is that of Nb<sub>3</sub>Sn superconductors, 12 tesla. The field in the plasma is lower, and an important efficiency factor is the ratio of the average magnetic field strength in the plasma to maximum field at a coil. A typical value for this ratio is 0.5. See SUPERCONDUCTING DEVICES. Allen H. Boozer

### Inertial Confinement Fusion

The basic idea of inertial confinement fusion is to use a laser or particle beam (often called a driver) to deliver sufficient energy, power, and intensity [ $\sim 5 \times 10^6$  joules,  $10^{15}$  W in about 8 to 10 nanoseconds, and  $10^{14}$  W/cm<sup>2</sup> to a capsule several millimeters in diameter, containing several milligrams of a mixture of deuterium and tritium (D-T)], to produce an implosion. This implosion is designed to compress and heat the D-T mixture to a density of about 200 g/cm<sup>3</sup> (200 times that of water or approximately 1000 times the liquid density of D-T) and a temperature of 3–5 keV. With this range of drive energy and under these conditions, 1–10 mg of D-T could be ignited and, while inertially confined, with a burn efficiency of about 30%, could release 100–1000 megajoules of fusion energy, with energy gains, depending upon specific target and driver conditions, from 10 to greater than 100. An energy of

**TABLE 2. Large-scale inertial confinement fusion facilities**

Driver	Location	Type	Power, TW	Energy, kJ	Wavelength, $\mu\text{m}$	Number of beams
<b>Laser</b>						
Nova	Lawrence Livermore National Laboratory, United States	Neodymium:glass	120	120	1.06	10
			30–60	50–80	0.53	
			20–40	40–70	0.33	
Aurora	Los Alamos National Laboratory, United States	Krypton fluoride	1–23	[5–10]*	0.25	48
Omega	University of Rochester, United States	Neodymium:glass	12	4	1.06	24
			6	2	0.33	
Phebus	Centre d'Etudes de Limeil, France	Neodymium:glass	24	24	1.06	2
Vulcan	Rutherford-Appleton Laboratory, United Kingdom	Neodymium:glass	4	1.0	1.06	12
Gekko XII	Osaka University, Japan	Neodymium:glass	2	0.4	0.53	12
			55	30	1.06	
Asterix III	Max Planck Institute for Quantum Physics, Germany	Iodine	5	2	0.53	1
					1.3	
<b>Particle beams</b>						
PBFA II	Sandia National Laboratory, United States	Lithium	[100]*	[2000]†		
Kalif	KFK, Karlsruhe Physics, Germany	Hydrogen	0.7	20‡		
SIS/ESR	GSI, Darmstadt Physics, Germany	Neon	45 MW	0.045		
		Xenon	0.036	2.5		

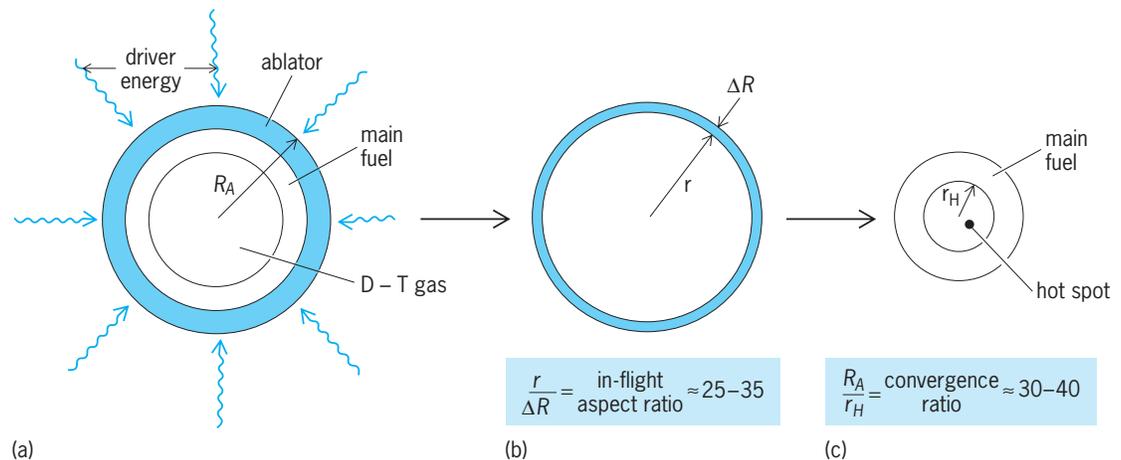
\*[ ] denotes design goals, not yet achieved. †15-ns pulse. ‡50-ns pulse.

100 MJ is equivalent to about 50 lb (25 kg) of TNT, or about 0.5 gallon (2 liters) of oil, an amount of energy tractable under laboratory conditions and potentially useful for a variety of applications, including basic and applied science, weapons physics and weapons effects, energy production, and production of special nuclear materials and nuclear fuels.

Drivers with light ions, heavy ions, krypton fluoride (KrF) lasers, and neodymium:glass lasers are being developed. Although they are at widely different stages of development, all four of these technologies could probably provide a driver capable of delivering 5–10 MJ. It is not clear, however, that they can all deliver this energy while meeting the additional requirements of flexible temporal pulse shape,  $10^{15}$  W peak power, greater than  $10^{14}$  W/cm<sup>2</sup> inten-

sity, and focusing of the beam over the standoff distance. Neodymium:glass laser technology is probably the most developed of the four. **Table 2** lists some representative research facilities.

The two fundamental approaches to most inertial confinement fusion research are differentiated by the mechanism chosen to drive the process. In the direct-drive approach, laser (or ion) beams are arrayed around the target in a near-uniform pattern and aimed directly at the D-T fuel capsule. In the indirect-drive approach, the driver energy is deposited (absorbed) in an intermediate medium that heats up to temperatures high enough to emit x-rays. The fuel capsule is contained inside this x-ray converter (hohlraum), and the x-rays are used to drive the capsule.



**Fig. 6. Direct-drive inertial confinement fusion target.** The radial elements of the (a) initial, (b) in-flight, and (c) final fuel configurations are identified. Additional requirements for a successful implosion include peak pressure of  $\sim 100$  Mbar (10 TPa), flux uniform to about 2%, x-ray and electron preheat to the fuel of less than a few electronvolts, accurate pulse shaping, and very smooth target surface (roughness of  $\sim 0.1 \mu\text{m}$ ).

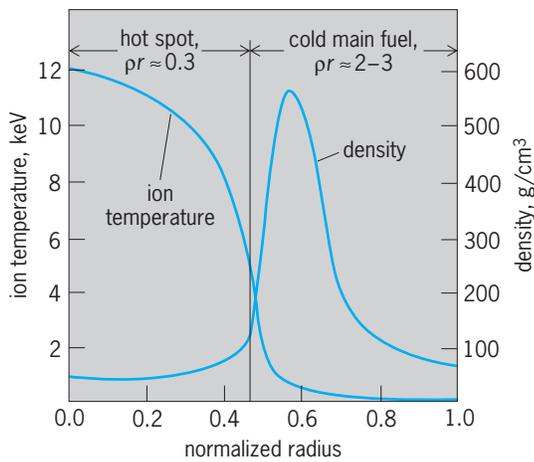


Fig. 7. Deuterium-tritium fuel ion temperature and density at ignition, plotted against normalized radius, for a typical, direct-drive high-gain capsule. Hot-spot value of  $\rho r$  must exceed alpha-particle range for effective self-heating from 5 keV.

The details of indirectly driven targets are classified. Therefore, in order to give a coherent description of the capsule physics, only the requirements for directly driven targets will be discussed.

**Implosion physics.** The events leading to a capsule burn are as follows. The driver deposits its energy by electron conduction (directly driven) in the outer layers of the fuel-containing capsule (Fig. 6). This energy deposition in turn ablates the surface of the capsule, and the ablation acts as a rocket exhaust to drive the implosion of the capsule. In the outer portions of the capsule, the ablation generates pressures from 1 megabar (100 gigapascals) to about 100 Mbar (10 terapascals) over a period of about 10 ns, and this accelerates the material of the capsule inward to velocities of  $\sim 10^6$  ft/s ( $3 \times 10^5$  m/s). The inward acceleration continues until the internal pressures exceed the pressures of ablation. At that time, the rapidly converging shell begins to decelerate, compresses the central portion of the fuel to densities of the order of 200 g/cm<sup>3</sup>, and then heats the core to about 5 keV, whereupon thermonuclear burn is initiated.

The final high fuel densities for inertial confinement fusion are required to burn the D-T efficiently. In inertial confinement fusion, the product of fuel density  $\rho$  and fuel radius  $r$  is an important parameter, analogous to the  $n\tau_E$  product in magnetic fusion. The burnefficiency  $\eta$  for inertial confinement fusion



Fig. 8. Nova inertial confinement research facility at Lawrence Livermore National Laboratory. Laboratory building containing neodymium:glass laser system is separated into four clean-room areas: master oscillator room on lower level where initial pulse light is generated; laser bay containing 10 main laser amplifier chains; optical switchyard in which the 29-in.-diameter (74-cm) beams are diagnosed and reflected into the target room; and target area that houses the 15-ft-diameter (4.6-m) aluminum target chamber. (Lawrence Livermore National Laboratory)

depends on  $\rho r$ ; specifically, it is given by Eq. (9),

$$\eta = \frac{\rho r}{\rho r + 6} \quad (9)$$

where  $\rho r$  is expressed in  $\text{g}/\text{cm}^2$ . A  $\rho r$  value of  $3 \text{ g}/\text{cm}^2$  using liquid-density D-T can be achieved only with a large amount of D-T (approximately 6 lb or 3 kg). The fusion yield of that much D-T, however, is about 70 kilotons, which cannot be contained in a practical way. Instead, if the fuel is compressed to 1000-times-liquid-density D-T, then about 5 mg of D-T is sufficient to obtain the same  $\rho r$ , and the yield is then about 250 lb (125 kg) of TNT, which can be readily contained. This requirement is known as high compression.

Although high compression is necessary, it is not sufficient. To achieve high gain, inertial confinement fusion also relies upon central ignition and propagation of the burn into the compressed D-T. If the fuel is compressed to a  $\rho r$  of  $\sim 3 \text{ g}/\text{cm}^2$  and heated to 5 keV, then 5 mg of D-T releases about 500 MJ. However, to heat the entire fuel mass to 5 keV requires too large a driver in terms of size and cost. Instead, a small portion of the fuel mass—less than 5%—at the center of the capsule is heated to ignition temperatures, while the rest of the fuel remains cold but compressed to  $\sim 200 \text{ g}/\text{cm}^3$ . **Figure 7** depicts such a configuration, where the hot-spot value of  $\rho r$  is just large enough to stop the alpha particles generated from fusion reactions in the hot-spot ignition region, starting the self-heating process and propagating the burn into the main fuel.

This sequence of events is complicated by some facts of nature and the need to build drivers (either lasers or particle beams) whose energy, power, and focusability are limited by costs and technology. These make it necessary to use a driver with a maxi-

mum energy output of  $\sim 10 \text{ MJ}$ , in  $\sim 10 \text{ ns}$ , and peak power of  $10^{15} \text{ W}$ . The basic facts of nature are as follows.

1. In order to achieve the final fuel densities of  $\sim 200 \text{ g}/\text{cm}^3$  with the 100-Mbar (10-TPa) peak pressures available with laser intensities in the  $10^{14}$ – $10^{15} \text{ W}/\text{cm}^2$  range, it is necessary to start with the D-T fuel in a liquid or solid shell.

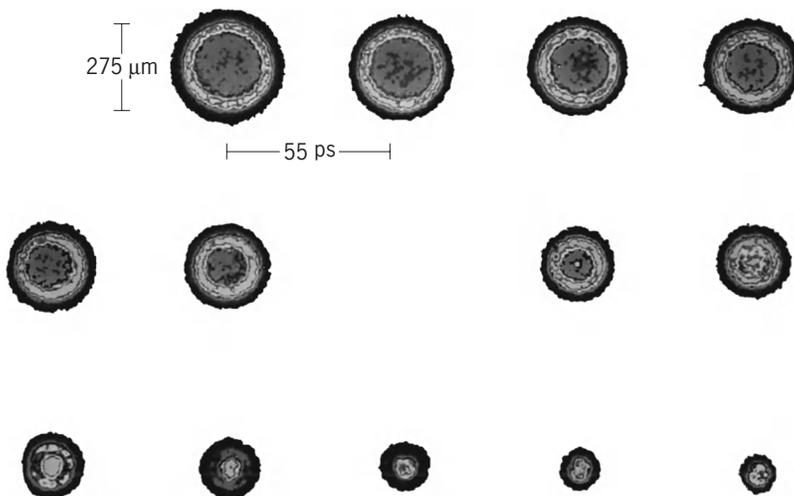
2. The implosion must be carried out in a reasonably isentropic manner until a late stage of the implosion, or else the continual heating of the fuel will make a high compression too demanding on driver energy.

3. Potential hydrodynamic instabilities demand both target surface finishes that are nearly perfect and pressures that are uniform to a few percent, to avoid mixing of the outer portion of the shell (the ablator) and minimize mixing of the inner portion of the shell (the pusher) with the fuel and subsequent degradation of the burn efficiency.

The critical issues for the success of inertial confinement fusion are whether a sufficiently uniform capsule can be made, whether sufficiently uniform pressure pulse can be produced, and whether there is full understanding of the hydrodynamics and physics of the D-T burn on which the models are based. The information now available indicates that high gain can be achieved with a 5–10-MJ driver.

**Target technology.** A free-standing, liquid D-T shell would be the ideal direct-drive target, but it is still not known how to make one. However, a free-standing solid layer of D-T is nearly as good, and it appears that the natural heat generated by the tritium beta decay in a D-T mixture may offer a natural process to produce the required thicknesses of solid D-T layers with the required uniformity, a technique called beta layering. Uniform spherical D-T layers up to several millimeters in thickness have been produced. Spherical plastic shells to contain the D-T and provide the ablator for the implosion have been made in the 0.5-mm-diameter range with the required surface finishes of less than 100-nm thickness. Another option is to use a shell of low-density plastic foam to wick up liquid D-T and make an analog D-T liquid layer. With a plastic foam of density  $50 \text{ mg}/\text{cm}^3$ , the target would be 80% liquid D-T and 20% low-density foam. If it were made with a sufficiently uniform surface finish, it would be a very close approximation to the liquid D-T layer shell. Such low-Z, porous foams can be machined or molded to a surface finish of less than  $1\text{-}\mu\text{m}$  thickness.

**Target experimental facilities.** The Nova laser facility (**Fig. 8**) is a good example of the very large sophisticated, and flexible laboratory facilities that have been developed for studying inertial confinement to fusion. Its 10 arms produce laser pulses that together can provide the greatest energy ( $10^5 \text{ J}$ ) and power ( $10^{14} \text{ W}$ ) of any driver, and deliver it to a target in  $10^{-9} \text{ s}$ . The neodymium: glass laser produces infrared



**Fig. 9.** X-ray images with a framing camera with 80 ps exposure time (developed at Lawrence Livermore National Laboratory) from a directly driven inertial confinement fusion capsule at the University of Rochester's Omega Facility. Camera has 80 ps exposure time and 55 ps time between frames; because exposure time exceeds time between frames, there is some temporal overlap. (Data from missing frame was unavailable.)

light at a wavelength of about 1.06 micrometers, and the system was built with the capability to convert the frequency of that laser light to the second (green) or third (blue) harmonics. These shorter wavelengths greatly improve the coupling of the light to the target and help in achieving high compression.

The laser power must be delivered with a precise temporal pulse shape, ranging in duration from 1 to 10 ns. Nova and other solid-state lasers can produce a variety of complex pulse shapes from simple gaussian-shaped pulses to pulses with a monotonically increasing temporal profile, or combinations with so-called picket fences that approximate the desired monotonically increasing drive pulse.

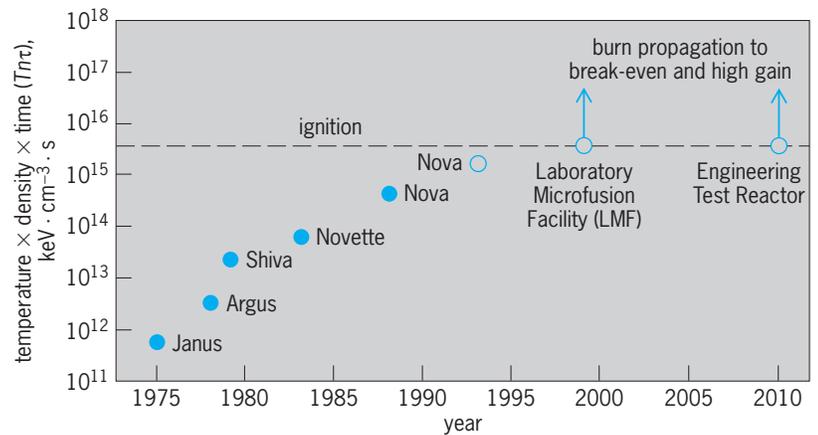
**Diagnostic technologies.** Sophisticated implosion diagnostics have been developed. **Figure 9** shows a series of x-ray images of an imploding capsule taken with an x-ray framing camera with  $\sim 80$  picosecond exposure time. Other diagnostics include neutron pinhole cameras with about  $20\text{-}\mu\text{m}$  spatial resolution, neutron time-resolving detectors with about 50-ps resolution, and a variety of x-ray optics such as x-ray microscopes with  $1\text{-}2\text{-}\mu\text{m}$  resolution, normal-incidence x-ray mirrors with reflection up to 50%, and a variety of x-ray gratings and beam splitters.

**Experimental progress.** The primary focus of the program at Nova is the indirect-drive approach with targets of a design that scales to the high-density regime. With temporally shaped pulses and appropriately scaled values for the drive temperature, values of pressure, preheat, and symmetry have been measured that meet or exceed those required for high gain. Values of  $n\tau_E$  of about  $2\text{-}5 \times 10^{14} \text{ cm}^{-3} \cdot \text{s}$  and fuel ion temperatures of  $1.5\text{-}2.0 \text{ keV}$  have been obtained, which are less than a factor of 10 away from the values needed to achieve fusion ignition (**Fig. 10**). These results have provided stringent tests of theoretical modeling capabilities and confirm that indirectly driven capsules with a 5-10-MJ driver should achieve the  $n\tau$  values and fuel conditions required for ignition and high gain.

Significant advances have also been made in the direct-drive approach. There has been progress in understanding and controlling illumination uniformity which is critical to direct-drive fusion and also important to the energy-to-x-ray conversion efficiency of indirect drive.

By using an improved beam-smoothing technique, excellent implosion performance has been demonstrated using gas-filled glass shells. If similar improvements can be obtained for target designs that have the higher in-flight aspect ratios (ratio of the shell thickness to the shell radius) ultimately required, it will mean that direct-drive targets also hold the promise of high gain for drive energies in the 5-10-MJ range.

These results and others have led the U.S. Department of Energy to undertake planning of the Laboratory Microfusion Facility (LMF), which would provide a yield of up to 1000 MJ. In addition to military



**Fig. 10.** Progress of inertial confinement fusion experiments in achieving conditions required for high gain. The product of ion temperature  $T$ , number density  $n$ , and confinement time  $\tau$  is plotted against calendar year. Solid circles have been achieved; open circles represent future goals.

applications, this facility would provide the basis for a decision to initiate engineering development aimed toward practical fusion power. Eric Storm

### Aneutronic Fusion

Certain reactions of light nuclei other than the fusion of deuterium and tritium, the reaction studied in conventional nuclear fusion research, are attractive because they are aneutronic and nonradioactive, meaning that they generate negligible amounts of neutrons and radioactive nuclei. A revival of interest in this energy source in the mid-1980s was stimulated by the introduction of new concepts borrowed from particle physics research.

While aneutronic fusion has many potential advantages, the concept also has many difficulties, mostly associated with the very high energies required for fusion. Since fusion research has concentrated on D-T systems, quantitative theoretical information and experimental data on aneutronic fusion are relatively meager. Without radically new experimental results, this form of fusion, therefore, must be regarded more as a concept for exploration than as a technology.

**Advantages of aneutronic power.** The absence of neutrons and radioactivity obviates the need for shielding. This is particularly significant for aerospace applications, since the weight of shielding in a nuclear fission or fusion reactor is at least 100 times greater than that of the reactor itself.

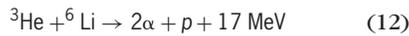
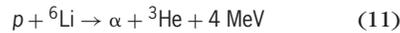
An aneutronic reactor also offers the advantages of nonradioactive fuel and nonradioactive waste. Furthermore, since all nuclear energy released in aneutronic reactions is carried by charged particles, if these particles could be directed into a beam a flow of electric charge would result, and nuclear energy could be converted directly into electrical energy, with no waste heat.

An aneutronic reactor could be small, producing 1-100 MW of electric power, and mass production might be possible. Finally, aneutronic reactors cannot breed plutonium for nuclear weapons.

**Definitions.** A nuclear reaction is defined as aneutronic if its neutronicism  $N$ , the power carried by the neutrons as a fraction of the total power released in the reaction, is less than 0.01. A reaction is defined as nonradioactive if its radionuclide content  $R$ , the sum of the number of radioactive nuclei before and after the reaction as a fraction of the total number of nuclei involved, is less than 0.01. Since  $R$  is proportional to  $N$ , an aneutronic reactor usually is also nonradioactive.

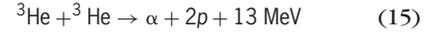
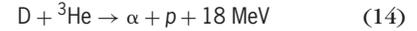
Even in a reactor based on aneutronic reactions, parallel or secondary reactions will always produce some neutrons. The objective of aneutronic energy research is to design a reactor that would suppress the neutronicism to 1% or below.

**Reactions.** Aneutronic reactions can be divided into two classes: fission of light metals by protons, or helium-3 ( $^3\text{He}$ ) nuclei, such as reactions (10)–



(13); and fusion involving helium-3, such as reactions

(14) and (15). Reaction (15) has been called the ulti-

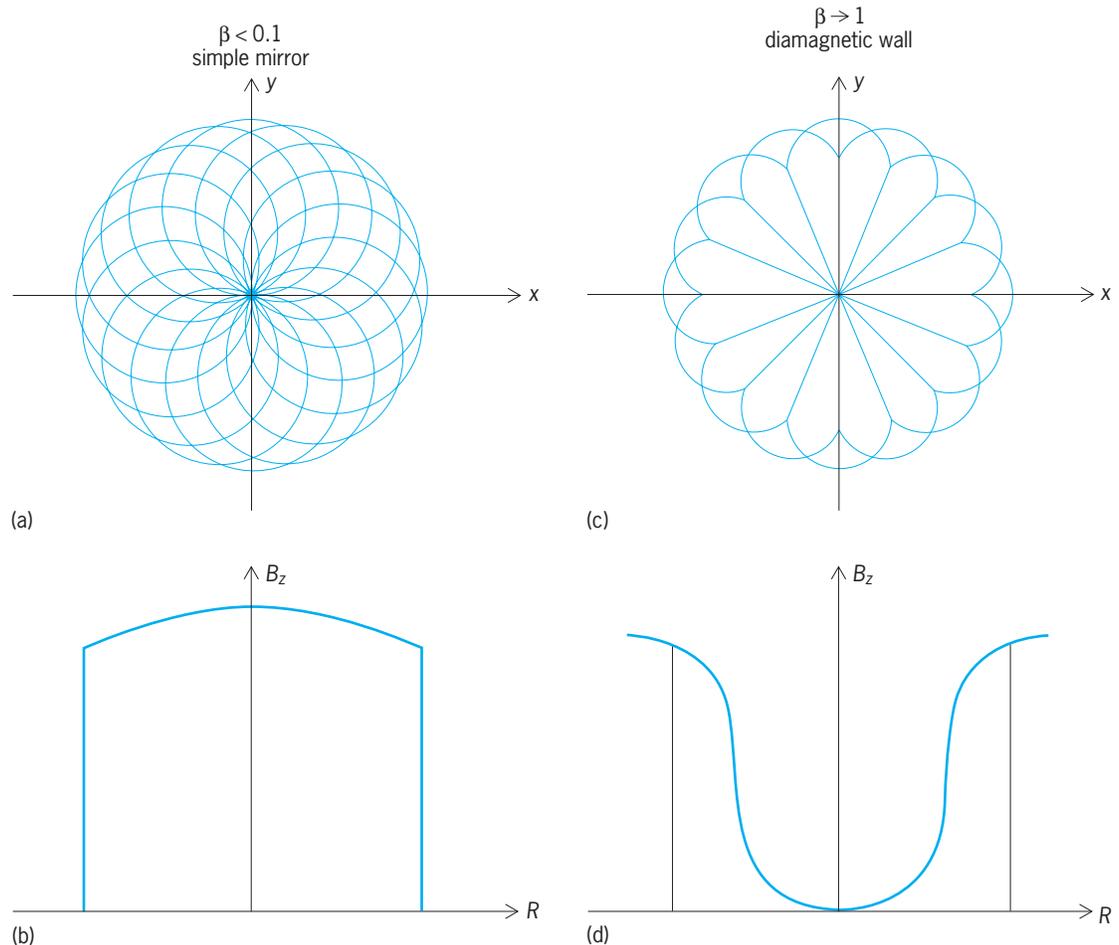


mate aneutronic reaction: both its  $N$  and  $R$  are nearly zero.

**Critical issues.** The primary obstacle to using aneutronic reactions as power sources is the lack of means to energize the fuel ions by plasma heating to the high energies required to initiate such reactions. Their reactivity approaches that of D-T fusion only at energies 10–100 times higher (Fig. 1).

The next important issue is how to increase by a factor of 10 the parameter beta, in order to suppress large electromagnetic power losses (bremsstrahlung and synchrotron radiation) from the reacting plasmas. These losses, proportional to  $Z^2$ , are negligible in hydrogen plasmas, but aneutronic fuels have ions with  $Z$  equal to 2, 3, or 4. See BREMSSTRAHLUNG; SYNCHROTRON RADIATION.

Studies in the early 1980s concluded that Maxwellian plasmas are incompatible with aneutronic fuels because of the large ion energies required and the electromagnetic power losses.



**Fig. 11. Migma-plasma devices.** (a) Ion orbits in simple-mirror magnetic field configuration at low ion density ( $n < 10^{12} \text{ cm}^{-3}$ ;  $\beta < 0.1$ ). (b) Graph of magnetic field component  $B_z$  perpendicular to orbit plane versus radial distance  $R$  in this configuration. (c) Ion orbits when higher ion densities ( $n \sim 10^{14} \text{ cm}^{-3}$ ;  $\beta \rightarrow 1$ ) create a diamagnetic well. (d) Graph of  $B_z$  versus  $R$  in this case.

**New concepts.** New concepts, however, may solve or bypass these obstacles.

*Direct heating.* Potential solutions to the problem of raising ions to the required energies came from colliding-beam technology and beam-storage physics, introduced in particle physics research in the 1970s. Compared to plasma ion confinement, colliding beams are characterized by many-orders-of-magnitude greater particle energies and confinement times, very large beta (near or above 1), and orders-of-magnitude lower vacuum pressure. The so-called classical Coulomb interaction processes that have limited plasma confinement times are inversely proportional to the  $^{3/2}$  power of the ion energy, and directly proportional to the vacuum pressure; thus, classical confinement becomes many orders of magnitude easier to achieve at the higher ion energies and ultrahigh vacuum that characterize colliding beams. Colliding beams are not neutralized, however, and the particle densities are orders of magnitude lower, limited by the electrostatic repulsion known as the space-charge limit. See PARTICLE ACCELERATOR.

Self-colliding beams can be neutralized in a migma-plasma, a synthetic type of plasma that can be made without heating, consisting of large-orbit, megaelectronvolt ions and thermal, ambient electrons. This concept was demonstrated in 1982. Accelerated ions were injected into a 6-T magnetic field in a so-called simple-mirror configuration to form a self-colliding orbit rosette pattern (Fig. 11a and b) in the presence of the electrons that neutralized the stored beam.

The Self-Collider IV or Migma IV migma-plasma device (Fig. 12) achieved a stable confinement of 0.7-MeV ions, with an energy confinement time of 25 s, and density of  $10^{10} \text{ cm}^{-3}$ . None of the disrup-

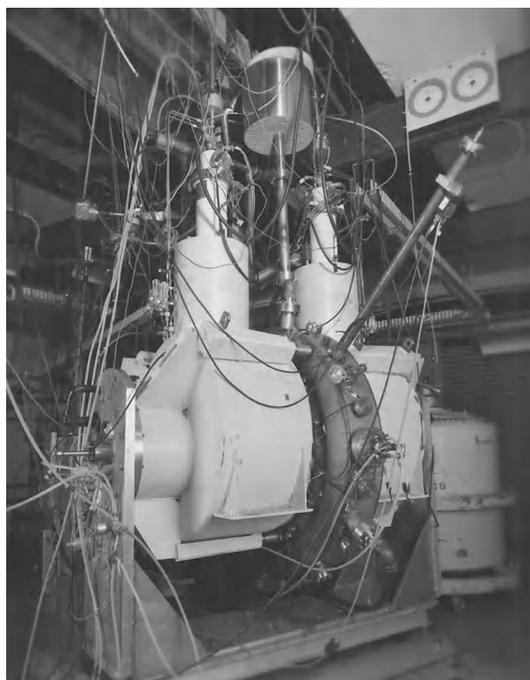


Fig. 12. Self-Collider IV or Migma IV migma-plasma device.

tive instabilities observed in or predicted for thermal plasmas at these and lower densities have been observed in migma-plasma. The figure of merit known as the triple product (temperature  $\times$  confinement time  $\times$  density) reached  $4 \times 10^{14} \text{ keV} \cdot \text{cm}^{-3} \cdot \text{s}$ , approaching that of the best thermal plasma devices at that time.

*Indirect heating via D-T reaction.* B. Coppi proposed achieving the 100-keV temperatures required to ignite reaction (13) in a compact toroidal system by first igniting the D-T fuel and then replacing the tritium with helium-3. He also predicted a so-called second stability region for the tokamak in which  $\beta = 0.3$  could be achieved. A field-reversed mirror configuration (discussed above) has also been proposed to achieve high beta values to ignite reaction (13) via D-T ignition.

**Avoiding radiation losses.** As the plasma density increases to the values that would be required in a reactor (the parameter beta approaches unity), the diamagnetic field produced by the migma ions will push out the magnetic field from in the central zone to the periphery and create a diamagnetic well (Fig. 12). (Such a 90% diamagnetic depression was observed in the 2X-IIB mirror.) The orbit trajectories in the free-field region become linear so that synchrotron radiation is not emitted there. The diamagnetic well, in addition to making the mirror ratio large (which is crucial for ion confinement), suppresses by orders of magnitude the synchrotron radiation loss.

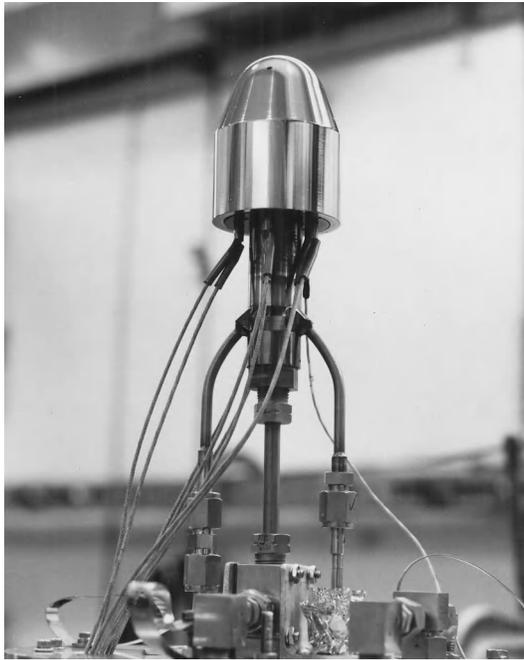
**Reactor simulation.** Extensive studies carried out on supercomputers support the concept of a reactor based on reaction (13). Preliminary results on the so-called ultimate aneutronic reaction (14) in a migma indicate that the theoretical energy balance can be reached, but the reaction is not ignited and must be driven.

Bogdan Maglich

### Muon-Catalyzed Fusion

Nuclear fusion is difficult because of the electrostatic (Coulomb) repulsion between nuclei. In traditional approaches, such as magnetic confinement and inertial confinement fusion, this barrier is overcome by energetic collisions at extraordinarily high temperatures,  $\sim 10^8 \text{ K}$ . In muon-catalyzed fusion, the barrier is effectively removed by binding the nuclei in an exotic molecule with a very short bond length. Fusion then occurs quickly at normal temperatures. (The process is often called cold fusion, not to be confused with the claim of cold fusion in condensed matter.) Muon-catalyzed fusion was discovered experimentally by L. A. Alvarez and collaborators in 1956 but had been hypothesized by F. C. Frank in 1947. Interest in the process as a possible energy source was rekindled by theoretical predictions in the late 1970s and experimental findings in the early 1980s.

In the normal deuterium molecule  $\text{D}_2$  or molecular ion  $\text{D}_2^+$ , with bond lengths  $\sim 10^{-10} \text{ m}$ , fusion is unobservably slow (calculated to take  $\sim 10^{64} \text{ s}$ ). However, when the electron of  $\text{D}_2^+$  is replaced by a negative muon, the bond distance is reduced by approximately the ratio of muon to electron masses



**Fig. 13.** Target and connections used for muon-catalyzed fusion experiments at the Los Alamos Meson Physics Facility. The deuterium-tritium mixture is inside the high-strength stainless steel vessel [outer diameter 3 in. (7.6 cm), interior volume 2.26 in.<sup>3</sup> (37 cm<sup>3</sup>)], and the muons are incident at the rounded nose where the vessel is thinnest. The target is filled through the center stem and can be cooled by liquid helium flowing through the side tubes or heated by the four resistive heaters. The white wire goes to a temperature sensor. In the experiment, the target is surrounded by five neutron detectors. (Idaho National Engineering Laboratory)

( $m^\mu/m_e$ ), which is 207. The muon behaves like a heavy electron except that it decays with an average lifetime  $\tau_0 = 2 \times 10^{-6}$  s. The resulting small molecular ion is usually denoted  $dd^\mu$  (analogously,  $D_2^+$  could be denoted  $dde$ ). Though the average distance be-

tween deuterons is still large compared with the distance where fusion can occur ( $\sim 4 \times 10^{-15}$  m), the nuclei can readily quantum-mechanically tunnel to the shorter distance. For  $dd^\mu$  and  $dt^\mu$  ( $d = \text{deuteron} = {}^2\text{H}$  and  $t = \text{triton} = {}^3\text{H}$ ), the times required for fusion ( $2 \times 10^{-9}$  and  $8 \times 10^{-13}$  s, respectively) are short compared with the lifetime of the muon, though the times required to form these molecules are much longer. Muon-catalyzed fusion of nuclei with charges greater than 1 is impractical since formation of a molecule containing more than one muon is extremely improbable.

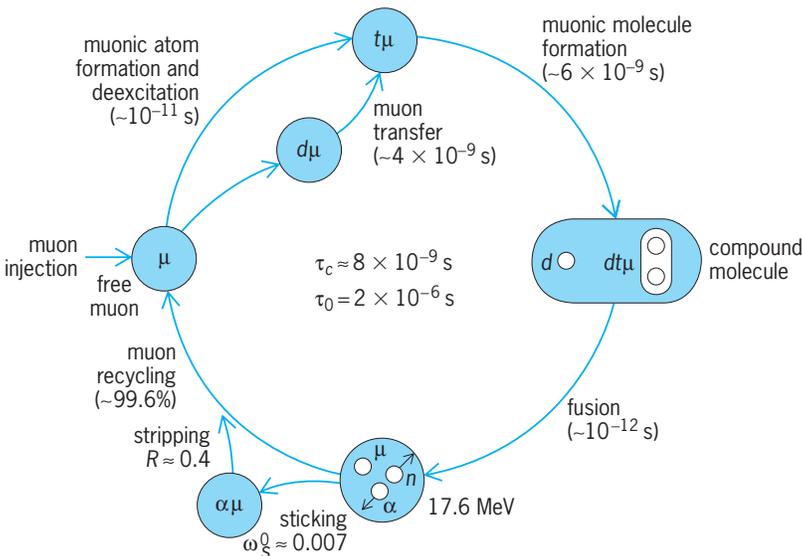
The  $dt^\mu$  molecule is of greatest interest for possible applications because fusion catalysis with it is most rapid and has least losses. A target used for high-pressure gas or liquid  $d-t$  mixtures is shown in **Fig. 13**. Muons from an accelerator are given energies such that they penetrate the steel vessel and are stopped in the fluid. The subsequent fusion neutrons also penetrate the vessel and are detected outside.

**Fusion cycle.** The  $d-t$  muon-catalyzed fusion cycle (**Fig. 14**) starts with a free muon (directly from the accelerator or recycled from a previous fusion) being captured by a deuterium or tritium atom (capture occurs by displacing an electron). If initially captured by deuterium, it transfers to tritium, where it is more strongly bound. At this point an unusual mechanism allows the  $t^\mu$  atom to combine rapidly with a  $D_2$  (or  $D-T$ ) molecule to form  $dt^\mu$ . A resonance is made possible by the fortuitous existence of a state of  $dt^\mu$  so weakly bound that its binding energy can go into vibrational and rotational excitation of the electronic molecule. In this process the positively charged  $dt^\mu$  is so small that it acts essentially as a mass-5 hydrogenic nucleus in the electronic molecule to form a compound molecule (**Fig. 14**). After the  $dt^\mu$  is formed, nuclear fusion rapidly ensues, and an energetic neutron and alpha particle ( ${}^4\text{He}$ ) are emitted. Since the total time required for these events is designated  $\tau_c$ , during the lifetime  $\tau_0$  of the muon there is time for  $\tau_0/\tau_c$  cycles. See MOLECULAR STRUCTURE AND SPECTRA.

**Yield.** Usually the muon is left behind by the fusion products to begin another catalytic cycle. However, sometimes the muon will be lost as a catalyst before it decays. This loss is primarily due to the probability  $\omega_s$  that the muon sticks to a fusion alpha particle to form a muonic helium atomic ion ( $\text{He}-\mu^+$ ). Hence, the average number of fusions (yield)  $Y_n$  that can be catalyzed by a single muon is given by Eq. (15).

$$Y_n = \frac{1}{\frac{\tau_c}{\tau_0} + \omega_s} \quad (15)$$

The optimum value of  $\tau_0/\tau_c$  thus far attained by experiments at temperatures up to 80°F (300 K) is  $\sim 250$  with a 60:40  $d:t$  mixture at liquid density. The measured value of  $\omega_s$  is 0.004; Eq. (15) then gives  $Y_n \approx 125$ . Further improvement would require reduction of both  $\tau_c$  and  $\omega_s$ . Experiments indicate that  $\tau_c$  is shorter at higher temperatures, and theory predicts a minimum at about 1700°F (1200 K). Application of intense electromagnetic fields to enhance



**Fig. 14.** Schematic diagram of the muon-catalyzed  $d-t$  fusion cycle. The approximate times indicated are for liquid-hydrogen density and tritium fraction 0.4. The muon, with lifetime  $2 \times 10^{-6}$  s, can decay at any point in the cycle.  $R$  is the fraction of initially sticking muons that are reactivated by stripping.

stripping of the stuck muon has been proposed to reduce  $\omega_s$  (the indicated value already reflects some collisional stripping that occurs naturally).

**Energy balances.** The energy generated by the number of fusions achieved thus far is not yet sufficient to recover the energy expended to make the muon. Each fusion yields 17.6 MeV of energy so that the total energy release per muon is  $\sim 2200$  MeV. This is considerably greater than 106 MeV, the rest mass energy of a muon, but is still short of the  $\sim 8000$  MeV required to make the muon by using present accelerator technology. Energy production would require either using less energy to make muons or more fusions per muon.

James S. Cohen

**Bibliography.** A. H. Boozer, Physics of magnetically confined plasmas, *Rev. Mod. Phys.*, 76:1071-1141, 2004; K. Brueckner et al., *Inertial Confinement Fusion*, 1992; R. G. Craxton, R. L. McCrory, and J. M. Soures, Progress in laser fusion, *Sci. Amer.*, 255(2): 68-79, August 1986; J. I. Dunderstadt and G. A. Moses, *Inertial Confinement Fusion*, 1982; Fusion special issue, *Euro-physics News*, vol. 29, no. 6, November/December 1998; S. E. Jones, J. Rafelski, and H. J. Mankorst (eds.), *Muon-Catalyzed Fusion*, 1989; B. Maglich and J. Norwood (eds.), Proceedings of the 1st International Symposium on Feasibility of Aneutronic Power, *Nucl. Instrum. Meth.*, A271:1-240, 1988; B. Maglich and J. Norwood (eds.), *Proceedings of the 2d International Symposium on Aneutronic Power; Fusion Technology*, 1990; K. Niu, *Nuclear Fusion*, 1989; J. Rafelski and S. E. Jones, Cold nuclear fusion, *Sci. Amer.*, 257:84-89, 1987; E. Storm et al., *High Gain Inertial Confinement Fusion: Recent Progress and Future Prospects*, Lawrence Livermore Nat. Lab. Rep. UCRL-99383, 1989; J. Wesson, *Tokamaks*, 3d ed., Clarendon Press, Oxford, 2004.

## Nuclear hormone receptors

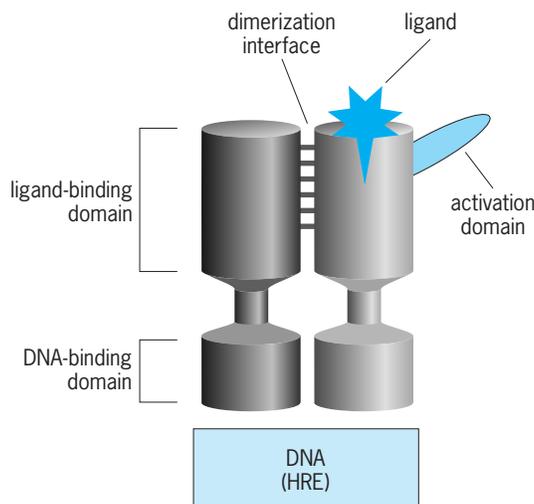
Specialized proteins in the nucleus that are involved in the regulation of gene activity in response to hormones and other signals. It has been estimated that approximately 30,000 genes are expressed in the human genome. The transcription of these genes into messenger ribonucleic acid (mRNA) and their subsequent translation into protein determines the identity and function of each cell in the human body. The balance between sickness and health requires that transcription of these genes be precisely regulated in response to both physiological and environmental changes. A variety of signaling strategies have evolved to coordinate the body's transcriptional response to such events. The nuclear hormone receptors represent one of the largest family of proteins that directly regulate transcription in response to hormones and other ligands (chemical entities that bind to and activate the receptor).

Nuclear hormone receptors have been identified in many species ranging from *Caenorhabditis elegans* (worms) to humans. Nearly 50 distinct nuclear receptor genes have been identified in the human

genome. These include receptors for the steroid hormones (glucocorticoids, mineralocorticoids, estrogens, progestins, and androgens), thyroid hormones, vitamin D, and retinoic acid. An even larger number of nuclear receptor proteins have been found for which no known hormone or ligand has yet been identified. These proteins have been termed orphan receptors. Orphan receptors hold considerable promise, as they provide the first clues toward the identification of novel regulatory molecules and new drug therapies. See GENE; HORMONE.

**Molecular view.** Although each of the nuclear receptors mediates distinct biological effects, their activities at the molecular level are remarkably similar. Nuclear receptors contain two functional modules that characterize this family of proteins: the deoxyribonucleic acid (DNA)-binding domain and the hormone-binding or ligand-binding domain (**Fig. 1**).

**DNA-binding domain.** In order to regulate gene transcription, nuclear receptors must first bind specific DNA sequences adjacent to their target genes. These sequences are referred to as hormone response elements (HREs). The DNA-binding domain is the portion of the receptor that binds to these sequences. As each receptor controls only a specific subset of all genes, the DNA-binding domain must discriminate between its hormone response elements and other closely related sequences. Structural analyses demonstrate that the DNA-binding domain of nuclear receptors folds into a three-dimensional structure containing three  $\alpha$ -helical segments. The first of these helices makes direct contact with nucleotides in the major groove of DNA, while the third helix contacts phosphate groups in the minor groove. These specific interactions help determine the precise DNA sequence that each receptor recognizes. Some receptors bind to DNA by themselves (as monomers); however, the majority bind to DNA in association with



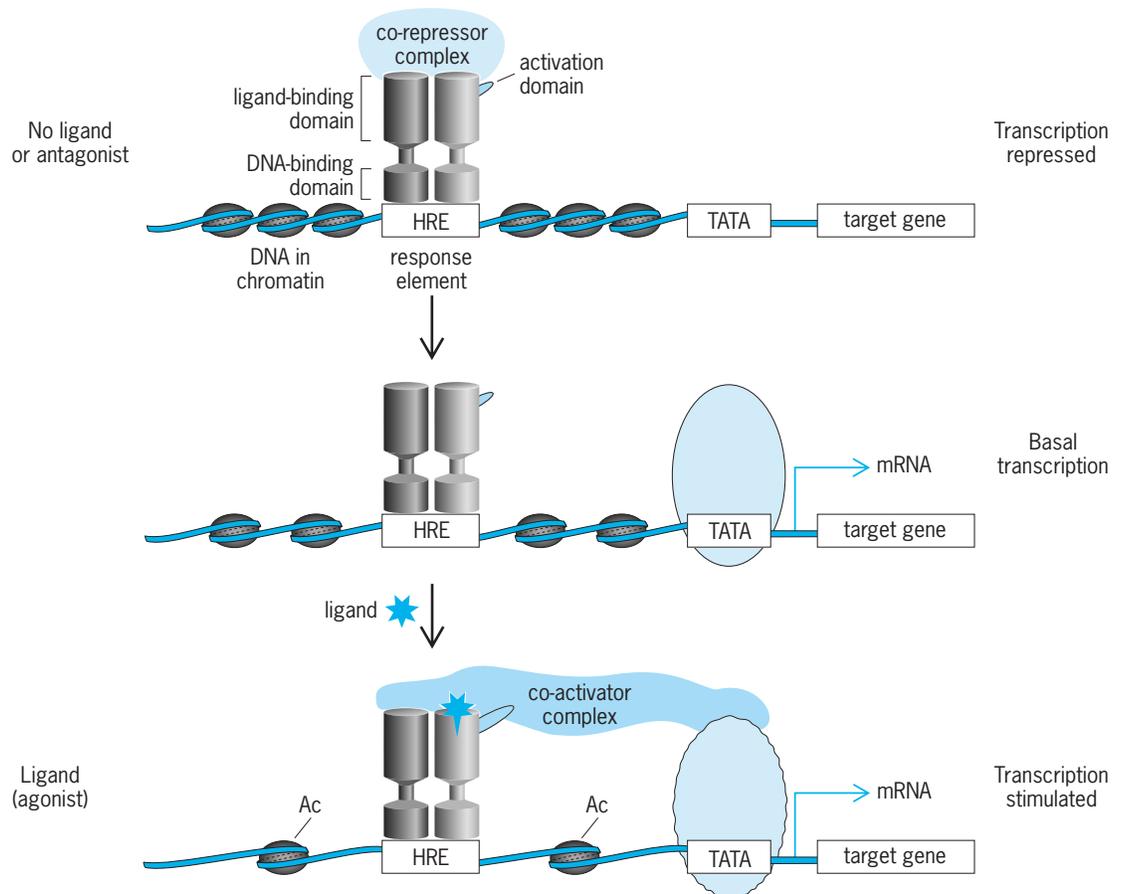
**Fig. 1.** Schematic illustration of a nuclear receptor dimer bound to DNA. The DNA-binding domains are shown contacting the hormone response element (HRE). The receptor at the right is shown bound to its ligand. In addition to binding ligand, the ligand-binding domain contains subdomains for transcriptional activation and dimerization.

a partner receptor (Fig. 1). The formation of such dimers adds another level of selectivity to the DNA-receptor interaction, as each receptor subunit may recognize different DNA sequences. Moreover, the relative orientation of the two DNA-binding domains may vary from one dimer to another. This forces each dimer pair to recognize hormone response elements containing sequence gaps of different sizes in the center of the element. The combination of these features allows individual nuclear receptors to regulate a specific set of genes. *See* DEOXYRIBONUCLEIC ACID (DNA); NUCLEIC ACID.

**Ligand-binding domain.** While the DNA-binding domain selects the target genes, it is the hormone- or ligand-binding domain that determines which hormones activate these genes. The ligand-binding domain contains a cavity or pocket that specifically recognizes the cognate ligand. When a hormonal ligand is bound within this pocket, the receptor undergoes a conformation change that reorients the position of the activation domain (Fig. 1). This reorientation switches the receptor into a transcriptionally active

state. Thus, the ligand-binding domain senses the presence of specific ligands in the body, and in the presence of that ligand it allows the receptor to activate gene transcription. In addition to serving as the hormonal sensor, the ligand-binding domain contains a dimerization domain that determines which receptors can interact with each other.

**Regulation of target genes.** In the nucleus, DNA is wrapped around histone proteins in a structure known as chromatin. DNA that is in a more compact chromatin structure is less transcriptionally active. Indeed, one way in which nuclear receptors activate transcription is to modulate chromatin structure. In the absence of ligand, some receptors (such as the thyroid hormone and retinoic acid receptors) associate with a corepressor complex. This association is facilitated by a groove on the surface of the receptor that binds to corepressor (or coactivator) proteins. Formation of receptor-corepressor complexes is critical, as the corepressor complex exhibits an enzymatic activity that removes acetyl groups from histones (Fig. 2, top). The recruitment of corepressors



**Fig. 2.** Schematic illustration of the mechanism of transcriptional regulation by nuclear receptors. Nuclear receptor dimer binds to its HRE in the absence of ligand [or in the presence of an antagonist] (top). Receptors in this state interact with corepressor complexes that deacetylate histones. Hence, chromatin is compact and transcription is repressed. In the presence of ligand, corepressor is released and transcription can proceed at low or basal levels (middle). Note that many receptors do not recruit corepressors; hence in the absence of ligand their target genes are transcribed at basal levels (middle). When ligand is present, coactivators are recruited to the complex (bottom); chromatin is then acetylated (Ac) and/or methylated (Me), becomes less compact, and the rate of transcription is stimulated (bottom). In the case of the estrogen receptor, target genes are in the basal state in the absence of ligand. Estrogen agonists promote coactivator recruitment and transcriptional activation (bottom), whereas antagonists promote corepressor association and transcriptional repression (top).

allows the receptor to specifically deacetylate the surrounding chromatin, and, since hypoacetylated chromatin is transcriptionally repressed, thereby repress transcription of its target genes.

Ligand binding to the receptor results in the reorientation of the activation domain so that it now sits in the coregulator groove. This displaces the corepressor from the receptor and results in a loss of transcriptional repression (Fig. 2, middle). Genes that are in this state are neither repressed nor activated; they are actively transcribing mRNA at low or basal levels. This basal rate of transcription is determined by RNA polymerase II, which functions as part of a complex of proteins that bind the TATA sequence [in eukaryotes, a short sequence of base pairs that is rich in adenine (A) and thymidine (T) residues and is located 25–30 nucleotides upstream of the transcriptional initiation site] within the promoters of these genes. Thus, one role of ligand is to relieve repression, thereby raising activity to basal levels. Note that many nuclear receptors are unable to recruit corepressors; genes regulated by these receptors are already in the basal state in the absence of ligand (Fig. 2, middle).

In addition to relieving repression, the hormonal ligands activate transcription above and beyond the basal state; the same ligand-induced conformation change that displaces the corepressor also allows the receptor to associate with coactivator complexes (Fig. 2, bottom). One class of coactivator complex functions in a manner opposite to that of corepressors; specifically, it acetylates (Ac) and/or methylates (Me) histones leading to a more transcriptionally active chromatin structure. A second type of coactivator complex directly links the receptor to the RNA polymerase II complex. Taken together, the recruitment of these protein assemblies allows nuclear receptor ligands to simultaneously modify chromatin and increase RNA polymerase II-dependent transcription. *See* CELL NUCLEUS; GENE ACTION; NUCLEOPROTEIN; NUCLEOSOME.

#### **Estrogen receptors, breast cancer, and osteoporosis.**

The estrogen receptor provides an interesting example of how receptors can be exploited to develop selective drug therapies.

Following menopause, estrogen production declines, causing changes in gene transcription that increase the risk of developing osteoporosis (bone loss), atherosclerosis (heart attacks), and hot flashes. These adverse effects can be delayed by the administration of ligands that activate the estrogen receptor (agonists). However, estrogens also promote the proliferation of breast and uterine cancer cells. Thus, for patients at risk for these cancers, it may be more important to inhibit estrogen action. Indeed, synthetic drugs have been identified that bind to estrogen receptors but fail to activate the receptor. These drugs are known as antagonists because they function in a manner opposite to that of the natural hormone, that is, they displace coactivators and/or recruit corepressors. Such estrogen receptor antagonists are currently used for both the treatment and prevention of breast cancer.

If a given estrogen receptor ligand either activates or antagonizes the receptor, the physician must carefully evaluate whether an individual patient is more likely to benefit by having their estrogen receptors “turned-on” or “turned-off.” Fortunately, nuclear receptors are more flexible and can be modulated in ways that are more complex than this simple on/off scenario. The reason for this flexibility is that the ligand does not control receptor activity by itself; transcriptional activity is determined by the type of coregulator protein that associates with the ligand-bound receptor (Fig. 2). An important advance in understanding nuclear receptors has been the realization that these receptors interact with a variety of coactivators and corepressor proteins. Thus, a more useful drug for the estrogen receptor may be one that preferentially recruits a coactivator in bone cells, has no effect in uterine cells, and recruits a corepressor in breast cancer cells. Indeed, several drugs have been identified that can inhibit estrogen action in the breast, prevent osteoporosis by activating estrogen in bone (tamoxifen, raloxifene), and fail to activate transcription and promote cancer in the uterus (raloxifene). These types of ligands are known as selective estrogen receptor modulators (SERMs), as they can activate certain genes while inhibiting others.

Another level of flexibility in the estrogen response has emerged with the identification of a second receptor. Estrogen receptor  $\beta$  binds to similar DNA sequences as estrogen receptor  $\alpha$ , but it is found in different tissues and can bind different synthetic ligands. Additional variations in response are possible, as these two receptors may form  $\alpha/\beta$  heterodimers. Preliminary results suggest that the positive cardiovascular effects of estrogens result from activation of estrogen receptor  $\beta$ , not estrogen receptor  $\alpha$ . *See* BREAST DISORDERS; CANCER (MEDICINE); ESTROGEN; HEART DISORDERS; MENOPAUSE; OSTEOPOROSIS.

**PPAR $\gamma$  and diabetes.** The identification of the orphan receptor peroxisome proliferator-activated receptor gamma (PPAR $\gamma$ ) as a factor required for fat cell formation led to the exciting suggestion that a yet-to-be discovered ligand may control this process. Indeed, a fatty-acid derivative known as 15-deoxy- $\Delta^{12,14}$ -prostaglandin J<sub>2</sub> promotes fat cell formation (adipogenesis) by binding to and activating PPAR $\gamma$ . Activation of PPAR $\gamma$  in diabetic patients results in a shift of lipids from muscle to fat, thus allowing the muscle to function properly. It is now clear that PPAR $\gamma$  promotes adipogenesis by acting as a master regulator of a network of fatty acid storage genes. This network includes genes involved in fatty acid release from serum lipoproteins, uptake of fatty acids into the fat cell, synthesis and uptake of glycerol, and ultimately conversion of fatty acids and glycerol into triglycerides—the major lipid stored in the fat cell. These findings explain how synthetic PPAR $\gamma$  agonists (thiazolidinediones) function as antidiabetic agents in people with type II diabetes. *See* ADIPOSE TISSUE; DIABETES; LIPID METABOLISM.

**Orphan receptors for the future.** Recent data have identified two orphan receptors (PXR and CAR) that

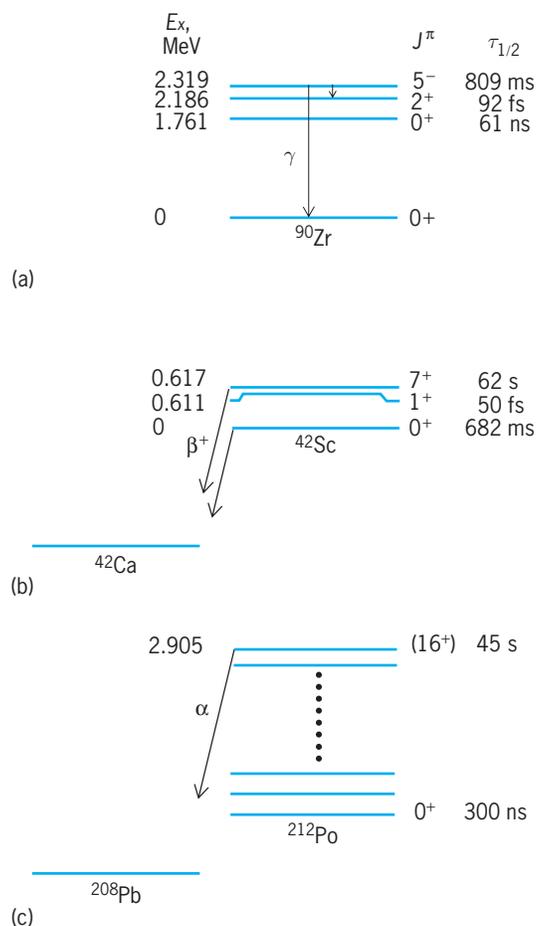
define a new paradigm central to “chemical immunity” (protection from toxic compounds). These receptors regulate genes involved in toxin clearance, but they do so in response to a highly diverse array of endogenous and exogenous toxins. Other orphan receptors have been identified that play critical roles in cholesterol homeostasis, Parkinson’s disease, vision, neuronal development, and other processes. The biological function of many other orphan receptors remains to be elucidated. Thus, these proteins will continue to provide tools for the identification of novel transcriptional regulatory pathways and important drug therapies. See TOXIN. Barry Marc Forman Bibliography. M. W. Draper, The role of selective estrogen receptor modulators (SERMs) in postmenopausal health, *Ann. NY Acad. Sci.*, 997:373–377, 2003; V. Laudet and H. Gronemeyer, *The Nuclear Receptor Factsbook*, Academic Press, 2001; J. M. Olefsky, Nuclear receptor minireview series, *J. Biol. Chem.*, 276(40):36863–36864, 2001.

### Nuclear isomerism

The existence of excited states of atomic nuclei with unusually long lifetimes. A nucleus may exist in an excited quantum state with well-defined excitation energy ( $E_x$ ), spin ( $J$ ), and parity ( $\pi$ ). Such states are unstable and decay, usually by the emission of electromagnetic radiation (gamma rays), to lower excited states or to the ground state of the nucleus. The rate at which this decay takes place is characterized by a half-life ( $\tau_{1/2}$ ), the time in which one-half of a large number of nuclei, each in the same excited state, will have decayed. If the lifetime of a specific excited state is unusually long, compared with the lifetimes of other excited states in the same nucleus, the state is said to be isomeric. The definition of the boundary between isomeric and normal decays is arbitrary; thus the term is used loosely. See EXCITED STATE; PARITY (QUANTUM MECHANICS); SPIN (QUANTUM MECHANICS).

**Spin isomerism.** The predominant decay mode of excited nuclear states is by gamma-ray emission. The rate at which this process occurs is determined largely by the spins, parities, and excitation energies of the decaying state and of those to which it is decaying. In particular, the rate is extremely sensitive to the difference in the spins of initial and final states and to the difference in excitation energies. Both extremely large spin differences and extremely small energy differences can result in a slowing of the gamma-ray emission by many orders of magnitude, resulting in some excited states having unusually long lifetimes and therefore being termed isomeric.

**Occurrence.** Isomeric states have been observed to occur in almost all known nuclei. However, they occur predominantly in regions of nuclei with neutron numbers  $N$  and proton numbers  $Z$  close to the so-called magic numbers at which shell closures occur. This observation is taken as important evidence for the correctness of the nuclear shell model which predicts that high-spin, and therefore iso-



Examples of nuclear isomerism in (a)  $^{90}\text{Zr}$ , (b)  $^{42}\text{Sc}$ , and (c)  $^{212}\text{Po}$ . (After C. M. Lederer and V. S. Shirley, *Table of Isotopes*, 7th ed., John Wiley and Sons, 1978)

meric, states should occur at quite low excitation energies in such nuclei. See MAGIC NUMBERS.

**Examples.** Three examples of isomeric states are shown in the illustration. In the case of  $^{90}\text{Zr}$  (illus. a), the 2.319-MeV,  $J^\pi = 5^-$  state has a half-life of 809 milliseconds, compared to the much shorter lifetimes of 93 femtoseconds, and 61 nanoseconds for the 2.186-MeV and 1.761-MeV states, respectively. The spin difference of 5 between the 2.319-MeV state and the ground and 1.761  $J^\pi = 0^+$  states, together with the spin difference of 3 and small energy difference between the 2.319-MeV state and the 2.186-MeV,  $J^\pi = 2^+$  state, produce this long lifetime.

For  $^{42}\text{Sc}$  (illus. b), the gamma decay of the  $J^\pi = 7^+$  state at 0.617 MeV is so retarded by the large spin changes involved that the intrinsically much slower  $\beta^+$  decay process can take place, resulting in a half-life of 62 s.

Yet another process takes place in the case of the high-spin isomer in  $^{212}\text{Po}$  (illus. c), which decays by alpha-particle emission with a half-life of 45 s rather than gamma decays.

The common feature of all these examples is the slowing of the gamma-ray emission process due to the high spin of the isomeric state.

**Other mechanisms.** Not all isomers are the spin isomers described above. Two other types of isomers

have been identified. The first of these arises from the fact that some excited nuclear states represent a drastic change in shape of the nucleus from the shape of the ground state. In many cases this extremely deformed shape displays unusual stability, and states with this shape are therefore isomeric. A particularly important class of these shape isomers is observed in the decay of heavy nuclei by fission, and the study of such fission isomers has been the subject of intensive effort. The possibility that nuclei may undergo sudden changes of shape at high rotational velocities has spurred searches for isomers with extremely high spin which may also be termed shape isomers. See NUCLEAR FISSION.

A more esoteric form of isomer has also been observed, the so-called pairing isomer, which results from differences in the microscopic motions of the constituent nucleons in the nucleus. A state of this type has a quite different character from the ground state of the nucleus, and is therefore also termed isomeric. See NUCLEAR STRUCTURE. Russell Betts

Bibliography. H. A. Enge and R. P. Redwine, *Introduction to Nuclear Physics*, 2d ed., 1995; K. S. Krane, *Introductory Nuclear Physics*, 1987; C. M. Lederer and V. S. Shirley, *Table of Isotopes*, 7th ed., 1978; K. Siegbahn (ed.), *Alpha, Beta and Gamma-Ray Spectroscopy*, vols. 1 and 2, 1965.

## Nuclear magnetic resonance (NMR)

A phenomenon exhibited when atomic nuclei in a static magnetic field absorb energy from a radio-frequency field of certain characteristic frequencies. Nuclear magnetic resonance is a powerful analytical tool for the characterization of molecular structure, quantitative analysis, and the examination of dynamic processes. It is based on quantized spectral transitions between nuclear Zeeman levels of stable isotopes, and is unrelated to radioactivity. See ZEE-MAN EFFECT.

The format of nuclear magnetic resonance data is a spectrum that contains peaks referred to as resonances. The resonance of an isotope is distinguished by the transition frequency of the nucleus. The intensity of the resonance is directly proportional to the number of nuclei that produce the signal. Although the majority of nuclear magnetic resonance spectra are measured for samples in solution or as neat liquids, it is possible to measure nuclear magnetic resonance spectra of solid samples. Nuclear magnetic resonance is a nondestructive technique that can be used to measure spectra of cells and living organisms.

**Nuclear magnetic properties.** In order to understand the basic phenomena that give rise to the nuclear magnetic resonance spectrum, it is necessary to examine the magnetic resonance of nuclei. The nuclei of most atoms possess an intrinsic nuclear angular momentum. The classical picture of nuclear angular momentum is a spherical nucleus rotating about an axis in a manner analogous to the daily rotation of the Earth. Nuclear angular momentum, like

most other atomic quantities, can be expressed as a series of quantized levels.

The transitions that give rise to the nuclear magnetic resonance spectrum are produced when the nuclei are placed in a static magnetic field. The external or applied magnetic field defines a geometric axis, denoted as the  $z$  axis. The external magnetic field orients the  $z$  components of nuclear angular momentum and the magnetic moment with respect to the  $z$  axis. The nuclear magnetic resonance spectrum is produced by spectral transitions between different spin states and is therefore dependent on the value of the nuclear spin. Nuclei for which the nuclear spin value is 0 have a magnetic spin-state value of 0. Therefore, these nuclei do not give rise to a nuclear magnetic resonance spectrum under any circumstances. Many important elements in chemistry have zero values of nuclear spin and are inherently nuclear magnetic resonance inactive, including the most abundant isotopes of carbon-12, oxygen-16, and sulfur-32. Nuclear magnetic resonance spectra are measured most often for nuclei that have nuclear spin values of  $1/2$ . Chemically important spin- $1/2$  nuclei include the isotopes hydrogen-1, phosphorus-31, fluorine-19, carbon-13, nitrogen-15, silicon-29, iron-57, selenium-77, cadmium-113, silver-107, platinum-195, and mercury-199. Nuclear magnetic resonance spectra of nuclei that have nuclear spin values of greater than  $1/2$ , known as quadrupolar nuclei, can be measured; but the measurements are complicated by faster rates of nuclear relaxation. Nuclei that fall into this class include the isotopes hydrogen-2, lithium-6, nitrogen-14, oxygen-17, boron-11, chlorine-35, sodium-23, aluminum-27, and sulfur-33.

The transition frequency in nuclear magnetic resonance depends on the energy difference of the spin states. The intensity of the resonance is dependent on the population difference of the two spin states, which in turn depends directly on the magnitude of the energy difference. For spin- $1/2$  nuclei,  $\Delta E$  is the difference in energy between the  $+1/2$  and  $-1/2$  values of the magnetic spin state. In contrast to other spectroscopic methods, the nuclear magnetic resonance frequency is variable, depending on the strength of the magnet used to measure the spectrum. Larger magnetic fields produce a greater difference in energy between the spin states. This translates into a larger difference in the population of the spin states and therefore a more intense resonance in the resulting nuclear magnetic resonance spectrum.

**Pulsed-Fourier-transform NMR spectroscopy.** Early nuclear magnetic resonance spectrometers were continuous-wave instruments that scanned through the magnetic field until the resonance frequency of each group of nuclei was reached. Continuous-wave spectrometers have virtually disappeared and have been replaced by pulsed-Fourier-transform instruments. In the pulsed-Fourier-transform experiment, a secondary oscillating magnetic field perpendicular to the applied magnetic field is produced by the application of a pulse of radio-frequency radiation.

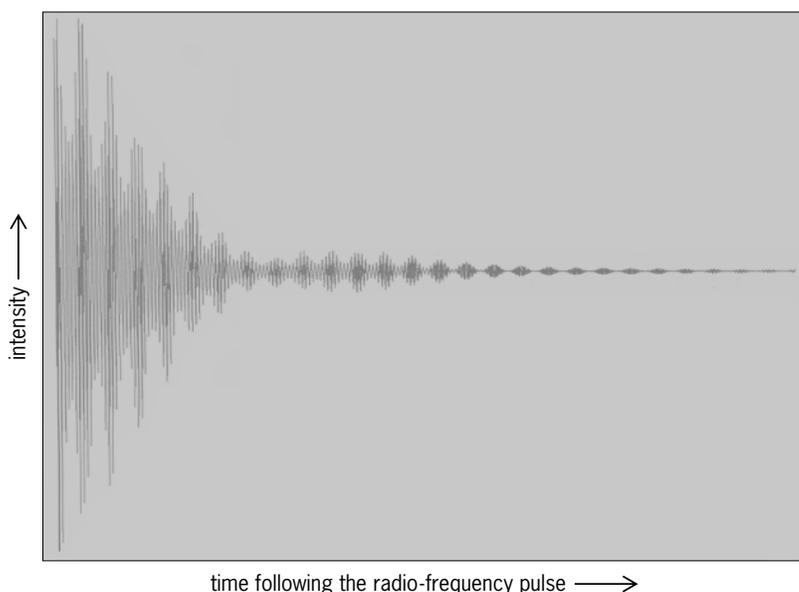


Fig. 1. Proton free induction decay measured for a solution of ethyl acetate dissolved in deuterated chloroform. The free induction decay is the detected signal which contains all of the information of the nuclear magnetic resonance spectrum, including the resonance frequency, intensity, and linewidth of each chemically distinct proton in the sample.

Absorption of the radio-frequency energy involves flipping the magnetic moment, effectively inverting the population of the spin states. The nuclear magnetic resonance signal generated as result of the radio-frequency pulse is a transient response called the free induction decay (Fig. 1). The response is transient because the nuclei return to their equilibrium populations usually within a few seconds by nonradiative processes. The intensity of the free induction decay is a maximum immediately following the radio-frequency pulse and decays logarithmically as a function of time. The nuclear magnetic resonance spectrum (Fig. 2) is produced by Fourier transformation of the free induction decay. See FOURIER SERIES AND TRANSFORMS; INTEGRAL TRANSFORM.

**Chemical shift.** Individual nuclei of the same isotope in a molecule have transition frequencies that

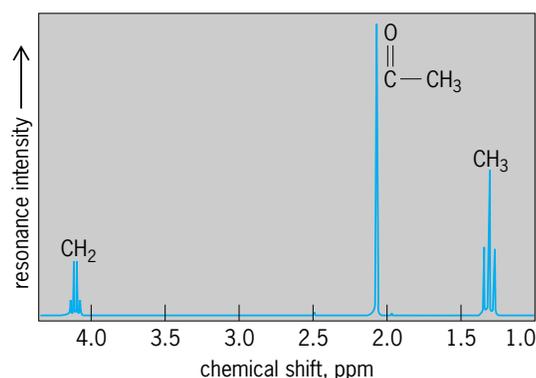


Fig. 2. Proton spectrum of ethyl acetate ( $\text{CH}_3\text{COOCH}_2\text{CH}_3$ ) obtained by Fourier transformation of the free induction decay shown in Fig. 1. Chemical shift is relative to the protons of tetramethylsilane. The integrated intensity of the resonances corresponds to the relative number of protons which give rise to each resonance.

differ depending on their chemical environment. This phenomenon, called chemical shift, occurs because the effective magnetic field at a particular nucleus in a molecule is less than the applied magnetic field due to shielding by electrons.

An example of resonance chemical shift is observed in the  $^1\text{H}$  nuclear magnetic resonance spectrum of ethyl acetate ( $\text{CH}_3\text{COOCH}_2\text{CH}_3$ ; Fig. 2). Resonances at several frequencies are observed in this spectrum. The methylene ( $\text{CH}_2$ ) protons are affected by the electron-withdrawing oxygen atoms of the neighboring ester group, and as a result the chemical shift of the methylene proton resonance is significantly different from the chemical shift of the resonances of the protons of the methyl ( $\text{CH}_3$ ) groups of ethyl acetate. The two methyl groups are in different chemical environments and therefore give rise to resonances that have different chemical shifts. Because of the dependence of the transition frequency of a nucleus on its chemical environment, chemical shift is diagnostic of the functional group containing the nucleus of interest. Nuclear magnetic resonance spectroscopy is a frequently employed tool in chemical synthesis studies, because the nuclear magnetic resonance spectrum can confirm the chemical structure of a synthetic product.

In most nuclear magnetic resonance spectra, rather than frequency units, the spectra are plotted in units of chemical shift expressed as part per million (ppm). The ppm scale calculates the ratio of the resonance frequency in hertz (Hz) to the Larmor frequency of the nucleus at the magnetic field strength of the measurement. The ppm scale allows direct comparison of nuclear magnetic resonance spectra acquired by using magnets of differing field strength. This unit of nuclear magnetic resonance chemical shift should not be confused with the concentration units of ppm (mg/kg), often referred to by analytical chemists in the field of trace analysis.

The chemical shift (either in hertz or ppm) of a resonance is assigned relative to the chemical shift of a standard reference material. The nuclear magnetic resonance community has agreed to arbitrarily set the chemical shift of certain standard compounds to 0 ppm. For  $^1\text{H}$  and  $^{13}\text{C}$  nuclear magnetic resonance, the accepted standard is tetramethylsilane, which is defined to have a chemical shift of 0 ppm. However, any molecule with a resonance frequency in the appropriate chemical shift region of the spectrum that does not overlap with the resonances of the sample can be employed as a chemical shift reference. The use of a chemical shift reference compound other than the accepted standard is particularly common for nuclei that have a very large chemical shift range such as fluorine-19 or selenium-77, since it may be impossible to measure the spectrum of the accepted chemical shift reference compound and the analyte of interest simultaneously because of their very different frequencies.

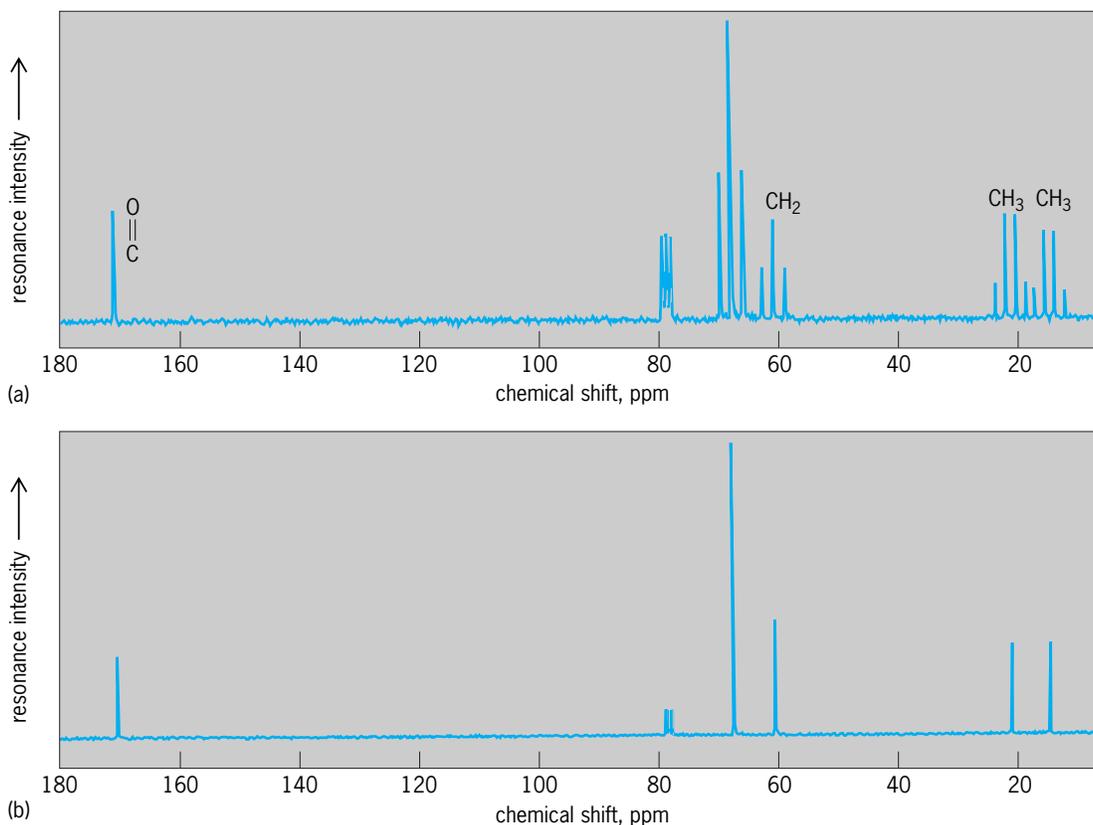
**Spin-spin coupling.** In addition to the differences in chemical shift, resonances in a given spectrum, for example for ethyl acetate (Fig. 2), also differ in the number of signals composing the resonance

detected for each group of protons. Instead of one resonance resulting from a single transition between two spin states, the methylene group ( $\text{CH}_2$ ) resonance is actually a quartet composed of four lines. Similarly, although the acetate  $\text{CH}_3$  resonance is a single resonance, the ethylene group ( $\text{CH}_2$ ) resonance consists of three lines. This splitting of some resonances, called spin-spin or scalar coupling, arises from interactions between nuclei through their bonding electrons rather than through space. Scalar coupling is a short-range interaction that usually occurs for nuclei separated by one to three bonds.

**Spin decoupling.** The effects of spin-spin coupling can be removed from the spectrum by application of a low-strength magnetic field to one of the coupled spins. The effect of this secondary field is to equalize the population of the coupled transitions and to remove the effects of spin-spin coupling. Decoupling can be homonuclear or heteronuclear. Homonuclear decoupling is useful for assigning the resonances of coupled spin systems in the spectrum of a compound or complex mixture. For example, in ethyl acetate, decoupling at the frequency of the methylene group ( $\text{CH}_2$ ) protons would remove the effects of spin-spin coupling, collapsing the ethylene group ( $\text{CH}_2$ ) resonance into a singlet and confirming the resonance assignments.

The previous discussion of chemical shift and spin-spin coupling has focused on  $^1\text{H}$  nuclear magnetic resonance spectroscopy. When detecting nuclei other than protons ( $^1\text{H}$ ), for example carbon-13 ( $^{13}\text{C}$ ), the effects of heteronuclear spin-spin coupling between the protons and the carbon-13 can be removed by decoupling the protons during acquisition of the carbon free induction decay. This is called heteronuclear decoupling, and it is commonly used to simplify nuclear magnetic resonance spectra. The majority (98.9%) of the protons are bound to  $^{12}\text{C}$  nuclei, which have zero spin. However, 1.1% of the protons are bound to  $^{13}\text{C}$  nuclei. The resonances of protons bound to  $^{13}\text{C}$  nuclei are split into a doublet with a coupling constant,  $J_{\text{CH}}$ , of 120–160 Hz. These weak proton-coupled  $^{13}\text{C}$  resonances can be detected in proton nuclear magnetic resonance spectra when the signal-to-noise ratio is sufficiently high. In the  $^{13}\text{C}$  nuclear magnetic resonance spectrum, the  $^{13}\text{C}$  resonances are split into multiplets reflecting the number of directly bound protons.

In the proton-coupled  $^{13}\text{C}$  spectrum of ethyl acetate (Fig. 3a), the quartet resonances of the two methyl carbons occur at 13.9 and 20.3 ppm. The  $\text{CH}_2$  carbon resonance occurring at 59.9 ppm is split into a triplet by the two attached protons. This spectrum indicates that the carboxyl carbon occurs at 170.2 ppm; it is a singlet since it is not directly



**Fig. 3. Spin decoupling.** (a) The proton-coupled natural abundance carbon-13 ( $^{13}\text{C}$ ) spectrum of ethyl acetate. The splitting of the  $^{13}\text{C}$  resonances arises from spin-spin coupling to the protons bound to their respective carbon atoms. (b) The  $^{13}\text{C}$  spectrum of ethyl acetate measured with proton decoupling. With the effects of heteronuclear spin-spin coupling removed, the  $^{13}\text{C}$  spectrum is greatly simplified and the signal-to-noise ratio is increased relative to the proton coupled spectrum. Spectra of a and b are plotted at different vertical scales.

bound to any protons. Notably, no  $^{13}\text{C}$ - $^{13}\text{C}$  scalar coupling is observed in such a spectrum because of the low probability that two  $^{13}\text{C}$  nuclei will be directly bonded at natural abundance levels.

In the proton-decoupled  $^{13}\text{C}$  nuclear magnetic resonance spectrum of ethyl acetate (Fig. 3*b*), the heteronuclear spin-spin coupling between the protons and the carbon nuclei can be eliminated by decoupling over the entire proton chemical shift region during acquisition of the free induction decay. This type of decoupling is known as broadband decoupling, because a wide band of proton frequencies are decoupled. The only  $^{13}\text{C}$  resonances that are unaffected by proton decoupling are those not directly bonded to protons. The proton-decoupled  $^{13}\text{C}$  spectrum (Fig. 3*b*) is much simpler than the proton-coupled spectrum (Fig. 3*a*) since it consists of all singlets. In addition to simplifying the spectrum, proton decoupling produces a higher signal-to-noise spectrum, because all the  $^{13}\text{C}$  resonance intensity is concentrated into a single resonance rather than spread over a multiplet.

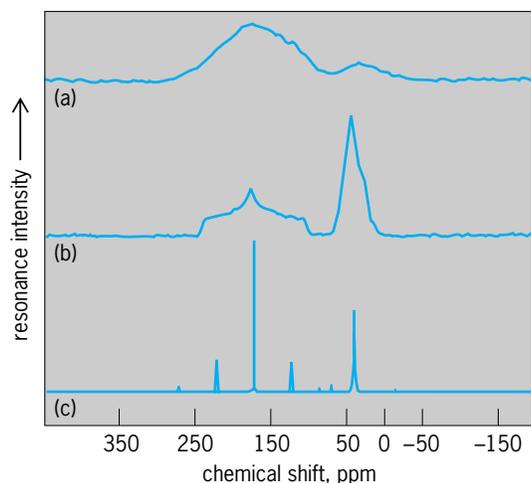
**Solid-state NMR spectroscopy.** Nuclear magnetic resonance spectra of solid-phase materials have the disadvantage that the resonances are very broad, a feature observed in the  $^{13}\text{C}$  spectrum of solid glycine (Fig. 4*a*). Since this compound is structurally very simple, the resonances of the individual carbon atoms can be distinguished in its spectrum, even though the individual resonances are quite broad. In the liquid phase, dipolar coupling between nuclei on different molecules is not observed because of the rapid reorientation of the molecules. In the solid phase, where molecular reorientation does not readily occur, direct dipolar magnetic interactions be-

tween the nuclei of adjacent molecules are relatively large and contribute significantly to the linewidth of resonances in the nuclear magnetic resonance spectra of solids. The removal of the effects of dipolar coupling by proton decoupling results in much narrower resonances (Fig. 4*b*).

An additional contribution to the broad resonances obtained for solid-phase samples comes from chemical shift anisotropy, an orientation dependence of the chemical shift. Chemical shift anisotropy is not usually a problem in nuclear magnetic resonance spectra of liquid-phase samples, because its effects are averaged by the rapid rotation of the molecules. In solids, both the dipolar coupling and the chemical shift anisotropy have an angular dependence. To minimize the effects of dipolar interactions and chemical shift anisotropy, solid-state samples are spun rapidly (that is, 3 kHz) at the angle  $55^\circ 44'$ , known as the magic angle. Dramatic improvement is obtained by using both proton decoupling and magic angle spinning to narrow the  $^{13}\text{C}$  resonances (Fig. 4*c*). Solid-state nuclear magnetic resonance is particularly useful for measuring spectra of insoluble materials such as synthetic and natural polymers. Deuterium (hydrogen-2) nuclear magnetic resonance is widely used in the solid state to measure nuclear magnetic resonance spectra of biological membranes. Other applications of solid-state nuclear magnetic resonance have been the use of silicon-29 and aluminum-27 nuclear magnetic resonance to study zeolites and the measurement of nuclear magnetic resonance spectra of xenon-129 adsorbed on surfaces. *See* DEUTERIUM.

**Quantitative analysis.** Quantitative analysis using nuclear magnetic resonance makes use of the direct relationship between the number of nuclei and the resonance intensity. Complex mixtures can be analyzed quantitatively by nuclear magnetic resonance without separating the components of the mixture. Most spectroscopic techniques require that a standard of the pure analyte be used for calibration, because the response of the instrument is different for each analyte. In nuclear magnetic resonance spectroscopy, the integrated intensity of a resonance is directly proportional to the concentration of nuclei that give rise to the resonance. Therefore, when the experimental parameters (particularly the repetition time) are carefully chosen, concentrations can be determined relative to an internal standard that may be chemically unrelated to the compound of interest. This is a particular advantage when pure standards of the material being analyzed are not available. *See* QUANTITATIVE CHEMICAL ANALYSIS.

**Qualitative analysis.** Nuclear magnetic resonance spectroscopy is one of the best methods available for the structural analysis of molecules. Integrals indicate the relative numbers of nuclei that give rise to the resonances in the nuclear magnetic resonance spectrum. Coupled spin systems can be elucidated by spin-decoupling experiments in which individual multiplets are irradiated in succession. Although these methods have been used to successfully assign the structure of numerous molecules, there is a



**Fig. 4. Solid-state nuclear magnetic resonance spectroscopy.** Narrowing the resonances in nuclear magnetic resonance is accompanied by a large increase in the resonance intensity; hence the sensitivity of the measurement. Vertical axes are plotted at different scales. (a) The static carbon-13 ( $^{13}\text{C}$ ) spectrum of solid glycine ( $+\text{NH}_3-\text{CH}_2-\text{COO}^-$ ). (b) The  $^{13}\text{C}$  spectrum of solid glycine measured with proton decoupling to remove the effects of dipolar coupling between the proton and carbon nuclei of adjacent molecules. (c) The  $^{13}\text{C}$  spectrum of solid glycine measured by using both proton decoupling and magic angle spinning. (Martine Zillox, Bruker Instruments, Inc.)

fundamental limitation in the complexity of the nuclear magnetic resonance spectrum that is amenable to analysis by these simple one-dimensional techniques. See QUALITATIVE CHEMICAL ANALYSIS.

**Two-dimensional NMR experiments.** The introduction of two-dimensional nuclear magnetic resonance experiments has greatly expanded both the complexity of molecules that can be analyzed by nuclear magnetic resonance spectroscopy and the repertoire of experiments available to the spectroscopist. In general, two-dimensional nuclear magnetic resonance experiments can be analyzed as containing four separate periods: preparation, evolution, mixing, and detection. The preparation period usually contains a delay during which the magnetization is allowed to recover from the previous acquisition, followed by a pulse that initiates the experiment. In most two-dimensional experiments, the first pulse is a  $90^\circ$  pulse that generates transverse magnetization. The transverse magnetization is allowed to precess during the evolution period, which consists of an incremented delay period. During the evolution period, the magnetization is encoded with the information of the two-dimensional experiment. In some experiments, a mixing period consisting of a pulse, a delay period, or some combination of the two is necessary to convert the encoded information to a detectable form. The free induction decay is acquired during the detection period. This approach has been extended to three- and four-dimensional nuclear magnetic resonance properties experiments by introduction of additional evolution and mixing periods.

Two-dimensional nuclear magnetic resonance spectra are acquired as a series of free induction decays collected at incremented values of the incremented delay period. The two-dimensional spectrum is produced by double Fourier transformation in the dimensions of both the incremented delay period and the detection period. Two-dimensional nuclear magnetic resonance spectra are defined by two nuclear magnetic resonance frequency axes and actually contain a third dimension, resonance intensity. By using the available arsenal of multidimensional nuclear magnetic resonance experiments, three-dimensional solution structures of complex molecules such as proteins have been determined that are comparable in quality to those produced by x-ray crystallography.

The utility of multidimensional nuclear magnetic resonance experiments for structural elucidation is demonstrated for a relatively simple molecule, thymidine. This discussion is not meant to be a complete description, but an illustration of the power of this methodology. Often, the purpose of the nuclear magnetic resonance study is determination of the conformation of molecular structure. However, proton nuclear magnetic resonance spectra can be relatively complicated even for a simple molecule such as thymidine (Fig. 5). Therefore, two-dimensional nuclear magnetic resonance experiments are employed to assign the resonances in the nuclear magnetic resonance spectrum and to confirm the structure of the molecule.

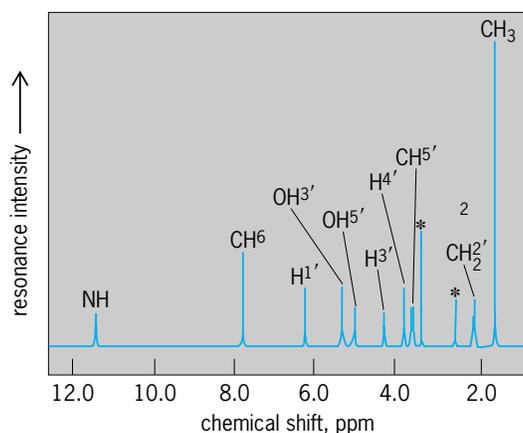


Fig. 5. Proton nuclear magnetic resonance spectrum of thymidine in deuterated dimethylsulfoxide. The thymidine protons giving rise to each resonance are indicated by the labels above each resonance. The two resonances marked \* are not due to thymidine protons but result from impurities in the solution. Chemical shifts are relative to dimethylsilylpropionate, sodium salt.

The two-dimensional nuclear magnetic resonance experiment, correlated spectroscopy (COSY), produces a spectrum (Fig. 6) that identifies protons that are spin-spin coupled. Since spin-spin coupling is produced by through-bond interactions of the nuclei, the correlated spectroscopy spectrum identifies protons on adjacent carbon atoms. The two frequency dimensions of the COSY spectrum are proton chemical shift. The presentation format of the COSY spectrum is a contour plot, in which relative resonance intensity is distinguished by the number of contours. A contour plot is analogous to the depiction of a mountain range in a topographical map (Fig. 6).

The COSY spectrum is characterized by a diagonal that runs from the bottom left-hand corner of

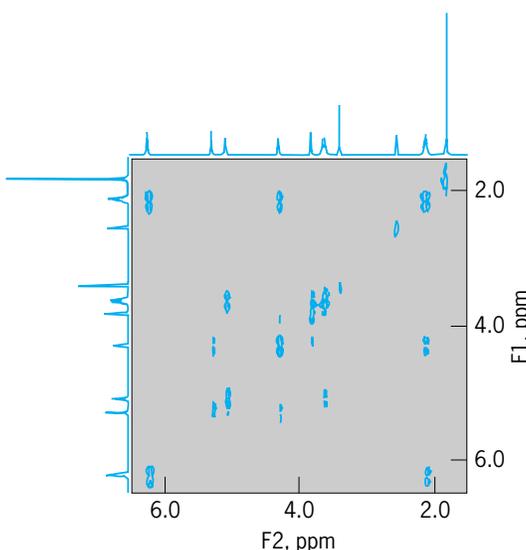


Fig. 6. Portion of the contour plot of the proton correlated spectroscopy (COSY) spectrum of thymidine. The off-diagonal elements or cross-peaks connect resonances of scalar coupled protons. The one-dimensional proton nuclear magnetic resonance spectrum of thymidine is plotted along each frequency axis (F1 and F2).

the spectrum to the upper right-hand corner; this diagonal corresponds to the one-dimensional proton spectrum. The information in a COSY spectrum is provided by the off-diagonal elements or cross-peaks. In a correlated spectroscopy experiment, during the evolution period the magnetization evolves under the effects of homonuclear spin-spin coupling. The cross-peaks in the COSY spectrum connect diagonal resonances for protons that are spin-spin coupled. If the identity of one resonance is known, the other resonances of the molecule can be assigned by tracing through the coupling patterns defined by the cross-peaks.

Cynthia K. Larive

**Bibliography.** F. A. Bovey, *Nuclear Magnetic Resonance Spectroscopy*, 1986; H. Friebolin, *Basic One- and Two-Dimensional Nuclear Magnetic Resonance Spectroscopy*, 2d ed., 1993; G. E. Martin and A. S. Zektzer, *Two-Dimensional Nuclear Magnetic Resonance Methods for Establishing Molecular Connectivity: A Chemist's Guide to Experiment Selection, Performance and Interpretation*, 1988; J. K. M. Saunders and B. K. Hunter, *Modern Nuclear Magnetic Resonance Spectroscopy: A Guide for Chemists*, 2d ed., 1993; K. Wüthrich, The development of nuclear magnetic resonance spectroscopy as a technique for protein structure determination, *Acc. Chem. Res.*, 22:36-44, 1986.

## Nuclear medicine

A subspecialty of medicine based on the use of radioactive substances in medical diagnosis, treatment, and research. Cyclotron-produced radioactive materials were introduced in the early 1930s, but the invention of the nuclear reactor during World War II made carbon-14, hydrogen-3, iodine-131, and later technetium-99m available in large quantities. Today most biomedical research and the care of many patients depend on the use of radioactive materials.

The most widely used radionuclides are technetium-99m, iodine-123, carbon-11, and fluorine-18. The latter two require a cyclotron near the site of radiotracer production because of their rapid radioactive decay (carbon-11 decays with a half-life of 20 min, fluorine-18 with a half-life of 110 min). The short half-life of the radioactive tracers makes it possible to administer the radiotracers to individuals without the harmful effects of radiation.

**Tracer principle.** The most fundamental principle in nuclear medicine is the tracer principle, invented in 1912 by Nobel laureate G. Hevesy, who found that radioactive elements had identical chemical properties to the nonradioactive form and therefore could be used to trace chemical behavior in solutions or in the body. One of the important consequences of the use of tracers was to establish the principle of the dynamic state of body constituents. Prior to the development of radioactive tracer methods, the only way to study biochemical processes within the body was to measure the input and output of dietary constituents and examine concentrations of molecules at autopsy. With radioactive tracers, the movement of

labeled molecules could be followed from the processes within organs, to excretion. In essence, the use of radioactive tracers makes it possible to study the biochemistry within the various organs of the human body. *See* RADIOACTIVE TRACER.

According to the principle of the constancy of the internal environment, the concentration of chemical constituents in body fluids is usually kept within a very narrow range, and disturbances of these values result in disease. This concept has been one of the foundations of modern biochemistry. Nuclear medicine makes it possible to examine regional physiology and biochemistry in ways that at times surpass the perception of surgeons during an operation or pathologists during an autopsy. Imaging methods make it possible to measure regional as well as overall organ function, and to portray the results in the form of functional or biochemical pictures of the body in health and in disease. Such pictures enhance the information about structure that is obtained by other imaging methods, such as computerized tomography (CT) or magnetic resonance imaging (MRI), often providing unique, objective evidence of disease long before structural changes are seen. *See* COMPUTERIZED TOMOGRAPHY; NUCLEAR MAGNETIC RESONANCE (NMR).

The physiological and biochemical orientation of nuclear medicine provides a better approach to understanding disease. The body is viewed as a complex array of coordinated chemical and physical processes that can become impaired before signs of disease develop. This has led to the concept of chemical reserve. It is now known that chemical abnormalities, such as a reduced rate of glucose metabolism in Huntington's disease or a marked deficiency of a neurotransmitter such as dopamine in Parkinson's disease, can occur long before the onset of symptoms. This makes possible the detection of disease far earlier than symptoms or structural abnormalities can. In focal epilepsy, for example, chemical abnormalities are often detectable before structural changes occur. *See* MOLECULAR PATHOLOGY.

**Diagnosis.** In a typical examination, a radioactive molecule is injected into an arm vein, and its distribution at specific time periods afterward is imaged in certain organs of the body or in the entire body. The images are created by measuring the gamma-ray photons emitted from the organs or regions of interest within the body. Nuclear medicine imaging procedures differ from ordinary x-rays in that the gamma rays are emitted from the body rather than transmitted across the body, as in the case of x-rays. As in most modern imaging, the principle of tomography is used, that is, the person is viewed by radiation detectors surrounding the body, or by rotation of a gamma camera around the body. Such procedures include single-photon emission computed tomography (SPECT), based on the use of iodine-123 or technetium-99m, and positron emission tomography (PET), based on the use of carbon-11 and fluorine-18. The latter two elements are short-lived (carbon-11 half-life is 20 min; fluorine-18 half-life is 110 min) and therefore must be produced near the site where

the studies are performed. See RADIOISOTOPE (BIOLOGY).

The nature of the injected material, called a radiopharmaceutical, determines the information that will be obtained. In most cases, either blood flow or biochemical processes within an organ or part of an organ are examined. The essence of a nuclear medicine examination is measurement of the regional chemistry of a living human body.

Examples of commonly used procedures in nuclear medicine using the tracer principle are examination of the blood flow to regional heart muscle with thallium-201- or technetium-99m-labeled radiopharmaceuticals, imaging the regional movements of the ventricles of the heart, detection of blood clots in the lung or impaired lung function, detection of breast and prostate tumors, detection of acute inflammation of the gallbladder, and examination of practically all organs of the body.

Positron emission tomography and single-photon emission computed tomography are used to study regional blood flow, substrate metabolism, and chemical information transfer. In the last category, positron emission tomography has been used to establish the biological basis of neurological and psychiatric disorders, and may help improve the drug treatment of depression, Parkinson's disease, epilepsy, tardive dyskinesia, Alzheimer's disease, and substance abuse. Advances in PET and SPECT and the use of simple detector systems may help in the monitoring of the response of an individual to drug treatment, and perhaps reduce the incidence of side effects. These methods can also provide information on the physiologic severity of coronary stenosis and myocardial viability, especially after thrombolytic therapy or other forms of treatment.

One of the most important areas of research in nuclear medicine is the study of recognition sites, that is, the mechanisms by which cells communicate with each other. For example, some tumors possess recognition sites, such as estrogen receptors.

Another area is in assessment of the availability of receptors that are the primary site of action of many medications. Specific effects of a drug begin by the binding of the drug to specific chemical receptors on specific cells of the body. For example, the finding that Parkinson's disease involves the neurotransmitter dopamine led to the development of L-DOPA treatment, which relieves many of the symptoms of the disease. Measurement of abnormalities of pre-dopaminergic neurons makes it possible to characterize the abnormalities of pre-synaptic neurons in individuals with Parkinson's disease early in their illness at a time when the progress of the disease might be halted.

**Treatment.** In some diseases, radiation can be used to produce a biological effect. An example is the use of radioactive iodine to treat hyperthyroidism or cancer of the thyroid. The effects of treatment can be assessed with nuclear medicine techniques as well. For example, the metabolism of pituitary tumors can be used as an index of the effectiveness of chemotherapy with drugs that stimulate dopamine receptors.

See MEDICAL IMAGING; RADIATION THERAPY; RADIOLOGY.  
Henry N. Wagner, Jr.

**Bibliography.** H. N. Wagner, Jr., A new era of certainty (Highlights of the 1995 Society of Nuclear Medicine Annual Meeting, Minneapolis), *J. Nucl. Med.*, 36:13N-28N, 1995; H. N. Wagner, Jr., Z. Szabo, and J. W. Buchanan (eds.), *Principles of Nuclear Medicine*, 2d ed., 1995.

## Nuclear molecule

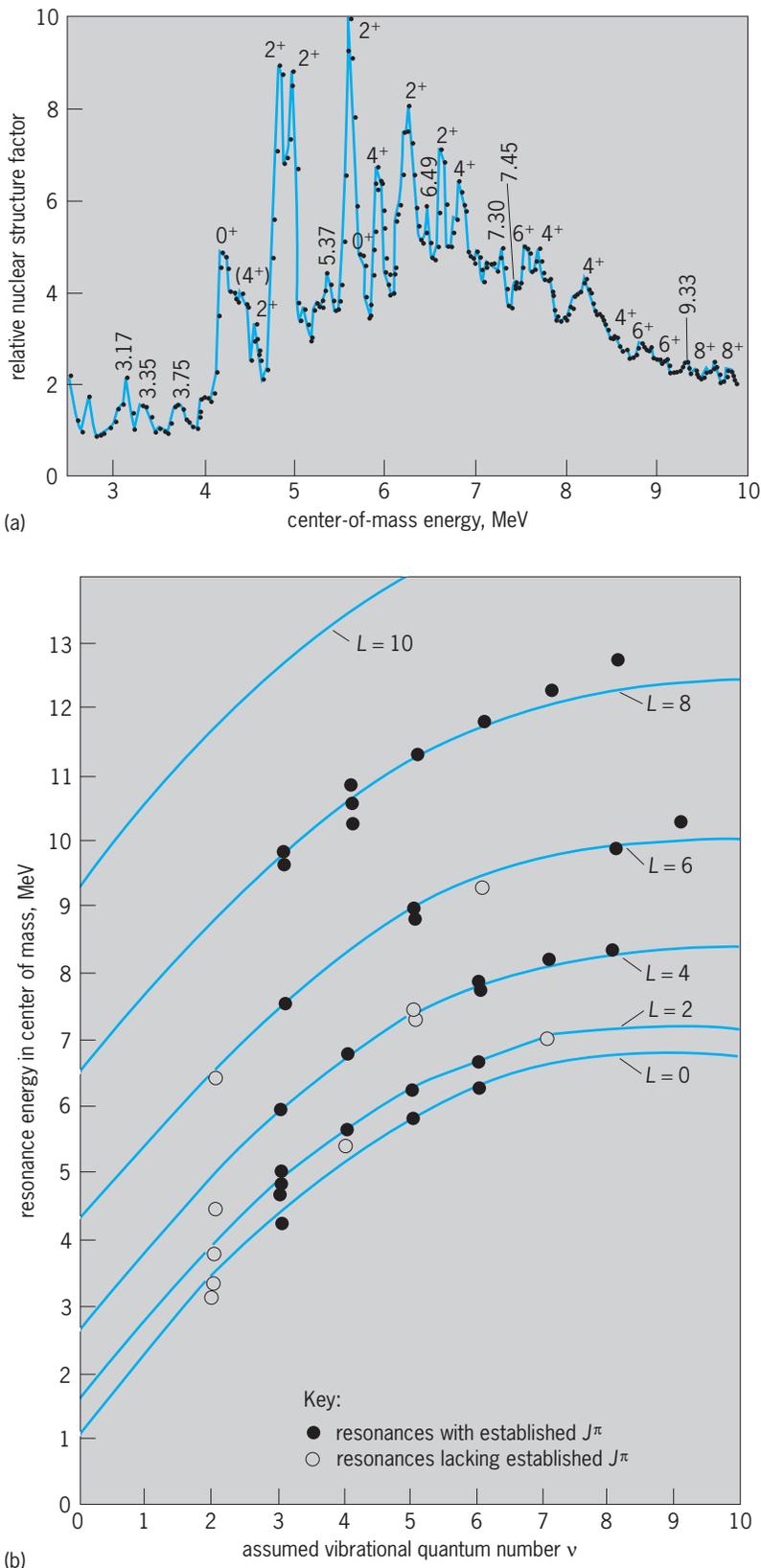
A quasistable entity of nuclear dimensions formed in nuclear collisions and comprising two or more discrete nuclei that retain their identities and are bound together by strong nuclear forces. Whereas the stable molecules of chemistry and biology consist of atoms bound through various electronic mechanisms, nuclear molecules do not form in nature except possibly in the hearts of giant stars; this simply reflects the fact that all nuclei carry positive electrical charges, and that under all natural conditions the long-range electrostatic repulsion prevents nuclear components from coming within the grasp of the short-range attractive nuclear force which could provide molecular binding. But in energetic collisions this electrostatic repulsion can be overcome.

Nuclear molecules were first suggested, in 1960, by D. A. Bromley, J. A. Kuehner, and E. Almqvist to explain very surprising experimental results obtained in the first studies on collisions between carbon nuclei carried out under conditions of high precision. In the original discovery, three sharp resonances appeared in the region of the Coulomb barrier corresponding to states in the  $^{24}\text{Mg}$  compound nucleus near 20 MeV of excitation. These were completely unexpected and indicated a new mode of behavior for the 24-body  $^{24}\text{Mg}$  system. Although such phenomena were known only in this  $^{24}\text{Mg}$  system for almost 10 years, they have subsequently been shown to be ubiquitous features of both nuclear structure and nuclear dynamics. The exact mechanisms leading to the appearance of these molecular configurations are not yet known, however.

Thus far, attention has been focused on dinuclear molecular systems; the search for polynuclear configurations could provide important new insight into the behavior of many-body systems, generally.

**$^{12}\text{C} + ^{12}\text{C}$  system.** In 1975 interest in molecular phenomena was renewed by discovery by Y. Abe that the original theoretical models proposed to explain the 1960 discovery were much richer in terms of predicted phenomena than had been realized. Experimental studies rapidly increased the number of resonances in the  $^{12}\text{C} + ^{12}\text{C}$  system from the original 3 to a total of 39 (Fig. 1a), established the spins and parities of many of them, and further demonstrated that the wave functions of the states in  $^{24}\text{Mg}$  that corresponded to the resonances had a marked molecular structure in which the  $^{12}\text{C}$  nuclei retained their identity.

**Models of molecular phenomena.** Bromley, Kuehner, and Almqvist had noted that the effective potential,



**Fig. 1. Resonances in the  $^{12}\text{C} + ^{12}\text{C}$  system. (a) Total cross section for  $^{12}\text{C} + ^{12}\text{C}$  interactions as a function of the center-of-mass energy and expressed as a nuclear structure factor to remove penetrability effects. Numbers with superscript indicate spin ( $J$ ) and parity ( $\pi$ ). (b) Resonances correlated in terms of a  $U(4)$  symmetry and the corresponding analytic vibration-rotation spectrum. Curves are obtained from the equation in the text with  $D = 0.34$  MeV;  $a = 1.44$  MeV;  $b = 0.08$  MeV;  $c = 0.0757$  MeV.**

composed of the sum of nuclear, Coulomb, and centrifugal components, involved in the  $^{12}\text{C} + ^{12}\text{C}$  collisions closely resembled a familiar Morse molecular one. In 1981 F. Iachello demonstrated that in such a potential it was possible, by using group theoretical techniques, to obtain an analytic expression for the allowed quantum states of the form of the equation below, where  $v$  and  $L$  are the vibrational and rota-

$E(v, L)$

$$= -D + a(v + 1/2) - b(v + 1/2)^2 + cL(L + 1)$$

tional quantum numbers and  $D$ ,  $a$ ,  $b$ , and  $c$  are constants. The governing group here is  $U(4)$ . This simple expression provides a remarkably successful and complete description of all the experimental data (Fig. 1b). See MOLECULAR STRUCTURE AND SPECTRA.

The fact that this model predicts molecular dissociation at around 7 MeV suggested that the  $0^+$  state in  $^{12}\text{C}$  at 7.66 MeV might play an essential role in the molecular interaction rather than the  $2^+$  at 4.43 MeV which had been assumed to be critical to all earlier models. By using a constrained time-dependent Hartree-Fock (TDHF) approach, the density distribution for  $^{12}\text{C}$  was first calculated in the ground (Fig. 2a) and excited (Fig. 2b)  $0^+$  states; the latter shows a linear, three-alpha-particle structure suggested several years previously. Assuming that one of the  $^{12}\text{C}$  nuclei is inelastically excited to this state in an early stage of the collision, the TDHF model allows the collision to be followed in microscopic detail. At appropriately chosen relative energies, the excited carbon oscillates through the unexcited one, with a period of  $2.3 \times 10^{-21}$  s in the case shown in Fig. 2c. This is still a very crude model, but is suggestive of the kind of simple relative motions that are present in the molecular configurations.

In the  $^{12}\text{C} + ^{16}\text{O}$  system, measurements at energies at, and below, the Coulomb barrier show that simple dinuclear molecular configurations are present in the form of rotational bands, just as in  $^{12}\text{C} + ^{12}\text{C}$ , but at higher energies, although the resonances persist, their structure is substantially more complex. Below the barrier, a  $U(4)$  group description is again extremely successful.

**Phenomena in heavier systems.** Sharp, long-lived resonant states appear in the scattering of  $^{28}\text{Si}$  on  $^{28}\text{Si}$  up to the highest energies studied, 75 MeV of excitation in the  $^{56}\text{Ni}$  compound system. Measurements of the angular distributions of the scattered nuclei (Fig. 3) show that these resonances have very high angular momentum; measurements have extended to  $44\hbar$  (where  $\hbar$  is Planck's constant divided by  $2\pi$ ), the highest angular momentum established in any quantum system. This is particularly interesting inasmuch as simple calculations suggest that at  $50\hbar$  the  $^{56}\text{Ni}$  nucleus should be torn apart by centrifugal forces. Resonance structure also appears in the  $^{24}\text{Mg} + ^{24}\text{Mg}$  system, but not in  $^{28}\text{Si} + ^{30}\text{Si}$  and  $^{30}\text{Si} + ^{30}\text{Si}$ . Potential energy contours have been calculated for the compound nuclei involved in all these systems, at the angular momenta involved, assuming a generalized Nilsson shell model. It is suggestive that

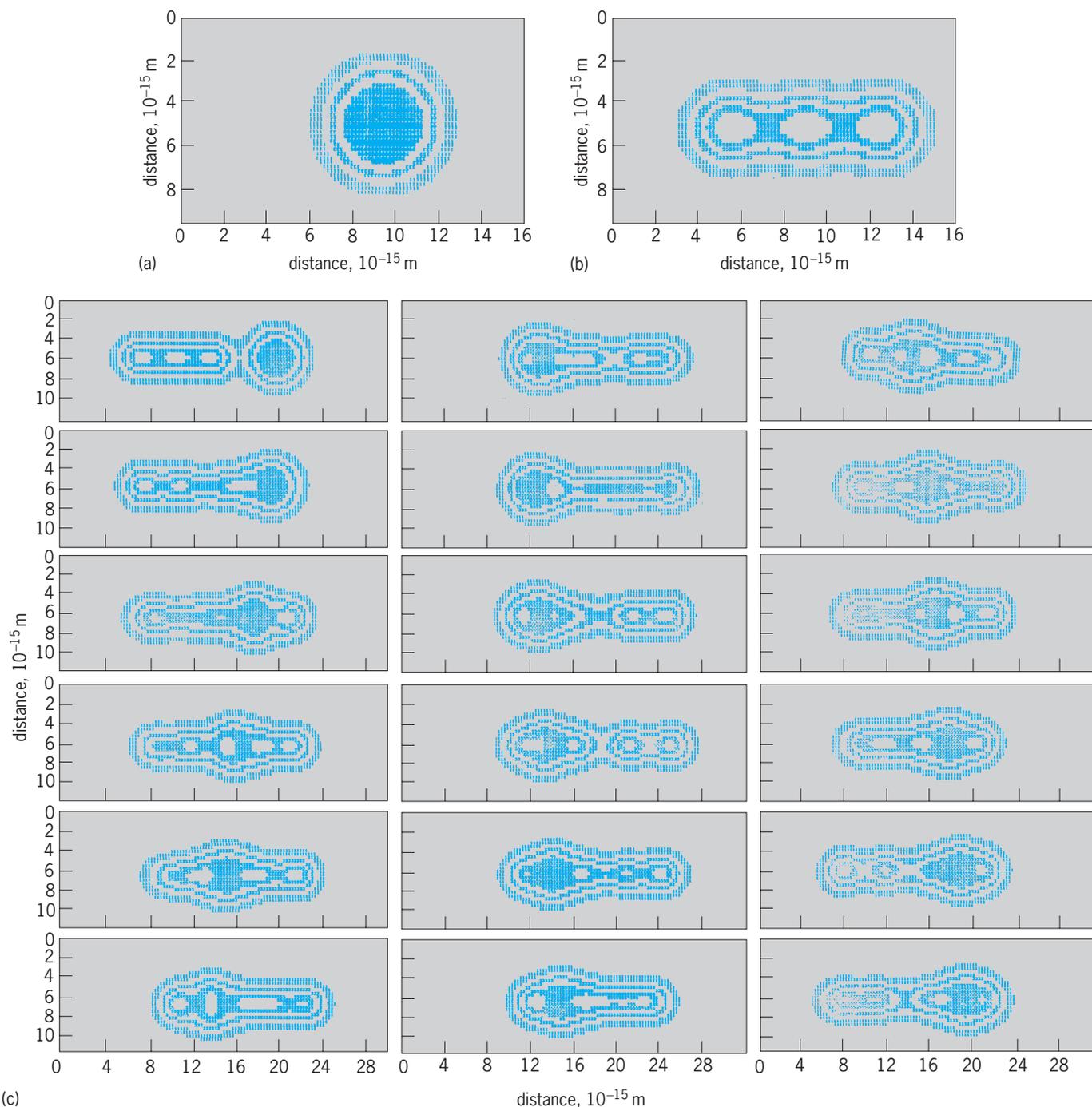


Fig. 2. Density contours calculated by using a constrained, time-dependent Hartree-Fock (TDHF) model. (a) Ground state of  $^{12}\text{C}$ . (b) First excited  $O^+$  state of  $^{12}\text{C}$ . (c) Time sequence for the interaction of two carbon nuclei, one in each of these configurations.

those systems which show a potential minimum at large prolate deformation appear to be those that show pronounced resonance structure. A quantitative theory that would allow calculation of the molecular phenomena from first principles, however, is still far from available.

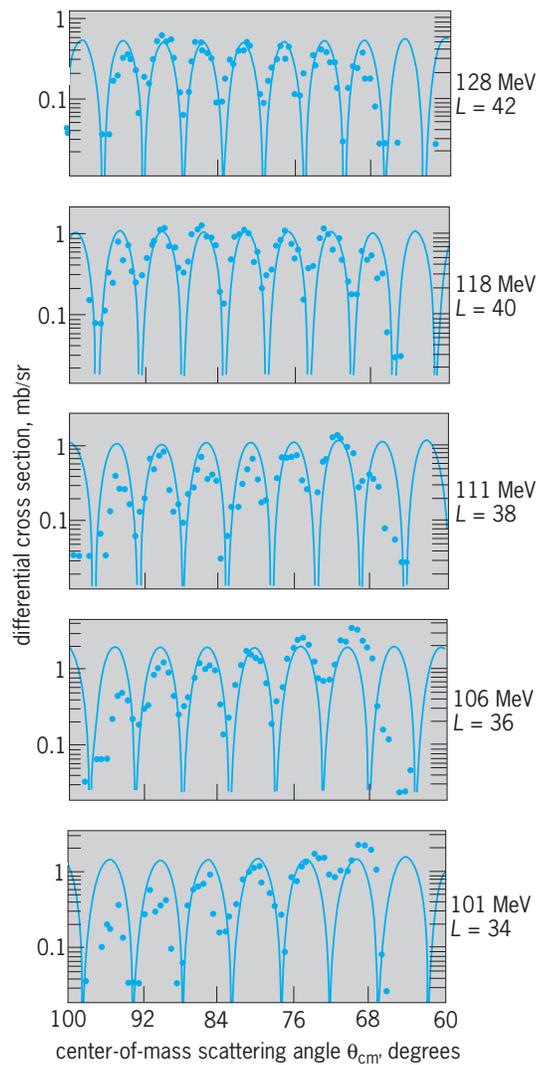
Positron spectra have been measured from collisions of very heavy (uranium + uranium, uranium + thorium, uranium + curium) at energies in the region of the Coulomb barrier. A completely unexpected feature of these spectra is a sharp line at about

320 keV. If this is to arise from spontaneous positron creation in the supercritical Coulomb field created in the vicinity of the colliding ions, its width requires that some mechanism exist to maintain this supercritical field for much longer than would be the case in a simple Coulomb scattering. It has been suggested that this would occur if the interacting ions formed a molecular configuration and preliminary calculations show that this would be most probable in collisions in which the ends of the long prolate axes of these nuclei just came into contact. This

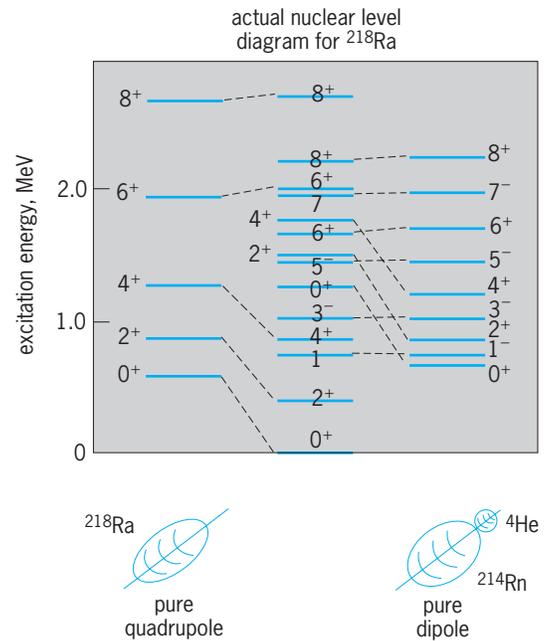
configuration has not been established, but all available evidence points to the existence of superheavy nuclear molecular states in these interactions. See POSITRON; QUASIATOM; SUPERCRITICAL FIELDS.

In all heavy-ion interactions yet studied, it has been the case that when adequate experimental energy resolution has been available to resolve sharp resonance structure corresponding to molecular phenomena, such structure has appeared. It will be an important challenge for higher-energy, high-precision nuclear accelerators to establish how high in excitation such sharp resonance structures can persist.

**Radioactive decay.** Indirect evidence supporting very asymmetric molecular configurations in heavy nuclei comes from the discovery of spontaneous radioactive decay of radium nuclei in which  $^{14}\text{C}$ , and perhaps  $^{16}\text{C}$ , nuclei are emitted. The probability of such decay is approximately  $6 \times 10^{-10}$  of that involv-



**Fig. 3.** Angular distributions of  $^{28}\text{Si}$  nuclei elastically scattered from  $^{28}\text{Si}$  at indicated energies of resonances. The solid curves are obtained by squaring the Legendre polynomial  $P_L$  of the indicated order  $L$ , giving the angular momentum of the resonance in units of  $\hbar$  (Planck's constant divided by  $2\pi$ ).  $1 \text{ mb} = 10^{-31} \text{ m}^2$ .



**Fig. 4.** Mixing of rotational bands based on pure quadrupole and dipole configurations in low-lying states of  $^{218}\text{Ra}$ .

ing emission of an alpha particle, and the ratio of the probabilities is very closely equal to that calculated for the penetrability of the two products through the Coulomb barriers involved. This suggests that there is a substantial probability for both an alpha particle and  $^{14}\text{C}$  to preexist in a molecular configuration. See RADIOACTIVITY.

**Dipole nuclear collectivity.** That alpha particles do indeed preexist in these heavy nuclei has been suggested frequently on the basis of observed alpha-particle radioactivity. Rather striking confirmatory evidence has been obtained from measurements on the structure and electromagnetic deexcitation of bound states in radium nuclei. For example, in  $^{218}\text{Ra}$  (Fig. 4) the presence of low-lying negative parity states and of very strong electric dipole transitions involving them provides strong evidence for the presence of a molecular dipole configuration involving an alpha particle and a  $^{214}\text{Rn}$  nucleus. States having this configuration would be expected to mix with the normal quadrupole states except in the case of negative parity ones for which no low-lying quadrupole partner exists. Similar strong dipole effects provide evidence for the existence of alpha-particle molecular configurations in nuclei as light as  $^{18}\text{O}$  where the molecular participants are  $^4\text{H}$  and  $^{14}\text{C}$ . See NUCLEAR STRUCTURE; SCATTERING EXPERIMENTS (NUCLEI).

D. Allan Bromley  
Bibliography. Y. Akaishi et al., *Developments of Nuclear Cluster Dynamics*, 1989; D. A. Bromley, Nuclear molecules, *Sci. Amer.*, 239(6):58-69, 1978; N. Cindro (ed.), *Nuclear Molecular Phenomena*, 1982; K. A. Erb and D. A. Bromley, Nuclear molecular resonances in heavy-ion collisions, *Phys. Today*, 32(1): 34-42, 1979; W. Greiner, W. Scheid, and J. Y. Park, *Nuclear Molecules*, 1995; J. S. Lilley and M. A. Nagarajan (eds.), *Nuclear Cluster Phenomena*, 1984.

## Nuclear moments

Intrinsic properties of atomic nuclei; electric moments result from deviations of the nuclear charge distribution from spherical symmetry; magnetic moments are a consequence of the intrinsic spin and the rotational motion of nucleons within the nucleus. The classical definitions of the magnetic and electric multipole moments are written in general in terms of multipole expansions. *See* NUCLEAR STRUCTURE; SPIN (QUANTUM MECHANICS).

Parity conservation allows only even-rank electric moments and odd-rank magnetic moments to be nonzero. The most important terms are the magnetic dipole, given by Eq. (1), and the electric monopole, quadrupole, and hexadecapole, given by Eq. (2), for

$$\vec{\mu} = \int \vec{M}(\vec{r}) dv \quad (1)$$

$$Q = \frac{1}{e} \int r^l Y_{lm}(\theta, \phi) \rho(\vec{r}) dv \quad (2)$$

$l = 0, 2, 4$ . Here  $m$  is the projection of the orbital angular momentum  $l$  on a  $z$  axis appropriately chosen in space,  $\vec{M}\vec{r}$  is the magnetization density of the nucleus and depends on the space coordinates  $\vec{r}$ ,  $e$  is the electronic charge,  $\rho\vec{r}$  is the charge density in the nucleus, and  $Y_{lm}$  are normalized spherical harmonics that depend on the angular coordinates  $\theta$  and  $\phi$ . *See* MAGNETIZATION; PARITY (QUANTUM MECHANICS); SPHERICAL HARMONICS.

Quantum-mechanically only the  $z$  components of the effective operators have nonvanishing values. Magnetic dipole and electric quadrupole moments are usually expressed in terms of the nuclear spin  $I$  through Eqs. (3) and (4), where  $g$ , the nuclear gy-

$$\frac{\mu}{\mu_N} = gI \quad (3)$$

$$Q(m_I) = \frac{[3m_I^2 - I(I+1)]Q}{I(2I-1)} \quad (4)$$

romagnetic factor, is a measure of the coupling of nuclear spins and orbital angular momenta,  $\mu_N = eh/4\pi M_p c = 5.0508 \times 10^{-24}$  erg/gauss =  $5.0508 \times 10^{-27}$  joule/tesla, is the nuclear magneton,  $M_p$  is the proton mass,  $h$  is Planck's constant,  $e$  is the electron charge, and  $c$  is the speed of light.  $Q(m_I)$  is the effective quadrupole moment in the state  $m_I$ , and  $Q$  is the quadrupole moment for the state  $m_I = I$ . All angular momenta are expressed in units of  $h/2\pi$ . The magnitude of  $g$  varies between 0 and 1.8, and  $Q$  is of the order of  $10^{-25}$  cm<sup>2</sup>. *See* ANGULAR MOMENTUM; MAGNETON.

In special cases nuclear moments can be measured by direct methods involving the interaction of the nucleus with an external magnetic field or with an electric field gradient produced by the scattering of high-energy charged particles. In general, however, nuclear moments manifest themselves through the hyperfine interaction between the nuclear moments and the fields or field gradients produced by either the atomic electrons' currents and spins, or

the molecular or crystalline electronic and lattice structures. *See* HYPERFINE STRUCTURE.

**Effects of nuclear moments.** In a free atom the magnetic hyperfine interaction between the nuclear spin  $\vec{I}$  and the effective magnetic field  $H\vec{r}_e$  associated with electronic angular momentum  $\vec{J}$  results in an energy  $W = \vec{\mu} \cdot \vec{H}_e = ba\vec{I} \cdot \vec{J}$ , which appears as a splitting of the energy levels of the atom. The magnetic field at the nucleus due to atomic electrons can be as large as 10-100 teslas for a neutral atom. The constant  $a$  is of the order of 1000 MHz. *See* ATOMIC STRUCTURE AND SPECTRA.

The electric monopole moment is a measure of the nuclear charge and does not give rise to hyperfine interactions. The quadrupole moment  $Q$  reflects the deviation of the nuclear charge distribution from a spherical charge distribution. It is responsible for a quadrupole hyperfine interaction energy  $W_Q$ , which is proportional to the quadrupole moment  $Q$  and to the spatial derivative of the electric field at the nucleus due to the electronic charges, and is given by Eq. (5), where  $q$  is the average of expression (6). In

$$W_Q = \frac{e^2 Q q}{2I(2I-1)J(2J-1)} \times [3(\vec{I} \cdot \vec{J})^2 + \frac{3}{2}(\vec{I} \cdot \vec{J}) - I(I+1)] \quad (5)$$

$$\sum_i \frac{3 \cos^2 \theta_i - 1}{r_i^3} \quad (6)$$

this expression,  $r_i$  is the radius vector from the nucleus to the  $i$ th electron, and  $\theta_i$  is the angle between  $r_i$  and the  $z$  axis. *See* MÖSSBAUER EFFECT.

In free molecules the hyperfine couplings are similar to those encountered in free atoms, but as the charge distributions and the spin coupling of valence electrons vary widely, depending on the nature of the molecular bonding, a greater diversity of magnetic dipole and quadrupole interactions is met. *See* MOLECULAR STRUCTURE AND SPECTRA.

In crystals the hyperfine interaction patterns become extremely complex, because the crystalline electric field is usually strong enough to compete with the spin orbit coupling of the electrons in the ion. Nevertheless, the energy-level structure can often be resolved by selective experiments at low temperatures on dilute concentrations of the ion of interest. *See* ELECTRON PARAMAGNETIC RESONANCE (EPR) SPECTROSCOPY; MAGNETIC RELAXATION; MAGNETIC RESONANCE.

**Measurement.** The hyperfine interactions affect the energy levels of either the nuclei or the atoms, molecules, or ions, and therefore can be observed either in nuclear parameters or in the atomic, molecular, or ionic structure. The many different techniques that have been developed to measure nuclear moments can be grossly grouped in three categories: the conventional techniques based mostly on spectroscopy of energy levels, the methods based on the detection of nuclear radiation from aligned excited nuclei, and techniques involving the interactions of fast ions with matter or of fast ions with laser beams.

*Hyperfine structure of spectral lines.* The hyperfine interaction causes a splitting of the electronic energy levels which is proportional to the magnitude of the nuclear moments, to the angular momenta  $I$  and  $J$  of the nuclei and their electronic environment, and to the magnetic field or electric field gradient at the nucleus. The magnitude of the splitting is determined by the nuclear moments, and the multiplicity of levels is given by the relevant angular momenta  $I$  or  $J$  involved in the interaction. The energy levels are identified either by optical or microwave spectroscopy.

Optical spectroscopy (in the visible and ultraviolet) has the advantage of allowing the study of atomic excited states and of atoms in different states of ionization. Furthermore, optical spectra provide a direct measure of the monopole moments, which are manifested as shifts in the energy levels of atoms of different nuclear isotopes exhibiting different nuclear radii and charge distributions. Optical spectroscopy has a special advantage over other methods in that the intensity of the lines often yields the sign of the interaction constant  $a$ . See ISOTOPE SHIFT; SPECTROSCOPY.

Microwave spectroscopy is a high-resolution technique involving the attenuation of a signal in a waveguide containing the absorber in the form of a low-pressure gas. The states are identified by the observation of electric dipole transitions of the order of 20,000 MHz. The levels are split by quadrupole interactions of the order of 100 MHz. Very precise quadrupole couplings are obtained, as well as vibrational and rotational constants of molecules, and nuclear spins. See MICROWAVE SPECTROSCOPY.

*Atomic and molecular beams and nuclear resonance.* Atomic and molecular beams passing through inhomogeneous magnetic fields are deflected by an amount depending on the nuclear moment. However, because of the small size of the nuclear moment, the observable effect is very small. The addition of a radio-frequency magnetic field at the frequency corresponding to the energy difference between hyperfine electronic states has vastly extended the scope of the technique. For nuclei in solids, liquids, or gases, the internal magnetic fields and gradients of the electric fields may be quenched if the pairing of electrons and the interaction between the nuclear magnetic moment and the external field dominate. The molecular beam apparatus is designed to detect the change in orientation of the nuclei, while the nuclear magnetic resonance system is designed to detect absorbed power (resonance absorption) or a signal induced at resonance in a pick-up coil around the sample (nuclear induction). The required frequencies for fields of about 0.5 tesla are of the order of 1–5 MHz. The principal calibration of the field is accomplished in relation to the resonant frequency for the proton whose  $g$ -factor is accurately known. Sensitivities of 1 part in  $10^8$  are possible under optimum experimental conditions. The constant  $a$  for  $^{133}\text{Cs}$  has been measured to 1 part in  $10^{10}$ , and this isotope is used as a time standard. See ATOMIC CLOCK; MOLECULAR BEAMS.

The existence of quadrupole interactions produces a broadening of the resonance line above the natural width and a definite structure determined by the value of the nuclear spin. In some crystals the electric field gradient at the nucleus is large enough to split the energy levels without the need for an external field, and pure quadrupole resonance spectra are observed. This technique allows very accurate comparison of quadrupole moments of isotopes.

*Atomic and molecular beams with radioactive nuclei.* The conventional atomic and molecular beam investigations can be applied to radioactive nuclei if the beam current measurement is replaced by the much more sensitive detectors of radiations emitted in a radioactive decay. Moments of nuclei with half-lives down to the order of minutes have been determined. See RADIOACTIVITY.

*Perturbed angular correlations.* The angular distribution and the polarization of radiation emitted by nuclei depend on the angle between the nuclear spin axis and the direction of emissions. In radioactive sources in which the nuclei have been somewhat oriented by a nuclear reaction or by static or dynamic polarization techniques at low temperatures, the ensuing nonisotropic angular correlation of the decay radiation can be perturbed by the application of external magnetic fields, or by the hyperfine interaction between the nuclear moment and the electronic or crystalline fields acting at the nuclear site. Magnetic dipole and electric quadrupole moments of ground and excited nuclear states with half-lives as short as  $10^{-9}$  have been measured. See DYNAMIC NUCLEAR POLARIZATION.

*Techniques involving interactions of fast ions.* Techniques involving the interaction of intense light beams from tuned lasers with fast ion beams have extended the realm of resonance spectroscopy to the study of exotic species, such as nuclei far from stability, fission isomers, and ground-state nuclei with half-lives shorter than minutes. The hyperfine interactions in beams of fast ions traversing magnetic materials result from the coupling between the nuclear moments and unpaired polarized  $s$ -electrons, and are strong enough ( $H_e$  is of the order of  $2 \times 10^3$  T) to extend the moment measurements to excited states with lifetimes as short as  $10^{-12}$  s. Progress in atomic and nuclear technology has contributed to the production of hyperfine interactions of increasing strength, thus allowing for the observation of nuclear moments of nuclei and nuclear states of increasing rarity.

Noémie Koller

Bibliography. B. Castel and I. S. Towner, *Modern Theories of Nuclear Moments*, 1990; N. F. Ramsey, *Molecular Beams*, 1956, reprint 1990.

## Nuclear orientation

The directional ordering of an assembly of nuclear spins  $I$  with respect to some axis in space. Under normal conditions nuclei are not oriented; that is, all directions in space are equally probable. For a system of nuclear spins with rotational symmetry about an

axis, the degree of orientation is completely characterized by the relative populations  $a_m$  of the  $2I + 1$  magnetic sublevels  $m$  ( $m = I, I - 1, \dots, -I$ ). There are just  $2I$  independent values of  $a_m$ , since they are normalized to unity, namely,

$$\sum_m a_m = 1$$

Rather than specify these populations directly, it turns out to be more useful to form the moments

$$\sum_m m^v a_m$$

since these occur in the theoretical calculations. There are  $2I$  independent linear combinations of these moments which are called orientation parameters,  $f_k(I)$ , and are defined by the equation below. Here  $f_0(I) = 1$  and all  $f_k(I)$  with  $k \geq 2I + 1$  are zero.

$$f_k(I) = \left(\frac{2k}{k}\right)^{-1} I^{-k} \sum_m \sum_{\nu=0}^k (-1)^\nu \times \frac{(I-m)!(I+m)!}{(I-m-\nu)!(I+m-k+\nu)!} \binom{k}{\nu}^2 a_m$$

**Nuclear polarization and alignment.** Nuclear polarization is said to be present when one or more  $f_k(I)$  with  $k$ -odd is not zero, regardless of the even  $f_k(I)$  values. In this case the nuclear spin system is polarized. If all the  $f_k(I)$  for  $k$ -odd are zero and at least one  $f_k(I)$  for  $k$ -even is not zero, nuclear alignment is said to be present; that is, the nuclear spin system is aligned. Simply stated, if the  $z$  axis is the axis of quantization of the nuclear spin system, polarization represents a net moment along the  $z$  axis, whereas alignment does not. Unfortunately, the term nuclear polarization is usually associated with  $f_1(I)$ , and nuclear alignment with  $f_2(I)$ , although their meanings are in fact much more general. There are other definitions of nuclear orientation parameters; they are mathematically equivalent to the one above. If the nuclear spin system does not have cylindrical symmetry, a more general definition of nuclear orientation is needed leading to the statistical tensors. See SPIN (QUANTUM MECHANICS).

**Production.** Nuclear orientation can be achieved in various ways. The most obvious way is to modify the energies of the  $2I + 1$  magnetic sublevels so as to remove their degeneracy and thereby change the populations of these sublevels. The spin degeneracy can be removed by a magnetic field interacting with the nuclear magnetic dipole moment, or by an inhomogeneous electric field interacting with the nuclear electric quadrupole moment. Significant differences in the populations of the sublevels can be established by cooling the nuclear sample to low temperatures  $T$  such that  $T$  is in the region around  $\Delta E/k$ , where  $\Delta E$  is the energy separation of adjacent magnetic sublevels of energy  $E_m$ , and  $k$  is the Boltzmann constant. If the nuclear spin system is in thermal equilibrium, the

populations  $a_m$  are given by the normalized Boltzmann factor

$$\frac{\exp(-E_m/kT)}{\sum_m \exp(-E_m/kT)}$$

This means of producing nuclear orientation is called the static method. In contrast, there is the dynamic method, which is related to optical pumping in gases. There are other ways to produce oriented nuclei; for example, in a nuclear reaction such as the capture of polarized neutrons (produced by magnetic scattering) by unoriented nuclei, the resulting compound nuclei could be polarized. In addition to polarized neutron beams, polarized beams of protons, deuterons, tritons, helium-3, lithium-6, and other nuclei have been produced. See BOLTZMANN STATISTICS; DYNAMIC NUCLEAR POLARIZATION; OPTICAL PUMPING.

**Applications.** Oriented nuclei have proved to be very useful in various fields of physics. They have been used to measure nuclear properties, for example, magnetic dipole and electric quadrupole moments, spins, parities, and mixing ratios of nuclear states. Oriented nuclei have been used to examine some of the fundamental properties of nuclear forces, for example, nonconservation of parity in the weak interaction. Measurement of hyperfine fields, electric-field gradients, and other properties relating to the environment of the nucleus have been made by using oriented nuclei. Nuclear orientation thermometry is one of the few sources of a primary temperature scale at low temperatures. Oriented nuclear targets used in conjunction with beams of polarized and unpolarized particles have proved very useful in examining certain aspects of the nuclear force. See HYPERFINE STRUCTURE; LOW-TEMPERATURE THERMOMETRY; NUCLEAR MOMENTS; NUCLEAR STRUCTURE; PARITY (QUANTUM MECHANICS).

With the advent of helium-3/helium-4 dilution refrigerators, large superconducting magnets, and new methods of producing oriented nuclei and polarized beams, the field of nuclear orientation physics is expected to grow substantially. See CRYOGENICS; SUPERCONDUCTING DEVICES.

Harvey Marshak

**Bibliography.** A. O. Barut et al., *Polarization Dynamics in Nuclear and Particle Physics*, 1993; K. Siegbahn (ed.), *Alpha-, Beta- and Gamma Ray Spectroscopy*, 1968; S. Stringari (ed.), *Spin Polarized Quantum Systems*, 1989; W. J. Thompson and T. B. Clegg, Physics with polarized nuclei, *Phys. Today*, pp. 32-39, February 1979.

## Nuclear physics

The discipline involving the structure of atomic nuclei and their interactions with each other, with their constituent particles, and with the whole spectrum of elementary particles that is provided by very large accelerators. The nuclear domain occupies a central position between the atomic range of forces and sizes and those of elementary-particle physics,

characteristically within the nucleons themselves. As the only system in which all the known natural forces can be studied simultaneously, it provides a natural laboratory for the testing and extending of many fundamental symmetries and laws of nature. *See* ATOMIC NUCLEUS; ATOMIC STRUCTURE AND SPECTRA; ELEMENTARY PARTICLE; SYMMETRY LAWS (PHYSICS).

Containing a reasonably large, yet manageable number of strongly interacting components, the nucleus also occupies a central position in the universal many-body problem of physics, falling between the few-body problems, characteristic of elementary-particle interactions, and the extreme many-body situations of plasma physics and condensed matter, in which statistical approaches dominate; it provides the scientist with a rich range of phenomena to investigate—with the hope of understanding these phenomena at a microscopic level. *See* PLASMA (PHYSICS); STATISTICAL MECHANICS.

Activity in the field centers on three broad and interdependent subareas. The first is referred to as classical nuclear physics, wherein the structural and dynamic aspects of nuclear behavior are probed in numerous laboratories, and in many nuclear systems, with the use of a broad range of experimental and theoretical techniques. Second is higher-energy nuclear physics (referred to as medium-energy physics in the United States), which emphasizes the nuclear interior and nuclear interactions with mesonic probes. Third is heavy-ion physics, internationally the most rapidly growing subfield, wherein accelerated beams of nuclei spanning the periodic table are used to study previously inaccessible nuclear phenomena.

Nuclear physics is unique in the extent to which it merges the most fundamental and the most applied topics. Its instrumentation has found broad applicability throughout science, technology, and medicine; nuclear engineering and nuclear medicine are two very important areas of applied specialization. *See* NUCLEAR ENGINEERING; NUCLEAR RADIATION (BIOLOGY); RADIOLOGY.

Nuclear chemistry, certain aspects of condensed matter and materials science, and nuclear physics together constitute the broad field of nuclear science; outside the United States and Canada elementary particle physics is frequently included in this more general classification. *See* ANALOG STATES; COSMIC RAYS; FUNDAMENTAL INTERACTIONS; ISOTOPE; NUCLEAR CHEMISTRY; NUCLEAR FISSION; NUCLEAR FUSION; NUCLEAR ISOMERISM; NUCLEAR MOMENTS; NUCLEAR REACTION; NUCLEAR REACTOR; NUCLEAR SPECTRA; NUCLEAR STRUCTURE; PARTICLE ACCELERATOR; PARTICLE DETECTOR; RADIOACTIVITY; SCATTERING EXPERIMENTS (NUCLEID); WEAK NUCLEAR INTERACTIONS.

D. Allan Bromley

## Nuclear power

Power derived from fission or fusion nuclear reactions. More conventionally, nuclear power is interpreted as the utilization of the fission reactions in a nuclear power reactor to produce steam for electric

power production, for ship propulsion, or for process heat. Fission reactions involve the breakup of the nucleus of high-mass atoms and yield an energy release which is more than a millionfold greater than that obtained from chemical reactions involving the burning of a fuel. Successful control of the nuclear fission reactions utilizes this intensive source of energy. *See* NUCLEAR FISSION.

Fission reactions provide intensive sources of energy. For example, the fissioning of an atom of uranium yields about 200 MeV, whereas the oxidation of an atom of carbon releases only 4 eV. On a weight basis, this  $50 \times 10^6$  energy ratio becomes about  $2.5 \times 10^6$ . Uranium consists of several isotopes, only 0.7% of which is uranium-235, the fissile fuel currently used in reactors. Even with these considerations, including the need to enrich the fuel to several percent uranium-235, the fission reactions are attractive energy sources when coupled with abundant and relatively cheap uranium ore.

### Nuclear Fuel Cycle

Although the main process of nuclear power is the release of energy in the fission process which occurs in the reactor, there are a number of other important processes, such as mining and waste disposal, which both precede and follow fission. Together they constitute the nuclear fuel cycle. (The term cycle may not be exactly appropriate when the cycle is not closed, as is the case at present.) *See* NUCLEAR FUEL CYCLE.

The only fissionable material now found in nature is uranium-235. (However, prehistoric natural reactors operated in Gabon, Africa, and possibly also in Colorado in the United States. These produced plutonium that has since decayed to uranium-235.) Plutonium becomes part of the fuel cycle when the fertile material uranium-238 is converted into fissile plutonium-239. Thus, the dominant fuel cycle is the uranium-plutonium cycle. If thorium-232 is used in a reactor to produce fissile uranium-233, another cycle emerges, the thorium-uranium cycle. Since the applications of the latter have been limited to small-scale tests, and a commercial thorium-uranium cycle has not been developed, the discussion below will concentrate on the uranium-plutonium cycle. *See* NUCLEAR FUELS; PLUTONIUM; THORIUM; URANIUM.

**Fuel preparation.** Fissionable materials must be prepared for use in a nuclear reactor through mining, milling, conversion, enrichment, and fabrication.

*Mining.* The nuclear fuel cycle begins with the mining of uranium or thorium. Uranium is estimated to be present in the Earth's crust to the extent of 4 parts per million. The concentration of thorium is nearly three times greater. However, uranium and thorium are so widely distributed that significant concentrations in workable deposits are more the exception than the rule. Exploitable deposits have on average a concentration of 0.1–0.5% of uranium oxide ( $U_3O_8$ ) by weight.

Large deposits of uranium-rich minerals are found in many places: in central Africa and around the gold-mining areas of South Africa, in Canada's Great Bear

Lake region in Ontario, and in Australia. Lower-grade ores have been mined extensively on the Colorado Plateau in the United States. There are other deposits being worked in west-central France, in the western mountains of the Czech Republic, in southwestern Hungary, in Gabon, West Africa, and in Madagascar. Also in Africa, rich deposits have been found in the Republic of Niger. Uranium concentration exceeds 15% in a deposit in the Cigar Lake area of Canada.

The chief sources of thorium are coastal sands rich in monazite found at Travancore near the southern tip of India, and on the coast of Brazil. Monazite sands have also been found on the shores of Florida's panhandle.

*Millng.* After uranium ore has been mined, it is crushed and the host rock separated from the core, usually by a flotation process. The uranium is milled and concentrated as a uranium salt, ammonium diuranate, which is generally known in the industry as yellowcake because of its color.

*Conversion.* The yellowcake is then shipped to a conversion plant where it is oxidized to uranium oxide and then is fluorinated to produce the uranium hexafluoride ( $UF_6$ ). This is a convenient form for the gaseous diffusion enrichment process because the uranium hexafluoride sublimates (passes directly from the solid phase to the gaseous phase without liquefying) at  $127^\circ F$  ( $53^\circ C$ ).

*Enrichment.* The uranium hexafluoride, familiarly called hex, is shipped in tanklike containers to one of the three United States gaseous diffusion enrichment plants or to one of several other enrichment plants throughout the world. Gas centrifuges are widely used for enrichment outside the United States. See ISOTOPE SEPARATION.

After enrichment, the two resulting streams—enriched uranium, and depleted uranium—part company. The depleted uranium is stored adjacent to the diffusion plant, and the enriched material is converted back to an oxide—this time uranium dioxide ( $UO_2$ )—and sent to a fuel fabrication plant.

*Fabrication.* At these plants, uranium dioxide intended for light-water reactor fuel is shaped into cylindrical pellets about 0.6 in. (15 mm) long and 0.4 in. (10 mm) in diameter. The pellets are sintered, that is, baked, to obtain a hard, dense consistency. After polishing, they are loaded into tubes made of Zircaloy, an alloy of zirconium. A number of tubes are put together with appropriate tie plates, fittings, and spacers to form fuel assemblies. See ZIRCONIUM.

**Fission in power reactors.** From about 1955 to 1965, numerous United States companies explored or planned power reactor product lines, and almost every combination of feasible fuel, coolant, and moderator was suggested.

Power reactors in the United States are the light-water-moderated and -cooled reactors (LWRs), including the pressurized-water reactor (PWR) and the boiling-water reactor (BWR), represent 100% of capacity in operation. The high-temperature gas-cooled reactor (HTGR), and the liquid-metal-cooled fast breeder reactor (LMFBR) have reached a high

level of development but are not used for commercial purposes.

Of these concepts, the light-water reactor and the high-temperature gas-cooled reactor are thermal reactors; that is, they operate by using moderated neutrons slowed down to "thermal" velocities (so called because their speed is determined by the thermal, or kinetic, energy of the substance in which the fuel is placed, namely the moderator). The third type, the fast breeder, operates on fast neutrons—unmoderated neutrons that have the high velocities near to those with which they are released from a fissioning nucleus. See NEUTRON; THERMAL NEUTRONS.

The pressurized-water reactor is in use in France. Canada has developed the CANDU natural-uranium-fueled and heavy-water-moderated and -cooled reactor. The Soviet Union and its successor republics have developed and built considerable numbers of two types of water-cooled reactors. One is a conventional pressurized-water reactor; the other is a tube-type, graphite-moderated reactor.

*Boiling-water reactor (BWR).* In one boiling water reactor, designed to produce about 3580 MW-thermal and 1220 MW net electric power, the reactor vessel is 238 in. (6 m) in inside diameter, 6 in. (15 cm) thick, and about 71 ft (22 m) in height. The active height of the core containing the fuel assemblies is 148 in. (4 m). Each fuel assembly typically contains 63 fuel rods, and 732 fuel assemblies are used. The diameter of the fuel rod is 0.5 in. (12.5 mm). The reactor is controlled by cruciform-shape (in cross section) control rods moving up from the bottom of the reactor in spaces between the fuel assemblies (177 control rods are provided). The water coolant is circulated up through the fuel assemblies by 20 jet pumps at about 70 atm (7 megapascals), and boiling occurs within the core. The steam is fed through four 26-in.-diameter (66-cm) steamlines to the turbine. As is typical for steam power cycles, about one-third of the energy released by fission is converted into mechanical work, and then to electrical energy, by the turbine-generator, and the remaining heat is removed in the condenser. The condenser operates in the same manner in fossil and nuclear power plants, with the heat it removes having to be dissipated to the environment. Some limited use of the energy rejected by the condenser is possible. The steam produced in the plant's nuclear steam supply system will be at lower temperatures and pressures than that from fossil plants, and thus the efficiency of the nuclear plant in producing electric power is somewhat less, leading to somewhat greater heat rejection to the environment per kilowatthour produced. See GENERATOR; STEAM CONDENSER; STEAM TURBINE.

Shielding is provided to reduce radiation levels, and pools of water are used for fuel storage and when access to the core is necessary for fuel transfers. Among the engineered safety features that minimize the consequences of reactor accidents is the primary containment building (Fig. 1). The function of the containment building is to cope with the energy released by depressurization of the coolant system should a failure occur in the primary piping, and to provide a secure enclosure to minimize

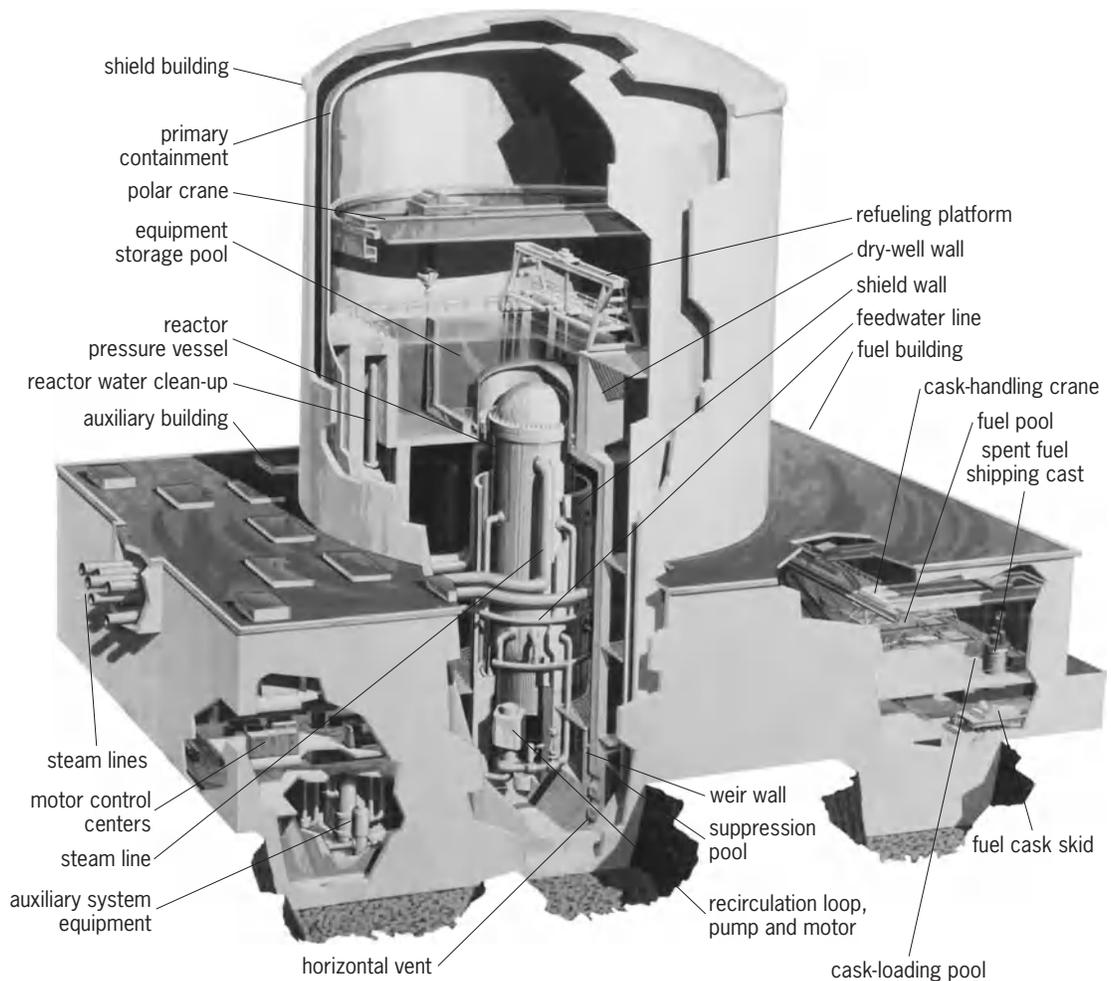


Fig. 1. Mark III containment building for a boiling-water reactor, which illustrates safety features designed to minimize the consequences of reactor accidents. (General Electric Co.)

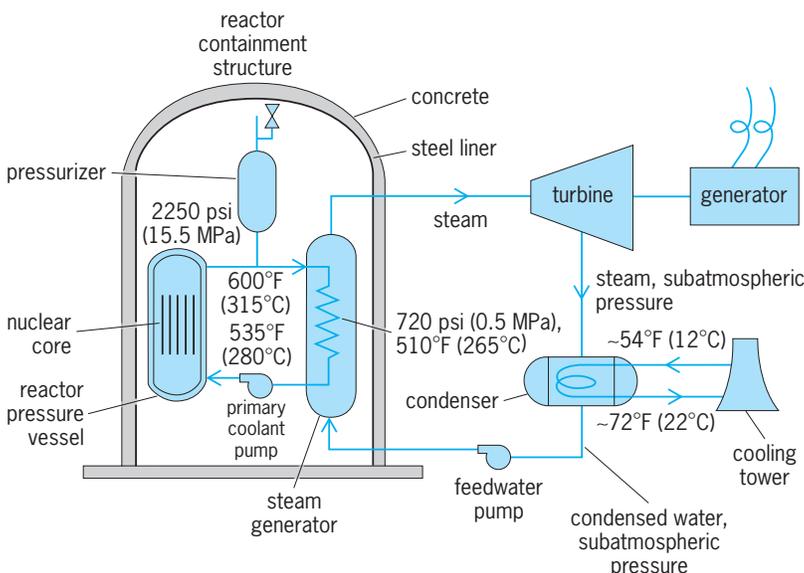


Fig. 2. Schematic of a pressurized-water reactor power plant. Heat transfer occurs within the containment building. (After F. J. Rahn et al., *A Guide to Nuclear Power Technology*, Krieger Publishing, 1992).

leakage of radioactive material to the surroundings. The boiling-water reactor utilizes a pool of water (called a suppression pool) to condense the steam produced by the depressurization of the primary coolant system. Various arrangements have been used for the suppression pool. Other engineered safety features include the emergency core-cooling system, the containment spray system, and the high-efficiency particulate filters for removing radioactivity from the containment building's atmosphere.

*Pressurized-water reactor (PWR).* Whereas in the boiling-water reactor a direct cycle is used in which steam from the reactor is fed to the turbine, the pressurized-water reactor employs a closed system (Fig. 2). The water coolant in the primary system is pumped through the reactor vessel, transporting the heat to a steam generator, and is recirculated in a closed primary system. A separate secondary water system is used on the shell side of the steam generator to produce steam, which is fed to the turbine, condensed, and recycled to the steam generator. A pressurizer is used in the primary system to maintain about 150 atm (15 MPa) pressure to suppress boiling in the primary coolant. Up to four loops have been used.

The reactor pressure vessel is about 44 ft (13.5 m) high and about 15 ft (4.4 m) in inside diameter, and has wall thickness exceeding 8.5 in. (22 cm). The active length of the fuel assemblies may range from 12 to 14 ft (about 4 m), and different configurations are used by the manufacturers. For example, one type of fuel assembly contains 264 fuel rods, and 193 fuel assemblies are used for the 3411-MW-thermal, four-loop plant. The outside diameter of the fuel rods is 0.4 in. (9.5 mm). For this arrangement, the control rods are grouped in clusters of 24 rods, with 61 clusters provided. In pressurized-water reactors, the control rods enter from the top of the core. Reactor operations are carried out by using both the control rods and a system to increase or decrease the boric acid content of the primary coolant. The latter controls the gross reactivity of the core, while the former allows fine tuning of the nuclear reaction. A typical prestressed concrete containment building (Fig. 3) is designed for a 4 atm (400 kilopascals) rise in pressure, with an inside diameter of about 116 ft (35.3 m) and height of 208 ft (64 m). The walls are 45 in. (1.3 m) thick. The containments have cooling and radioactive absorption systems as part of the engineered safety features. Another design (Fig. 4) uses a spherical steel containment (about 200 ft, or 60 m, in diameter), surrounded by a reinforced concrete shield building (Fig. 5). See NUCLEAR REACTOR.

*Breeder reactor.* In a breeder reactor, more fissile fuel is generated than is consumed. For example, in the fissioning of uranium-235, the neutrons released by fission are used both to continue the neutron chain reaction which produces the fission and to react with uranium-238 to produce uranium-239. The uranium-239 in turn decays to neptunium-239 and then to plutonium-239. Uranium-238 is called a fertile fuel, and uranium 235, as well as plutonium-239, is a fissile fuel and can be used in nuclear power reactors. The reactions noted can be used to convert most of the uranium-238 to plutonium-239 and thus provide about a 60-fold extension in the available uranium energy source relative to its use in nonbreeder reactors. Breeder power reactors can be used to generate electric power and to produce more fissile fuel. Breeding is also possible by using the fertile material thorium-232, which in turn is converted to the fissile fuel uranium-233. An almost inexhaustible energy source becomes possible with breeder reactors. The use of breeder reactors would decrease the long-range needs for enrichment and for mining more uranium.

Prototype breeder reactors are in use in France, England, Japan, and Russia. A full-scale (1250 MW-electric) commercial breeder, the Super Phenix, was constructed in France and operated commercially between 1986 and 1999.

*Advanced light-water reactor (ALWR).* The most advanced nuclear power reactor designs, referred to as advanced light-water reactors, have evolved from operating experience with light-water reactors. Advanced light-water reactors represent highly engineered designs that emphasize higher reliability,



Fig. 3. Oconee Nuclear Power Station (Greenville, South Carolina) containment structures.



Fig. 4. Spherical containment design for a pressurized-water reactor. (Combustion Engineering, Inc.)

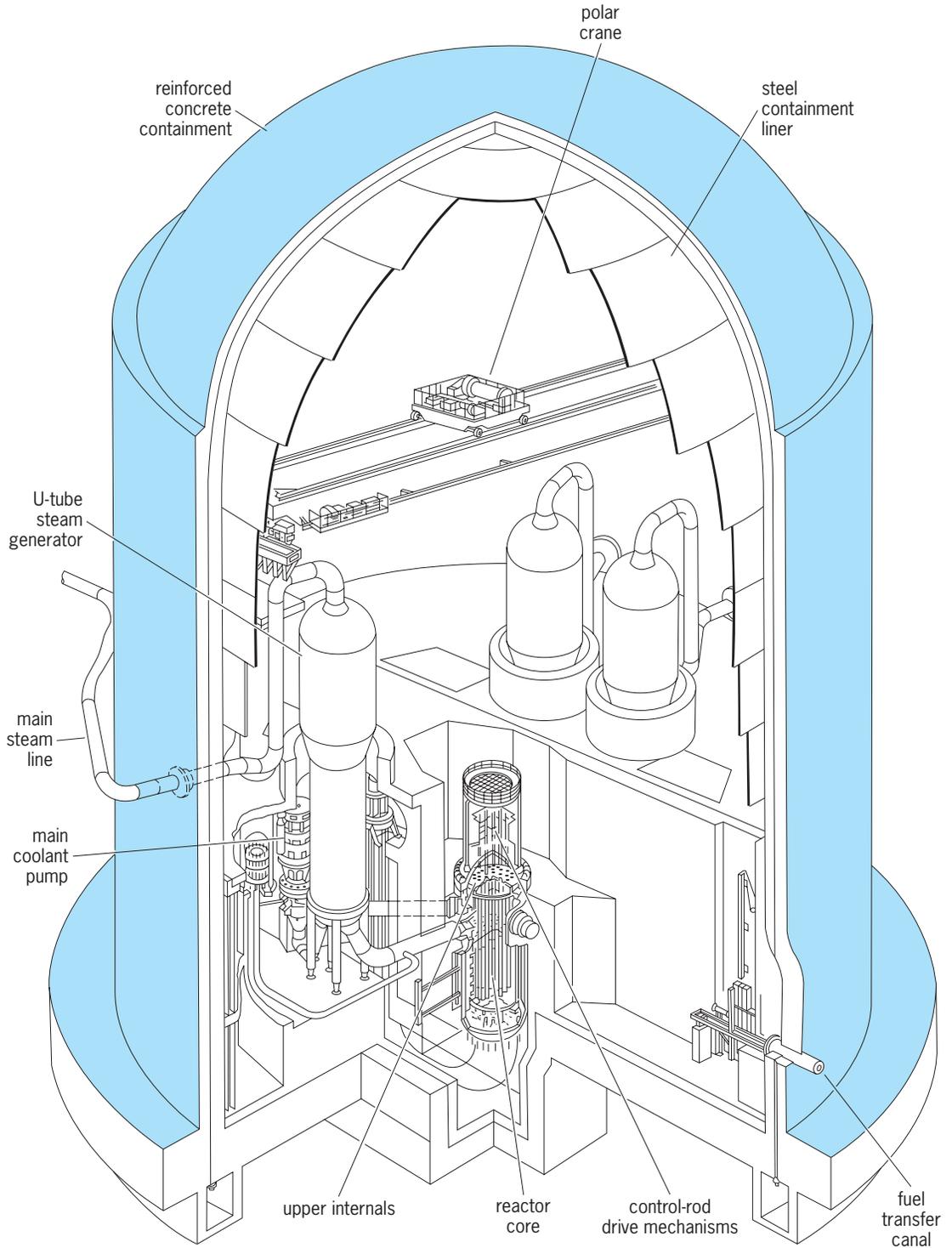


Fig. 5. Cutaway view of containment building of a typical pressurized-water reactor system. (Westinghouse Electric Corp.)

have few components and complex systems, are easier to build and maintain than current reactors, and, in particular, are safer and less sensitive to transients. The higher degree of safety is made possible by the use of redundant, passive systems that do not require power for actuation, and the low sensitivity to transients is achieved through low power density in the reactor core and a large inventory of coolant and containment water. Higher reliability depends

on the use of systems and components that are proven in service. The designs feature construction periods of as little as 4 years from the start of site excavation, low operating and maintenance costs, long fuel cycles, and simplified maintenance. Excellent fuel-cycle economics are due to high fuel burn-up and a flexible operating cycle. The average occupational radiation dose equivalent is less than 0.5 sievert per year. The plants require increased

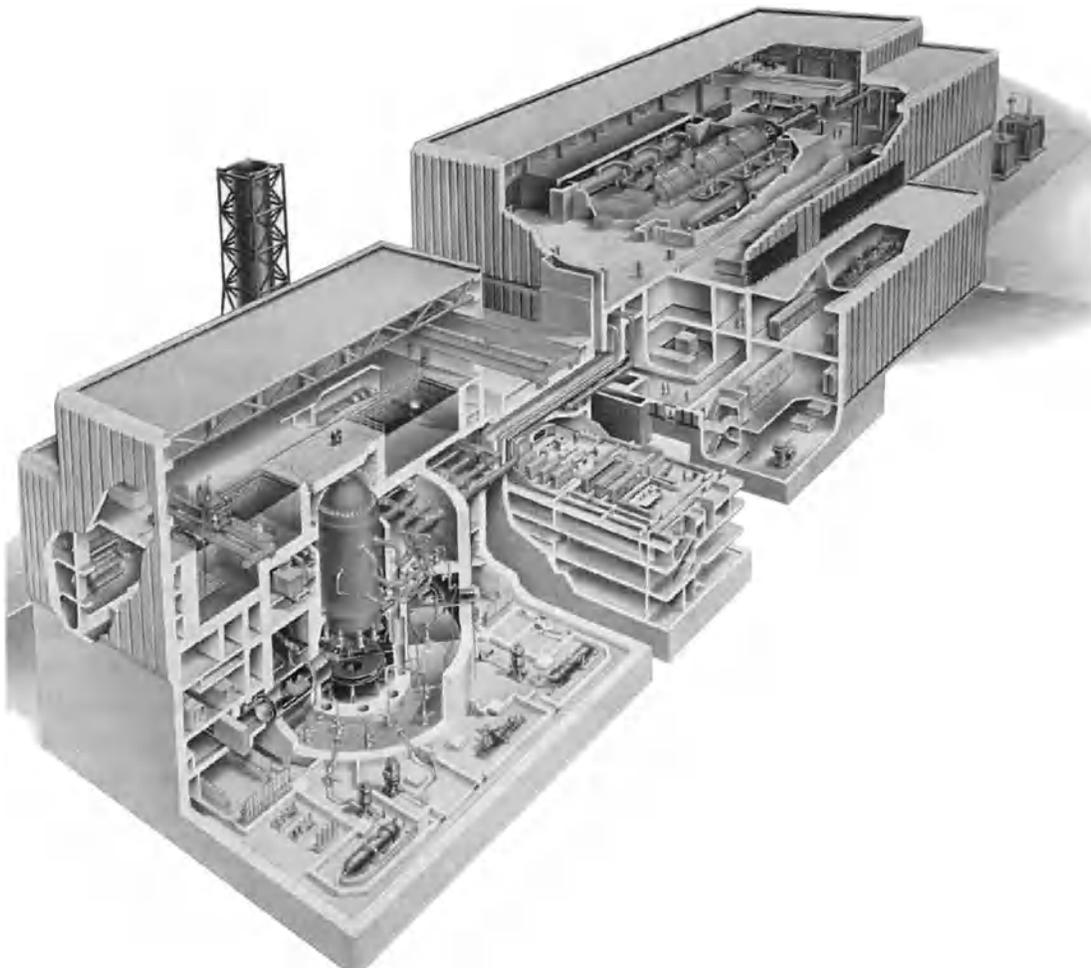
Advanced light-water reactor (ALWR) designs	
Name	Features
Advanced boiling-water reactor (ABWR)	1300-MW-electric reactor with simplified safety systems, a completely internal recirculation system, and improved controls
System 80+	1350-MW-electric pressurized-water reactor with greater design margins, more redundancy and diversity, and better operability and maintainability
AP-600	600-MW-electric midsize pressurized-water reactor with simplified design and safety features that allow 50% fewer valves, 35% fewer pumps, and 80% less safety-grade piping
SWR-1000	977-MW-electric boiling-water reactor with proven component design, no need for operator intervention in case of accidents, low power density, and 60-year design lifetime
Simplified boiling-water reactor (SBWR)	600-MW-electric midsize reactor employing natural water circulation for power generation, with gravity-driven emergency core cooling from an elevated suppression pool

spent fuel storage to accommodate all the fuel discharged over the plant lifetime.

International participation has been a critical factor in the development of advanced light-water reactors. The **table** summarizes the plants that are offered. Due to the lengthy design, licensing, and prototype development required for new-generation reactors, the majority of the plants constructed in the

period through 2005 are expected to be those in the table.

The first advanced nuclear plant to be constructed was a twin-unit advanced boiling-water reactor (ABWR; **Fig. 6**) at the Kashiwazaki-Kariwa site in Japan. The plant produced power for the electric grid 58 months after groundbreaking. Another advanced reactor concept is the System 80+, under



**Fig. 6.** Advanced boiling-water reactor (ABWR), the first advanced light-water reactor to be deployed commercially. (*General Electric*)

construction in Korea. In 1998, following a 12-year design program, the AP-600 plant became the first advanced light-water reactor to be certified by the U.S. Nuclear Regulatory Commission.

**Fuel loading and reactor operations.** Completed fuel assemblies are shipped from the fuel fabrication plant to the power plant in criticality-proof containers and on arrival are stored in a clean fuel vault.

Bolting or unbolting the head of a reactor vessel, to close or open it, is a complex undertaking. A ring of bolts each 2 ft (60 cm) long and 3 in. (8 cm) in diameter is used, with nuts of corresponding magnitude that are fastened in place with mechanical bolt-tighteners. At refueling time, the nuts are removed, and a polar crane at the top of the reactor containment building eases the vessel head off its seating and hoists it to one side onto a pad. The reactor cavity—the concrete pit in which the reactor vessel is moored—is filled with water to a height such that full-length (usually 12 ft or 4 m long) fuel assemblies lifted out of the core can swing clear of the lip of the reactor vessel without coming out of the water. The purpose of the water is to shield the workers from the radiation of the fuel assemblies. Those fuel assemblies removed are lowered into a canal that connects the reactor cavity with the spent-fuel pool by a bridge crane that spans the pool and runs on rails its full length, and as far as the reactor cavity. In this manner, the discharged assemblies are moved to an open position in the racks on the floor of the spent-fuel pool. These racks are made of a metal containing enough neutron poison material to make certain that the spent assemblies in the pool cannot go critical there.

The assemblies not ready for discharge are shuffled, that is, their position is rearranged in the core, and the fresh clean assemblies are then loaded into the core grid spaces freed. Underwater periscopes are provided for use in the reactor vessel and in the spent-fuel pool as required. Assemblies discharged are inspected visually by means of these periscopes to check their condition, and a log is kept, recording the serial number of each assembly moved into or out of the core. The radioactive content of the water is also checked to make certain that none of the fuel rods has developed any pinhole leaks that permit release of fission products into the water. See PERISCOPE.

After the reactor has been “buttoned up” again, that is, the vessel head secured back in place after completion of reloading, the director of reloading operations turns the reactor back to the operating staff for startup. In the process of starting up, a startup source—a neutron emitter—is used as there are not enough spontaneous neutrons emitted to permit close monitoring of the startup process by external neutron detectors. A neutron source (usually an antimony or a polonium-beryllium capsule) which emits a large number of neutrons spontaneously is used for this purpose, and is lowered into the core through a guide tube while the control rods are slowly and carefully retracted.

Radiation monitors around the core report re-

motely to the control room the neutron flux as it increases. Flux is the number of neutrons crossing a unit area per unit time and is proportional to the fission reaction rate. Once the reactor has become critical, the control rods are retracted further to increase power until the reactor is at 100% power.

**Spent-fuel storage and transportation.** The predominant fueling strategy requires that either one-third or one-fourth of the fuel assemblies be discharged from the core each year. While fresh assemblies are placed in the core to replenish its reactivity, the spent fuel is transferred to an adjacent pool where it is stored for several years. Spent fuel is highly radioactive, and the water serves as shielding and as a cooling medium, to remove the heat produced by the decaying fission products. After several years of cooling in the spent-fuel pool, the assemblies are suitable for transportation to a reprocessing plant. However, this option is not available in the United States, and spent fuel must be stored in the spent-fuel pool at the reactor. Since the amounts of stored fuel increases every year, and because utilities must maintain enough space in the pool for a full core discharge, spent fuel has required changes in storage. These include: redesigning or modifying the pools to allow for denser storage patterns in the same total space available; building more storage capacity at the reactor site; allowing transfers between reactor sites or between utilities to complement needs and available space; and building centralized storage capacity “away from reactor” (AFR) to accommodate the excess quantities. The Department of Energy was required by law to accept spent fuel for permanent storage or disposal, starting in the 1990s, but has not yet done so pending resolution of technical and legal issues.

Transportation of spent nuclear fuel and nuclear wastes has received special attention. With increasing truck and train shipments and increased probabilities for accidents, the protection of the public from radioactive hazards is achieved through regulations and implementation which provide transport packages with multiple barriers to withstand major accidents. For example, the cask used to transport irradiated fuel is designed to withstand severe drop, puncture, fire, and immersion tests. Actual train and truck collision tests have been done to demonstrate the integrity of the casks.

**Reprocessing and refabrication.** At the reprocessing center, the spent-fuel rods are stripped of cladding, and the spent-fuel pellets are dropped in a pool of nitric acid in which they dissolve. The solution is then fed to countercurrent extraction systems. Usually, in the first extraction cycle about 99% of the fission waste products are removed. Then further purification and separation of the plutonium from the uranium is performed. The end products of this step are usually uranium dioxide and plutonium dioxide ( $\text{PuO}_2$ ) which can be recycled. The separation is a straightforward chemical process that is carried out by the United States government for weapons material and for spent fuel from nuclear-propelled naval vessels. The reprocessing of commercial fuel is done in order to return the unfissioned fuel material to the

inventory of material to be used for fuel fabrication. Commercial nuclear fuel is reprocessed in France, Great Britain, Russia, and on a lesser scale in Belgium, Germany, Japan, India, and Italy. Commercial fuel reprocessing activities were discontinued in the United States in 1976. See NUCLEAR FUELS REPROCESSING.

**Radioactive waste management.** The fission waste products are removed from a reprocessing plant and disposed of in various ways. High-level waste can be concentrated into a glassy bead form, and either buried in salt beds deep in the earth or shipped to a heavily guarded disposal site. Low-level wastes are stored in liquid or solid form.

Critics of nuclear power consider the radioactive wastes generated by the nuclear industry to be too great a burden for society to bear. They argue that since the high-level wastes will contain highly toxic materials with long half-lives, such as a few tenths of one percent of plutonium that was in the irradiated fuel, the safekeeping of these materials must be assured for time periods longer than social orders have existed in the past. Nuclear proponents answer that the time required for isolation is much shorter, since only 500 to 1000 years is needed before the hazard posed by nuclear waste falls below that posed by common natural ore deposits in the environment (Fig. 7). The proposed use of bedded salts, for example, found in geologic formations that have prevented access of water and have been undisturbed for millions of years provides one of the options for assurance that safe storage can be engineered. A relatively small area of several hundred acres (a few hundred hectares) would be needed for disposal of projected wastes.

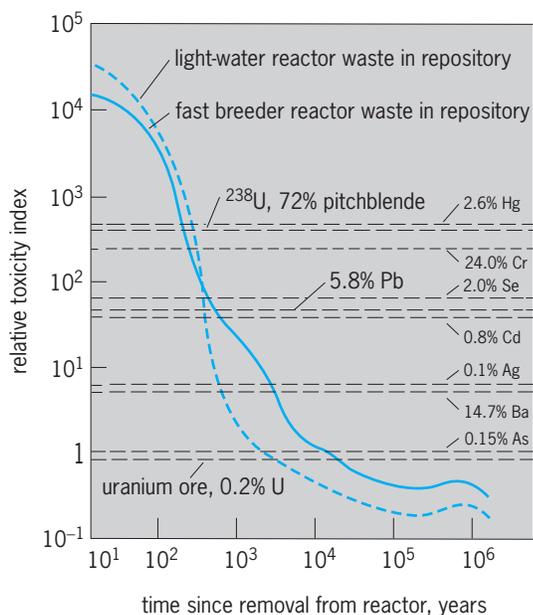


Fig. 7. Relative toxicity of nuclear waste over time, compared with that of average mineral ores of toxic elements. (After K. A. Tonnessen and J. J. Cohen, *Survey of naturally occurring hazardous materials in deep geologic formations: A perspective on the relative hazard of deep burial of nuclear wastes*, University of California and Lawrence Livermore Laboratories, UCRL-52199, 1977)

Management of low-level wastes generated by the nuclear energy industry requires use of burial sites for isolation of the wastes while they decay to innocuous levels. Other radioactive wastes, such as those from medical procedures and industrial applications, equal or exceed these from nuclear power plants. Operation of the commercial burial sites is subject to regulations by federal and state agencies.

Routine operations of nuclear power stations result in very small releases of radioactivity in the gaseous and water effluents. The NRC has adopted the principle that all releases should conform to the "as low as reasonably achievable" (ALARA) standard. ALARA guidance has been extended to other portions of the nuclear fuel cycle. See RADIOACTIVE WASTE MANAGEMENT.

**Decommissioning of nuclear plants.** There are several light-water reactors in the United States and many light-water reactors and gas-cooled reactors in Europe that have been decommissioned.

Decommissioning requires several steps: assessments and characterization of the hazards, environmental review, engineering, dismantling and decontamination of the plant, and final and remediation. The process requires that the plant be maintained cost-effectively, safely and securely while it awaits decommissioning, and that the site and buildings be made available for reuse following decommissioning (if practical). Some nuclear plants have been considered for conversion to a fossil-fuel power station.

In the United States, there are essentially three options for decommissioning: prompt DECON, delayed DECON, and SAFSTOR. Prompt DECON requires that the equipment, structures, and portions of a facility and site containing radioactive contaminants be removed or decontaminated to a level that permits unrestricted use. Delayed DECON is essentially the same as prompt DECON, but includes a delay for on-site spent fuel storage (usually 10–20 years) to allow sufficient time for the U.S. Department of Energy to develop a spent-fuel repository. SAFSTOR requires that the facility be placed and maintained in a condition that allows it to be safely stored and later decontaminated for unrestricted use, generally within about 60 years of reactor shutdown.

### Safety

Nuclear power facilities present a potential hazard rarely encountered with other facilities; that is, radiation. A major health hazard would result if, for instance, a significant fraction of the core inventory of a power reactor were released to the atmosphere. Such a release of radioactivity is clearly unacceptable, and steps are taken to assure it could never happen. These include use of engineered safety systems, various construction and design codes (for example, standards of the American Society for Testing and Materials), regulations on reactor operation, and periodic maintenance and inspection. In the last analysis, however, the ultimate safety of any facility depends on the ability of its designers to use the forces of nature so that a large release of radioactivity

is not possible. To help them, various techniques are employed, including conservative design margins, the use of safety equipment, and reliance on various physical barriers to radiation release in case all else fails.

It is the practice, in the United States and elsewhere, for regulatory bodies to establish licensing procedures for nuclear facilities. These procedures set design requirements, construction practices, operational limits, and the siting of such facilities. All power reactors built in the United States (and overseas except in the former Soviet Union) have a containment building and are sited in generally low or moderate population areas with exclusion areas defined by regulation.

All reactors have engineered safety features, both active and passive, designed to prevent serious accidents and mitigate them if they occur. A nuclear plant's safety is achieved through a concept of defense in depth. This provides a series of protective barriers with redundancy at each level and for each active component.

Every reactor has four main barriers to radioactivity release in the event of an accident:

1. *Fuel matrix.* The exceptionally high melting point (5000°F or 2760°C) and chemical stability of uranium dioxide prevent the escape of fission products except in extreme accident conditions. Although the fission process creates large amounts of radioactivity in the fuel rods, the ceramic pellets of uranium dioxide fuel retain more than 98% of this radioactivity. Without fuel melting and subsequent release of fission products, a nuclear reactor accident involves only hazards to the general public comparable with conventional power plant accidents.

2. *Fuel cladding.* The Zircaloy clad surrounding the fuel pellets retains any radioactivity released from the uranium dioxide. Fuel cladding behavior is of importance to the safety of a nuclear plant primarily because the fuel contains the major part of the radioactive products in the plant. The cladding is protected through use of design criteria which limit the heat in the core.

3. *Reactor primary coolant system.* Boundary integrity is assured by the thick steel vessel and piping up to 8 in. (20 cm) thick, and the continual inspection of these components. Licensing requirements specify that the design of the safety systems must accommodate ruptures in the reactor coolant loop, including a break of the largest coolant pipe. This constitutes the design basis for protection against loss-of-coolant accidents.

4. *Reactor containment building.* The reactor containment building generally consists of a 4-ft-thick (1.2-m) concrete shell lined with steel, and is heavily reinforced with steel bars. This steel, embedded in the concrete, gives a reactor containment building great strength to withstand forces that might occur in a reactor accident. See REINFORCED CONCRETE.

Each of these barriers is capable of retaining the hazardous radioactivity contained in a reactor core. Only if all were simultaneously breached would a

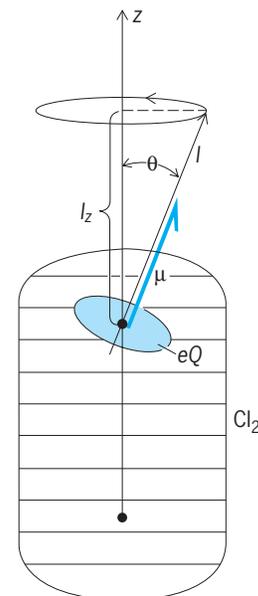
release to the environment be possible. To prevent breaches, and to attenuate any radioactivity, various other engineered safety features are employed. Their objectives are threefold: shut down the reactor (in the event that the normal control rod mechanisms fail), cool the core (to prevent overheating and melt-down), and safeguard the integrity of the barriers (such as limiting the pressure in a containment building). Frank J. Rahn

**Bibliography.** J. G. Collier and G. F. Hewitt, *Introduction to Nuclear Power*, 2d ed., 2000; S. Glasstone and A. Sesonske, *Nuclear Reactor Engineering*, 2 vols., 4th ed., 1993; R. L. Murray and J. A. Powell, *Understanding Radioactive Waste*, 4th ed., 1994; F. J. Rahn et al., *A Guide to Nuclear Power Technology*, 1984, reprint 1992; S. C. Stulz and J. B. Kitto, *Steam: Its Generation and Use*, 40th ed., Babcock & Wilcox Co., 1992.

## Nuclear quadrupole resonance

A selective absorption phenomenon observable in a wide variety of polycrystalline compounds containing nonspherical atomic nuclei when placed in a magnetic radio-frequency field. Nuclear quadrupole resonance (NQR) is very similar to nuclear magnetic resonance (NMR), and was originated in the late 1940s by H. G. Dehmelt and H. Krüger as an inexpensive (no stable homogeneous large magnetic field is required) alternative way to study nuclear moments. See MAGNETIC RESONANCE; NUCLEAR MAGNETIC RESONANCE (NMR).

**Principles.** In the simplest case, for example,  $^{35}\text{Cl}$  in solid  $\text{Cl}_2$ , NQR is associated with the precession of the angular momentum of the nucleus, depicted in the **illustration** as a flat ellipsoid of rotation, around the symmetry axis (taken as the  $z$  axis) of the  $\text{Cl}_2$  molecule fixed in the crystalline solid. (The direction



Interaction of  $^{35}\text{Cl}$  nucleus with the electric field of a  $^{2}\text{Cl}$  molecule.

of the nuclear angular momentum coincides with those of the symmetry axis of the ellipsoid and of the nuclear magnetic dipole moment  $\mu$ .) The precession, with constant angle  $\theta$  between the nuclear axis and symmetry axis of the molecule, is due to the torque which the inhomogeneous molecular electric field exerts on the nucleus of electric quadrupole moment  $eQ$ . This torque corresponds to the fact that the electrostatic interaction energy of the nucleus with the molecular electric field depends on the angle  $\theta$ . The interaction energy is given by Eq. (1), where  $\rho$

$$E = \int \phi \rho \, dV \quad (1)$$

is the nuclear charge density distribution and  $\phi$  is the potential of the molecular electric field. Its dependence on  $\theta$  is given by Eq. (2), where  $\phi_{zz}$  is the

$$E = \frac{eQ\phi_{zz}(3\cos^2\theta - 1)}{8} \quad (2)$$

axial gradient of the (approximately) axially symmetric molecular electric field. The quantum-mechanical analog of this expression is Eq. (3), where  $I$  and  $m$

$$E_m = \frac{eQ\phi_{zz}[3m^2 - I(I+1)]}{4I(2I-1)} \quad (3)$$

denote the quantum numbers of the nuclear angular momentum and its  $z$  component  $I_z$ . The absorption occurs classically when the frequency of the rf field  $\nu$  and that of the precession motion of the angular momentum coincide, or quantum-mechanically when Eq. (4) is satisfied, where  $m$  and  $m'$  are given by Eqs. (5) and  $m' - m = +1$ , corresponding to the

$$h\nu = |E_{m'} - E_m| \quad (4)$$

$$m, m' = 0 \pm 1, \pm 2 \dots \pm I \quad \text{for integer } I \geq 1 \quad (5)$$

$$m, m' = \pm 1/2, \pm 3/2 \dots \pm I \quad \text{for half - integer } I \geq 3/2$$

magnetic dipole transitions. See NUCLEAR MOMENTS.

It is not necessary that the rf field direction is perpendicular to  $z$ ; a nonvanishing perpendicular component suffices. This eliminates the necessity of using single crystals and makes it practical, unlike in the NMR of solids, to use polycrystalline samples of unlimited mass and volume. In fact, a polycrystalline natural sulfur sample of 3.18 quarts (3 liters) volume was used in work on the rare (0.74% abundance)  $^{33}\text{S}$  isotope.

**Techniques.** The original NQR work was done on approximately 3 in.<sup>3</sup> (50 cm<sup>3</sup>) of frozen *trans*-dichloroethylene submerged in liquid air using a superregenerative detector. The oscillator incorporated a vibrating capacitor driven by the power line to sweep a frequency band about 50 kHz wide over the approximately 10-kHz-wide  $^{35}\text{Cl}$  and  $^{37}\text{Cl}$  resonances near 30 MHz, and an oscilloscopic signal display was used. This work demonstrated the good sensitivity and rugged character of these simple, easily tunable circuits which are capable of combining high rf power levels with low noise. Their chief disadvantage is the occurrence of side bands spaced by the quench frequency which may confuse the line shape.

For nuclear species of low abundance it becomes important to use nuclear modulation. In the  $^{33}\text{S}$  work zero-based magnetic-field pulses periodically smearing out the absorption line proved satisfactory.

**Application.** NQR spectra have been observed in the approximate range 1-1000 MHz. Such a range clearly requires more than one spectrometer. Most of the NQR work has been on molecular crystals. While halogen-containing (Cl, Br, I) organic compounds have been in the forefront since the inception of the field, NQR spectra have also been observed for K, Rb, Cs, Cu, Au, Ba, Hg, B, Al, Ga, In, La, N, As, Sb, Bi, S, Mn, Re, and Co isotopes. For molecular crystals the coupling constants  $eQ\phi_{zz}$  found do not differ very much from those measured for the isolated molecules in microwave spectroscopy. The most precise nuclear information which may be extracted from NQR  $eQ\phi_{zz}$  data are quadrupole moment ratios of isotopes of the same element, since one may assume that  $\phi_{zz}$  is practically independent of the nuclear mass. As far as  $\phi_{zz}$  values may be estimated from atomic fine structure data, for example, for  $\text{Cl}_2$  where a pure  $p$ -bond is expected and the molecular nature of the solid is suggested by a low boiling point and so forth, fair  $Q$  values may be obtained. However, it has also proved very productive to use the quadrupole nucleus as a probe of bond character and orientation and crystalline electric fields and lattice sites, and a large body of data has been accumulated in this area.

See MICROWAVE SPECTROSCOPY. Hans Dehmelt

Bibliography. I. P. Biryukov, M. G. Voronkov, and I. A. Safin, *Tables of Nuclear Quadrupole Resonance Frequencies*, 1969; H. Chihara and N. Nakamura, *Landolt-Bornstein Numerical Data and Functional Relationships in Science and Technology*, vol. 31, subvol. B: *Nuclear Quadrupole Resonance Spectroscopy Data*, 1993; T. P. Das and E. L. Hahn, *Nuclear Quadrupole Resonance Spectroscopy*, 1958; H. G. Dehmelt, Nuclear quadrupole resonance (in solids), *Amer. J. Phys.*, 22:110-120, 1954, and *Faraday Soc. Discuss.*, 19:263-274, 1955; G. K. Semin, T. A. Babushkina, and G. G. Yakobson, *Nuclear Quadrupole Resonance in Chemistry*, 1975; J. A. S. Smith (ed.), *Advances in Nuclear Quadrupole Resonance*, vols. 1-5, 1974-1983.

## Nuclear radiation

All particles and radiations emanating from an atomic nucleus due to radioactive decay and nuclear reactions. Thus the criterion for nuclear radiations is that a nuclear process is involved in their production. The term was originally used to denote the ionizing radiations observed from naturally occurring radioactive materials. These radiations were alpha particles (energetic helium nuclei), beta particles (negative electrons), and gamma rays (electromagnetic radiation with wavelength much shorter than visible light). See BETA PARTICLES; GAMMA RAYS.

Nuclear radiations have traditionally been considered to be of three types based on the manner in which they interact with matter as they pass through

it. These are the charged heavy particles with masses comparable to that of the nuclear mass (for example, protons, alpha particles, and heavier nuclei), electrons (both negatively and positively charged), and electromagnetic radiation. For all of these, the interactions with matter are considered to be primarily electromagnetic. (The neutron, which is also a nuclear radiation, behaves quite differently.) The behavior of mesons and other particles is intermediate between that of the electron and heavy charged particles. *See* CHARGED PARTICLE BEAMS.

A striking difference in the absorption of the three types of radiations is that only heavy charged particles have a range. That is, a monoenergetic beam of heavy charged particles, in passing through a certain amount of matter, will lose energy without changing the number of particles in the beam. Ultimately, they will be stopped after crossing practically the same thickness of absorber. The minimum amount of absorber that stops a particle is its range. The greatest part of the energy loss results from collisions with atomic electrons, causing the electrons to be excited or freed. The energy loss per unit path length is the specific energy loss, and its average value is the stopping power of the absorbing substance.

For electromagnetic radiation (gamma rays) and neutrons, on the other hand, the absorption is exponential; that is, the intensity decreases in such a way that the equation below is valid, where  $I$  is the in-

$$-\frac{dI}{I} = \mu dx$$

tensity of the primary radiation,  $\mu$  is the absorption coefficient, and  $dx$  is the thickness traversed. The difference in behavior reflects the fact that charged particles are not removed from the beam by individual interactions, whereas gamma radiation photons (and neutrons) are. Three main types of phenomena involved in the interaction of electromagnetic radiation with matter (namely, photoelectric absorption, Compton scattering, and electron-positron production) are responsible for this behavior.

Electrons exhibit a more complex behavior. They radiate electromagnetic energy easily because they have a large charge-to-mass ratio and hence are subject to violent acceleration under the action of the electric forces. Moreover, they undergo scattering to such an extent that they follow irregular paths. *See* ELECTRON.

Whereas in the case of the heavy charged particles, electrons, or gamma rays the energy loss is mostly due to electromagnetic effects, neutrons are slowed down by nuclear collisions. These may be inelastic collisions, in which a nucleus is left in an excited state, or elastic collisions, in which the colliding nucleus acquires part of the energy (of the order of 1 MeV) to excite the collision partner. With less kinetic energy, only elastic scattering can slow down the neutron, a process which is effective down to thermal energies (approximately 1/40 keV). At this stage the collision, on the average, has no further effect on the energy of the neutron. *See* NEUTRON.

As noted previously, the other nuclear radiations

such as mesons have behaviors which are intermediate between that of heavy charged particles and electrons. Another radioactive decay product is the neutrino; because of its small interaction with matter, it is not ordinarily considered to be a nuclear radiation. *See* MESON; NEUTRINO; NUCLEAR REACTION.

Dennis G. Kovar

## Nuclear radiation (biology)

Nuclear radiations are used in biology because of their common property of ionizing matter. This makes their detection relatively simple, or makes possible the production of biological effects in any living cell. Nuclear radiations originate in atomic nuclei, either spontaneously, as in radioactive substances, or through interactions with neutrons, photons, and so on. Gamma radiation originates in atomic nuclei and constitutes one kind of nuclear radiation, but it is otherwise indistinguishable in its effects from x-radiation produced by extranuclear reactions. Because x-rays have been readily available for many years, they have been used more extensively in biology and medicine than gamma rays. Therefore, x-rays must be included in any discussion of the biological and medical uses of nuclear radiations.

**Ionizing radiation.** Ionizing radiation is any electromagnetic or particulate radiation capable of producing ions, directly or indirectly, in its passage through matter.

*Electromagnetic radiations.* X-rays and gamma rays are electromagnetic radiations, traveling at the speed of light as packages of energy called photons. They ionize indirectly by first ejecting electrons at high speed from the atoms with which they interact; these secondary electrons then produce most of the ionization associated with the primary radiation. *See* GAMMA RAYS; PHOTON; X-RAYS.

*Particulate radiation.* Fast neutrons are particulate radiation consisting of nuclear particles of mass number 1 and zero charge, traveling at high speed. They ionize indirectly, largely by setting in motion charged particles from the atomic nuclei with which they collide. Slow or thermal neutrons ionize indirectly by interacting with nuclei, in a process known as neutron capture, to produce ionizing radiation. *See* NEUTRON.

Alpha rays are particulate radiation consisting of helium nuclei traveling at high speed. Since they are charged particles, they ionize directly. Alpha particles are emitted spontaneously by some radioactive nuclides or may result from neutron capture; for example, neutron capture by boron-10 produces lithium-7 and an alpha particle. The energy of alpha particles emitted by radioactive substances is of the order of a few megaelectronvolts, but alpha particles of very much higher energy may be produced in cyclotrons or other particle accelerators from helium-ion beams. With such machines, other ionizing particles of very high energy, such as protons, deuterons, and so on, may also be produced.

Beta rays are particulate radiation consisting of electrons or positrons emitted from a nucleus during beta decay and traveling at high speed. Since they are charged particles, that is,  $-$  or  $+$ , they ionize directly. Electron beams of very high energy may be produced by high-voltage accelerators, but in that case they are not called beta particles. A pair consisting of one electron and one positron may be formed by one high-energy (1.022-MeV) photon where it traverses a strong electric field, such as that surrounding a nucleus or an electron. Subsequently, the positron and another electron react, and their mass is transformed into energy in the form of two photons traveling in opposite directions. This is called the annihilation process and is the inverse of the pair-production process that is mentioned above. Ionizing radiations, such as protons, deuterons, alpha particles, and neutrons, may be produced simultaneously by spallation when a very high-energy particle collides with an atom. In a photographic emulsion or in a cloud chamber, the ionizing particles originating from a common point form stars. *See* BETA PARTICLES.

Fission occurs in certain heavy nuclei spontaneously or through interaction with neutrons, charged particles, or photons, and it results in the division of the nucleus into two approximately equal parts. These fragments produced by fission are endowed with very large amounts of kinetic energy, carry large positive charges, and produce very dense ionization in their short passage through matter. *See* NUCLEAR FISSION.

Primary cosmic rays probably consist of atomic nuclei, mainly protons, with extremely high energies which interact with nuclei and electrons in the atmosphere and produce secondary cosmic rays, consisting mainly of mesons, protons, neutrons, electrons, and photons of lower energy. *See* COSMIC RAYS; MESON; PROTON.

*Biological effects.* All ionizing radiations produce biological changes, directly by ionization or excitation of the atoms in the molecules of biological entities, such as in chromosomes, or indirectly by the formation of active radicals or deleterious agents, through ionization and excitation, in the medium surrounding the biological entities. Ionizing radiation, having high penetrating power, can reach the most vulnerable part of a cell, an organ, or a whole organism, and is thus very effective. In terms of the energy absorbed per unit mass of a biological entity in which an effect is produced, some ionizing radiations are more effective than others. The relative biological effectiveness (RBE) depends in fact on the density of ionization (also termed the specific ionization or linear energy transfer, LET) along the path of the ionizing particle rather than on the nature of the particle itself. Relative biological effectiveness depends also on many other factors. *See* LINEAR ENERGY TRANSFER (BIOLOGY); RADIATION BIOLOGY.

**Use in medicine.** The medical uses of nuclear radiations may be divided into three distinct classes:

1. The radiations, which are principally x-rays, are used to study the anatomical configuration of body organs, usually for the purpose of detecting

abnormalities as an aid in diagnosis.

2. The radiations are used for therapeutic purposes to produce biological changes in such tissues as tumors.

3. The radiations are used as a simple means of tracing a suitable radioactive substance through different steps in its course through the body, in the study of some particular physiological process. *See* RADIOLOGY.      *Gioacchino Failla; Edith H. Quimby*

**Use in biological research.** The radiations emitted by radioactive isotopes of the various elements are used in biological research. The most useful ones in biological research are the isotopes of the elements which are important in metabolism and in the structural materials of cells. These include carbon, hydrogen, sulfur, and phosphorus. Unfortunately, nitrogen and oxygen do not have usable radioisotopes. Isotopes of calcium, iodine, potassium, sodium, iron, and a few others have more limited usefulness in biological research. In addition, the radioactive metals like cobalt-60, radium, and others can be used to produce radiations for external application to cells and tissues. *See* RADIOISOTOPE (BIOLOGY).

Most of the isotopes mentioned emit beta particles when they decay, and a few emit gamma rays. Therefore, they can be easily detected by various means. If the radiations emitted are highly penetrating, like the gamma rays and the high energy beta particles from phosphorus-32, the presence of the isotope may be detected with Geiger counters or scintillation counters applied to the surface of the biological material. Likewise, application of photographic emulsions to the surface or to cut surfaces of cells or tissues may serve to locate the isotope. When the biological material can be broken down and destroyed, particular compounds or larger components of cells may be isolated, and the isotope determined by the various types of radiation detectors. These procedures are sometimes used in the study of the movement of elements or their compounds in plants and animals. They are frequently used for tracing the sequence of reactions in metabolism. The great advantage of radioisotopes for such studies is that small amounts or concentrations may be readily detected and measured, either in the presence of other isotopes of the same element or when mixed with a variety of other elements. *See* AUTORADIOGRAPHY.

In addition to the radiations emitted, some of the elements change to a different element when they decay. Phosphorus-32 changes to sulfur, sulfur-35 changes to chlorine, and tritium (hydrogen-3) changes to helium when they decay by the emission of an electron, a beta particle. Therefore, in addition to the radiation produced, the transmutation of the element affects the molecule and the cell of which it is a part. Attempts have been made to evaluate the effects of the two factors by giving the fungus *Neurospora* the same amount of radiation, and by having different proportions of the radioactive isotope incorporated into the molecules of the cell. The incorporation of the isotope was regulated by having the same concentration of the radioisotope in two cultures, but in one, the radiosulfur or radiophosphorus

was diluted greatly with the nonradioactive isotope. A difference in lethal effect was demonstrated.

In other experiments, the decay of phosphorus-32 has been used to give information on the nature and importance of the phosphorus-containing molecules to the survival or reproduction of a cell or virus particle. When bacterial viruses are allowed to grow in the presence of phosphate containing phosphorus-32, they incorporate the radioisotope into their genetic material, deoxyribonucleic acid (DNA). If these virus particles are then stored under conditions in which no growth occurs, the decay of the radioisotope is very effective in inactivating the viruses. About 1 in 10 atoms which decay inactivates 1 virus particle. From such experiments, biologists are able to learn something about the importance of the molecule as well as something of its size and possible organization. Similar experiments have been carried out with the cells of ameba, but the lethal effect was observed, not on the stored cells, but on their offspring. From this experiment, deductions concerning the organization of genetic material and its mutability were drawn. See DEOXYRIBONUCLEIC ACID (DNA).

Nuclear radiations have also proved useful in many studies of the nature and the mechanisms of the effects of radiations on cells and cell constituents. The radiations with low penetrating power, for example, alpha particles and low-energy beta particles, can be most effective when an element which will emit the particles is placed inside the cells to be studied. Radon, which is a gas, can be used in this way for the production of alpha particles. Likewise, a variety of the heavy metals like thorium, uranium, and polonium emit alpha particles. Various beta emitters, such as phosphorus-32, carbon-14, sulfur-35, and tritium, can also be used for producing radiations inside the cell. One of the most interesting of this group is tritium, which emits very soft beta particles. Their maximum range is 6 micrometers in water and about the same in tissues, but the average range is much less, about 1 micrometer or  $1/25$  the diameter of a medium-sized cell. Tritium can be put into the cell as water or in many other compounds. Many of these are unselective and are equally distributed to all parts of the cell. However, there is one substance, thymidine, which selectively labels the DNA which is restricted to the cell nucleus. If the cell is a relatively large one compared to its nuclear size, nearly all of the radiation will be absorbed by the nucleus, while the other part of the cell is irradiated hardly at all. Experiments have shown that tritium given to cells in the form of tritium-thymidine is about 1000 times as effective as tritium-water in producing radiation effects. See HISTORADIOGRAPHY; SCINTILLATION COUNTER.

J. Herbert Taylor

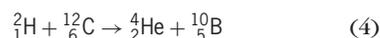
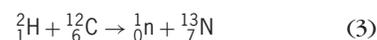
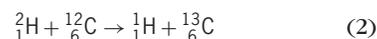
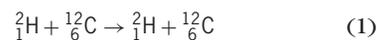
## Nuclear reaction

A process that occurs as a result of interactions between atomic nuclei when the interacting particles approach each other to within distances of the order of nuclear dimensions ( $\sim 10^{-12}$  cm). While nuclear

reactions occur in nature, understanding of them and use of them as tools have taken place primarily in the controlled laboratory environment. In the usual experimental situation, nuclear reactions are initiated by bombarding one of the interacting particles, the stationary target nucleus, with nuclear projectiles of some type, and the reaction products and their behaviors are studied. The study of nuclear reactions is the largest area of nuclear and subnuclear (or particle) physics; the threshold for producing pions has historically been taken to be the energy boundary between the two fields.

**Types of nuclear interaction.** As a generalized nuclear process, consider a collision in which an incident particle strikes a previously stationary particle to produce an unspecified number of final products. If the final products are the same as the two initial particles, the process is called scattering. The scattering is said to be elastic or inelastic, depending on whether some of the kinetic energy of the incident particle is used to raise either of the particles to an excited state. If the product particles are different from the initial pair, the process is referred to as a reaction.

The most common type of nuclear reaction, and the one which has been most extensively studied, involves the production of two final products. Such reactions can be observed, for example, when deuterons with a kinetic energy of a few megaelectronvolts are allowed to strike a carbon nucleus of mass 12. Protons, neutrons, deuterons, and alpha particles are observed to be emitted, and reactions (1)-(4) are responsible. In these equations the nuclei



are indicated by the usual chemical symbols; the subscripts indicate the atomic number (nuclear charge) of the nucleus, and the superscripts the mass number of the particular isotope. These reactions are conventionally written in the compact notation  ${}^{12}\text{C}(d,d){}^{12}\text{C}$ ,  ${}^{12}\text{C}(d,p){}^{13}\text{C}$ ,  ${}^{12}\text{C}(d,n){}^{13}\text{N}$ , and  ${}^{12}\text{C}(d,\alpha){}^{10}\text{B}$ , where  $d$  represents deuteron,  $p$  proton,  $n$  neutron, and  $\alpha$  alpha particle. In each of these cases the reaction results in the production of an emitted light particle and a heavy residual nucleus. The  $(d,d)$  process denotes the elastic scattering as well as the inelastic scattering processes that raise the  ${}^{12}\text{C}$  nucleus to one of its excited states. The other three reactions are examples of nuclear transmutation or disintegration where the residual nuclei may also be formed in their ground states or one of their many excited states. The processes producing the residual nucleus in different excited states are considered to be the different reaction channels of the particular reaction. If the residual nucleus is formed in an excited state, it will subsequently emit this excitation energy in

the form of gamma rays or, in special cases, electrons. The residual nucleus may also be a radioactive species, as in the case of  $^{13}\text{N}$  formed in the  $^{12}\text{C}(d,n)$  reaction. In this case the residual nucleus will undergo further transformation in accordance with its characteristic radioactive decay scheme. See GAMMA RAYS; RADIOACTIVITY.

**Nuclear cross section.** In general, one is interested in the probability of occurrence of the various reactions as a function of the bombarding energy of the incident particle. The measure of probability for a nuclear reaction is its cross section. Consider a reaction initiated by a beam of particles incident on a region which contains  $N$  atoms per unit area (uniformly distributed), and where  $I$  particles per second striking the area result in  $R$  reactions of a particular type per second. The fraction of the area bombarded which is effective in producing the reaction products is  $R/I$ . If this is divided by the number of nuclei per unit area, the effective area or cross section  $\sigma = R/IN$ . This is referred to as the total cross section for the specific reaction, since it involves all the occurrences of the reaction. The dimensions are those of an area, and total cross sections are expressed in either square centimeters or barns ( $1 \text{ barn} = 10^{-24} \text{ cm}^2$ ). The differential cross section refers to the probability that a particular reaction product will be observed at a given angle with respect to the beam direction. Its dimensions are those of an area per unit solid angle (for example, barns per steradian).

**Requirements for a reaction.** Whether a specific reaction occurs and with what cross section it is observed depend upon a number of factors, some of which are not always completely understood. However, there are some necessary conditions which must be fulfilled if a reaction is to proceed.

**Coulomb barrier.** For a reaction to occur, the two interacting particles must approach each other to within the order of nuclear dimensions ( $\sim 10^{-12} \text{ cm}$ ). With the exception of the uncharged neutron, all incident particles must therefore have sufficient kinetic energy to overcome the electrostatic (Coulomb) repulsion produced by the intense electrostatic field of the nuclear charge. The kinetic energy must be comparable to or greater than the so-called Coulomb barrier, whose magnitude is approximately given by the expression  $E_{\text{Coul}} \approx Z_1 Z_2 / (A_1^{1/3} + A_2^{1/3}) \text{ MeV}$ , where  $Z$  and  $A$  refer to the nuclear charge and mass number of the interacting particles 1 and 2, respectively. It can be seen that while, for the lightest targets, protons with kinetic energies of a few hundred kiloelectronvolts are sufficient to initiate reactions, energies of many hundreds of megaelectronvolts are required to initiate reactions between heavier nuclei. In order to provide energetic charged particles, to be used as projectiles in reaction studies, particle accelerators of various kinds (such as Van de Graaff generators, cyclotrons, and linear accelerators) have been developed, making possible studies of nuclear reactions induced by projectiles as light as protons and as heavy as  $^{208}\text{Pb}$ . See ELECTROSTATICS; PARTICLE ACCELERATOR.

Since neutrons are uncharged, they are not re-

pelled by the electrostatic field of the target nucleus, and neutron energies of only a fraction of an electronvolt are sufficient to initiate some reactions. Neutrons for reaction studies can be obtained from nuclear reactors or from various nuclear reactions which produce neutrons as reaction products. There are two other means of producing nuclear reactions which do not fall into the general definition given above. Both electromagnetic radiation and high-energy electrons are capable of disintegrating nuclei under special conditions. However, both interact much less strongly with nuclei than nucleons or other nuclei, through the electromagnetic and weak nuclear forces, respectively, rather than the strong nuclear force responsible for nuclear interactions. See NEUTRON.

**Q value.** For a nuclear reaction to occur, there must be sufficient kinetic energy available to bring about the transmutation of the original nuclear species into the final reaction products. The sum of the kinetic energies of the reaction products may be greater than, equal to, or less than the sum of the kinetic energies before the reaction. The difference in the sums is the  $Q$  value for that particular reaction. It can be shown that the  $Q$  value is also equal to the difference in the masses (rest energies) of the reaction products and the masses of the initial nuclei. Reactions with a positive  $Q$  value are called exoergic or exothermic reactions, while those with a negative  $Q$  value are called endoergic or endothermic reactions.

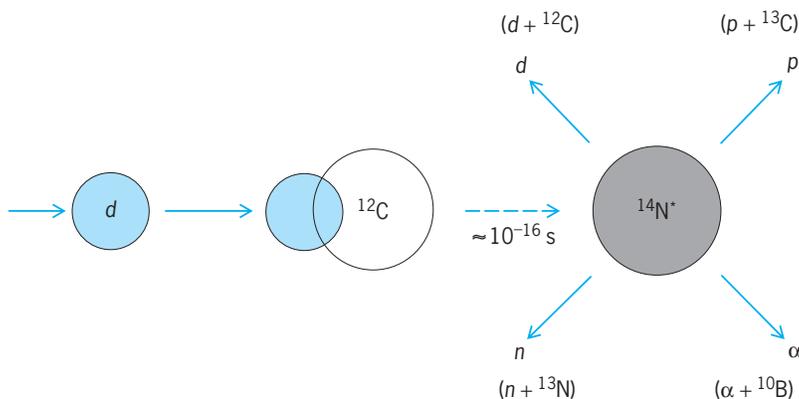
In reactions (1)–(4), where the residual nuclei are formed in their ground states, the  $Q$  values are  $^{12}\text{C}(d,d)^{12}\text{C}$ ,  $Q = 0.0 \text{ MeV}$ ;  $^{12}\text{C}(d,p)^{13}\text{C}$ ,  $Q = 2.72 \text{ MeV}$ ;  $^{12}\text{C}(d,n)^{13}\text{N}$ ,  $Q = -0.28 \text{ MeV}$ ; and  $^{12}\text{C}(d,\alpha)^{10}\text{B}$ ,  $Q = -1.34 \text{ MeV}$ . For reactions with a negative  $Q$  value, a definite minimum kinetic energy is necessary for the reaction to take place. While there is no threshold energy for reactions with positive  $Q$  values, the cross section for the reactions induced by charged particles is very small unless the energies are sufficient to overcome the Coulomb barrier. A nuclear reaction and its inverse are reversible in the sense that the  $Q$  values are the same but have the opposite sign [for example, the  $Q$  value for the  $^{10}\text{B}(\alpha,d)^{12}\text{C}$  reaction is  $+1.39 \text{ MeV}$ ].

**Conservation laws.** It has been found experimentally that certain physical quantities must be the same both before and after the reaction. The quantities conserved are electric charge, number of nucleons, energy, linear momentum, angular momentum, and in most cases parity. Except for high-energy reactions involving the production of mesons, the conservation of charge and number of nucleons allow one to infer that the numbers of protons and neutrons are always conserved. The conservation of the number of nucleons indicates that the statistics governing the system are the same before, during, and after the reaction. Fermi-Dirac statistics are obeyed if the total number is odd, and Bose-Einstein if the number is even. The conservation laws taken together serve to strongly restrict the reactions that can take place, and the conservation of angular momentum and parity in particular allow one to establish spins and parities

of states excited in various reactions. See ANGULAR MOMENTUM; CONSERVATION LAWS (PHYSICS); PARITY (QUANTUM MECHANICS); QUANTUM STATISTICS; SYMMETRY LAWS (PHYSICS).

**Reaction mechanism.** What happens when a projectile collides with a target nucleus is a complicated many-body problem which is still not completely understood. Progress made in the last decades has been in the development of various reaction models which have been extremely successful in describing certain classes or types of nuclear reaction processes. In general, all reactions can be classified according to the time scale on which they occur, and the degree to which the kinetic energy of the incident particle is converted into internal excitation of the final products. A large fraction of the reactions observed has properties consistent with those predicted by two reaction mechanisms which represent the extremes in this general classification. These are the mechanisms of compound nucleus formation and direct interaction.

*Compound nucleus formation.* As originally proposed by N. Bohr, the process is envisioned to take place in two distinct steps. In the first step the incident particle is captured by (or fuses with) the target nucleus, forming an intermediate or compound nucleus which lives a long time ( $\sim 10^{-16}$  s) compared to the approximately  $10^{-22}$  s it takes the incident particle to travel past the target. During this time the kinetic energy of the incident particle is shared among all the nucleons, and all memory of the incident particle and target is lost. The compound nucleus is always formed in a highly excited unstable state, is assumed to approach thermodynamic equilibrium involving all or most of the available degrees of freedom, and will decay, as this second step, into different reaction products, or through so-called exit channels. In most cases, the decay can be understood as a statistical evaporation of nucleons or light particles. In the examples of reactions (1)–(4), the compound nucleus formed is  $^{14}\text{N}$ , and four possible exit channels are indicated (**Fig. 1**). In reactions involving heavier targets (for example,  $A \approx 200$ ), one of the exit channels may be the fission channel where the compound nucleus splits into two large fragments. See NUCLEAR FISSION.



**Fig. 1.** Formation of the compound nucleus  $^{14}\text{N}$  after capture of the deuteron by  $^{12}\text{C}$ . Four exit channels are indicated.

The essential feature of the compound nucleus formation or fusion reaction is that the probability for a specific reaction depends on two independent probabilities: the probability for forming the compound nucleus, and the probability for decaying into that specific exit channel. While certain features of various interactions cannot be completely explained within the framework of the compound nucleus hypothesis, it appears that the mechanism is responsible for a large fraction of reactions occurring in almost all projectile-target interactions. Fusion reactions have been extremely useful in several kinds of spectroscopic studies. Particularly notable have been the resonance studies performed with light particles, such as neutrons, protons, deuterons, and alpha particles, on light target nuclei, and the gamma-ray studies of reactions induced by heavy projectiles, such as  $^{16}\text{O}$  and  $^{32}\text{S}$ , on target nuclei spanning the periodic table. These studies have provided an enormous amount of information regarding the excitation energies and spins of levels in nuclei. See NUCLEAR SPECTRA.

*Direct interactions.* Some reactions have properties which are in striking conflict with the predictions of the compound nucleus hypothesis. Many of these are consistent with the picture of a mechanism where no long-lived intermediate system is formed, but rather a fast mechanism where the incident particle, or some portion of it, interacts with the surface, or some nucleons on the surface, of the target nucleus. Models for direct processes make use of a concept of a homogeneous lump of nuclear matter with specific modes of excitation, which acts to scatter the incident particle through forces described, in the simplest cases, by an ordinary spherically symmetric potential. In the process of scattering, some of the kinetic energy may be used to excite the target, giving rise to an inelastic process, and nucleons may be exchanged, giving rise to a transfer process. In general, however, direct reactions are assumed to involve only a very small number of the available degrees of freedom.

Most direct reactions are of the transfer type where one or more nucleons are transferred to or from the incident particle as it passes the target, leaving the two final partners either in their ground states or in one of their many excited states. Such transfer reactions are generally referred to as stripping or pick-up reactions, depending on whether the incident particle has lost or acquired nucleons in the reaction. The  $(d,p)$  reaction is an example of a stripping reaction, where the incident deuteron is envisioned as being stripped of its neutron as it passes the target nucleus, and the proton continues along its way (**Fig. 2**).

The properties of the target nucleus determine the details of the reaction, fixing the energy and angular momentum with which the neutron must enter it. The energy of the outgoing proton is determined by how much of the deuteron's energy is taken into the target nucleus by the neutron, and indeed serves to identify the final state populated by the  $Q$  value of the reaction as a whole. The angular distribution of the differential cross sections will, at appropriate bombarding energies, not be smooth but rather will

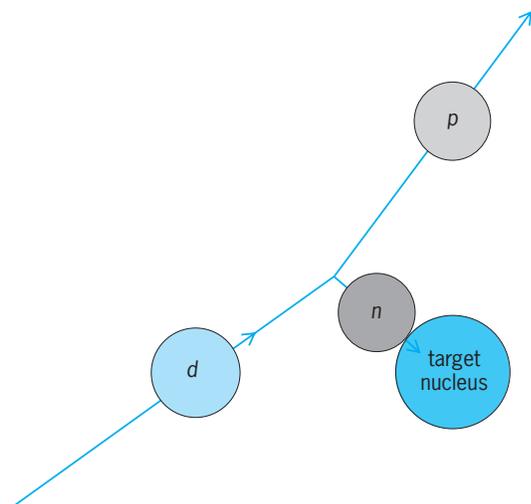


Fig. 2. A  $(d, p)$  transfer reaction.

show a distinct pattern of maxima and minima which are indicative of the spin and parity of the final state. The cross section for populating a specific final state in the target nucleus depends on the nuclear structure of the nuclei involved. This sensitivity has been used in studies of single-nucleon, two-nucleon, and four-nucleon transfer reactions, such as  $(d, p)$ ,  $(p, d)$ ,  $(^3\text{He}, d)$ ,  $(t, p)$ , and  $(^7\text{Li}, d)$ , to establish the validity and usefulness of the shell-model description of nuclei. Multinucleon transfer reactions with heavier projectiles have been powerful tools for reaching nuclei inaccessible by other means and in producing new isotopes.

Inelastic scattering is also a direct reaction whose angular distribution can provide information about the spin and parity of the excited state. Whereas the states preferentially populated in transfer reactions are those of specific single-particle or shell-model structure, the states preferentially excited in inelastic scattering are collective in nature. The states are most easily understood in the framework of macroscopic descriptions in which they are considered to be oscillations in shape about a spherical mean (vibrations) or to be the rotations of a statically deformed shape. The cross section for inelastic scattering is related to the shape or deformation of the target nucleus in its various collective excitations. Inelastic excitation can be caused by both a nuclear interaction and an electromagnetic interaction known as Coulomb excitation, where the target nucleus interacts with the rapidly changing electric field caused by the passage of the charged particle. Coulomb excitation is an important process at low bombarding energies in interactions involving heavy projectiles. Studies in inelastic scattering to low-lying states of nuclei across the periodic table have provided graphic demonstrations of the collective aspects of nuclei. See COULOMB EXCITATION.

Elastic scattering is the direct interaction which leaves the interacting particles unchanged. For charged particles at low bombarding energies, the elastic scattering is well described in terms of the inverse-square force law between two electrically

charged bodies. In this case, the process is known as Rutherford scattering. At higher energies the particles come into the range of the nuclear force, and the elastic scattering deviates from the inverse-square behavior. See NUCLEAR STRUCTURE; SCATTERING EXPERIMENTS (NUCLEI).

*More complex reaction mechanisms.* Processes intermediate between direct and compound nucleus formation do occur. The best example of such a process is the so-called preequilibrium emission, where light particles are emitted before the kinetic energy has been shared among all the nucleons in the compound nucleus. Another example is seen in the interaction of two massive nuclei, such as  $^{84}\text{Kr} + ^{209}\text{Bi}$ , where the probability for formation of the compound nucleus is found to be very small. The experimental observations indicate that the nuclei interact for a short time and then separate in what appears to be a direct reaction. Although the interaction times for these so-called strongly damped or deep inelastic collisions are very small compared to that for compound nucleus formations, they are sufficiently long for considerable mass transfer and loss of relative kinetic energy to occur.

**Nuclear reaction studies.** In most instances the study of nuclear reactions is directed toward the long-range goal of obtaining information about the properties and structure of nuclei. Such studies usually proceed in two stages. In the first stage the attention focuses on the mechanism of the reaction and on establishing the dependence on the nuclei involved. The specific models proposed continue to be modified and improved, as they are confronted with more experimental results, until their predictions become reliable. At that point the second stage is entered, in which the focus of the effort is on extraction of information about nuclei.

There are other studies which are focused on reaction cross-section behaviors for other purposes. Examples of this are the neutron-capture reactions on heavy target nuclei that fission, and the  $^3\text{H}(d, n)^4\text{He}$  reaction, important in thermonuclear processes, which continue to be studied because of their application as energy sources. Studies of the interactions of light nuclei at low energies are intensely studied because of their implications for astrophysics and cosmology. See NUCLEAR FUSION; NUCLEOSYNTHESIS; THERMONUCLEAR REACTION.

Dennis G. Kovar

**Collisions of very high energy nuclei.** The protons and neutrons which make up nuclei are not elementary particles. They are each clusters of quarks, tightly bound together by the exchange of gluons. One central characteristic of quantum chromodynamics, the theory which describes the interactions of quarks and gluons, is that the force between quarks becomes weaker as the quarks interact at higher energies. During most nuclear reactions, the force between quarks is so strong that free quarks cannot exist; quarks only occur confined in composite particles called hadrons. Protons and neutrons are hadrons, and there are many other types, such as pions and  $J/\psi$  particles, which are unstable. However, at temperatures above a critical temperature value

of about 150 MeV (about  $10^{12}$  K, or about  $10^5$  times the temperature at the center of the Sun), quantum chromodynamics predicts that the force between the quarks will be so weakened that hadrons will fall apart. The transition between hadronic matter and a plasma of quarks and gluons is called the quantum chromodynamics phase transition.

For the first few microseconds after the big bang, the whole universe was hotter than 150 MeV and was filled with quark-gluon plasma. The quantum chromodynamics phase transition occurred throughout the universe as it cooled through the critical temperature. The extreme temperatures required to observe the quantum chromodynamics phase transition are difficult to obtain in a laboratory. The only method known is to collide large nuclei at very high energies. See BIG BANG THEORY; PARTICLE ACCELERATOR.

Detailed theoretical calculations which simulate the quantum chromodynamics interactions of the thousands of quarks and gluons produced in such collisions lend support to the expectation that under these conditions quark-gluon plasma will be created for the first time since the big bang. This plasma will occur in a region several times  $10^{-13}$  cm in size, and will exist for several times  $10^{-22}$  s before the quantum chromodynamics phase transition occurs and ordinary hadrons are reformed.

The distribution and energy of the particles emerging from the collision should carry signatures of the fleeting presence of a quark-gluon plasma. One such signature involves the  $J/\psi$  particle, which is made of a heavy quark and a heavy antiquark. In a quark-gluon plasma, the  $J/\psi$  (like any hadron) dissociates and the heavy quark and antiquark separate. When the plasma cools through the phase transition, heavy quarks are much more likely to form hadrons in combination with the far more numerous light quarks than to reunite to form  $J/\psi$  hadrons. The expected reduction in the number of  $J/\psi$  particles produced in a collision has in fact been detected. However,  $J/\psi$  hadrons can also be broken apart by repeated collisions in a hot, dense hadronic environment. Therefore, the observed  $J/\psi$  suppression establishes that such an environment has been created, but does not in itself demonstrate the creation of a quark-gluon plasma. See J/PSI PARTICLE.

In an idealized model, as the hadronic state of matter is reestablished after an abrupt quantum chromodynamics phase transition, pion waves are amplified and coherent emission of pions occurs. Such pion lasers would be easily observable, because they would occasionally emit only electrically neutral pions or only electrically charged ones, whereas ordinarily different types of pions are uniformly mixed. This effect would be an unambiguous signature for the brief presence of a quark-gluon plasma in a collision of very high energy nuclei. See GLUONS; PHASE TRANSITIONS; QUANTUM CHROMODYNAMICS; QUARK-GLUON PLASMA; QUARKS; RELATIVISTIC HEAVY-ION COLLISIONS.

Krishna Rajagopal

Bibliography. H. Feshbach, *Theoretical Nuclear Physics*, vol. 2: *Nuclear Reactions*, 1992; R. Hwa (ed.), *Quark-Gluon Plasma*, 1990; S. Mukherjee,

M. K. Pal, and R. Shyam (eds.), *Nuclear Reaction Mechanisms*, 1989; K. Rajagopal and F. Wilczek, Emergence of long wavelength oscillations after a quench: Application to QCD, *Nucl. Phys.*, B399:395-425, 1993; G. R. Satchler, *Introduction to Nuclear Reactions*, 2d ed., 1990; A. G. Sitenko and O. D. Kocherga (eds.), *Theory of Nuclear Reactions*, 1990.

## Nuclear reactor

A system utilizing nuclear fission in a controlled and self-sustaining manner. Neutrons are used to fission the nuclear fuel, and the fission reaction produces not only energy and radiation but also additional neutrons. Thus a neutron chain reaction ensues. A nuclear reactor provides the assembly of materials to sustain and control the neutron chain reaction, to appropriately transport the heat produced from the fission reactions, and to provide the necessary safety features to cope with the radiation and radioactive materials produced by its operation. See CHAIN REACTION (PHYSICS); NUCLEAR FISSION.

Nuclear reactors are used in a variety of ways as sources for energy, for nuclear irradiations, and to produce special materials by transmutation reactions. Since the first demonstration of a nuclear reactor, made beneath the west stands of Stagg Field at the University of Chicago on December 2, 1942, many hundreds of nuclear reactors have been built and operated in the United States. Extreme diversification is possible with the materials available, and reactor power may vary from a fraction of a watt to thousands of megawatts. The size of a nuclear reactor core is governed by its power level, time between refueling, fuel environment, the factors affecting the control of the neutron chain reaction, and the effectiveness of the coolant in removing the fission energy released.

The generation of electrical energy by a nuclear power plant makes use of heat to produce steam or to heat gases to drive turbogenerators. Direct conversion of the fission energy into useful work is possible, but an efficient process has not yet been realized to accomplish this. Thus, in its operation the nuclear power plant is similar to the conventional coal-fired plant, except that the nuclear reactor is substituted for the conventional boiler as the source of heat.

The rating of a reactor is usually given in kilowatts (kW) or megawatts-thermal [MW(th)], representing the heat generation rate. The net output of electricity of a nuclear plant is about one-third of the thermal output. Significant economic gains have been achieved by building improved nuclear reactors with outputs of about 3300 MW(th) and about 1000 MW-electrical [MW(e)]. See ELECTRIC POWER GENERATION; NUCLEAR POWER.

### Fuel and Moderator

The fission neutrons are released at high energies and are called fast neutrons. The average kinetic energy is 2 MeV, with a corresponding neutron speed of  $1/15$

the speed of light. Neutrons slow down through collisions with nuclei of the surrounding material. This slowing-down process is made more effective by the introduction of materials of low atomic weight, called moderators, such as heavy water (deuterium oxide), ordinary (light) water, graphite, beryllium, beryllium oxide, hydrides, and organic materials (hydrocarbons). Neutrons that have slowed down to an energy state in equilibrium with the surrounding materials are called thermal neutrons, moving at 0.0006% of the speed of light. The probability that a neutron will cause the fuel material to fission is greatly enhanced at thermal energies, and thus most reactors utilize a moderator for the conversion of fast neutrons to thermal neutrons. This permits using smaller amounts and lower concentrations of fissile materials. *See* NEUTRON; THERMAL NEUTRONS.

With suitable concentrations of the fuel material, neutron chain reactions also can be sustained at higher neutron energy levels. The energy range between fast and thermal is designated as intermediate. Fast reactors do not have moderators and are relatively small.

Reactors have been built in all three categories. The first fast reactor was the Los Alamos (New Mexico) assembly called Clementine, which operated from 1946 to 1953. The fuel core consisted of nickel-coated rods of pure plutonium metal, contained in a 6-in.-diameter (15-cm) steel pot. Coolants for fast reactors may be steam, gas, or liquid metals. Current fast reactors utilize liquid sodium as the coolant and are being developed for breeding and power. An example of an intermediate reactor was the first propulsion reactor for the submarine USS *Seawolf*. The fuel core consisted of enriched uranium with beryllium as a moderator; the original coolant was sodium, and the reactor operated from 1956 to 1959. Examples of thermal reactors, currently the basis of commercial nuclear power production, are given below.

**Fuel composition.** Only three isotopes—uranium-235, uranium-233, and plutonium-239—are feasible as fission fuels, but a wide selection of materials incorporating these isotopes is available.

*Uranium-235.* Naturally occurring uranium contains only 0.7% of the fissionable isotope uranium-235, the balance being essentially uranium-238. Uranium with higher concentrations of uranium-235 is called enriched uranium.

Uranium metal is susceptible to irradiation damage, which limits its operating life in a reactor. The life expectancy can be improved somewhat by heat treatment, and considerably more by alloying with elements such as zirconium or molybdenum. Uranium oxide exhibits better irradiation damage resistance and, in addition, is corrosion-resistant in water. Ceramics such as uranium oxide have lower thermal conductivities and lower densities than metals, which are disadvantageous in certain applications. *See* ALLOY.

Current light-water-cooled nuclear power reactors utilize uranium oxide as a fuel, with an enrichment of several percent uranium-235. Cylindrical rods are the most common fuel-element configuration. They

can be fabricated by compacting and sintering cylindrical pellets which are then assembled into metal tubes which are sealed.

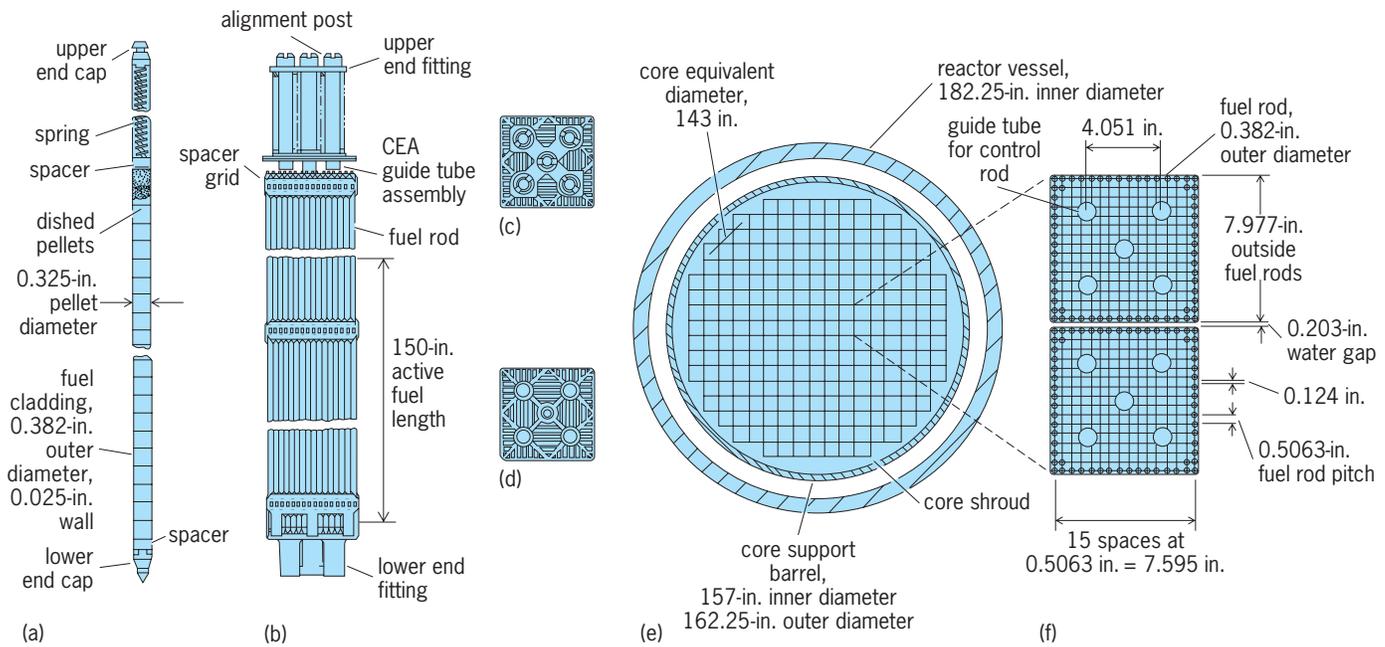
Developmental programs for attaining long-lived solid-fuel elements include studies with uranium oxide, uranium carbide, and other refractory uranium compounds. *See* URANIUM.

*Plutonium-239.* Plutonium-239 is produced by neutron capture in uranium-238. It is a by-product in power reactors and is becoming increasingly available as nuclear power production increases. For example, Japan has produced over 10 tons of plutonium solely as the result of its commercial reactor program. However, plutonium as a commercial fuel is still at a demonstration stage of development in Europe and Japan. The commercial recycle of plutonium from processed spent fuel was deferred indefinitely in the United States by the Carter administration in April 1977.

Plutonium is hazardous to handle because of its biological toxicity and, like radium, is a carcinogen if ingested. It is fabricated in glove boxes to ensure isolation from operating personnel. It can be alloyed with other metals and fabricated into various ceramic compounds. It is normally used in conjunction with uranium-238; alloys of uranium-plutonium, and mixtures of uranium-plutonium oxides and carbides, are of most interest as fuels. Except for the additional requirements imposed by plutonium toxicity and proliferation safeguards, much of the uranium technology is applicable to plutonium. For light-water nuclear power reactors, the oxide fuel pellets are contained in a zirconium alloy (Zircaloy) tube. Stainless steel tubes are used for containing the oxide fuel for the fast breeder reactors. *See* PLUTONIUM.

*Uranium-233.* Uranium-233, like plutonium, does not occur naturally, but is produced by neutron absorption in thorium-232, a process similar to that by which plutonium is produced from uranium-238. Interest in uranium-233 arises from its favorable nuclear properties and the abundance of thorium. However, studies of this fuel cycle are at a relatively early stage. Uranium-233 also imposes special handling problems because of proliferation concerns and the radiological toxicity of the daughter products of another uranium isotope (uranium-232) present in the fuel cycle. It does not, however, introduce new metallurgical problems. Thorium is metallurgically different, but it has favorable properties both as a metal and as a ceramic. *See* NUCLEAR FUELS; THORIUM.

**Fuel configurations.** Fuel-moderator assemblies may be homogeneous or heterogeneous. Homogeneous assemblies include the aqueous-solution-type water boilers and molten-salt-solution dispersions, slurries, and suspensions. The few homogeneous reactors built have been used for limited research and for demonstration of the principles and design features. In heterogeneous assemblies, the fuel and moderator form separate solid or liquid phases, such as solid-fuel elements spaced either in a graphite matrix or in a water phase. Most power reactors utilize an arrangement of closely spaced, solid fuel rods, about  $\frac{1}{2}$  in. (13 mm) in diameter and 12 ft (3.7 m)



**Fig. 1.** Arrangement of fuel in the core of a pressurized-water reactor, a typical heterogeneous reactor. (a) Fuel rod; (b) side view (CEA = control element assembly), (c) top view, and (d) bottom view of fuel assembly; (e) cross section of reactor core showing arrangement of fuel assemblies; (f) cross section of two adjacent fuel assemblies, showing arrangement of fuel rods. 1 in. = 25 mm. (Combustion Engineering, Inc.)

long, in water. In the arrangement shown in **Fig. 1**, fuel rods are arranged in a grid pattern to form a fuel assembly, and over 200 fuel assemblies are in turn arranged in a grid pattern in the reactor core.

**Heat Removal**

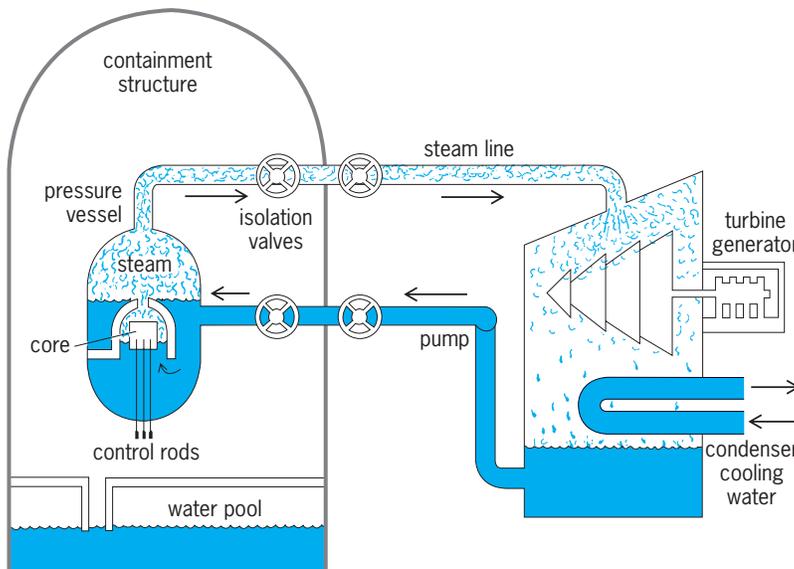
The major portion of the energy released by the fissioning of the fuel is in the form of kinetic energy of the fission fragments, which in turn is converted into heat through the slowing down and stopping of the fragments. For the heterogeneous reactors this heating occurs within the fuel elements. Heating also

arises through the release and absorption of the radiation from the fission process and from the radioactive materials formed. The heat generated in a reactor is removed by a primary coolant flowing through it.

Heat is not generated uniformly in a reactor core. The heat flux generally decreases axially and radially from a peak near the center of the reactor. In addition, local perturbations in heat generation occur because of the arrangement of the reactor fuel, its burn-up, various neutron “poisons” used to shape the power distribution, and inhomogeneities in the reactor structure. These variations impose special considerations in the design of reactor cooling systems, including the need for establishing variations in coolant flow rate through the reactor core to achieve uniform temperature rise in the coolant, avoiding local hot-spot conditions, and avoiding local thermal stresses and distortions in the structural members of the reactor.

Nuclear reactors have the unique thermal characteristic that heat generation continues after shutdown because of fission and radioactive decay of fission products. Significant fission heat generation occurs only for a few seconds after shutdown. Radioactive-decay heating, however, varies with the decay characteristics of the mixture of fission products and persists at low but significant power levels for many days. See RADIOACTIVITY.

Accurate analysis of fission heat generation as a function of time immediately after reactor shutdown requires detailed knowledge of the speed and reactivity worth of the control rods. The longer-term fission-product-decay heating, on the other hand, depends upon the time and power level of prior reactor operation and the isotopic composition of the fuel. Typical



**Fig. 2.** Boiling-water reactor. (Atomic Industrial Forum, Inc.)

values of the total heat generation after shutdown (as percent of operating power) are 10–20% after 1 s, 5–10% after 10 s, approximately 2% after 10 min, 1.5% after 1 h, and 0.7% after 1 day. These rates are important in reactor safety since 0.7% of the thermal power (2564 MW) of a 1000-MW(e) commercial nuclear power plant is approximately 18 MW of heat still being generated 1 day after the reactor is shut down.

**Reactor coolants.** Coolants are selected for specific applications on the basis of their heat-transfer capability, physical properties, and nuclear properties.

**Water.** Water has many desirable characteristics. It was employed as the coolant in many of the first production reactors, and most power reactors still utilize water as the coolant. In a boiling-water reactor (BWR; Fig. 2), the water boils directly in the reactor core to make steam that is piped to the turbine. In a pressurized-water reactor (PWR; Fig. 3), the coolant water is kept under increased pressure to prevent boiling. It transfers heat to a separate stream of feed water in a steam generator, changing that water to steam. Figure 4 shows the relation of the core and heat removal systems to the condenser, electric power system, and waste management system in the Prairie Island (Minnesota) nuclear plant, which is typical of plants using pressurized-water reactors. Coolant intake water is pumped through hundreds of 1-in.-diameter (25-mm) tubes in the condenser, and the warm water from the condenser is then pumped over cooling towers and returned to the plant. See COOLING TOWER; RADIOACTIVE WASTE MANAGEMENT; VAPOR CONDENSER.

For both boiling-water and pressurized-water reactors, the water serves as the moderator as well as the coolant. Both light water and heavy water are excellent neutron moderators, although heavy water (deuterium oxide) has a neutron-absorption cross section approximate  $1/500$  that for light water that makes it possible to operate reactors using heavy water with natural uranium fuel.

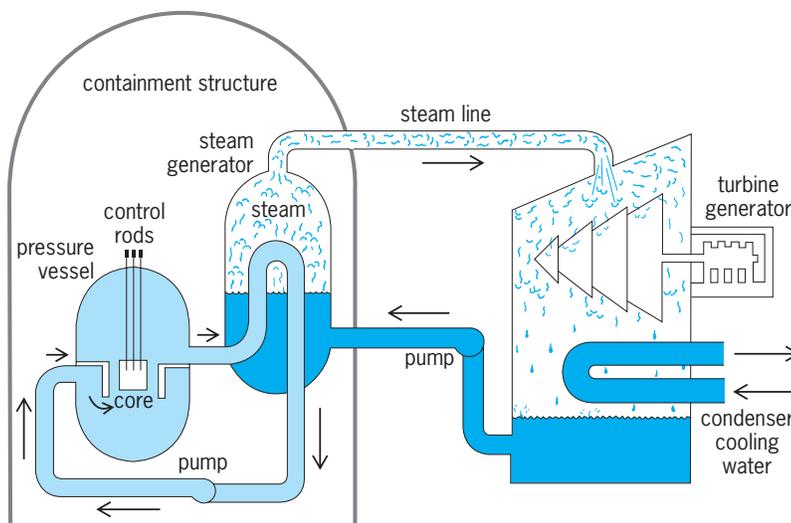


Fig. 3. Pressurized-water reactor. (Atomic Industrial Forum, Inc.)

There is no serious neutron-activation problem with pure water;  $^{16}\text{N}$ , formed by the  $(n,p)$  reaction with  $^{16}\text{O}$  (absorption of a neutron followed by emission of a proton), is a major source of activity, but its 7.5-s half-life minimizes this problem since the radioactivity, never very high to begin with, quickly decays away. The most serious limitation of water as a coolant for power reactors is its high vapor pressure. A coolant temperature of  $550^\circ\text{F}$  ( $288^\circ\text{C}$ ) requires a system pressure of at least 1100 psi (7.3 megapascals). This temperature is below fossil-fuel power station practice, for which steam temperatures near  $1000^\circ\text{F}$  ( $538^\circ\text{C}$ ) have become common. Lower thermal efficiencies result from lower temperatures. Boiling-water reactors operate at about 70 atm (7 MPa), and pressurized-water reactors at 150 atm (15 MPa). The high pressure necessary for water-cooled power reactors determines much of the plant design. This will be discussed below. See NUCLEAR REACTION.

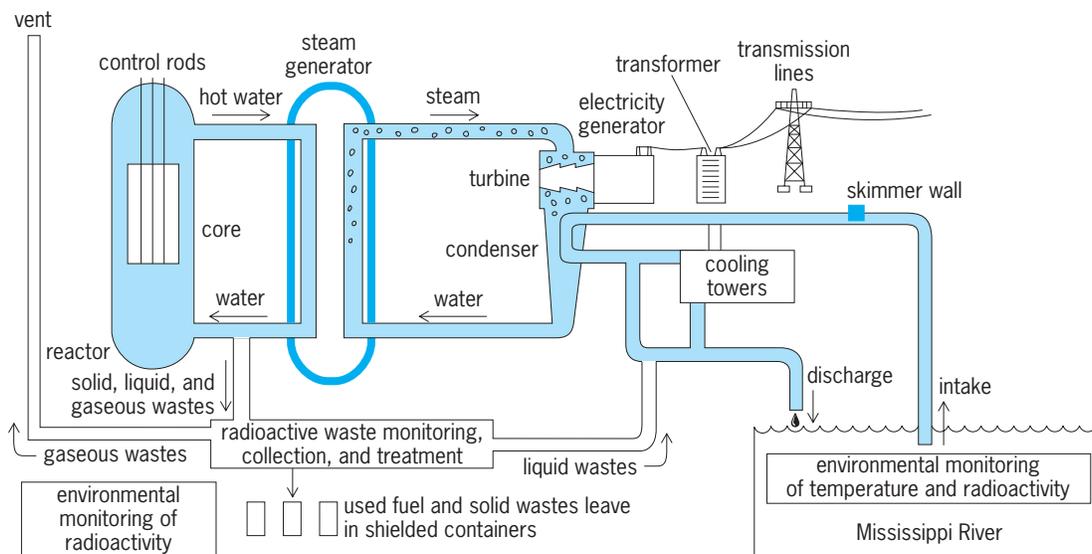


Fig. 4. Nuclear plant, using pressurized-water reactors. (Northern States Power Company)

*Gases.* Gases are inherently poor heat-transfer fluids as compared with liquids because of their low density. This situation can be improved by increasing the gas pressure; however, this introduces other problems and costs. Helium is the most attractive gas (it is chemically inert and has good thermodynamic and nuclear properties) and has been selected as the coolant for the development of high-temperature gas-cooled reactor (HTGR) systems (Fig. 5), in which the gas transfers heat from the reactor core to a steam generator. The British advanced gas reactor (AGR), however, uses carbon dioxide (CO<sub>2</sub>). Gases are capable of operation at extremely high temperature, and they are being considered for special process applications and direct-cycle gas-turbine applications. Hydrogen was used as the coolant for the reactor developed in the Nuclear Engine Rocket Vehicle Ap-

plication (NERVA) Program, now terminated. Heated gas discharging through the nozzle developed the propulsive thrust.

*Liquid metals.* The alkali metals, in particular, have excellent heat-transfer properties and extremely low vapor pressures at temperatures of interest for power generation. Sodium vapor pressure is less than 17.6 lb/in.<sup>2</sup> (120 kilopascals) at 1650°F (900°C). Sodium is attractive because of its relatively low melting point (208°F or 98°C) and high heat-transfer coefficient. It is also abundant, commercially available in acceptable purity, and relatively inexpensive. It is not particularly corrosive, provided low oxygen concentration is maintained. Its nuclear properties are excellent for fast reactors. In the liquid-metal fast breeder reactor (LMFBR; Fig. 6), sodium in the primary loop collects the heat generated in the core and transfers it to a secondary sodium loop in the heat exchanger, from which it is carried to the steam generator in which water is boiled to make steam.

Sodium-24, formed by the absorption of a neutron, is an energetic gamma emitter with a 15-h half-life. The primary system is surrounded by biological shielding, and approximately 2 weeks is required for decay of <sup>24</sup>Na activity prior to access to the system for repair or maintenance. Another sodium isotope, <sup>22</sup>Na, has a 2.6-year half-life and builds up slowly with plant operation until eventually the radioactivity reaches a level where it is necessary to drain the sodium before maintenance can be performed.

Sodium does not decompose, and no makeup is required. However, sodium reacts violently if mixed with water. This requires extreme care in the design and fabrication of sodium-to-water steam generators and backup systems to cope with occasional leaks. The poor lubricating properties of sodium and its reaction with air further specify the mechanical design requirements of sodium-cooled reactors. Nevertheless, sodium-cooled reactors have operated with good reliability and relatively high operating availability. The other alkali metals exhibit similar characteristics but appear to be less attractive than sodium. The eutectic alloy of sodium with potassium (NaK), however, has the advantage that it remains liquid at room temperature, but adversely affects the properties of steel used in system components. Mercury has also been used as a coolant but its overall properties are less favorable than sodium.

**Plant balance.** The nuclear chain reaction in the reactor core produces energy in the form of heat, as the fission fragments slow down and dissipate their kinetic energy in the fuel. This heat must be removed efficiently and at the same rate it is being generated in order to prevent overheating of the core and to transport the energy outside the core, where it can be converted to a convenient form for further utilization. The energy transferred to the coolant, as it flows past the fuel element, is stored in it in the form of sensible heat and pressure and is called the enthalpy of the fluid. In an electric power plant, the energy stored in the fuel is further converted to kinetic energy (the energy of motion) through a device called a prime mover which, in the case of

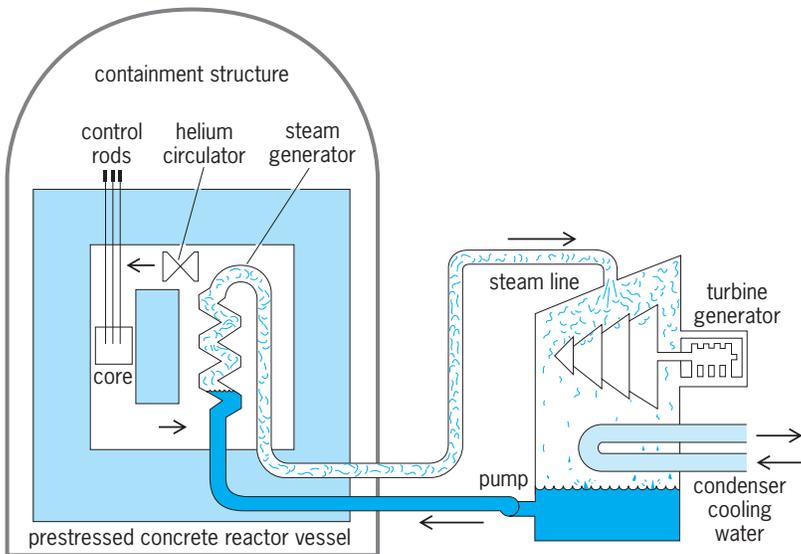


Fig. 5. High-temperature gas-cooled reactor. (Atomic Industrial Forum, Inc.)

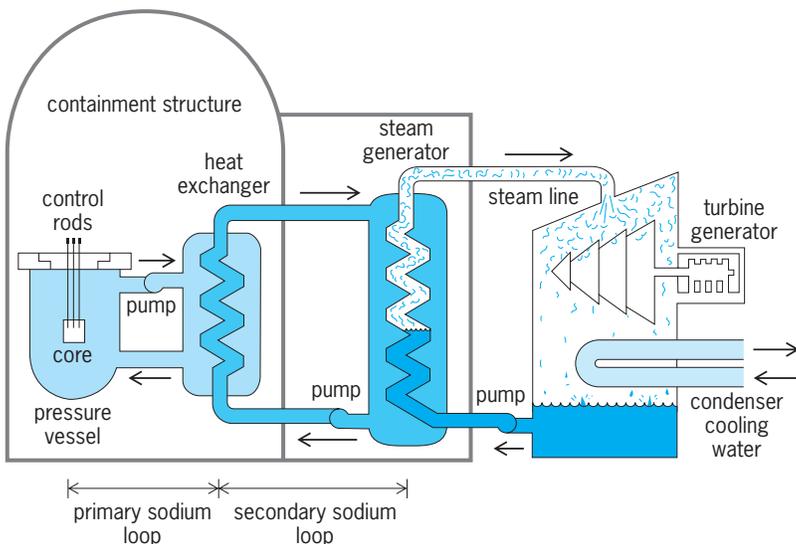


Fig. 6. Loop-type liquid-metal fast breeder reactor. In some designs, referred to as pool-type, the heat exchanger and primary sodium pump are located with the core inside the pressure vessel. (Atomic Industrial Forum, Inc.)

nuclear reactors, is predominantly a steam turbine. Another conversion takes place in the electric generator, where kinetic energy is converted into electric power as the final energy form to be distributed to the consumers through the power grid and distribution system. *See* ENTHALPY; GENERATOR; PRIME MOVER; STEAM TURBINE.

After the steam expands to spin the rotor in the turbine, it exits into the condenser where it is cooled to produce condensate water, which is then fed back to the core or to the steam generator, and the cycle is repeated. The condenser is a very large and important part of the plant. Roughly twice the amount of heat that has been converted in the turbine for the production of electric power is removed in the condenser for further rejection to the environment. The second law of thermodynamics defines the fraction of useful energy or work that can be attained by a thermal engine. Large amounts of water circulate through the condenser to carry the waste heat to its ultimate sink, which may be the sea, a river, a lake, or the atmosphere itself through cooling towers. *See* COOLING TOWER; STEAM CONDENSER; THERMODYNAMIC PRINCIPLES.

The energy conversion part of the power plant (that is, excluding the reactor itself with its main components and system) is often called balance of plant (BOP). It incorporates a large number of components and systems. It represents over four-fifths of the plant's total cost, and is important for the efficient and safe operation of the plant.

**Fluid flow and hydrodynamics.** Because heat removal must be accomplished as efficiently as possible, considerable attention must be given to fluid-flow and hydrodynamic characteristics of the system. *See* FLUID FLOW; HYDRODYNAMICS.

The heat capacity and thermal conductivity of the fluid at the temperature of operation have a fundamental effect upon the design of the reactor system. The heat capacity determines the mass flow of the coolant required. The fluid properties (thermal conductivity, viscosity, density, and specific heat) are important in determining the surface area required for the fuel—in particular, the number and arrangement of the fuel elements. These factors combine to establish the pumping characteristics of the system because the pressure drop and coolant temperature rise in the core are directly related. *See* CONDUCTION (HEAT); HEAT CAPACITY; VISCOSITY.

Secondary considerations include other physical properties of the coolant, particularly its vapor pressure. If the vapor pressure is high at the operating temperature, local or bulk boiling of the fluid may occur, unless the system pressure is maintained at a higher pressure at all times. This, in turn, must be considered in establishing the heat transfer coefficient for the fluid. *See* VAPOR PRESSURE.

Because the coolant absorbs and scatters neutrons, variations in coolant density also affect reactor performance and control. This is particularly significant in reactors in which the coolant exists in two phases—for example, the liquid and vapor phases in boiling systems. Gases, of course, do not

undergo phase change, nor do liquids operating at temperatures well below their boiling point; however, the fuel density does change with temperature and may have an important effect upon the reactor.

Power generation and, therefore, the heat removal rate are not uniform throughout the reactor. If the mass flow rate of the coolant is uniform through the reactor core, then unequal temperature rise of the coolant results. This becomes particularly significant in power reactors in which it is desired to achieve the highest possible coolant outlet temperature to attain maximum thermal efficiency of the power cycle. The performance limit of the coolant is set by the temperature in the hottest region or channel of the reactor. Unless the coolant flow rate is adjusted in the other regions of the reactor, the coolant will leave these regions at a lower temperature and thus will reduce the average coolant outlet temperature. In power reactors, this effect is reduced by adjusting the rate of flow to each region of the reactor in proportion to its heat generation rate. This involves very careful design and analysis of the system. In the boiling-type reactor, this effect upon coolant temperature does not occur because the exit temperature of the coolant is constant at the saturation temperature for the system. However, the variation in power generation in the reactor is reflected by a difference in the amount of steam generated in the various zones, and orificing is still required to achieve most effective use of coolant flow.

In some power reactors, the flow rate and consequent pressure drop of the coolant are sufficient to create large mechanical forces in the system. It is possible for the pressure drop through the fuel assemblies to exceed the weight of the fuel elements in the reactor, with a resulting hydraulic lifting force on the fuel elements. Often this requires a design arrangement to hold the fuel elements down. Although this problem can be overcome by employing downward flow through the system, it is often undesirable to do so because of shutdown-cooling considerations. It is desirable in most systems to accomplish shutdown cooling by natural-convection circulation of the coolant. If downflow is employed for forced circulation, then shutdown cooling by natural-convection circulation requires a flow reversal, which can introduce new problems. Therefore, hydraulic forces are overcome by use of support plates, seismic restraints, and fuel spacers.

**Thermal stress considerations.** The temperature of the reactor coolant increases as it circulates through the reactor core. This increase in temperature is constant at steady-state conditions. Fluctuations in power level or in coolant flow rate result in variations in the temperature rise. These are reflected as temperature changes in the coolant core exit temperature, which in turn result in temperature changes in the coolant system.

A reactor is capable of very rapid changes in power level, particularly reduction in power level, which is a safety feature of the plant. Reactors are equipped with mechanisms (reactor scram systems) to ensure

rapid shutdown of the system in the event of leaks, failure of power conversion systems, or other operational abnormalities.

Therefore, reactor coolant systems must be designed to accommodate the temperature transients that may occur because of rapid power changes. In addition, they must be designed to accommodate temperature transients that might occur as a result of a coolant system malfunction, such as pump stoppage. The consequent temperature stresses induced in the various parts of the system are superimposed upon the thermal stresses that exist under normal steady-state operations and produce conditions known as thermal shock or thermal cycling.

In some systems, it is not uncommon for the thermal stresses to be significant. In these cases, careful attention must be given to the transient stresses, and thermal shielding (such as thermal sleeves on pipes and baffles) is commonly employed in critical sections of the system. Normally, this consists of a thermal barrier which, by virtue of its heat capacity and insulating effect, delays the transfer of heat, thereby reducing the rate of change of temperature and protecting critical system components from thermal stresses.

Thermal stresses are also important in the design of reactor fuel elements. Metals that possess dissimilar thermal-expansion coefficients are frequently required. Heating of such systems gives rise to distortions, which in turn can result in flow restrictions in coolant passages. Careful analysis and experimental verification are often required to avoid such circumstances.

**Coolant system components.** The development of reactor systems has led to the development of special components for reactor component systems. Because of the hazard of radioactivity, leak-tight systems and components are a prerequisite to safe, reliable operation and maintenance. Special problems are introduced by many of the fluids employed as reactor coolants.

More extensive component developments have been required for sodium, which is chemically active and is an extremely poor lubricant. Centrifugal pumps employing unique bearings and seals have been specially designed. Sodium is an excellent electrical conductor and, in some special cases, electromagnetic-type pumps have been used. These pumps are completely sealed, contain no moving parts, and derive their pumping action from electromagnetic forces imposed directly on the fluid. *See* CENTRIFUGAL PUMP; ELECTROMAGNETIC PUMP.

In addition to the variety of special pumps developed for reactor coolant systems, there is a variety of piping system components and heating exchange components. As in all flow systems, flow-regulating devices such as valves are required, as well as flow instrumentation to measure and thereby control the systems. Here again, leak tightness has necessitated the development of valves with special seals such as metallic bellows around the valve stem to ensure system integrity. Measurement of flow and pressure has also required the development of sensing instru-

mentation that is reliable and leak-tight. *See* FLOW MEASUREMENT; PRESSURE MEASUREMENT; VALVE.

Many of these developments have borrowed from other technologies where toxic or flammable fluids are frequently pumped. In many cases, however, special equipment has been developed specifically to meet the requirements of the reactor systems. An example of this type of development involves the measurement of flow in liquid-metal piping systems. The simple principle of a moving conductor in a magnetic field is employed by placing a magnet around the pipe and measuring the voltage generated by the moving conductor (coolant) in terms of flow rate. Temperature compensation is required, and calibration is important.

Although the development of nuclear power reactors has introduced many new technologies, no method has yet displaced the conventional steam cycle for converting thermal energy to mechanical energy. Steam is generated either directly in the reactor (direct-cycle boiling reactor) or in auxiliary steam generation equipment, in which steam is generated by transfer of heat to water from the reactor coolant. These steam generators require a very special design, particularly when dissimilar fluids are involved. Typical of the latter problem is the sodium-to-water steam generators in which integrity is essential because of the potentially violent chemical reaction between sodium and water.

### Core Design and Materials

A typical reactor core for a power reactor consists of the fuel element rods supported by a grid-type structure inside a vessel (Fig. 1).

The primary function of the vessel is to contain the coolant. Its design and materials are determined by such factors as the nature of the coolant (corrosive properties) and quantity and configuration of fuel. The vessel has several large nozzles for coolant entrance and exit and smaller nozzles used for controlling reactor operation (control-rod drive mechanisms and instruments). The top of the vessel unbolts for refueling operations.

The pressure vessel design takes account of thermal stresses caused by temperature differences in the system. An exceptionally high degree of integrity is demanded of this equipment. Reactors are designed to permit removal of the internals from the vessel, which is then periodically inspected by automatic devices that can detect any cracks which might have developed during operation. These in-service inspections are required by codes and regulations on a fixed time schedule.

**Structural materials.** Structural materials employed in reactor systems must possess suitable nuclear and physical properties and must be compatible with the reactor coolant under the conditions of operation. Some requirements are established because of secondary effects; for example, corrosion limits may be established by the rate of deposition of coolant-entrained corrosion products on critical surfaces rather than by the rate of corrosion of the base material. *See* CORROSION.

The most common structural materials employed in reactor systems are stainless steel and zirconium alloys. Zirconium alloys have favorable nuclear and physical properties, whereas stainless steel has favorable physical properties. Aluminum is widely used in low-temperature test and research reactors; zirconium and stainless steel are used in high-temperature power reactors. Zirconium is relatively expensive, and its use is therefore confined to applications in the reactor core where neutron absorption is important. *See* ALUMINUM; STAINLESS STEEL; ZIRCONIUM.

The 18-8 series stainless steels have been used for structural members in both water-cooled reactors and sodium-cooled reactors because of their corrosion resistance and favorable physical properties at high temperatures. Type 304, 316, and 347 stainless steel have been used the most extensively because of their weldability, machinability, and physical properties, although other iron-nickel alloys, such as Inconel, are also used. To increase reliability and to reduce cost, heavy-walled pressure vessels are normally fabricated from carbon steels and clad on the internal surfaces with a thin layer of stainless steel to provide the necessary corrosion resistance. *See* IRON ALLOYS; NICKEL ALLOYS; STEEL.

As the size of power reactors has increased, it has become necessary in some instances to field-fabricate reactor vessels. A notable example is the on-site fabrication of the large pressure vessel required for the liquid-metal fast breeder reactor at Creys-Malville, France. This involved field-welding of wall sections and subsequent stress relieving. Both steel and prestressed concrete vessels for gas-cooled reactors are also field-fabricated.

Research reactors operating at low temperatures and pressures introduced special experimental considerations. The primary objective is to provide the maximum volume of unperturbed neutron flux for experimentation. It is desirable, therefore, to extend the experimental irradiation facilities beyond the vessel wall. This has introduced the need for vessels constructed of materials having a low cross section for neutron capture. Relatively large low-pressure aluminum reactor vessels with wall sections as thin as practicable have been manufactured for research reactors.

**Fuel cladding.** Reactors maintain a separation of fuel and coolant by cladding the fuel. The cladding is designed to prevent the release of radioactivity from the fuel. The cladding material must be compatible with both the fuel and the coolant.

The cladding materials must also have favorable nuclear properties. The neutron-capture cross section is most significant because the unwanted absorption of neutrons by these materials reduces the efficiency of the nuclear fission process. Aluminum is a very desirable material in this respect; however, its physical strength and corrosion resistance in water decrease very rapidly above about 300°F (149°C), and it is therefore used only in test and research reactors that are not used to produce power.

Zirconium has favorable neutron properties, and

in addition is corrosion-resistant in high-temperature water. It has found extensive use in water-cooled power reactors. The technology of zirconium production and the use of zirconium-based alloys, such as Zircaloy, have advanced tremendously under the impetus of the various reactor development programs.

Stainless steel is used for the fuel cladding in fast reactors, in some light-water reactors for which neutron captures are less important.

### Control and Instrumentation

A reactor is critical when the rate of production of neutrons equals the rate of absorption in the system plus the rate of leakage out of the core. The control of reactors requires the continuing measurement and adjustment of the critical condition. The neutrons are produced by the fission process and are consumed in a variety of ways, including absorption to cause fission, nonfission capture in fissionable materials, capture in fertile materials, capture in structure or coolant, and leakage from the reactor to the shielding. A reactor is subcritical (power level decreasing) if the number of neutrons produced is less than the number consumed. The reactor is supercritical (power level increasing) if the number of neutrons produced exceeds the number consumed. *See* REACTOR PHYSICS.

Reactors are controlled by adjusting the balance between neutron production and neutron consumption. Normally, neutron consumption is controlled by varying the absorption or leakage of neutrons; however, the neutron generation rate also can be controlled by varying the amount of fissionable material in the system.

It is necessary for orderly startup and control of a reactor that the neutron flux be sufficiently high to permit rapid power increase; in large reactors, too slow a startup is erratic and uneconomical. During reactor startup, a source of neutrons is useful, therefore, for control, and aids in the instrumentation of reactor systems. Neutrons are obtained from the photoneutron effect in materials such as beryllium. Neutron sources consist of a photon (gamma-ray) source and beryllium, such as antimony-beryllium. Antimony sources are particularly convenient for use in reactors because the antimony is reactivated by the reactor neutrons each time the reactor operates.

**Control drives and systems.** The reactor control system requires the movement of neutron-absorbing rods (control rods) in the reactor under carefully controlled conditions. They must be arranged to increase reactivity (increase neutron population) slowly and under good control. They must be capable of reducing reactivity, both rapidly and slowly. Power reactors have inherent negative temperature coefficients that make it practical to use stepwise control of rod motions but with smooth and limited changes in power level.

The control drives can be operated by the reactor operator or by automatic control systems. Reactor scram (rapid reactor shutdown) can be initiated automatically by a wide variety of system scram-safety

signals, or it can be started by the operator depressing a scram button in the control room.

Control drives are electromechanical or hydraulic devices that impart in-and-out motion to the control rods. They are usually equipped with a relatively slow-speed reversible drive system for normal operational control. Scram is usually effected by a high-speed overriding drive accompanied by disconnecting the main drive system. Allowing the control rods to drop into the core by means of gravity is one common method. To enhance reliability of the scram system, its operation can be initiated by deenergizing appropriate electrical circuits. This also automatically produces reactor scram in the event of a system power failure. Hydraulic or pneumatic drive systems, as well as a variety of electromechanical systems, have also been developed.

In addition to the actuating motions required, control-rod drive systems must also provide indication of the rod positions at all times. Various types of sensors, as well as arrangements of switch indicators, are employed as position indicators.

**Instrumentation.** Reactor control involves continuing measurement of the condition of the reactor. Neutron-sensitive ion chambers may be located outside the reactor core, and the flux measurements from the detectors are combined to measure a flux that is proportional to the average neutron density in the reactor. The chamber current is calibrated against a thermal power measurement and then is applied over a wide range of reactor power levels. The neutron-sensitive detector system is capable of measuring the lowest neutron flux in the system, including that produced by the neutron source when the reactor itself is subcritical. *See* IONIZATION CHAMBER.

Normally, several ranges of instrument sensitivities are required to cover the entire operating range. A range is required for low-level operation, beginning at the source level, whereas others are required for the intermediate- and high-power levels. Three ranges of detectors are common in power reactor systems, and some systems contain a larger number. The total range to be covered is 7–10 decades (factors of 10) of power level.

The chamber current of a neutron detector, suitably amplified, can be employed as a signal to operate automatic control system devices as well as to actuate reactor scram. In addition to power level, rate of change of power level is an important measurement which is recorded and employed to actuate various alarm and trip circuits. The normal range for the current ion chambers is approximately  $10^{-14}$  to  $10^{-4}$  A. This current is suitably amplified in logarithmic and period amplifiers. The power monitors are part of a reactor's plant protection system.

### Applications

Reactor applications include mobile, stationary, and packaged power plants; production of fissionable fuels (plutonium and uranium-233) for military and commercial applications; research, testing, teaching-demonstration, and experimental facilities; space

and process heat; dual-purpose design; and special applications. The potential use of reactor radiation or radioisotopes produced for sterilization of food and other products, steam for chemical processes, and gas for high-temperature applications has been recognized. *See* NUCLEAR FUEL CYCLE; NUCLEAR FUELS REPROCESSING; RADIOACTIVITY AND RADIATION APPLICATIONS; SHIP NUCLEAR PROPULSION.

Frank J. Rahn

Bibliography. J. Douglas, The nuclear option, *EPRI Journal*, December 1994; S. Glasstone and A. Sesonke, *Nuclear Reactor Engineering*, 4th ed., 1993; International Atomic Energy Association, *Nuclear Power Reactors in the World*, annually; R. A. Knief, *Nuclear Engineering: Theory and Technology of Commercial Nuclear Power*, 1992; A. Nero, *A Guidebook to Nuclear Reactors*, 1980; F. J. Rahn et al., *A Guide to Nuclear Power Technology*, 1984, reprint 1992; S. C. Stulz and J. B. Kitto, *Steam: Its Generation and Use*, 40th ed., 1992.

## Nuclear spectra

The distribution of the intensity of particles (or radiation) emitted in a nuclear process as a function of energy. The nuclear spectrum is a unique signature of the process.

For example, when very slow neutrons (with speeds less than 0.5% of the speed of light) hit nitrogen nuclei, there is a high probability that they will be captured and that the nuclear system which is formed will emit a set of gamma rays (electromagnetic radiation) of very precise energies. The 24 gamma rays have energies ranging from 1.68 to 10.83 MeV, and their relative intensities are well known. A spectrum of these gamma rays, that is, the number of gamma rays having a particular energy, versus that energy can provide a unique signature of the presence of nitrogen. An application is the passing of a beam of slow neutrons through luggage at an airport: the presence of unusual amounts of nitrogen indicates that a plastic explosive may be present. This testing for the presence of nitrogen is nondestructive: relatively few neutrons are needed to produce the characteristic spectrum, and the luggage and its contents are not harmed. *See* GAMMA RAYS; NONDESTRUCTIVE EVALUATION.

The process that leads to such a spectrum is well understood. The type (isotope) of nitrogen predominantly found in nature (over 99.6%) is  $^{14}\text{N}$ , where 7 is the number of protons and 14 is the total number of nucleons (neutrons + protons). When  $^{14}\text{N}$  captures a neutron,  $^{15}\text{N}$  is formed, but it is not formed in its state of lowest energy and most symmetric shape, the ground state. Rather it is created in an unstable form which has an energy that is 10.83 MeV greater than that of the ground state and which decays rapidly (in a time of less than  $10^{-11}$  s) to the ground state of  $^{15}\text{N}$  by emitting gamma rays. The decays to the ground state take place via excited states of  $^{15}\text{N}$ . When a number of  $^{15}\text{N}$  nuclei have decayed, gamma rays with 24 different but well-defined energies will have been

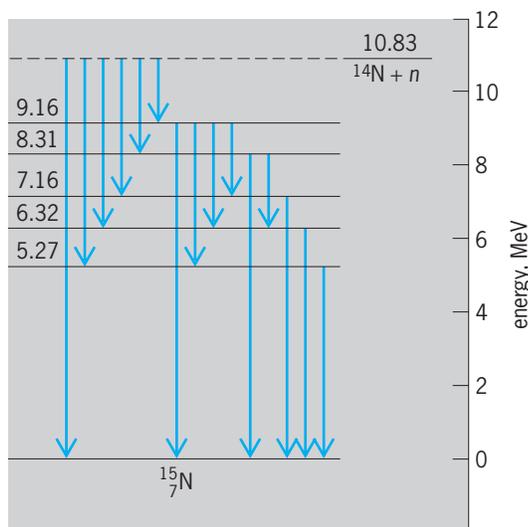


Fig. 1. Electromagnetic transitions (arrows) that result when  $^{14}\text{N}$  captures a slow neutron. Scale to the right measures energy above the ground state of  $^{15}\text{N}$ . Only some of the 24 gamma-ray transitions that can occur are shown.

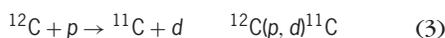
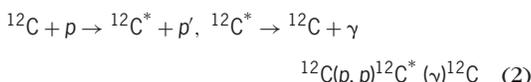
emitted (Fig. 1). Such a spectrum is called a discrete spectrum because the 24 energies form a discrete set of peaks.

**Natural radioactive decay.** Naturally occurring radioactive nuclei change (decay) by emitting negatively charged electrons ( $\beta^-$ ), positively charged alpha particles (a system composed of two protons and two neutrons), as well as uncharged gamma rays and neutrinos ( $\bar{\nu}_e$ ). In one sequence of nuclear decays found in nature, thorium-232 ( $^{232}_{90}\text{Th}$ ) is the progenitor of a family of nuclei that are sequentially formed in a series of alpha and  $\beta^-$  decays which terminate in the formation of lead-208 ( $^{208}_{82}\text{Pb}$ ), a stable nucleus. As  $^{232}\text{Th}$  is converted to  $^{208}\text{Pb}$  in this chain, bismuth-212 ( $^{212}_{83}\text{Bi}$ ) is produced. The decay of  $^{212}\text{Bi}$  illustrates the nature of the alpha, beta, and gamma radiations observed in radioactive chains. The mass of  $^{212}_{83}\text{Bi}$  is greater than the mass of polonium-212 ( $^{212}_{84}\text{Po}$ ) plus the mass of an electron, and it is also greater than the mass of thallium-208 ( $^{208}_{81}\text{Tl}$ ) plus the mass of an alpha particle (Fig. 2). When a physical system has a greater mass than that of a system to which it could decay (without breaking the fundamental conservation laws), it does so. For instance,  $^{212}\text{Bi}$  transforms into  $^{212}_{84}\text{Po}$  with the emission of a negative electron ( $\beta^-$ ) and a neutrino ( $\bar{\nu}_e$ ). The chance of this happening, statistically, is 2/3. With a chance of roughly 1/3, however,  $^{212}_{83}\text{Bi}$  transforms into  $^{208}_{81}\text{Tl}$  with the emission of an alpha particle. The nuclei  $^{212}\text{Po}$  and  $^{208}\text{Tl}$  are not always produced directly in their ground state. When these nuclei are formed in an excited state, that state decays to the ground state by emitting electromagnetic radiation in the form of gamma rays. The negative electron spectrum is continuous because the energy available in the transition is divided between three bodies:  $^{212}\text{Po}$ , the negative electron, and a neutrino. By studying the alpha and beta spectra emitted by  $^{212}\text{Bi}$  when it transforms into  $^{208}\text{Tl}$  and  $^{212}\text{Po}$ , the differences in the masses of these

three nuclei can be determined. When the spectrum of the gamma rays involved in the  $^{208}\text{Tl}$  beta decay to  $^{208}\text{Pb}$  is studied, the parameters of the excited states of  $^{208}\text{Pb}$  can be determined. See BETA PARTICLES; RADIOACTIVITY.

**Nuclear reactions.** When nuclear processes take place in the laboratory, not only can alpha particles, electrons, and gamma rays be produced, but most groupings of nucleons (neutrons and protons), leptons (electrons, muons, neutrinos), photons (electromagnetic radiation), and mesons (pions, kaons, and so forth) can be created, subject to conservation laws.

For instance, when the most abundant isotope of carbon,  $^{12}\text{C}$ , is bombarded by protons that have been given a kinetic energy in a particle accelerator, a large number of nuclear processes can take place. A few of these, reactions (1)–(4), are both written out and



given in compact notation. Reaction (1) is elastic scattering, with  $^{12}\text{C}$  remaining in its ground state. In Fig. 3, the group labeled  $p_0$  represents elastically scattered protons. Reaction (2) is inelastic scattering, with  $^{12}\text{C}$  first formed in an excited state, which then decays by gamma-ray emission to the ground state of  $^{12}\text{C}$ . The groups at lower energies in Fig. 3,  $p_1$  and  $p_2$ , represent inelastically scattered protons:  $^{12}\text{C}$  is then formed in its excited states at 4.4 and 7.7 MeV. In reaction (3),  $^{11}\text{C}$  is formed at the same time that a deuteron is emitted. The  $^{11}\text{C}$  nucleus then decays to  $^{11}\text{B}$  with the emission of a positive electron

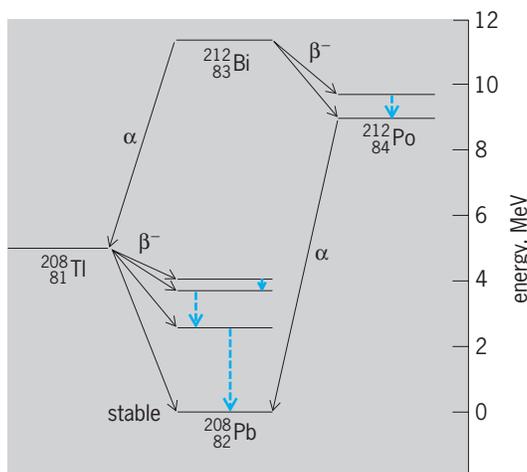
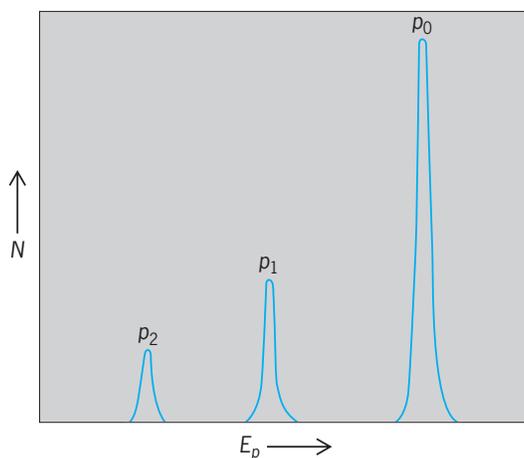


Fig. 2. Decay of  $^{212}\text{Bi}$  into  $^{208}\text{Pb}$  via  $^{212}\text{Po}$  and  $^{208}\text{Tl}$ . The vertical arrows show some of the observed decays of excited states of  $^{212}\text{Po}$  and  $^{208}\text{Pb}$  to the ground states of these nuclei, via electromagnetic transitions involving the emission of gamma rays. The scale on the right shows the energy differences between the states.



**Fig. 3.** Idealized spectrum of protons scattered from  $^{12}\text{C}$ .  $N$  is the number of scattered protons;  $E_p$  is their energy. The proton groups labeled  $p_0$ ,  $p_1$ , and  $p_2$  are identified in the text. The groups are shown as being broad (not monoenergetic) because the experimental equipment that measures them is not perfect.

( $\beta^+$ ) and of a neutrino ( $\nu_e$ ). In reaction (4), the role of the proton is to shock  $^{12}\text{C}$  into fissioning (splitting) into an alpha particle and  $^8\text{Be}$ . The nucleus is unstable and very rapidly breaks up into two alpha particles; thus  $^{12}\text{C}$  disappears and three alpha particles are created.

The minimum kinetic energies of the protons needed to create these processes vary. As the proton energy is increased, more exotic processes may take place. An example is the reaction  $\text{C}^{12}(p, \pi^+)\text{C}^{13}$ . Here the incident proton's kinetic energy permits a pion to emerge from the nucleus, and a  $^{13}\text{C}$  nucleus is created. The  $\pi^+$  then decays into two leptons: a muon ( $\mu^+$ ) and a neutrino ( $\nu_\mu$ ). By studying the spectra of the particles emitted in nuclear reactions (Fig. 3), the structure of the nuclei that are created can be determined. See NUCLEAR REACTION; NUCLEAR STRUCTURE.

**Fission.** A very important type of nuclear reaction is fission. The most common type of fission involves a very slow neutron that amalgamates with a heavy nucleus such as uranium-235 ( $^{235}\text{U}$ ). For a very short time, a composite nucleus,  $^{236}\text{U}$ , is formed, but it is so excited and its shape is so distended that it breaks up into several fragments. The larger fragments mostly decay by beta, gamma, and neutrino emission until stable end products are reached. The lightest fragments are fast neutrons, some of which can be slowed down to cause fission of another uranium nucleus. The spectrum of the neutrons released during the fission must be carefully studied to determine how best to moderate the velocities of the neutrons to make the probability of a chain reaction (continuously sustained fission) as high as possible. See NUCLEAR FISSION.

**Measurements.** The methods used to measure nuclear spectra depend on the nature of the particles (radiation) involved. The most accurate energy measurements are those of gamma rays. Gamma-ray spectra can be measured by determining the energy de-

posited by the gamma rays in a crystal, often made of sodium iodide, containing thallium impurities [ $\text{NaI}(\text{Tl})$ ], or of germanium, containing lithium impurities [ $\text{Ge}(\text{Li})$ ]. In a  $\text{NaI}(\text{Tl})$  detector, the gamma-ray energy is transferred to electrons within the crystal, and these charged particles in turn produce electromagnetic radiation with frequencies in the visible range. The crystal is surrounded by detectors (photomultipliers) that are sensitive to the visible light. The intensity of the signal in the photomultipliers is proportional to the energy of the gamma rays that entered the  $\text{NaI}(\text{Tl})$  crystal. The signal pulse is amplified electronically, and the pulse heights (pulse sizes) are displayed in a pulse-height multichannel analyzer in a histogram. Usually the number of pulses having a certain height (strength) is plotted versus the height. What results is a plot showing the number of gamma rays having a certain energy versus the energy of the gamma rays, a spectrum. Gamma-ray energy measurements with an uncertainty of less than 1 part in  $10^5$  have been reported. While the use of a  $\text{NaI}(\text{Tl})$  crystal is restricted to determining gamma-ray spectra, the procedure of looking at signals in bins, that is, in plotting histograms in which the number of signals corresponding to a certain energy is plotted versus that energy, is general to nuclear spectra measurements. See GAMMA-RAY DETECTORS; PHOTOMULTIPLIER.

Different types of nuclear particles (and radiation) are observed by detectors specific to the particles. For instance, neutron spectra are often determined by measuring their velocities. This is done by a time-of-flight technique in which an electronic timer measures the time interval between the emission of the neutron from a nucleus and its arrival at a detector a known distance away. This measurement uniquely determines the velocity, and thus the kinetic energy, of the neutrons. The detector is typically a plastic containing a large fraction of hydrogen. The incident neutrons interact with the hydrogen nuclei, protons, and transfer their kinetic energies to them. Protons, which are charged, then produce detectable electrical signals because they ionize the matter through which they move. The detection of gamma rays and of neutrons, both of which are uncharged, depends on their interaction with charged particles in the medium in which they move. It is the charged particles that then lead to the electronic signal that is measured. See NEUTRON SPECTROMETRY; TIME-OF-FLIGHT SPECTROMETERS.

Measurements of nuclear spectra involving charged particles, such as pions, protons and alpha particles, are often made by determining their momenta (mass  $\times$  velocity) and then calculating the corresponding kinetic energy. Momentum measurements are made by passing the beam of charged particles through a region in which a magnetic field exists. A magnetic field that is constant in time will not cause a change in a charged particle's speed, but it will cause a charged particle to deviate in its path. For instance, if the direction of a magnetic field of magnitude  $B$  is perpendicular to the path of a particle of charge  $q$ , mass  $m$ , and velocity  $v$ , the particle

will move in a circle of radius  $r$  given by Eq. (5).

$$r = \frac{mv}{qB} \quad (5)$$

Thus, by determining the radius of the circle, the momentum,  $mv$ , of the charged particle, and therefore its kinetic energy, can be determined. After the particle has been deviated from its path sufficiently so that  $r$  may be calculated, it strikes a detector. In earlier times, such detectors often consisted of photographic emulsions, in which ionizing particles left evidence of their passage. See PARTICLE ACCELERATOR.

Modern magnetic spectrometers use sophisticated counter telescopes and multiwire proportional counters, which permit not only the registering of the particles characterized by a certain value of the radius of curvature (and therefore of momentum) but enable the particular particle (proton, alpha particle, or whatever) that caused the signal to be identified. Contemporary magnetic spectrometer systems not only utilize complex arrangements of magnetic fields, detectors, and electronics but also generally require powerful computers to monitor and analyze the results. See PARTICLE DETECTOR.

Fay Ajzenberg-Selove

Bibliography. H. L. Anderson (ed.), *A Physicist's Desk Reference: Physics Vade Mecum*, 2d ed., 1989; F. Ajzenberg-Selove and E. K. Warburton, Nuclear spectroscopy, *Phys. Today*, 36(11):26-32, November 1983; K. S. Krane, *Introductory Nuclear Physics*, 1987; W. McHarris (ed.), *Exotic Nuclear Spectroscopy*, 1991; Y. Yoshiizawa, T. Otsuka, and K. Kusakari, *The Frontier of Nuclear Spectroscopy*, 1993.

## Nuclear structure

At the center of every atom lies a small, dense nucleus, which carries more than 99.97% of the atomic mass in less than  $10^{-12}$  of its volume. The nucleus is a tightly bound system of protons and neutrons which is held together by strong forces that are not normally perceptible in nature because of their extremely short range. The small size, strong forces, and many particles in the nucleus result in a highly complex and unique quantal system that at present defies exact analysis. The study of the nucleus and the forces that hold it together constitute the field of nuclear structure physics. See ATOMIC STRUCTURE AND SPECTRA; NEUTRON; PROTON; QUANTUM MECHANICS; STRONG NUCLEAR INTERACTIONS.

The protons of the nucleus, being positively charged, generate a spherically symmetric electric field in which the atomic electrons orbit. The strength of this electric field is proportional to the number of protons (the atomic number), and the quantized orbits of the electrons in this field define the material and chemical properties of each element. The cloud of negatively charged atomic electrons normally balances the positive nuclear charge, making the atom electrically neutral and screening

the nucleus from view. To penetrate the electron cloud and reach the nucleus generally requires high-energy probes, and so most research on nuclei requires large particle accelerators. See ELECTRON; PARTICLE ACCELERATOR.

A discussion of nuclear structure must begin with some definitions of specialized terms and by setting scales of energy, time, and distance for nuclear systems. The atomic number of protons is usually denoted by  $Z$  and the number of neutrons, which are electrically neutral, by  $N$ . The total number of protons and neutrons (or nucleons) is the mass number  $A = Z + N$ . Isotopes have the same atomic number,  $Z$ , and hence are forms of the same chemical element, having the same chemical properties, but they differ in neutron number; isotones have a common number of neutrons,  $N$ , and isobars have the same mass number,  $A$ . The normal scheme of notation of a nucleus is to use the chemical symbol with superscripts and subscripts for  $A$ ,  $Z$ , and  $N$  in the configuration  ${}^A_ZX_N$ ; for example,  ${}^{40}_{20}\text{Ca}_{20}$  and  ${}^{208}_{82}\text{Pb}_{126}$  for isotopes of calcium and lead. The mass of a proton is  $1.6726 \times 10^{-27}$  kg (approximately 1836 electron masses) and that of a neutron is  $1.6749 \times 10^{-27}$  kg. See ISOBAR (NUCLEAR PHYSICS); ISOTONE; ISOTOPE.

Nuclei have masses less than the sum of the constituents, the missing mass  $\Delta M$  being accounted for by the binding energy  $\Delta Mc^2$  (where  $c$  is the speed of light), which holds the nuclear system together. The characteristic energy scale is in megaelectronvolts (1 MeV =  $1.6 \times 10^{-13}$  joule). The internuclear forces generate an attractive potential field which holds the nucleus together and in which the nucleons orbit in highly correlated patterns. The strength of this field, or potential well, is of the order of 40 MeV. The average kinetic energy of a nucleon in a nucleus is about 20 MeV, giving an orbital time of  $\tau \simeq 10^{-21}$  s, and the average energy to separate the least-bound nucleon is about 8 MeV. The volume of nuclei increases approximately linearly with mass number  $A$ , and the radius is roughly  $R = 1.2 \times 10^{-15} \cdot A^{1/3}$  m. See NUCLEAR BINDING ENERGY.

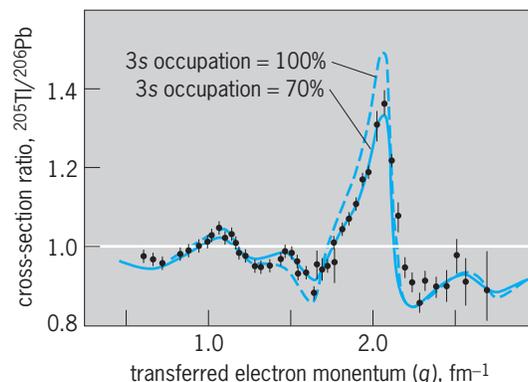
**Size, shape, and density distributions.** The existence of the atomic nucleus and the first estimates of its size came from experiments conducted by E. Rutherford and his collaborators in the period 1906-1911. They used energetic alpha particles emerging from the natural radioactive decays of very heavy elements to penetrate the cloud of atomic electrons and interact directly with the electric field caused by the nuclear protons. They were able to predict the alpha-particle scattering angular distribution and the dependence of the cross section on the number of protons,  $Z$ , in the atoms of the target material. However, for high-energy alpha particles and for light targets (with low  $Z$ ) the simple Rutherford scattering formula was found to break down as the alpha particles were able to overcome the Coulomb field and enter the domain of the strong (nuclear) forces at the nuclear surface. Thus, an estimate of nuclear size and its dependence on  $Z$  could be made. See ALPHA PARTICLES; SCATTERING EXPERIMENTS (NUCLEI).

This type of study has several shortcomings. Even

with the advent of particle accelerators, the deduced nuclear sizes were a composite of the nuclear size, the size of the alpha particle, and the range of the nuclear forces. Further, the measurements revealed information only about the position of the nuclear surface. For more precise estimates of the charge distribution, a variety of more sophisticated techniques have been developed, including electron scattering, the study of muonic atoms, and the laser spectroscopy of hyperfine atomic structure. Of these, electron scattering has produced some of the most spectacular and unambiguous results.

High-energy electrons interact with the nuclear charge distribution only through the electromagnetic force, which is a precisely understood phenomenon. These electrons can penetrate to the center of the nucleus, and the length scale to be probed can be varied according to the electron's energy. For example, by using electrons of 420-MeV energy, a length scale of 2.95 fermis (1 fermi =  $10^{-15}$  m) is probed, which reveals the distribution of nuclear charge without probing the nucleonic quark substructure. Many stable nuclei have been probed in this way. By examining the difference between scattering patterns from adjacent nuclei, the spatial probability distribution of particular protons can be investigated. This is a stringent test for nuclear models. **Figure 1** shows the difference in charge distributions of  $^{206}_{82}\text{Pb}_{126}$  and  $^{205}_{81}\text{Tl}_{126}$ , which can be ascribed to the one different proton. The probability distribution for that proton is close to that predicted by model calculations.

The nuclear charge distribution is an average property and is made up from the sum of contributions arising from the probability distributions of each individual proton. Oscillations in the mean charge density reflect the grouping of protons at certain radii, which is indicative of nuclear shells. However, an overall picture of the nuclear charge distributions emerges that has several defining features. First, the nuclear charge density saturates in the interior and



**Fig. 1.** Ratio of charge distributions of  $^{206}\text{Pb}$  and  $^{205}\text{Tl}$ , which differ by just one proton. The proton is calculated to lie in the  $3s_{1/2}$  shell, where it would have the probability distribution shown. The data (points and bars) indicate some mixing of configurations, and fit the distribution that is predicted when the proton occupies this state for only 70% of the time, as shown. (After B. Frois et al., *A precise determination of the 3s proton orbit*, *Nucl. Phys.*, A396:409–418, 1983)

has a roughly constant value in all but the lightest nuclei. The nucleus has a diffuse skin which is of nearly constant thickness. For spherical nuclei, the charge distribution can be approximated by the Woods-Saxon formula, Eq. (1), where  $\rho(r)$  is the density at

$$\rho(r) = \frac{\rho(0)}{1 + e^{(r-R)/a}} \quad (1)$$

$$R = (1.07 \pm 0.02) \times 10^{-15} \cdot A^{1/3}$$

$$a = 0.55 \pm 0.07 \times 10^{-15} \text{ m}$$

radius  $r$ ;  $\rho(0)$  is the mean density, approximately  $1.1 \times 10^{25}$  coulombs/m<sup>3</sup>;  $R$  the radius at which the density has fallen to  $\rho(0)/2$ ; and  $a$  the surface diffuseness.

Many nuclei are found to have nonspherical shapes. Unlike the atom, which has a spherically symmetric Coulomb field generated by the nucleus, the nuclear field is composed of a complicated superposition of short-range interactions between nucleons, and the most stable nuclear shape is the one that minimizes the energy of the system. In general, it is not spherical, and the nuclear shape is most simply described by a multipole power series, the most important term of which is the nuclear quadrupole moment. A positive quadrupole moment reflects the elongation of nuclei into a prolate or football-like shape, while a negative value reflects an oblate shape like that of Earth. *See* NUCLEAR MOMENTS.

An accurate determination of nuclear matter distributions, that is, the distribution of both protons and neutrons in nuclei, is harder to precisely ascertain. The moments and charge radii of long chains of isotopes have been investigated by using laser hyperfine spectroscopy. The nuclear charge distribution perturbs the atomic quantum states in a small but measurable way that subsequently can be measured by inducing transitions between states. The advantage of this technique over electron scattering is that since very small samples of short-lived radioactive isotopes can be measured, a vastly wider range of nuclei is accessible. **Figure 2** shows the deduced root-mean-square radii of a long chain of rubidium isotopes, clearly indicating that the assumption that the radius is proportional to  $A^{1/3}$  is only approximate and that a deviation from sphericity develops for the lightest rubidium isotopes. *See* HYPERFINE STRUCTURE; LASER SPECTROSCOPY.

**Nuclear masses and binding energies.** The masses of nuclei are measured relative to a standard. The standard is to define the mass of a neutral carbon-12 atom to be 12 atomic mass units or u. On an absolute energy scale 1 u = 931.50 MeV. Relative to this scale a proton has a mass 1.007276 u and a neutron 1.008665 u. Generally, the mass of a nucleus is expressed in terms of the mass of its neutral atom, that is, including the mass of the  $Z$  orbiting electrons.

The measured mass of a nucleus with  $Z$  protons and  $N$  neutrons is not the simple sum of the constituents but is less because of the energy used up in binding the system together. The binding energy of a nucleus is always positive and can be written

as in Eq. (2), where  $M$  stands for mass. Thus, the

Binding energy ( $A, Z, N$ )

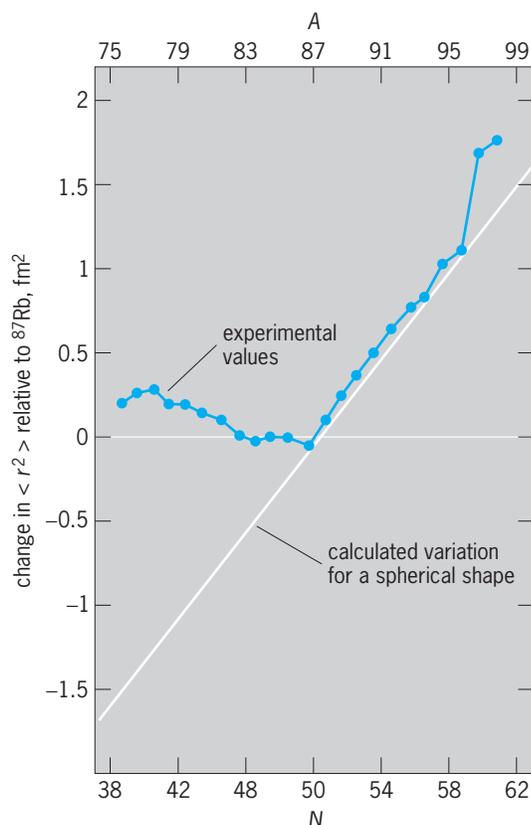
$$= M(\text{component protons, neutrons, and electrons})c^2 - M(\text{assembled nucleus})c^2 \quad (2)$$

binding energy is the energy required to separate all the components to infinity. Another expression of the energy involved in holding the nuclear system together is the mass excess, given by Eq. (3), which may be positive or negative.

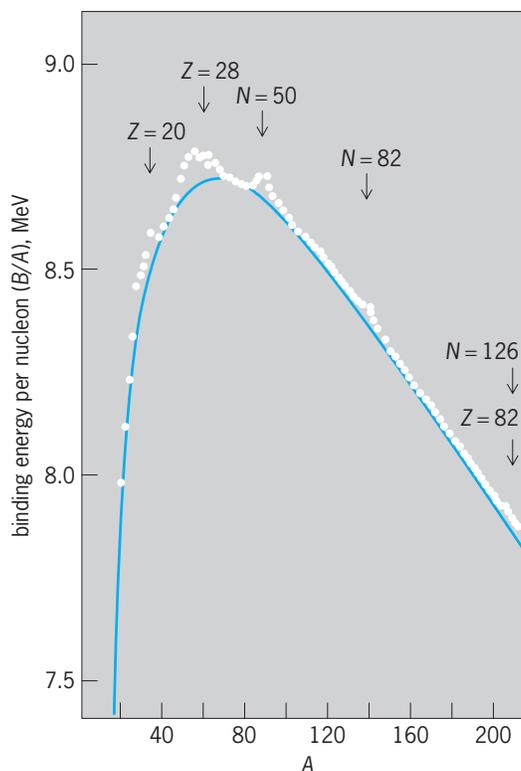
Mass excess ( $A, Z, N$ )

$$= M(\text{assembled nucleus})c^2 - Auc^2 \quad (3)$$

**Figure 3** shows the variation of average binding energy with mass number  $A$ . The shape of this curve can be reproduced by the Bethe-Weizsacker mass formula, which is noteworthy for its simplicity in reproducing the overall binding energy systematics. The formula is developed by modeling the nucleus on a liquid drop. By analogy with a drop of liquid, there is an attractive volume term, which depends on the number of particles ( $A$ ); a repulsive surface-tension



**Fig. 2.** Variation of the mean-square radius ( $r^2$ ) of rubidium isotopes (as measured by laser hyperfine spectroscopy) compared with the trend expected for spherical nuclei. The increase of mean radius due to deformation in the light nuclei is evident. (After C. Thibault et al., *Hyperfine structure and isotope shift of the  $D_2$  line of  $^{76-98}\text{Rb}$  and some of their isomers*, *Phys. Rev.*, C23:2720–2729, 1981)



**Fig. 3.** Binding energy  $B$  per nucleon plotted as a function of mass number  $A$ . The vertical axis is offset (it begins near 7.5 MeV instead of at 0 MeV) to highlight shell structure. (After W. N. Cottingham and D. A. Greenwood, *An Introduction to Nuclear Physics*, Cambridge University Press, 1986)

term (proportional to  $A^{2/3}$ , the area of the nuclear surface); and a term due to the mutual Coulomb repulsion of protons, which is responsible for the decrease in binding energy for heavy nuclei. To make the model reproduce experimental measurements requires further terms reflecting neutron-proton asymmetry [proportional to  $(N - Z)^2/A$ ] and reflecting odd-even pairing effects.

The model is spectacularly successful in reproducing the overall trends in nuclear binding energies, masses, and the energetics of nuclear fission, and in predicting the limits of stability where neutrons and protons become unbound. As in the case of predicting a mean nuclear shape, a comparison of the prediction of the Bethe-Weizsacker mass formula to measured masses shows periodic fluctuations with both  $N$  and  $Z$ . These are due to the quantum shell effects that cause deviations of binding energy from the smooth liquid-drop value and are clearly visible in Fig. 3 as irregularities at nuclear numbers 20, 28, 50, and 82. In order to incorporate these quantum effects a discussion of how nuclear systems carry energy and angular momentum is necessary. See NUCLEAR FISSION.

**Nuclear excited states.** The quantization of energy and angular momentum, which is apparent in all branches of subatomic physics, is drastically evident in nuclei. The small nuclear size and tightly bound nature impose very restrictive constraints on the orbits that protons and neutrons can undergo inside

the system. Thus, each nucleus has a series of quantum states that particles can occupy. Each state corresponds to a well-defined kinetic energy, rotational frequency (angular momentum), orbital orientation, and reflection symmetry (parity). These characteristics can be denoted by a series of quantum numbers that label the properties of the state. In addition to the quantum numbers of the states of the nucleus, the protons and neutrons have their own internal quantum numbers arising from their quark structure. These quantum numbers describe charge, internal angular momentum (spin), and magnetic moment. The Pauli principle requires that each particle have a unique set of quantum labels. Each nuclear state can then be filled with four particles: protons with internal angular momentum “up” and “down,” and likewise two neutrons. *See* ANGULAR MOMENTUM; ENERGY LEVEL (QUANTUM MECHANICS); EXCLUSION PRINCIPLE; PARITY (QUANTUM MECHANICS); QUANTUM NUMBERS; QUARKS; SPIN (QUANTUM MECHANICS).

A nucleus is most stable when all of its nucleons occupy the lowest possible states without violating this occupancy rule. This is called the nuclear ground state. During nuclear collisions the protons and neutrons can be excited from their most bound states and promoted to higher-lying unoccupied states. The process is usually very short-lived and the particles deexcite to their most stable configuration on a time scale of the order of  $10^{-12}$  s. The energy is usually released in the form of gamma rays of well-defined energy corresponding to the difference in energy of the initial and final nuclear states. Occasionally, gamma decay is not favored because of angular momentum selection rules, and long-lived nuclear isomers result. *See* GAMMA RAYS; NUCLEAR ISOMERISM; NUCLEAR SPECTRA; SELECTION RULES (PHYSICS).

**Nuclear models.** The detailed categorization of the excitation of protons and neutrons allows a mapping of the excited states of each nucleus and determination of its quantum numbers. These data are the essential information required for development of detailed models that can describe the motion of nucleons inside nuclei. The understanding of nuclear behavior advanced rapidly with the discovery and exploration of the wealth and variety of excited nuclear states, both in stable nuclei and in the many radioactive isotopes that could be produced in nuclear reactors and by particle accelerators. Many different modes of excitation were found that involved both collective (that is, involving all or most of the nucleons) and single-nucleon degrees of freedom. Unlike atomic molecules, where rotational, vibrational, and single-particle degrees of freedom involve different time scales and energies, the nucleus is highly complex, with rotation, vibration, and single-particle degrees of freedom being excited at similar energies and often strongly mixed. A full understanding of this behavior remains elusive, and the development of truly generalized nuclear models remains an unfulfilled goal. *See* MOLECULAR STRUCTURE AND SPECTRA.

The measurement of static electric and magnetic moments of nuclear states and of dynamic transi-

tion moments has provided a great deal of information. Electric moments have revealed a variety of enhanced collective modes of excitation, including elongated, flattened, and pear-shaped nuclear shapes. Magnetic moments have provided detailed information on the differences between excitations

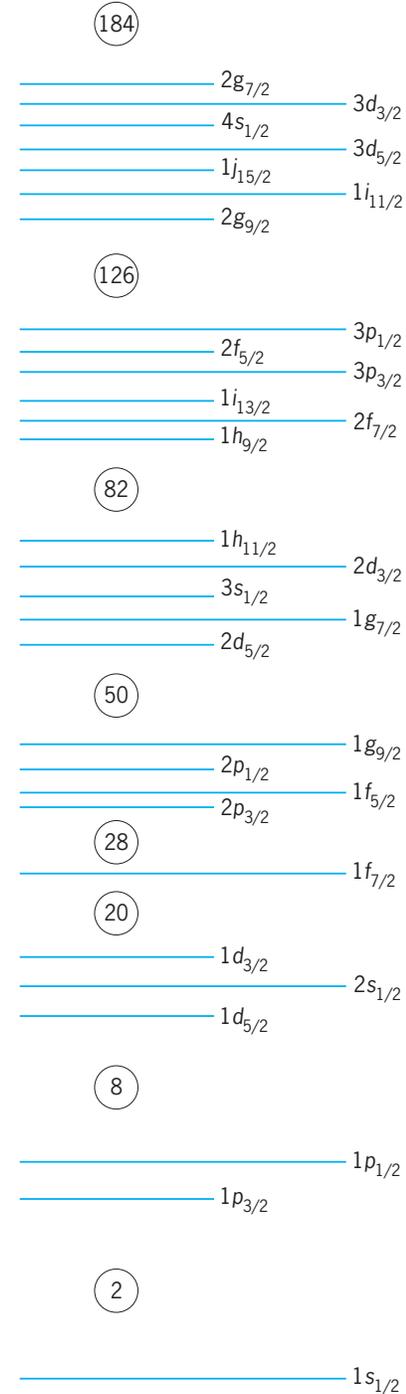


Fig. 4. Sequence of levels in a shell-model potential. The bunching of levels into groups with so-called magic gaps is evident. The numbers in these gaps are total numbers of protons or neutrons (magic numbers). The labeling of levels is according to their orbital angular momentum quantum numbers with  $\ell = 0, 1, 2, \dots$ , being denoted by the letters  $s, p, d, f, g, h, i, j$ . The subscripts denote the total angular momentum quantum number  $j$ . (After B.L. Cohen, *Concepts of Nuclear Physics*, McGraw-Hill, 1982)

involving neutrons (negative moments) and protons (positive moments).

In some particular regions of nuclei with certain well-defined neutron and proton numbers, the nuclear spectrum is drastically simplified when one mode of excitation dominates the quantum states. In these special regions, nuclear models were most successfully developed.

*Shell model.* For atoms, the solutions of the Schrödinger equation with a Coulomb potential lead to a reasonable prediction of the energies of quantized atomic states, as well as their spins, parities, and moments. Attempts to make the same progress for nuclei, using a variety of spherically symmetric geometric potentials of nuclear dimensions, failed to reproduce known properties until it was realized by M. G. Meyer and independently by J. H. Jensen in 1949 that an additional spin-orbit potential of the form given by Eq. (4) was required to reproduce

$$V_{so} = V_o(\vec{L} \cdot \vec{S}) \quad (4)$$

the known sequence of nuclear states, where  $\vec{L}$  is the orbital angular momentum and  $\vec{S}$  is the internal spin of the nucleons. A potential of this form binds states having the internal spin of the nucleons parallel to its orbital angular momentum more tightly than when they are antiparallel. The ensuing sequence of quantum shell gaps and subgaps was then correctly reproduced (Fig. 4). The shell model has evolved rapidly, and its domain of applicability has widened from the limited regions of sphericity near doubly magic nuclei to encompass most light nuclei with  $A < 60$  as well as enlarged regions around shell closures. This evolution has been closely linked with advances in computer technology that allow ever-increasing numbers of particles to be included in more configurations. The development of prescriptions to describe the residual interactions between the valence particles has also been very important.

Modern shell models can be spectacularly precise in their prediction of nuclear states and their wave functions. Calculations of gamma-ray and beta-particle transition rates can be made with great precision, and such detailed reproduction of nuclear wave functions presents the possibility of using the nucleus as a laboratory for searching for forbidden or exotic decay modes. See RADIOACTIVITY.

Large-basis shell models provide the most thorough description of nuclei that is available. However, the majority of nuclei still cannot be described by shell models because of the intractable computational problems presented by heavy nuclei, which have both larger numbers of basis states and many valence particles. To describe some of the other modes of nuclear excitation, a variety of models have been developed. Figure 5 shows the Segrè chart of nuclei plotted as a function of  $N$  and  $Z$ . The grid of lines is at the characteristic magic numbers where spherical shell effects are important and the shell model is most applicable. Away from these regions, in the areas denoted R, permanent deformation occurs and

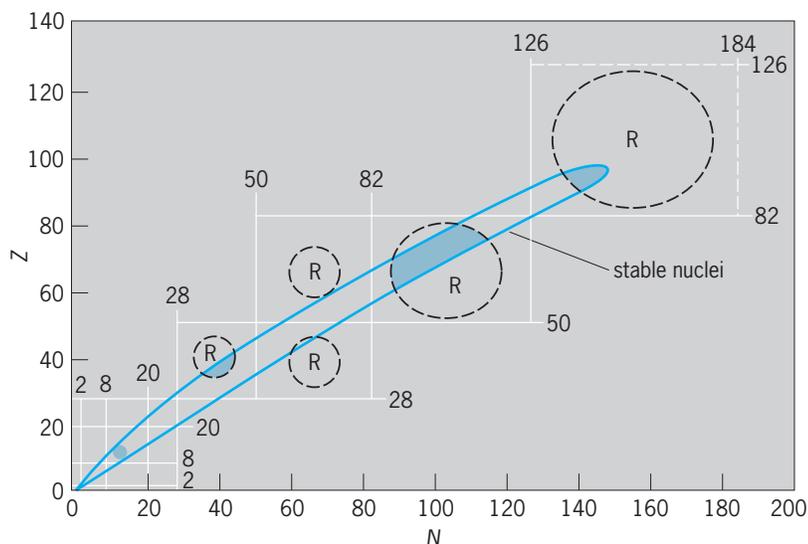


Fig. 5. Segrè chart of nuclides. The grid indicates the major spherical shell gaps. The regions denoted R are the areas where nuclei assume permanent deformation and rotational behavior dominates. The elongated diagonal band encloses the stable nuclei. (After E. Marshalek, *The systematics of deformations of atomic nuclei, Rev. Mod. Phys.*, 35:108–117, 1963)

rotational models are most appropriate. See MAGIC NUMBERS.

*Rotational models.* As valence particles are added to a closed core of nucleons, the mutual residual interactions can act coherently and polarize the nuclear system away from sphericity. The polarization effects are strongest when both valence protons and neutrons are involved. The deformed nuclear potential can then undergo collective rotation, which generally involves less energy than other degrees of freedom and thus dominates the spectrum of strongly deformed nuclei. The characteristic spectrum of a deformed quantum system, given by Eq. (5), encoun-

$$E = \frac{\hbar^2}{2\mathcal{I}} J(J+1) \quad (5)$$

tered, where  $\hbar$  is Planck's constant divided by  $2\pi$ ,  $J$  is the angular momentum of the deformed system, and  $\mathcal{I}$  is the moment of inertia, which is considerably less than that of a rigid body because of the highly correlated superfluid nuclear core. The deformation of most nuclear systems that lie far from both neutron and proton shell closures involves elongated, almost axially symmetric shapes, normally parametrized with a quadrupole deformation parameter,  $\beta_2$ . The elongated shape reflects the strong quadrupole correlations in the nucleon-nucleon residual interactions. In the transitional nuclei between shell closures and the highly deformed regions, there is great susceptibility to shape polarization. A wide variety of shapes are found that include triaxially deformed shapes and octupole (pear-shaped) deformations.

*Vibrational models.* Nuclei undergo collective vibrations about both spherical and deformed shapes. The degree of softness of these vibrations is characterized by the excitation energy required to populate states. The distinguishing feature of vibrational excited states is that they are grouped in nearly

degenerate angular momentum multiplets, each group being separated by a characteristic phonon energy.

*Generalized models.* It has been a goal of nuclear structure studies to develop models that incorporate all of the features described above in order to produce a unified nuclear picture. The aim is to reproduce and predict the properties of all bound nuclei of any  $N$  and  $Z$  as well as their evolution with angular momentum and excitation energy. The development of generalized nuclear models has relevance to other fields of physics. There are many isotopes that will never be accessible in the laboratory but may exist in stars or may have existed earlier in cosmological time. The evolution of generalized models greatly increases the power to predict nuclear behavior and provides information that is required for cosmological calculations. Considerable progress has been made in several areas, as discussed below. See COSMOLOGY; NUCLEOSYNTHESIS.

*Truncated shell models.* The expansion of the shell model to span all quantum states and their occupancy with protons and neutrons remains an intractable computational problem. However, a very successful approach to reducing the problem to tractable size has been made by A. Arima and F. Iachello, who considered a truncation incorporating pairs of particles coupled as bosons with angular momentum  $J = 0$  or  $2\hbar$ . The symmetries of recoupling these bosons are striking, and, in particular, the so-called limiting symmetries make it possible to predict properties of rotational, vibrational, and triaxial nuclei. The triaxial nuclei are of special interest since the model makes unique predictions about their nuclear deexcitation that have been experimentally

demonstrated. The power of this interacting boson approximation (IBA) model appears in its ability to predict the properties of transitional nuclei that lie between the limiting types of nuclear excitation and reproduce systematic trends found across the nuclear periodic table. See SYMMETRY LAWS (PHYSICS).

A development of the interacting boson approximation model that separately treats neutron and proton bosons is the IBA2 model, which has been useful in interpreting the very enhanced magnetic dipole (M1) electromagnetic transitions found in electron and gamma-ray fluorescent scattering experiments. One interpretation is that of a new type of collectivity, with deformed proton and neutron distributions oscillating in a so-called scissors mode. See MULTIPOLE RADIATION.

*Microscopic-macroscopic calculations.* A dramatically successful though less rigorous approach to the nuclear many-body problem has been applied to predicting high-spin phenomena, fission barriers, nuclear masses, and the shapes of nuclei far from stability. The deformation of a spherically symmetric potential causes many shell-model degeneracies to be lifted and the excitation energies of orbits to change according to their orientation with respect to the deformation axis. Initially, for small deformations, the level shifting and mixing destroys the normal shell gaps and magic numbers. However, at axis ratios approaching integer values (2:1, 3:1, 3:2, and so forth) and at some critical rotational frequencies, new symmetries appear and produce new quantum shell stabilization.

In order to calculate these new shell effects and their behavior with varying deformation and angular momentum, a combination of liquid droplet and deformed rotating shell models has been used. The approach has been to separate the total nuclear binding energy into two parts that are both shape-dependent but can be calculated independently. The overall bulk trends in binding energy and its evolution with  $N$ ,  $Z$ , shape, and angular momentum are incorporated into a semiclassical model based on the dynamics of a liquid drop. Separately, the energy of quantum states at the same conditions of  $N$ ,  $Z$ , shape, and spin are calculated, and the total binding energy from the single-particle energies is calculated. By using a formalism due to V. Strutinsky, the modulating effects of nuclear shell structure are calculated relative to the mean bulk properties, and the total binding energy can then be estimated. This method of calculation appears to provide a large advance in unifying nuclear models. One example of this type of calculation was used to predict the ground-state masses and shapes of all nuclei (more than 4000) predicted to be bound. Another version predicted the existence of so-called superdeformed nuclei with an axis ratio of 2:1, stabilized by deformed-shell effects. The observation of this exotic shape, through the characteristic gamma-ray patterns arising from its rotation (Fig. 6), provided a dramatic verification of this type of model.

The most stringent tests of these models appear to involve nuclei that are at the limits of nuclear

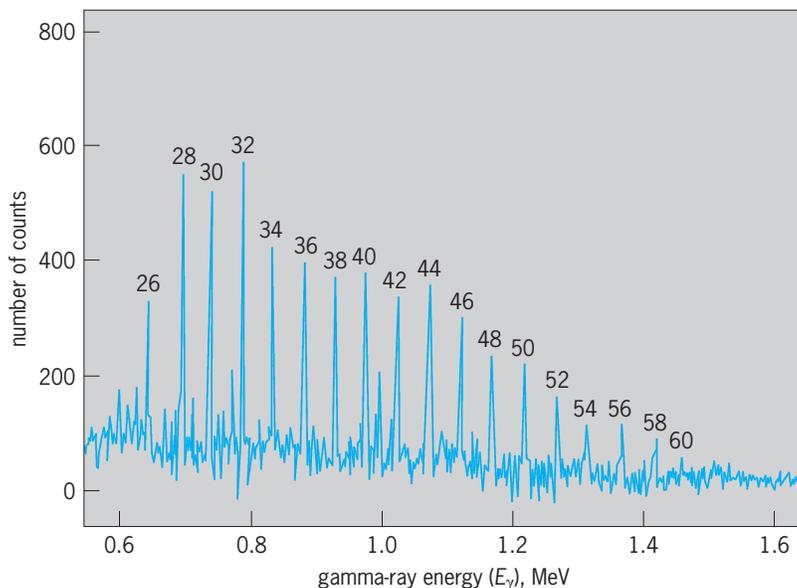


Fig. 6. Rotational sequence of gamma rays associated with superdeformation. The equal spacing between transitions is characteristic of an elongated nucleus with an axis ratio of 2:1. In this case, the rotating nucleus is dysprosium-152 ( $^{152}_{66}\text{Dy}$ ), which was formed in the fusion of a beam of calcium-48 ( $^{48}_{20}\text{Ca}$ ) ions with a palladium-108 ( $^{108}_{46}\text{Pd}$ ) target. Numbers above peaks indicate the total angular momentum quantum number  $J$  of the decaying state. (After P.J. Twin et al., *Observation of a discrete line superdeformed band up to  $60\hbar$  in  $^{152}\text{Dy}$* , *Phys. Rev. Lett.*, 57:811–814, 1986)

stability, either because of highly unstable neutron-to-proton ratios ( $N/Z$ ) or because of extremes of rotation (where Coriolis forces stress the nucleus) and temperature (where the highly organized correlations of nuclei break down and statistical behavior takes over). At these limits of stability, the subtle quantum effects can have a dominant influence on the total nuclear stability and can become the determining factor in maintaining stability. Much effort, both theoretical and experimental, is focused on these areas of marginal stability. Reactions involving the fusion of heavy ions can produce nuclei in a wide variety of states, including those of high angular momentum, neutron-to-proton ratio, and temperature. However, to attribute variations of nuclear behavior to any one of these degrees of freedom and to identify shell effects requires isolating specific isotopes and measuring their state of spin and temperature. To this end, large arrays of detectors subtending large solid angles and identifying all particles and photons emitted from nuclear reactions have been constructed in many laboratories. *See* EXOTIC NUCLEI; NUCLEAR REACTION.

**Nuclei at high excitation energies.** As nuclei are excited to ever higher excitation energies, it is anticipated that shell effects will be slowly replaced by statistical, or chaotic, behavior. The number of states per megaelectronvolt of excitation energy with each spin and parity rise exponentially with increasing excitation energy until the levels become sufficiently close that they overlap and mix strongly and so become a continuum of states. The extent and rate at which this mixing alters observable quantities such as electromagnetic moments is of considerable interest. The region of excitation that lies from a few megaelectronvolts to a few tens of megaelectronvolts above the yrast states, defined to be the lowest energy states of each spin, is under intense investigation.

Toward the top of this energy regime, new modes of nuclear collectivity become accessible. Giant resonance states can be excited that involve compression and oscillation of the nuclear medium with vibrations of the protons and neutrons in phase (isoscalar) or beating against each other (isovector). The excitation and decay of these giant resonances can provide information about shapes of nuclei at high excitation and about the compressibility of nuclear matter. Results from giant resonance studies indicate that the shell effect persists high into the region previously thought to be statistical. *See* GIANT NUCLEAR RESONANCES.

The semiclassical statistical and hydrodynamic behavior of hot nuclear matter and its experimental, theoretical, and astrophysical aspects are of great interest at the highest nuclear energies. The influence of compression and heat on the density of nuclear matter is being investigated in order to measure a nuclear equation of state in analogy with the properties of a classical fluid. It has been suggested that the nuclear matter may undergo phase changes under compression, with high-density condensates possibly providing a new metastable state. At the high-

est densities and temperatures, the nucleons themselves are forced to overlap and merge, leading to a plasma of quarks and gluons that are the nucleonic constituents. At present, experimental evidence for deconfined quark matter has not been found. A new generation of accelerators and detectors is undergoing development that will allow research to commence in this ultrahigh-energy domain that involves colliding beams of the very heaviest elements with energies of thousands of megaelectronvolts in order to reach this state of matter that may have previously only existed during the first few fractions of a second at the beginning of the universe. *See* QUARK-GLUON PLASMA; RELATIVISTIC HEAVY-ION COLLISIONS.

C. J. Lister

**Bibliography.** A. Bohr and B. Mottelson, *Nuclear Structure*, vol. 1, 1969, vol. 2, 1975; H. A. Enge and R. P. Redwine, *Introduction to Nuclear Physics*, 1995; K. S. Krane, *Introduction to Nuclear Physics*, 1987; J. D. Walecka, *Theoretical Nuclear and Subnuclear Physics*, 1995.

## Nucleation

The formation within an unstable, supersaturated solution of the first particles of precipitate capable of spontaneous growth into large crystals of a more stable solid phase. These first viable particles, called nuclei, may either be formed from solid particles already present in the system (heterogeneous nucleation), or be generated spontaneously by the supersaturated solution itself (homogeneous nucleation). *See* SUPERSATURATION.

Heterogeneous nucleation involves the adsorption of dissolved molecules onto the surface of solid materials such as dust, glass, and undissolved ionic substances. This adsorbed layer of solute molecules may then grow into a large crystal. Because the crystal lattice of the foreign solid is in general not the same as that of the solid to be precipitated, the first few layers are deposited in a lattice configuration which is strained, that is, less stable than the normal lattice of the precipitating material. The degree of lattice strain determines the effectiveness of a given heterogeneous nucleating agent. Thus, a material whose crystal structure is greatly different from that of the solid to be precipitated will not bring about precipitation unless the solution is fairly highly supersaturated, whereas, if the solution is seeded by adding small crystals of the precipitating substance itself, precipitation can occur at a concentration only slightly higher than that of the saturated solution.

If elaborate precautions are taken to exclude solid particles, it is possible to obtain systems in which the necessary precipitation nuclei are spontaneously generated within the supersaturated solution by the process of homogeneous nucleation. In a solution, ions interact with each other to form clusters of various sizes. These clusters in general do not act as nuclei, but instead, redissociate into ions. However, if the solution is sufficiently supersaturated so that its tendency to deplete itself by deposition of ions onto

the clusters overcomes the tendency of the clusters to dissociate, the clusters may act as nuclei and grow into large crystals. The rate at which suitable nuclei are generated within the system is strongly dependent upon the degree of supersaturation. For this reason, solutions which are not too highly supersaturated appear to be stable indefinitely, whereas solutions whose concentration is above some limiting value (the critical supersaturation) precipitate immediately.

Nucleation is significant in analytical chemistry because of its influence on the physical characteristics of precipitates. Processes occurring during the nucleation period establish the rate of precipitation, and the number and size of the final crystalline particles. See COLLOID; PRECIPITATION (CHEMISTRY).

David H. Klein; Louis Gordon

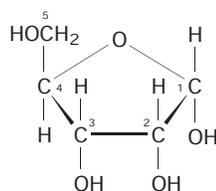
## Nucleic acid

An acidic, chainlike biological macromolecule consisting of multiply repeated units of phosphoric acid, sugar, and purine and pyrimidine bases. Nucleic acids as a class are involved in the preservation, replication, and expression of hereditary information in every living cell. There are two types: deoxyribonucleic acid (DNA) and ribonucleic acid (RNA).

### Deoxyribonucleic Acid

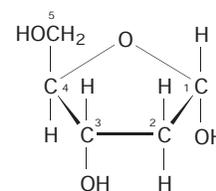
Each DNA strand is a long polymeric molecule consisting of many individual nucleotides linked end to end. The great size and complexity of DNAs are indicated by their molecular weights, which commonly range in the hundreds of millions. DNA is the chemical constituent of the genes of an organism, and is thus the ultimate biochemical object of the study of genetics. Information is contained in the DNA in the form of the sequence of nucleotide building blocks in the nucleic acid chain.

**Nucleotides.** The number of nucleotide building blocks in DNA is relatively small—only four nucleotides constitute the vast majority of DNA polymeric units. These are deoxyadenylic, deoxyguanylic, deoxycytidylic, and deoxythymidylic acids. For purposes of brevity, these nucleotides are symbolized by the letters A, G, C, and T, respectively. Each of these nucleotides consists of three fundamental chemical groups: a phosphoric acid group, a deoxyribose 5-carbon sugar group (**1b**), and a nitrogenous base which is a derivative of either purine (**2**) or pyrimidine (**3**). Some nucleotides contain the purine groups adenine (6-aminopurine; **4**) or guanine (2-amino-6-oxypurine; **5**), and some contain the pyrimidine groups cytosine (2-oxy-5-aminopyrimidine; **6**) or thymine (2,6-dioxy-6-methyl-pyrimidine; **7**). These are the only major bases found in most DNA, although in specific sequences certain methylated derivatives of these bases, such as 5-methyl cytosine (**8**) or *N*<sup>6</sup>-methyl adenine (**9**), can also be detected. In each nucleotide, these subunits are linked together in the following order: purine or pyrimidine base-ribose sugar-phosphoric acid (**10**)–(**12**).



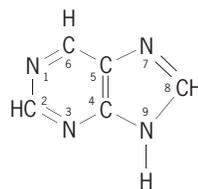
D-Ribose

(1a)

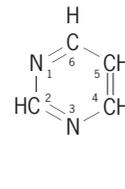


2-Deoxy-D-ribose

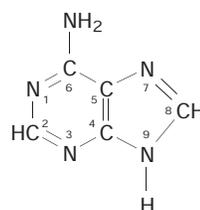
(1b)



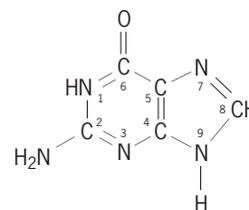
(2)



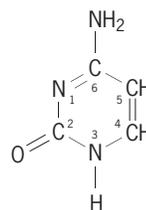
(3)



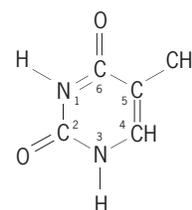
(4)



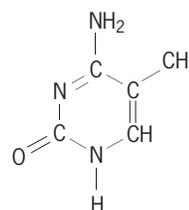
(5)



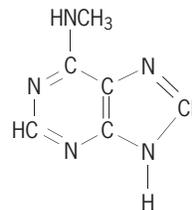
(6)



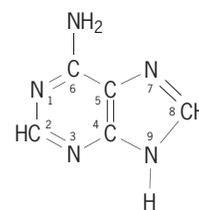
(7)



(8)

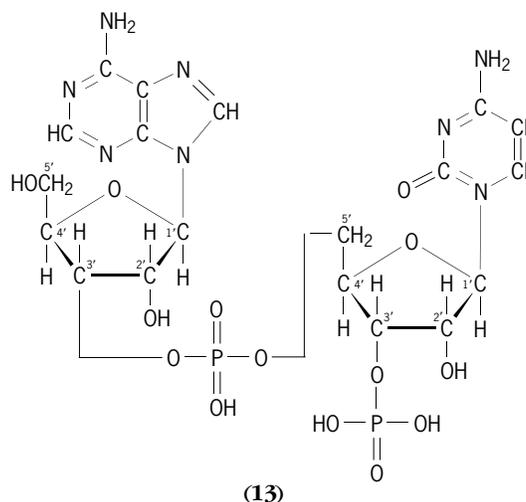
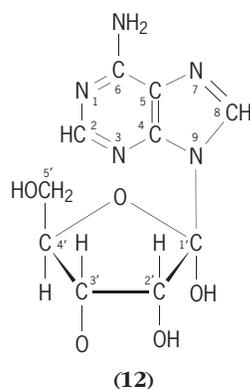
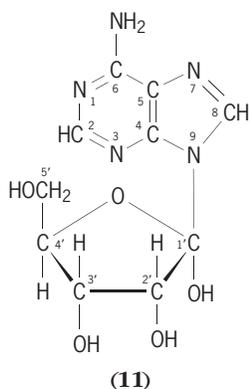


(9)



(10)

Removal of the phosphoric acid group leaves a base-sugar compound which is called a nucleoside (**11**). In each nucleoside the base is attached to the sugar through a bond from nitrogen at position 9 (for purines) or 3 (for pyrimidines) to carbon at position



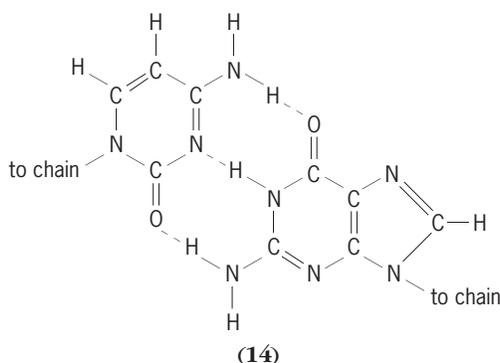
1 of the sugar ring. The nucleosides and nucleotides are named for the base they contain as follows:

Base	Nucleoside	Nucleotide
Adenine (10)	Adenosine (11)	Adenylic acid (12)
Guanine	Guanosine	Guanylic acid
Cytosine	Cytidine	Cytidylic acid
Thymine	Thymidine	Thymidylic acid
Uracil	Uridine	Uridylic acid

It is necessary to denote the position of the phosphoric acid residue when describing nucleotides. Nucleotides synthesized by cells for use as building blocks in nucleic acids all have phosphoric acid residues coupled to the 5' position of the ribose sugar ring, as shown for deoxyadenylic acid (deoxyadenosine-5'-phosphate; 12). Upon hydrolysis of DNA, however, nucleotides can be produced which have phosphoric acid coupled to the 3' position of the sugar ring.

When DNA is hydrolyzed by using the enzyme deoxyribonuclease I, prepared from bovine pancreas, the principal products are oligonucleotides ending with 3'-hydroxyl groups and 5'-phosphoric acid groups. In contrast, when DNA is hydrolyzed by using the enzyme micrococcal nuclease, prepared from *Staphylococcus aureus*, the principal products are nucleotides or oligonucleotides ending with 3'-phosphoric acid groups and 5'-hydroxyl groups. Studies such as these led very early to the conclusion that in intact DNA the nucleotides were linked via phosphoryl groups which join the 3' position of one sugar group to the 5' position of the next sugar group (13). This 5'-to-3' linkage of nucleotides imparts a polarity to each DNA strand which is an important factor in the ability of DNA to form the three-dimensional structures necessary for its ability to replicate and to serve as a genetic template. See NUCLEOTIDE.

**Helix.** In 1953 James Watson and Francis Crick proposed a double-helical structure for DNA based largely on x-ray diffraction studies of Maurice Wilkins and Rosalind Franklin. It had been known earlier that in most DNAs the ratios of A to T and of G to C are approximately 1:1. The Watson-Crick model attributes these ratios to the phenomenon of base pairing, in which each purine base on one strand of DNA is hydrogen-bonded to a complementary pyrimidine base in an opposing DNA strand (14). The x-ray



diffraction studies suggested that DNA can assume a structure called the B form, which is a right-handed helical configuration resembling a coiled spring. In most DNAs there are two single DNA strands which compose each helix. The strands wind about each other, with their sugar-phosphate chains forming the coil of the helix and with their bases extending inward toward the axis of the helix. The configuration of the bases allows hydrogen bonding between opposing purines and pyrimidines. Each of the base pairs lies in a plane at approximately right angles to the helix axis, forming a stack with the two sugar-phosphate chains coiled around the outside of the stack. In this configuration a wide (major) groove exists on the surface of the helix, winding parallel to a narrow (minor) groove (Fig. 1). In the Watson-Crick model the two opposing DNA strands have an opposite polarity; that is, they are antiparallel, with their 5' ends lying at opposite ends of each double-stranded molecule. The right-hand helix completes one turn every 10 nucleotides, and bases are spaced 0.34 nanometer apart. The width of the double-stranded helix in the B form averages about 2 nm. DNA can exist in helical structures other than the B form. One configuration, termed the Z form, is a left-handed helical structure with 12 nucleotides per turn. The Z form can exist in DNA sequences with alternating guanine and cytosine bases and may be functional in localized DNA regions, although the B form is thought to predominate in most biological systems.

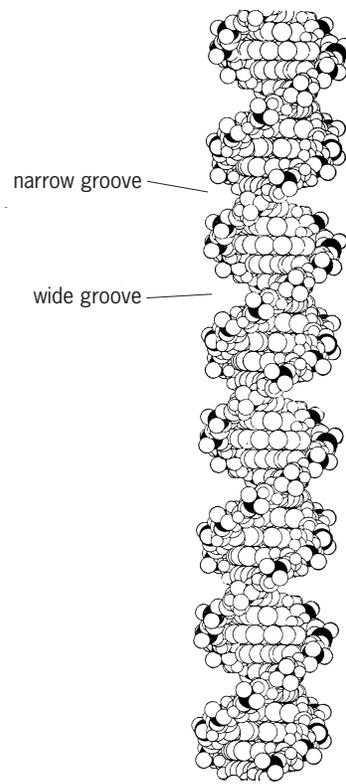


Fig. 1. Three-dimensional configuration of DNA.

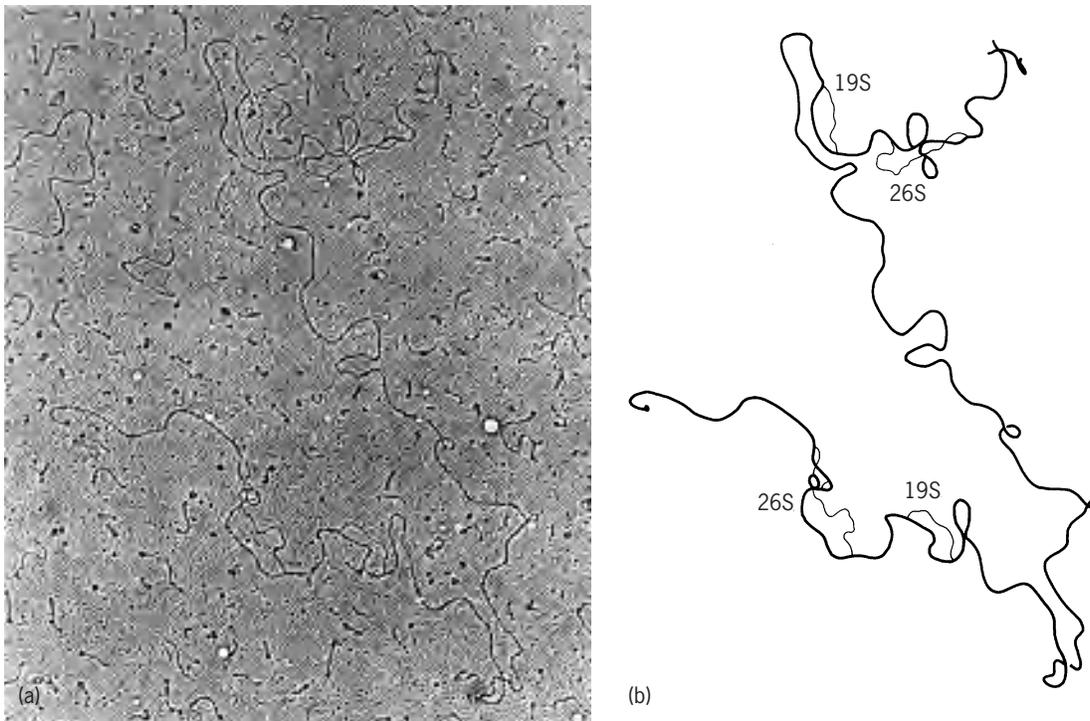
Due to the phenomenon of base pairing, the two DNA strands of a helix can be "melted" apart (denatured) or, once denatured, allowed to reanneal. Heating a DNA duplex in solution breaks the hydrogen bonds involved in base pairing and allows the two single strands of the helix to dissociate. This melting is dependent on the base sequence in the DNA, since there are two hydrogen bonds between A-T pairs and three hydrogen bonds between G-C pairs (14). In a solution of sheared DNA molecules, the rate of reannealing of denatured DNA depends on the extent to which the denatured DNA sequences are reiterated. Studies on the kinetics of renaturation of DNA from various sources have shown that different organisms differ widely in the complexity of sequences in their DNA. In general, cells of higher organisms contain more unique or seldom-repeated DNA sequences than do cells of lower organisms. This reflects in part the fact that higher organisms possess more individual genes required to code for the larger number of proteins they contain. Renaturation studies also show that many cells possess highly reiterated DNA sequences that may constitute a significant percentage of the total genome. When sheared cellular DNA sequences are separated according to their buoyant density, such as by centrifugation in density gradients of cesium chloride, many of these highly reiterated sequences can be detected as satellite DNA bands; that is, they form bands of DNA at positions in the gradient which differ from that of the bulk of cellular DNA. Several of the known satellite DNAs contain genes which are present in multiple copies in cells. For example, the genes coding for RNAs of

ribosomal subunits of cells are characteristically located in satellite DNA bands. At this point, however, the functional significance of most highly reiterated DNA sequences in cells is not known.

**C value.** The amount of DNA in the haploid genome of a given organism (called the C value) is only loosely correlated with the degree of evolutionary advancement of the organism. There are about  $2 \times 10^{-16}$  g of DNA in a bacteriophage, as compared to about  $10^{-14}$  g in the bacterium *Escherichia coli* and about  $3 \times 10^{-12}$  g in rat liver cells. Whereas mammalian cells contain about  $2-3 \times 10^9$  nucleotide pairs of DNA, amphibian cells vary widely, ranging from less than  $2 \times 10^9$  to about  $1.5 \times 10^{11}$  nucleotide pairs. Consideration of these figures leads to what is known as the C value paradox: generally cells of higher organisms contain much more DNA than is necessary to code for the number of proteins they contain. This paradox is not explained simply by different degrees of gene reiteration in different organisms. Instead, it is more likely that only a fraction of the total DNA codes for proteins in higher organisms and that the relative amount of noncoding DNA is highly variable. It is now known that many noncoding DNA sequences exist adjacent to RNA-coding sequences and serve in a regulatory capacity. In addition, noncoding DNA sequences, termed intervening sequences, may exist in the middle of sequences coding for certain RNA species.

**Sequences.** The sequence of nucleotide pairs in the DNA determines all of the hereditary characteristics of any given organism. The DNA acts as a template which is copied, through the process of transcription, to make RNA. The RNA in turn serves as a template in a process by which its encoded information is translated to determine the amino acid sequences of proteins. Each amino acid in a protein chain is specified by a triplet of nucleotides (in RNA) or nucleotide pairs (in DNA) known as a codon. The set of correlations between the amino acids and their specifying codons is called the genetic code. Each gene which codes for a protein thus contains a sequence of triplet codons which corresponds to the sequence of amino acids in the polypeptide. This sequence of codons may be interrupted by intervening DNA sequences so that the entire coding sequence is not continuous. In addition to coding sequences, there also exist regulatory sequences, which include promoter and operator sequences involved in initiating gene transcription and terminator sequences involved in stopping transcription. Regulatory sequences are not necessarily made up of triplets, as are the codons. In order to study the regulation of a given gene, it is necessary to determine its nucleotide sequence. See GENETIC CODE.

There are several methods for sequencing DNA. Most of these methods employ radioactive end-labeling of one or both DNA strands, followed by either cleavage or extension of labeled strands to produce end-labeled fragments which terminate at nucleotides with specific bases. For example, one commonly used method involves labeling the 5' end of a DNA strand with  $^{32}\text{P}$ -phosphate by using the



**Fig. 2.** Ribosomal RNA genes of *Physarum polycephalum*: (a) electron micrograph and (b) map, showing 19S and 26S RNA coding regions as R-loop hybrids. This DNA molecule is 60,000 base pairs long, or about  $20 \times 10^{-6}$  m. (Courtesy of G. R. Campbell, V. C. Littau, P. W. Melera, V. G. Allfrey, and E. M. Johnson)

enzyme polynucleotide kinase and  $\gamma$ - $^{32}$ P-adenosine triphosphate as a phosphate donor. Procedures are then used to induce base-specific chemical cleavage of the end-labeled strands at each of the four nucleotides in turn. Polyacrylamide gel electrophoresis is then used to size the radioactive fragments, and the sequence of nucleotides from the labeled 5' end can be deduced.

The complete DNA sequences of several different genes, together with adjacent regulatory sequences, are now known. The first entire genome of any organism to be sequenced was that of the single-strand DNA phage  $\phi$ X174. This sequence of 5386 nucleotides was worked out by Fred Sanger and coworkers. One interesting aspect of this sequence is that it reveals the presence of overlapping genes coding for proteins. A single nucleotide sequence can code for more than one amino acid sequence, depending on the phasing with which the sequence is grouped into triplet codons during the process of protein synthesis. Specific start and stop signal codons are required to specify the phasing.

**Figure 2a** shows an electron micrograph of a single DNA molecule containing the genes coding for ribosomal RNAs of the slime mold *Physarum polycephalum*. The coding sequences for two RNA species, 19S and 26S, are revealed by R-loop hybridization. (This is a process by which RNA is annealed with double-stranded DNA, thus displacing a single DNA strand.) In order to prepare the nucleic acids for visualization, they are first coated with a layer of electron-dense heavy metals—in this case platinum

and palladium. **Figure 3** shows the DNA sequence near the start point of transcription of the gene coding for the enzyme  $\beta$ -galactosidase in the bacterium *E. coli*. This sequence contains short segments which are important for binding and subsequent action of enzymes that polymerize RNA. Regions for binding of proteins (CAP protein, RNA polymerase, and repressor protein) are shown to the DNA map. The *i* gene codes for the repressor protein which binds to the operator region and regulates transcription. The CAP protein must be bound to the DNA in order for the polymerase to initiate properly. The CAP protein binds its site only in the presence of the compound cyclic adenosine monophosphate. RNA polymerase binds a site about 40 nucleotides long and, in the absence of repressor, begins transcription of messenger RNA. The *z* gene codes for the enzyme  $\beta$ -galactosidase. The DNA sequence of one strand of the polymerase binding region is shown at bottom. The entry site and Pribnow's box sequences are important recognition sequences for polymerase binding and initiation. See OPERON.

Enzymes called restriction endonucleases cleave DNA at specific sequences. It is possible, by using these enzymes and DNA ligases, to splice an exogenous gene segment into an existing DNA sequence. The resulting structure is known as recombinant DNA. When the recombinant includes the genome of a plasmid or phage, this spliced DNA molecule can be propagated in bacteria. This process is known as gene cloning. Through this process, genes which are present only in a few copies in mammalian cells can

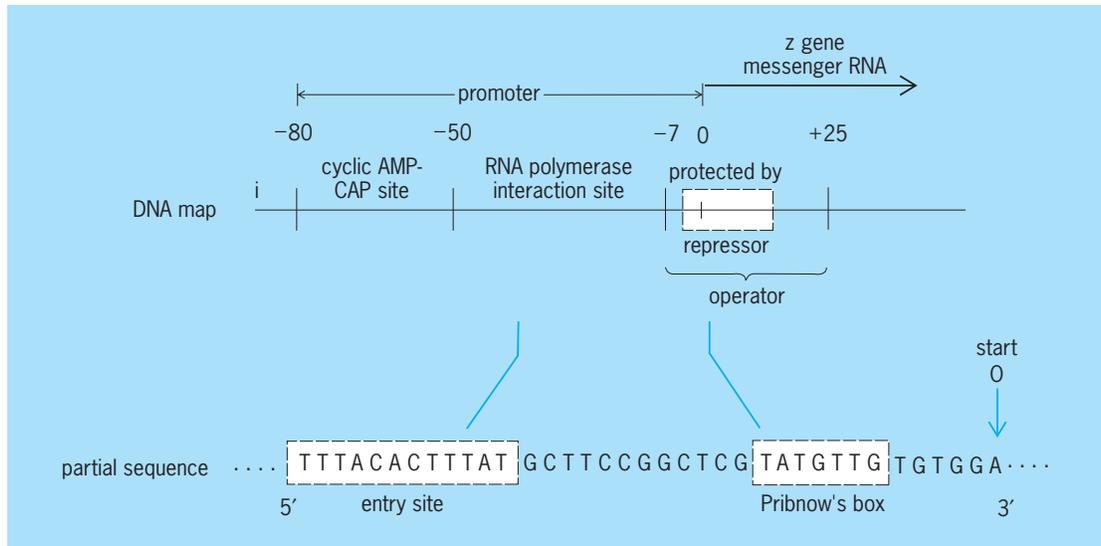
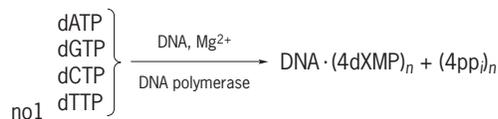


Fig. 3. DNA sequence of the promoter-operator region of the *lac* operon in *Escherichia coli*.

be grown in vast quantities in bacteria. Cloning is industrially important for the production of certain gene products, such as some hormones. See GENETIC ENGINEERING.

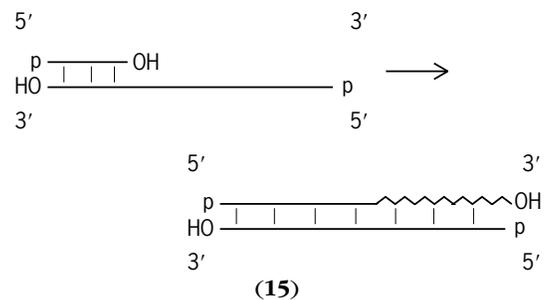
**Replication.** Since DNA is the substance containing the chemical code governing the heredity of all cells, it is clear that its biosynthesis and passage from generation to generation must be extremely precise. The way in which a new DNA molecule arises with the same nucleotides arranged in the same sequence as in the parent DNA molecule is one of the most intriguing problems in biochemistry.

In all cells, DNA is synthesized by using parent DNA as a template for a polymerization reaction which requires deoxyribonucleoside 5'-triphosphates as precursors and which is catalyzed by DNA polymerase enzymes. (The deoxyribonucleoside triphosphates are synthesized in cells by addition of three phosphate groups to each nucleoside.) The DNA synthetic reaction can be summarized as reaction (1), where XMP represents nucleoside



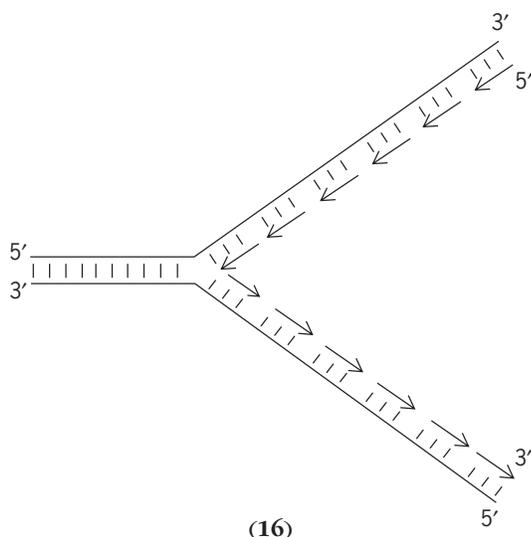
monophosphate and  $\text{pp}_i$  represents inorganic pyrophosphate. The reaction always proceeds in a 5'-to-3' direction. That is, each deoxyribonucleotide is attached through its 5'-phosphate to the 3'-OH group of an existing DNA strand, with the release of pyrophosphate. This 5'-to-3' directionality is true for all known nucleic acid biosynthetic reactions. All DNA polymerase enzymes require an existing strand of DNA as a template, with a base-paired segment to serve as a primer. The primer sequences may consist of ribonucleotides, but the template is DNA. No DNA polymerase is capable of synthesizing DNA de novo. The primer must possess a free 3'-OH group for synthesis to begin. Through base pairing, nucleotides

complementary to those in the template strand are added (15).



During DNA replication in cells, each strand of the helix serves as a template for synthesis. The resulting coupling of each new DNA strand with a preexisting strand is known as semiconservative replication. At a point along the double helix the strands unwind and, after attachment of primers, synthesis proceeds bidirectionally at what is termed the replication fork. When viral DNA is replicated in *E. coli*, the newly synthesized DNA, called nascent DNA, occurs in short fragments, now called Okazaki fragments (16). The action of enzymes known as DNA ligases is required to join the Okazaki fragments. In fact, it is now known that many enzymes in addition to DNA polymerase are involved in the process of DNA replication in cells. Studies on bacterial mutants defective in DNA synthesis have implicated more than a dozen such enzymes.

There are differences between DNA synthetic mechanisms in bacteria and higher organisms. In lower organisms such as bacteria or viruses, DNA replication begins at a single point in the genome and proceeds bidirectionally to replicate all the DNA. In higher organisms, replication begins at many points, thus forming numerous "eyes." The parental DNA strands are separated and replicated until these eyes meet. The DNA polymerases of bacteria comprise several different enzymes, each of which is involved in a separate aspect of replication. The DNA

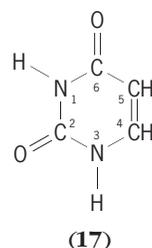


polymerases of higher organisms comprise a still different set of enzymes.

**Function.** In every living cell, as well as in certain viruses and subcellular organelles, the function of DNA is similar: it encodes genetic information and replicates to pass this information to subsequent generations. The nucleotide sequence of DNA in each organism determines the nature and number of proteins to be synthesized, as well as the organization of the protein-synthesizing apparatus. The first conclusive evidence that DNA alone can direct cell function was obtained in bacterial cells. In 1944 O. T. Avery, C. M. MacLeod, and M. McCarty showed that purified DNA from a strain of smooth, polysaccharide-encapsulated *Pneumococcus* could transform rough, nonencapsulated strains so that they and their offspring resemble the DNA donor. Subsequent studies on bacterial transformation showed that numerous other hereditary characteristics could be transferred between strains using pure DNA. G. Beadle and E. L. Tatum showed that mutations in DNA of the mold *Neurospora* resulted in alterations in the amino acid sequence of the protein. The entire process of gene expression, by which the flow of information proceeds from DNA to RNA to protein, remains one of the most fertile areas of molecular biological research. See DEOXYRIBONUCLEIC ACID (DNA).

### Ribonucleic Acid

RNAs are long polymeric chains of ribonucleotides joined end to end in a 5'-to-3' linkage. The primary chemical difference between RNA and DNA is in the structure of the ribose sugar of the individual nucleotide building blocks: RNA nucleotides possess a 2'-OH group, whereas DNA nucleotides do not (1). Another major chemical difference between RNA and DNA is the substitution of uridylic acid, which contains the base uracil (2,6-dioxyuracil; 17) for thymidylic acid as one of the four nucleotide building blocks. Thus incorporation of radioactive uridine can be used as a specific measure of RNA synthesis in cells, while incorporation of radioactive thymidine can be used as a measure of DNA synthesis.



Further modifications of RNA structure exist, such as the attachment of various chemical groups (for example, isopentenyl and sulfhydryl groups) to purine and pyrimidine rings, methylation of the sugars (usually at the 2' position), and folding and base pairing of sections of a single RNA strand to form regions of secondary structure. Unlike DNA, nearly all RNA in cells is single-stranded (except for regions of secondary structure) and does not consist of double-helical duplex molecules. Another distinguishing characteristic of RNA is its alkaline lability. In basic solutions, RNA is hydrolyzed to form a mixture of nucleoside 2'- and 3'-monophosphates. In contrast, DNA is stable to alkali.

**Classes.** Cellular RNA consists of classes of molecules widely divergent in size and complexity.

*Ribosomal RNA.* The most abundant class of RNA in cells is ribosomal RNA. This class comprises those molecular species that form part of the structure of ribosomes, which are components of the protein-synthesizing machinery in the cell cytoplasm. The predominant RNA molecules are of size 16S and 23S in bacteria (the S value denotes the sedimentation velocity of the RNA upon ultracentrifugation in water), and 18S and 28S in most mammalian cells. In most mammalian cells the 18S RNA is approximately 2100 nucleotides long, and the 28S RNA approximately 4200 nucleotides. In the large subunit of ribosomes, the 28S RNA is associated with a smaller 5S RNA species which is about 120 nucleotides long. In eukaryotic cells another small RNA species, 5.8S RNA, is also associated with 28S RNA. All of the ribosomal RNA species are relatively rich in G and C residues. For example, the 28S RNA of the newt *Triturus* contains 60.9% G and C residues. As they exist in ribosomes, the ribosomal RNAs contain considerable intrachain base pairing which helps maintain a folded three-dimensional structure. See RIBOSOMES.

*Messenger RNA.* Another prominent class of RNAs consists of the messenger RNA molecules and their synthetic precursors. Messenger RNAs are those species that code for proteins. They are transcribed from specific genes in the cell nucleus, and carry the genetic information to the cytoplasm, where their sequences are translated to determine amino acid sequences during the process of protein synthesis. The messenger RNAs thus consist primarily of triplet codons. Most messenger RNAs are derived from longer precursor molecules which are the primary products of transcription and which are found in the nucleus. These precursors undergo several steps known as RNA processing which result in production of cytoplasmic messenger molecules ready for

translation. Processing steps frequently include addition of a 7-methyl guanosine residue to the 5' terminal triphosphate group, splicing-out of intervening sequences in the coding regions, and addition of multiple adenylic acid residues (often as many as 200) to the 3' terminus of the polynucleotide chain. Certain messengers, such as those coding for histones, do not possess a polyadenylic acid tail. The messenger RNAs as a class possess a high degree of sequence complexity and size diversity, stemming from the fact that they are largely the products of uniquely represented structural genes. Most messenger RNAs range from 1500 to 3000 nucleotides in length. The relative size of their synthetic precursors varies widely, ranging from only slightly longer to more than twice as long as the final messenger RNA product.

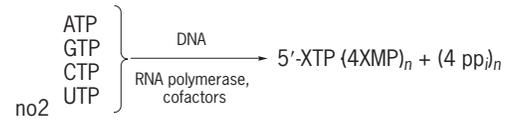
**Transfer RNA.** The third major RNA class is transfer RNA. There are approximately 80 transfer RNA molecular species in the bacterium *E. coli*. These are small RNA molecules 73–93 nucleotides long which possess a relatively high proportion of modified and unusual bases, such as methylinosine or pseudouridine. Each transfer RNA molecule possesses an anticodon and an amino acid binding site. The anticodon is a triplet complementary to the messenger RNA codon for a particular amino acid. A transfer RNA molecule bound to its particular amino acid is termed charged. The charged transfer RNAs participate in protein synthesis: through base pairing they bind to each appropriate codon in a messenger RNA molecule and thus order the sequence of attached amino acids for polymerization.

In addition to the three major RNA classes, numerous small RNA species are found in the cell nucleus. These RNAs frequently have a high content of unusual bases. Some of them are synthesized in the cell nucleolus. Their function is unknown. Some of the small nuclear RNAs arise from the splicing out of intervening sequences from the transcripts of large interrupted genes.

**Sequences.** Alanine transfer RNA (a transfer RNA recognizing the alanine codon) was the first RNA molecule to be sequenced in its entirety (18). The sequences of many other RNA molecules and segments of molecules are now known. Of particular interest are sequences near RNA transcription initiation or termination sites, since these provide information about gene-regulatory regions. Comparison of sequences of messenger RNAs and ribosomal RNAs has revealed that base pairing is important in attachment of messenger RNA to the ribosomes during initiation of protein synthesis. In general, RNA sequencing is carried out by cleavage of radioactively labeled RNA with nucleotide-specific nuclease enzymes, followed by two-dimensional separation and further analysis of the labeled cleavage products.

**Transcription.** The process of RNA biosynthesis from a DNA template is called transcription. Transcription requires nucleoside 5'-triphosphates as precursors, and is catalyzed by enzymes called RNA polymerases. Unlike DNA replication, RNA transcription is not semiconservative: only one DNA strand

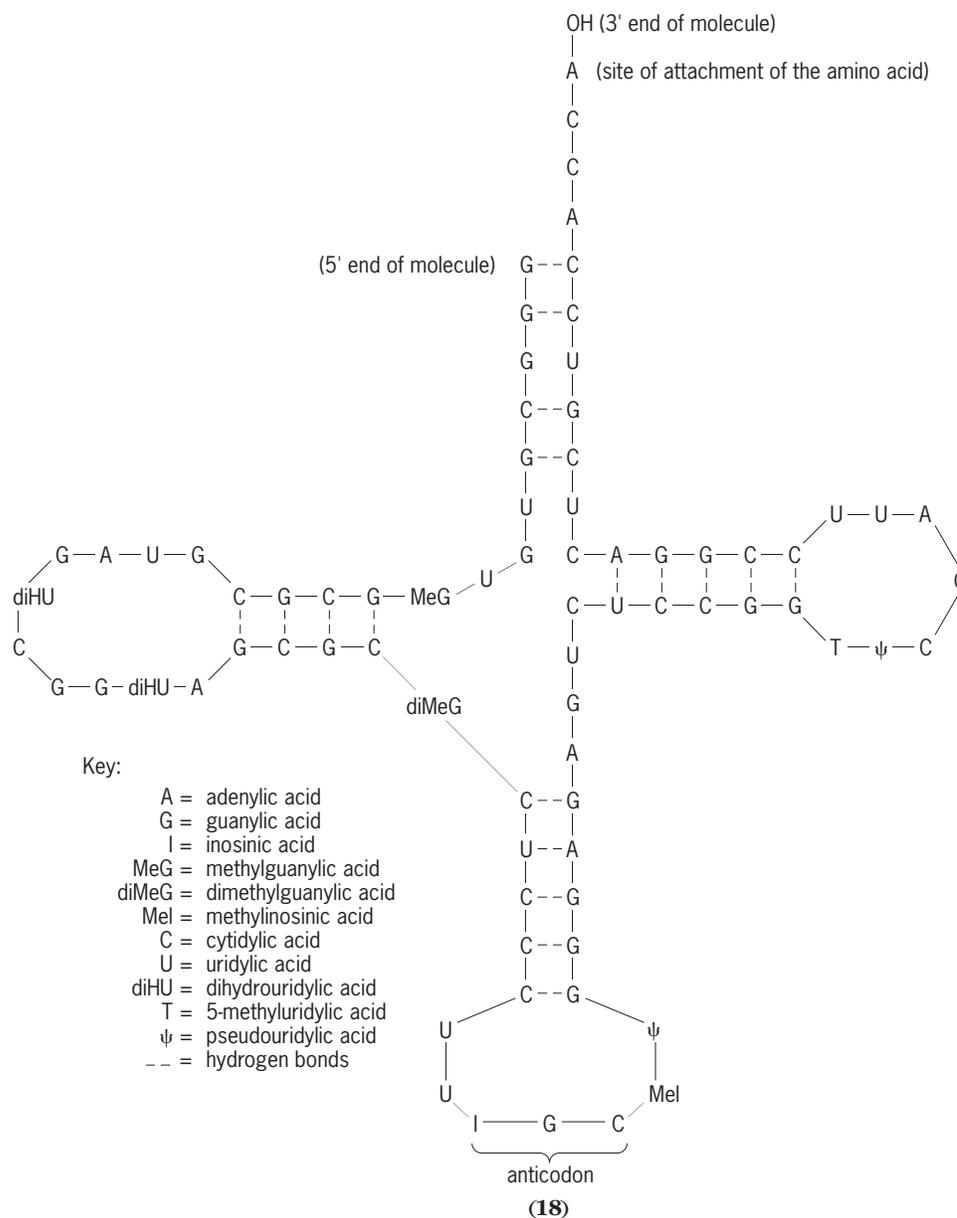
(the "sense" strand) is transcribed for any given gene, and the resulting RNA transcript is dissociated from the parental DNA strand. The RNA synthetic reaction can be summarized as reaction (2), where 5'-XTP rep-



resents nucleoside 5'-triphosphate, XMP represents nucleoside monophosphate, and  $\text{pp}_i$  represents inorganic pyrophosphate. As is true of DNA synthesis, RNA synthesis always proceeds in a 5'-to-3' direction. Since RNA synthesis requires no primer, as does DNA synthesis, the first nucleotide in any primary transcript retains its 5'-triphosphate group. Thus an important test to determine whether any RNA molecule is a primary transcript is to see whether it possesses an intact 5'-triphosphate terminus. As in DNA replication, base pairing orders the sequence of nucleotides during transcription. In RNA synthesis uridine, rather than thymidine, base-pairs with adenine.

There are profound differences between bacteria and higher organisms with regard to mechanisms of RNA synthesis and processing. In bacteria, RNA synthesis and translation are coupled; that is, nascent messenger RNA chains are attached directly to ribosomes for protein synthesis. In higher organisms, which possess a distinct cell nucleus, several processing and transport steps intervene between RNA synthesis (in the nucleus) and translation (in the cytoplasm). Bacterial RNA polymerization employs essentially one RNA polymerase enzyme. In contrast, eukaryotic RNA synthesis employs predominantly three RNA polymerases. Messenger RNA synthesis is catalyzed by one enzyme, 5S and transfer RNA synthesis by another, and ribosomal RNA synthesis by still another. All three of these eukaryotic RNA polymerases carry out the same 5'-to-3' phosphodiester linkage of nucleotides. They differ in their protein subunit composition and in the regulatory DNA sequences they recognize. The recognition of different regulatory sequences by different RNA polymerases under different conditions is one of the most interesting problems in molecular biology. It is this sequence-specific recognition that lies at the heart of the problem of differential gene activity and thus of cell differentiation.

It is known that many proteins cofunction with RNA polymerases to regulate DNA binding and RNA chain initiation. Among these regulatory proteins are several whose activities are influenced by hormones. For example, an estrogen-binding protein regulates transcription of the chick ovalbumin gene, and a cyclic adenosine monophosphate-binding protein regulates RNA polymerase activity at the *lac* operon in *E. coli* (16). In addition to these transcriptional regulatory factors, chromosomal proteins called histones exist in complexes with the DNA that can influence the rate of passage of polymerases. These histone-DNA complexes, or nucleosomes, are present on most eukaryotic cellular DNA, and are subject to several biochemical modifications which are



influenced by hormones. The interaction of RNA polymerases with DNA in nucleosomes is presently an active area of study.

**Functions.** The primary biological role of RNA is to direct the process of protein synthesis. The three major RNA classes perform different specialized functions toward this end. The 18S and 28S ribosomal RNAs of eukaryotes are organized with proteins and other smaller RNAs into the 45S and 60S ribosomal subunits, respectively. The completed ribosome serves as a minifactory at which all the components of protein synthesis are brought together during translation of the messenger RNA. The messenger RNA binds to the ribosome at a point near the initiation codon for protein synthesis. Through codon-anticodon base pairing between messenger and transfer RNA sequences, the transfer RNA molecules bearing amino acids are juxtaposed to allow formation of the first peptide bond between amino acids.

The ribosome then, in a presently unknown fashion, moves along the messenger RNA strand as more amino acids are added to the peptide chain. *See* PROTEIN.

RNA of certain bacterial viruses serves a dual function. In bacteriophages such as f2 and Q-beta, the RNA serves as a message to direct synthesis of viral-coat proteins and of enzymes needed for viral replication. The RNA also serves as a template for viral replication. Viral RNA polymerases which copy RNA rather than DNA are made after infection. These enzymes first produce an intermediate replicative form of the viral RNA which consists of complementary RNA strands. One of these strands then serves as the sense strand for synthesis of multiple copies of the original viral RNA. *See* BACTERIOPHAGE.

RNA also serves as the actively transmitted genomic agent of certain viruses which infect cells of higher organisms. For example, Rous sarcoma virus,

which is an avian tumor virus, contains RNA as its nucleic acid component. In this case the RNA is copied to make DNA by an enzyme called reverse transcriptase. The viral DNA is then incorporated into the host cell genome, where it codes for enzymes which are involved in altering normal cell processes. These enzymes, as well as the site at which the virus integrates, regulate the drastic transformation of cell functions, which induces cell division and the ultimate formation of a tumor. Transcription of the viral DNA results in replication of the original viral RNA. See CYCLIC NUCLEOTIDES; NUCLEOPROTEIN; RIBONUCLEIC ACID (RNA); TUMOR VIRUSES.

Edward Johnson

### In Situ Hybridization

In situ hybridization is a technique that permits identification of particular DNA or RNA sequences while these sequences remain in their original location in the cell. Single-stranded nucleic acids, both DNA and RNA, can bind to single strands of complementary nucleotide sequences to produce double-stranded molecules (nucleic acid hybrids). The stability of such hybrids is primarily dependent on the number of base pairs formed between the two strands. Under stringent hybridization conditions, only completely complementary sequences will form enough base pairs to remain associated. Thus a single-stranded nucleic acid probe can, by hybridization, identify a gene (DNA) or RNA transcribed from that gene.

In situ hybridization can be used to localize DNA sequences within chromosomes or nuclei and also to localize RNA within the nucleus or cytoplasm of the cell. Although most in situ hybridization studies are analyzed with the light microscope, it is also possible to perform in situ hybridization experiments on chromosomes prepared for analysis in the electron microscope.

The hybridization probe is the piece of single-stranded nucleic acid whose complementary sequence is to be localized within the cell. For example, the probe might be a recombinant DNA molecule carrying the gene that codes for a known

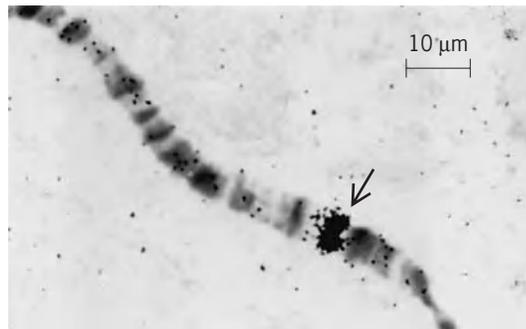


Fig. 4. Autoradiograph of part of a chromosome from *Drosophila melanogaster* hybridized with  $^3\text{H}$ -DNA carrying the genes for muscle tropomyosin. Autoradiographic grains (arrow) are over the region that has the genes for both tropomyosin I and tropomyosin II. (From V. L. Bautch et al., *Organization and expression of Drosophila tropomyosin genes*, *J. Mol. Biol.*, 162:231-250, 1982)

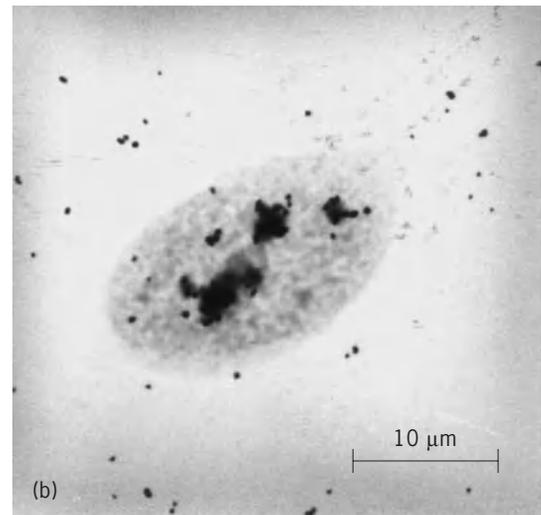
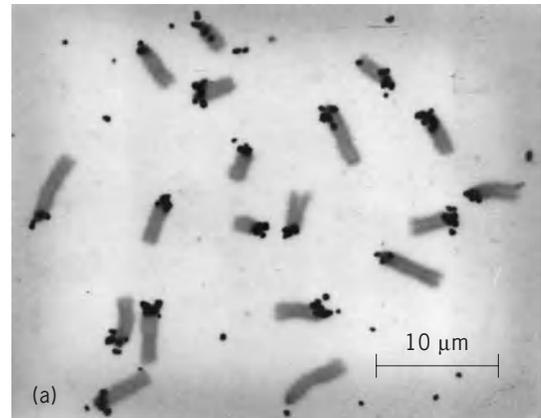
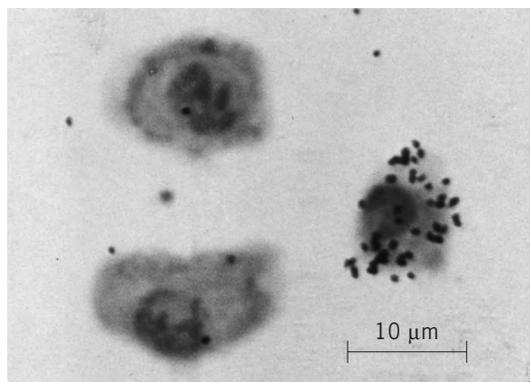


Fig. 5. Autoradiographs of mouse cytological preparations hybridized with  $^3\text{H}$ -RNA copied from mouse satellite DNA. (a) Metaphase chromosomes, each with the centromere at one end. The black autoradiographic grains indicate that the highly repeated simple sequences of satellite DNA are adjacent to the centromeres in most of the chromosomes. (b) Interphase nucleus of Sertoli cell of testis. The autoradiographic grains are found in only a few clusters, indicating that in nuclei of this particular cell type the centromeric regions of the chromosomes are tightly associated into a few small groups. (From C. D. Darlington and K. R. Lewis, eds., *Chromosomes Today*, vol. 3, Hafner Publishing, 1972)

protein. Nucleic acid probes can be labeled with the radioisotope tritium ( $^3\text{H}$ ) by several different techniques. Sulfur-35 can also be used to label probes. At the end of an experiment, the location of the hybridized radioactive probe is detected by autoradiography of the cell preparation, thus demonstrating the location of the complementary sequence or sequences in the cell or tissue (Fig. 4). It is also possible to use nonradioactive probes made with modified nucleotides. For those probes, the hybrid is detected by an antibody against the modified nucleotides. The antibodies used have been joined to fluorescent molecules or to enzymes to allow cytological detection at the end of the experiment. See AUTORADIOGRAPHY.

The cell preparation is made by treating the tissue to be studied with a fixative. Fixatives are agents that preserve the morphology of the cell as well as



**Fig. 6.** Autoradiograph of *Drosophila* cells showing use of in situ hybridization to detect histone messenger RNA. Cells from a nonsynchronized population were hybridized with a  $^3\text{H}$ -labeled probe complementary to histone messenger RNA, which is found only in cells that are synthesizing DNA. Thus, only the cell on the right, with autoradiographic grains indicating histone messenger RNA in the cytoplasm, appears to be actively replicating its DNA.

possible during subsequent steps. Fixatives usually act by cross-linking or denaturing proteins. After cells have been fixed, they are sectioned or squashed to yield preparations that are thin enough (a few micrometers) to be mounted on slides and analyzed in the microscope. If the experiment is designed to localize DNA in the cells, the DNA is denatured by treating the cells with a denaturing agent, such as alkali. The denaturation step also removes some protein from the DNA, but enough protein is left to maintain cellular morphology. Since RNA is, for the most part, single-stranded no denaturing step is used for hybridization to RNA in cells.

The probe is placed on the cytological preparation and held there under stringent conditions until hybrids have had sufficient time to form. The nonhybridized probe is then washed off the preparation. If the probe is radioactive, the preparation is covered with autoradiographic emulsion and exposed in the dark until tritium decay has produced grains in the emulsion over the labeled hybrid. The autoradiograph is then developed, and the location of grains over the cells is analyzed. When a nonradioactive probe is used, the hybrid detection is accomplished by incubating the preparation with an antibody that recognizes the modified nucleotide probe. If the antibody is coupled to a fluorescent molecule, the preparation is viewed in a fluorescence microscope to localize sites of hybridization. If the antibody is tagged with an enzyme, the preparation is incubated with a substrate from which the enzyme will produce a colored product, thus marking the site of the hybrid. See IMMUNOFLUORESCENCE.

In situ hybridization can be used to answer a number of biological questions. Hybridization to the DNA of condensed chromosomes can be used for gene mapping (Fig. 5a). Hybridization to DNA of interphase nuclei allows study of the functional organization of sequences in the diffuse chromatin of this stage of the cell cycle (Fig. 5b). Hybridiza-

tion to cellular RNA allows a precise analysis of the tissue distribution of any RNA (Fig. 6).

Mary Lou Pardue

Bibliography. R. L. Adams, J. T. Knowler, and D. P. Leader, *The Biochemistry of the Nucleic Acids*, 11th ed., 1992; G. Eckstein and D. M. Lilley, *Nucleic Acids and Molecular Biology*, 1994; G. Eckstein and D. M. Lilley (eds.), *Nucleic Acids and Molecular Biology*, vol. 2, 1988; A. R. Leitch, *In Situ Hybridization*, 1994; W. Saenger, *Principles of Nucleic Acid Structure*, 1993; P. Tijssen, *Hybridization with Nucleic Acid Probes*, 2 vols., 1993.

## Nucleon

The collective name for a proton or a neutron. These subatomic particles are the principal constituents of atomic nuclei and therefore of most matter in the universe. The proton and neutron share many characteristics. They have the same intrinsic spin, nearly the same mass, and similar interactions with other subatomic particles, and they can transform into one another by means of the weak interactions. Hence it is often useful to view them as two different states or configurations of the same particle, the nucleon. Nucleons are small compared to atomic dimensions and relatively heavy. Their characteristic size is of order 1/10,000 the size of a typical atom, and their mass is of order 2000 times the mass of the electron. See I-SPIN.

**Proton-neutron differences.** The proton and neutron differ chiefly in their electromagnetic properties. The proton has electric charge +1, the opposite of the electron, while the neutron is electrically neutral. They have significantly different intrinsic magnetic moments. Because the neutron is slightly heavier than the proton, roughly 1 part in 1000, the neutron is unstable, decaying into a proton, an electron, and an antineutrino with a characteristic lifetime of approximately 900 s. Although some unified field theories predict that the proton is unstable, no experiment has detected proton decay, leading to lower limits on its characteristic lifetime ranging from  $1.6 \times 10^{25}$  years to  $5 \times 10^{32}$  years, depending on the assumed decay channel.

**Structure and dynamics.** In the early 1900s the proton was identified as the nucleus of the hydrogen atom. After discovery of the neutron in the early 1930s, atomic nuclei were quickly understood to be bound states of nucleons. To explain most aspects of nuclear structure and dynamics, it suffices to treat the nucleons as pointlike particles interacting by means of complex forces that are parametrized numerically. The complex forces between nucleons and the discovery during the 1950s of many similar subatomic particles led physicists to suggest that nucleons might not be fundamental particles. During this period, hundreds of short-lived subatomic particles were discovered, characterized, and cataloged. Those that interact in strong and complex ways with nucleons are known as hadrons. Electron scattering

experiments during the 1950s indicated that nucleons and other hadrons are extended objects with a distributed charge density. Significant progress was made during the late 1960s and 1970s, when inelastic electron and neutrino scattering experiments indicated that nucleons are composed of pointlike particles with spin  $1/2$  and electric charges that are fractions of the charge on the electron. Particles with similar properties, named quarks, had been hypothesized in the early 1960s to explain other regularities among the properties of hadrons. In the early 1970s, it became clear that nucleons and other hadrons are indeed bound states of quarks. *See* HADRON; NUCLEAR STRUCTURE.

Quarks are believed to be fundamental particles without internal structure. The proton consists of two up-type quarks and one down-type quark (*uud*), while the neutron consists of *ddu*. Many of the properties and interactions of nucleons can be understood as consequences of their quark substructure. Quarks are bound into nucleons by strong forces carried by gluons. The nucleon contains ambient gluon fields in somewhat the same way that the atom contains ambient electromagnetic fields. Because quarks and gluons are much less massive than the nucleon itself, their motion inside the nucleon is relativistic, making quark-antiquark pair creation a significant factor. Thus the nucleon contains fluctuating quark-antiquark pairs in addition to quarks and gluons. A significant fraction (of order one-half) of the momentum of a rapidly moving nucleon is known to be carried by gluons. Likewise, only a small fraction of the nucleon's angular momentum resides on the spins of its constituent quarks. These facts indicate that the nucleon is a complex object characterized by highly nontrivial internal dynamics.

**Quantum chromodynamics.** The theory of quark-gluon interactions is known as quantum chromodynamics (QCD), in analogy to the quantum theory of electrodynamics (QED). Quantum chromodynamics has brought a deeper understanding to a previously bewildering subatomic world. Hundreds of other subatomic particles similar to nucleons are now understood to arise as different configurations of quarks and gluons. At the same time, the proton and neutron, once viewed as special because they are the constituents of ordinary matter, are now understood to be merely the most stable of a large number of composite quark-gluon bound states. Although quantum chromodynamics can be formulated in a simple and elegant manner, the dynamical equations that describe the nucleon bound state have so far resisted all attempts at quantitative solution. Thus a quantitative description of the nucleon from first principles has not yet been achieved. *See* ELEMENTARY PARTICLE; GLUONS; NEUTRON; PROTON; QUANTUM CHROMODYNAMICS; QUANTUM ELECTRODYNAMICS; QUARKS.

Robert L. Jaffe

Bibliography. R. K. Bhaduri, *Models of the Nucleon*, 1988; J. F. Donoghue, E. Golowich, and B. R. Holstein, *Dynamics of the Standard Model*, 1992; A. Pais, *Inward Bound*, 1986; R. G. Roberts, *The Structure of the Proton*, 1990.

## Nucleoprotein

A generic term for any member of a large class of proteins associated with nucleic acid molecules. Nucleoprotein complexes occur in all living cells and viruses, where they play vital roles in deoxyribonucleic acid (DNA) replication, transcription, ribonucleic acid (RNA) processing, and protein synthesis. The chemistry of the nucleoproteins has become one of the fastest-growing areas in cell and molecular biology. Developments in this field have already revolutionized our understanding of complex biological processes within a cell. These advances will have important consequences in the treatment of genetic disease, viral infections, and cancer.

Classification of nucleoproteins depends primarily upon the type of nucleic acid involved—DNA or RNA—and on the biological function of the associated proteins. DNA and RNA share very similar sugar-phosphate backbones, with nitrogenous bases attached to the ribose sugar. The major distinction is that DNA contains a deoxyribose sugar group, with the loss of a hydroxyl group at the 2' carbon. Deoxyribonucleoproteins (complexes of DNA and proteins) constitute the genetic material of all organisms and many viruses. They function as the chemical basis of heredity and are the primary means of its expression and control. Most of the mass of chromosomes is made up of DNA and proteins whose structural and enzymatic activities are required for the proper assembly and expression of the genetic information encoded in the molecular structure of the nucleic acid. *See* DEOXYRIBONUCLEIC ACID (DNA).

Ribonucleoproteins (complexes of RNA and proteins) occur in all cells as part of the machinery for protein synthesis. This complex operation requires the participation of messenger RNAs (mRNAs), amino acyl transfer RNAs (tRNAs), and ribosomal RNAs (rRNAs), each of which interacts with specific proteins to form functional complexes called polysomes, on which the synthesis of new proteins occurs. *See* RIBONUCLEIC ACID (RNA).

In simpler life forms, such as viruses, most of the mass of the viral particle is due to its nucleoprotein content. The material responsible for the hereditary continuity of the virus may be DNA or RNA, depending on the type of virus, and it is usually enveloped by one or more proteins which protect the nucleic acid and facilitate infections. *See* BACTERIOPHAGE; VIRUS.

Classically, the central dogma of molecular biology was that nucleic acids, the genetic material, are the transmissible hereditary material. According to the central dogma, DNA is transcribed into RNA, which is then translated into proteins that carry out various functions within the cell. With the discovery of noncoding RNAs and RNA viruses that use reverse transcriptase, it is increasingly clear that this central dogma needs further expansion. In addition, epigenetic modification of DNA (changes that do not alter DNA code, but can influence whether genes are expressed or repressed) represents an encryption not included in this dogma. The processes of reading the DNA code, transcription of RNA, and translation of

mRNA into protein all involve interaction of proteins with nucleic acids in some form to carry out cellular processes. Therefore, it is important to understand the composition and function of nucleoproteins. *See* GENE ACTION.

### Deoxyribonucleoproteins

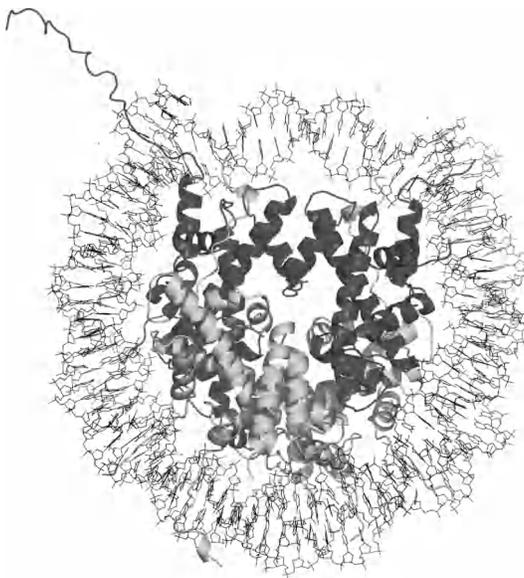
Those complexes in which the nucleic acid component contains the sugar 2-deoxyribose are primarily concerned with the storage and utilization of genetic information. This accounts for the localization of most of the deoxyribonucleoprotein in the cell nucleus as components of the chromosomes. *See* CHROMOSOME; NUCLEIC ACID.

The presence of DNA in the metaphase chromosomes of dividing cells and in the corresponding chromatin fibers (the complex of DNA and proteins) of interphase cells is revealed by direct chemical analysis of isolated chromosomes or chromatin, or by staining reactions that make DNA visible under the microscope. An important corollary of the chromosomal localization of DNA is the striking constancy in the amount of DNA per set of chromosomes in the different cells of a given organism. The nuclei of diploid somatic cells, which are body cells with two sets of chromosomes, have twice the DNA content of sperm cells, which contain only a single chromosome set. Cells in the process of division, at the time of chromosome duplication, show a corresponding increase in the amount of DNA which they contain. The direct demonstration of a genetic role of DNA came from experiments in bacterial transformation, when it was discovered that when DNA from one bacterial strain was transferred to a different strain, the recipient bacteria took on phenotypic characteristics of the DNA donor. The change induced by the DNA is inherited and persists through subsequent generations. *See* TRANSFORMATION (BACTERIA).

While the hereditary properties of cells and viruses are determined by their unique DNA structure (that is, by the nucleotide sequences which spell out the genetic code for protein structure), the ability to read the code and use such information as the basis of activity resides in the proteins that interact with DNA. In nature, the different types of DNA genomes, each encoding for a large number of functions, are associated with a complex array of different proteins. Many of these proteins are connected with generation of RNA molecules from DNA; other nuclear proteins are primarily concerned with the organization of chromosome structure. *See* GENETIC CODE.

**Role of histones in higher organisms.** The genetic complexity of higher organisms reflects a corresponding complexity in their DNA sequences. A typical human diploid nucleus, for example, contains enough DNA that, if fully extended, would be over 1 m (3.3 ft) long. However, the ability to compact this amount of DNA into a nucleus 0.01 mm in diameter is accomplished routinely by cells. The reduction in size is largely due to interactions between the DNA and sets of small basic proteins called histones.

All somatic cells of higher organisms contain five major histone classes, all of which are characterized



**Fig. 1. Histone core complex.** DNA is wrapped around a core octamer of histones. Two dimers of histones H2A/H2B form this complex with two copies of histones H3 and H4. (Rendered by Debra Ferraro)

by a high content of the basic amino acids arginine, lysine, and histidine (which are positively charged to interact with negatively charged DNA). The characterization of histones was originally based on their chromatographic or solubility properties, but emphasis is now placed on details of their structure and their capacity to organize DNA. The major classes are called H1, H2A, H2B, H3, and H4. Selective degradation of chromatin by staphylococcal nuclease releases subunits (nucleosome "core" particles) containing about 140 base pairs of DNA associated with eight molecules of histone. Analysis of the histones in the subunits reveals that only four of the five major histone classes are represented, with two copies of each per subunit particle. The histones form a core complex composed of two dimers of H2A/H2B and a tetramer of H3 and H4 (**Fig. 1**). Histone H1 does not appear to be part of the nucleosome core, but instead binds the linker DNA between cores, adding another level of DNA compaction. In general, the association of H1 prevents DNA sequences (genes) from being transcribed.

The presence of histone-DNA aggregates in chromosomes or in the chromatin of interphase cells can be visualized directly by electron microscopy (**Fig. 2**). The characteristic beads-on-a-string arrangement has now been observed in the chromosomes of organisms as diverse as slime molds, birds, and humans, and appears to be a general mechanism for the assembly and packaging of the genetic material.

While chromatin compaction and nucleosome arrangement provide stability and structure to DNA, chromatin must also be dynamic in the sense that selected DNA regions are called into action as templates for RNA synthesis at different times in the life of the cell. Two molecular mechanisms used to

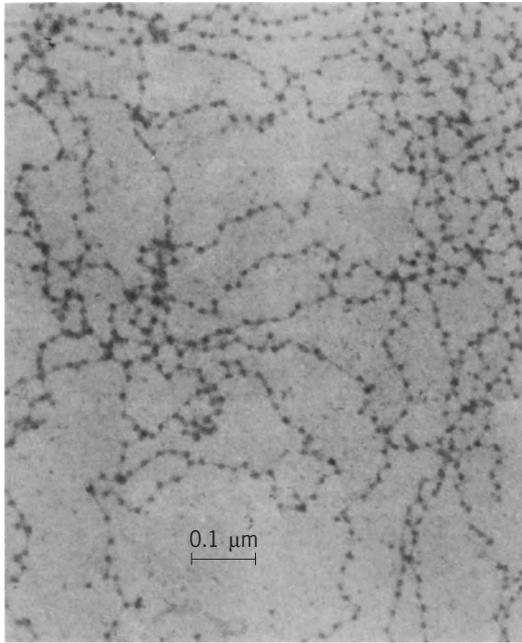


Fig. 2. Histone-DNA aggregates in chromosomes.

change chromatin assembly, properties, and function as needed are modifications of the histone proteins or their replacement with variant histones.

**Histone variants.** Histones H3 and H4 are among the most conserved proteins among species. As a whole, the nucleosome structure and proteins are also very conserved, hinting at a vital role. However, each of the five classes of histones contains variants that serve specific functions in the cell at different times. Histone variants are encoded by separate genes and contain core structures similar to their canonical family members. Changes in the size and the content of the N- and C-terminal domains, as well as single amino acid substitutions within the core, provide unique functions. While the major histones are ex-

pressed mainly during the DNA synthesis (S phase of the cell cycle), the variants are often used for specific functions at other stages of the cell cycle or development.

Histone H2A has the greatest number of variants. These variants range from the vertebrate-specific histone macroH2A which binds to the inactive X chromosome, to H2A.Z, a variant conserved from yeast to human that is involved in transcriptional activation and repression. The other histone classes have a number of members with specificities that provide specialized chromatin packaging in specific regions of genomic DNA.

**Histone modifications and the "histone code."** Biochemical alterations can modify specific amino acids in the histone polypeptide chain. Attachment of acetyl, phosphate, and methyl groups not only alters the structure and charge of the nucleosome but also regulates interactions between histone and nonhistone proteins (see **table**). These postsynthetic modifications of histones are sometimes referred to as an "epigenetic histone code." The code can be "translated" into various biological consequences, such as changes in transcriptional activity, chromatin condensation, or temporal and spatial regulation of gene activation. Numerous correlations have been noted between the extent of histone acetylation and RNA synthesis. Indeed, histone acetylation is enriched at very early stages in development when genes are activated in response to hormones, growth factors, or other stimuli. Furthermore, it has been observed that chromosome regions that are incapable of RNA synthesis are packaged with histones in the nonacetylated form.

**Nonhistone specialized packaging.** In sperm cells, particular mechanisms have evolved for the packaging and delivery of genetic material. Special DNA-binding proteins are synthesized in the course of sperm maturation. Among the simplest of these are the protamines which occur in the sperm cells of fish. These proteins are characterized by short chain

Examples of histone modifications and their effects on cellular functions\*

Histone	Modification (residue)	Associated function
Histone H1	Phosphorylation	Transcriptional activation
Histone H2A	Phosphorylated (S1) Phosphorylated (T119)	Transcriptional repression, chromatin condensation Cell cycle progression?
Histone H2B	Phosphorylated (S14) Phosphorylated (S33) Ubiquitinated (K119 or K123)	Apoptotic chromosome condensation Transcriptional activation Histone H3 methylation
Histone H3	Methylated (K4) Acetylated (K9) Methylated (K9) Phosphorylated (S10) Acetylated (K14) Methylated (K27) Methylated (K79)	Transcriptional activation Transcriptional activation Transcriptional repression Chromosome condensation and mitosis/meiosis Transcriptional activation Transcriptional repression Transcriptional activation, telomeric silencing
Histone H4	Phosphorylated (S1) Methylated (R3) Acetylated (K16)	Chromatin condensation during mitosis Enriched at developmentally regulated genes Transcriptional activation

\*Modifications are made to amino acid residues in the histones. Modified amino acids are abbreviated using a single letter code (serine, S; threonine, T; lysine, K; arginine, R) and are followed by the position of the amino acid within the protein, with the first amino acid of the protein designated 1.

lengths (about 33 amino acids) and a great preponderance of the basic amino acid arginine. In salmon, the protamine of salmon sperm, arginine alone accounts for 89–90% of the total protein nitrogen. It imparts a high positive-charge density to the polypeptide chain. The positively charged arginine residues occur in clusters, each member of which binds a phosphate group in DNA. The number of basic amino acids in nucleoprotamine complexes approximately equals the number of phosphate groups in the DNA, indicating almost complete neutralization of the negatively charged DNA molecule. This charge neutralization is accompanied by a supercoiling of the DNA helix, and is probably the mechanism by which DNA in mature sperm is kept tightly packed for delivery during fertilization. At fertilization, the mature sperm nucleus is incapable of RNA or DNA synthesis.

**DNA polymerases and ligases.** Replication of the genetic material requires an unerring synthesis of DNA molecules that are identical in nucleotide sequence to the parental DNA molecules. Enzymes which carry out this synthesis are called DNA polymerases. They have an absolute requirement for a DNA template and take instructions from that template as they move along it, adding nucleotides to the newly synthesized DNA chain. The synthesis of DNA from low-molecular-weight precursors (the deoxyribonucleoside triphosphates) proceeds rapidly at a rate approaching 1000 nucleotides per minute per molecule of enzyme. Some DNA polymerases have been isolated in pure form, but the initiation and accuracy of DNA replication involves the participation of many proteins in addition to the polymerase itself, and sometimes requires the presence of an RNA “primer” to initiate the process.

Biochemical studies of DNA replication in microorganisms and in animal cells in culture show that the earliest recognizable products of the DNA polymerase reaction are small fragments that are later joined to form the long DNA sequences found in the finished chromosome. The joining of these small DNA pieces involves the participation of other enzymes called DNA ligases. The ligases, by using adenosine triphosphate (ATP) as an energy source, join the smaller molecules by the formation of phosphodiester bonds to produce an uninterrupted polynucleotide chain.

DNA ligases also play an important role in DNA repair mechanisms. For example, if the genetic material is damaged by ultraviolet irradiation, the damaged areas can be excised. New nucleotides are then inserted by a DNA polymerase, and the repaired area is sealed to the rest of the DNA molecule by a DNA ligase.

**DNA unwinding proteins.** Bacteria, viruses, and animal cells contain proteins which destabilize the DNA double helix (helicases) or produce breaks in the DNA for coiling or untangling (topoisomerases).

**Helicases.** Helicases are a group of proteins that associate with both DNA and RNA and function to separate base pairing of the nucleic acids. These enzymes use the energy from ATP hydrolysis to separate the strands of DNA for a variety of functions, including

DNA replication, recombination, and transcription. Helicases often have multiple protein subunits that assemble to form a ring structure that encircles a single nucleic acid strand and then follows that strand, unwinding any secondary conformations it encounters.

**Topoisomerases.** DNA topoisomerases have been found in organisms from bacteria to humans and play critical roles in DNA replication, transcription, and repair. Topoisomerases function by generating single- or double-strand DNA breaks and then rotating or passing DNA through the break. Classifications of these proteins are based upon their product. Enzymes that make a break in one strand of DNA are type I topoisomerases; those that produce a break in both strands are called type II. The cleavage step leads to a phosphodiester bond between the topoisomerase protein and the DNA, allowing the protein to maintain contact with the broken strands during manipulation of the DNA. These proteins have gained importance as they are often targets of antibiotics or anticancer therapies.

**DNA repair.** Given the importance of DNA as the hereditary molecule, maintaining the correct DNA sequence is vital to a cell. However, insults to the hereditary material can occur from a variety of common sources, such as mistakes in DNA synthesis, exposure to ultraviolet light, and damage from harsh chemicals. Unfortunately, it is often impossible merely to fix the mistake where it is found; so the cell must find the error, degrade the DNA around the error, and then resynthesize that portion of DNA using the other DNA strand or even the homologous chromosome as a template. Many repair processes involve ribonucleoprotein complexes such as endonucleases, exonucleases, ligases, helicases, and polymerases.

At least five mechanisms exist to repair DNA based on the type of damage. The first mechanism is direct repair, involving proteins to directly modify the bases and revert them to original form. However, this is often an inefficient and ineffective method of repair. The second and third methods are base excision repair and nucleotide repair. Both processes involve the removal of the damaged bases and a few surrounding nucleotides from the DNA strand, followed by the resynthesis of the missing section with subsequent ligation of the new and old fragments. The difference between the two repair mechanisms lies in whether the sugar backbone is cut during or after the excision of the base. Defects in base excision repair severely limit the ability of cells to respond to thymine dimers (often caused by exposure to ultraviolet light) leading to the disease xeroderma pigmentosa, which is characterized by thinning of the skin, splotchy pigmentation, spider veins, and skin cancers.

The fourth mechanism is postreplication repair, which is used to correct double-strand breaks. This process often involves use of the homologous chromosome as a template for the synthesis and ligation of the two DNA ends, thus ensuring that no DNA bases were lost in the break and repair. This is very

similar to the type of homologous chromosome usage in meiotic recombination.

The fifth type of repair is known as mismatch repair. This process is often used during the replication of DNA during the cell cycle if DNA polymerases accidentally insert the incorrect base. The incorrect base is removed and the correct base is then inserted. Without this proofreading and repair mechanism, the error rate during replication can be 1000 times higher than normal, leading to high rates of mutagenesis.

**Nucleases and DNA-modifying enzymes.** Many of the enzymes that take part in the hydrolysis of DNA are known. Such proteins are called deoxyribonucleases and are classified in terms of their mode of action and source: endonucleases attack the internal structure of the polynucleotide chain, while exonucleases cleave the terminal nucleotides from the nucleic acid molecule. Of particular use are bacterial restriction enzymes, which cleave specific sequences in DNA molecules, enabling many techniques in molecular genetics. A battery of such enzymes is now available which provides DNA fragments suitable for DNA sequence analysis and recombinant DNA technology.

In a cell, endogenous deoxyribonucleases are used to recognize and degrade foreign DNA, among other functions. To determine what DNA is foreign to the cell, enzymes interact with endogenous DNA and modify the structures of the purine and pyrimidine bases, usually by the introduction of methyl groups at precise positions in the DNA molecule. Such structural modifications of DNA are a means of imparting species specificity among DNA molecules, as different sequences are methylated among different organisms. This allows for species-specific DNA endonucleases to degrade DNA foreign to the host. Enzymes that methylate DNA limit their activities on homologous DNA but may overmethylate (and cleave) DNA from another species.

In eukaryotes, DNA methylation marks regions of different transcriptional activity, as heavily methylated genes are transcriptionally silent. This is often thought of as a host-defense mechanism to guard against the spread of transposons (segments of DNA that can replicate and insert themselves elsewhere in the genome). A significant fraction of the genome within higher organisms is made of transposable elements that are methylated. Mutations that lead to demethylation of these elements cause genomic instability.

**Proteins involved in gene regulation.** Different cell types within higher organisms have equivalent DNA contents, but they utilize different portions of the DNA to direct their specialized functions. It follows that nuclear protein mechanisms exist for the selective utilization of DNA sequences. Because of the limited number of histone types and the relative uniformity in histone composition in different cells, it is likely that they are only part of the regulatory network of gene transcription.

Human (cultured cell) nuclei, for example, contain over 470 different protein classes that can be separated according to their molecular size and charge, with many capable of DNA binding. Others are en-

zymes involved in nucleic acid synthesis and repair, enzymes which modify histone structure, proteins involved in the processing and transport of mRNAs from the nucleus to the cytoplasm, and proteins destined to become components of ribosomes. The DNA-binding proteins are of primary interest for control mechanisms. Studies of RNA synthesis in isolated chromatin fractions have shown that the rate of RNA transcription depends on the source of the nonhistone nuclear proteins added to the DNA template. These proteins can have roles involving transcription initiation, elongation, and termination, or they can modify other proteins or mRNA by controlling stability, degradation, or interactions.

**Nuclear protein modifications and genetic control.** Many of the histone modifications, such as phosphorylation and methylation, are also found on other proteins. Additional modifications, such as glycosylation (the addition of sugars), ubiquitination (addition of ubiquitin, which targets a protein for degradation), and sumolation [addition of a small ubiquitin-related modifier (SUMO)] can also affect protein half-life and function. Effects of individual modifications are specific to the protein and amino acid residue, but the general idea is that these modifications add another level of complexity to the interaction of proteins and nucleic acids. For example, many of the chromosomal nonhistone proteins are subject to phosphorylation reactions that modify serine and threonine residues in the polypeptide chains. The result is an alteration of protein charge and structure which affects the interactions of regulatory proteins with each other and with the DNA template. The phosphorylation of nuclear proteins generally correlates in a positive way with increases in RNA synthesis when cells respond to hormones, growth factors, or other stimuli. Furthermore, the enzymatic dephosphorylation of the nuclear nonhistone proteins diminishes their capacity to stimulate RNA synthesis (transcription) in isolated chromatin fractions.

**RNA polymerases.** Among the most important of the proteins associated with DNA are the enzymes responsible for RNA synthesis. In yeast and higher organisms, there appear to be at least three different RNA polymerase activities in the nucleus. All are DNA-dependent, and their products are complementary to the DNA sequences copied by the enzyme. RNA polymerase I functions in the nucleolus, where it copies the ribosomal genes to produce rRNAs. A different polymerase, RNA polymerase II, transcribes mRNA sequences. RNA polymerase III directs the synthesis of small RNA molecules such as the amino acyl tRNAs used for protein synthesis. All three polymerases are complex enzymes that can be discriminated from one another by interactions with inhibitors. For example, polymerase II, but not polymerase I, is sensitive to inhibition by the mushroom poison  $\alpha$ -amanitin.

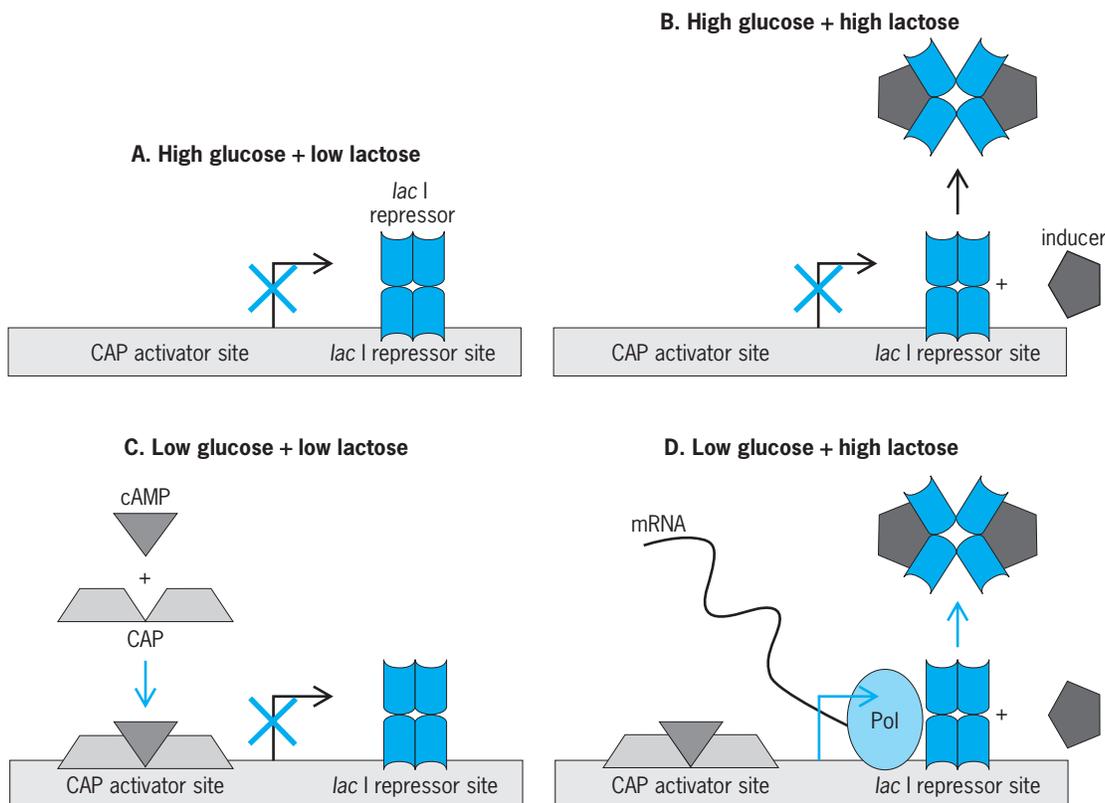
All of the RNA polymerases described above are DNA-dependent enzymes. However, other polymerases are known which copy RNA templates called RNA-dependent RNA polymerases. These polymerases are often used to replicate viruses that have an RNA genome.

There are at least three ways to control the ability of RNA polymerases to function, and the cellular machinery makes use of all of them. RNA transcription initiation, elongation, and termination each have specific machinery designed to add check-points that can modify transcriptional output. Such fine control allows the levels of RNA transcription to match the cellular needs in a small time frame.

*Control of RNA transcription initiation.* Mechanisms of genetic control in higher organisms remain an area of intense research interest. However, relevant models can be derived from the detailed analyses of genetic control elements in bacterial cells and viruses. The molecular basis of gene control is known in remarkable detail for the lactose (*lac*) operon in the bacterium *Escherichia coli* (Fig. 3). This segment of the bacterial chromosome directs the synthesis of three enzymes ( $\beta$ -galactosidase,  $\beta$ -galactoside permease, and thiogalactoside transacetylase) needed for lactose utilization. The genes for these three proteins are linked together on the bacterial chromosome and are directed by three major control elements located before the first gene in the sequence: (1) a repressor gene (the *i* gene), (2) a promoter sequence, and (3) an operator sequence. Transcription of the *lac*

operon is both negatively and positively controlled. Negative control is exerted by the *lac I* repressor, a protein product of the *i* gene which binds specifically to the operator region and thus prevents transcription (Fig. 3). Another region of the repressor molecule interacts with the low-molecular-weight substrate allolactose, which acts as an inducer of the *lac* operon. This interaction brings about a conformational change (an allosteric modification) in the repressor, which leads to its release from the operator DNA sequence. This allows RNA polymerase to transcribe the adjacent genes, whose products are necessary for galactoside metabolism.

An important point is that the removal of the repressor is not, in itself, sufficient to initiate transcription. Transcription also requires the participation of a positive control protein called the catabolite gene activator protein or the cyclic adenosine monophosphate (AMP) receptor protein (CAP protein). The CAP protein functions as a dimer that attaches to specific DNA sequences in the promoter region of the *lac* operon and activates transcription only when bound to the signaling molecule cAMP, which becomes available when glucose is not present. A significant clue to the specificity of interactions between



**Fig. 3.** Transcriptional regulation of the *lac* operon of *E. coli*. Transcription of the *lac* operon is both negatively and positively controlled. Negative control is exerted by the *lac I* repressor, which binds to a repressor site and prevents mRNA transcription unless an inducer is produced in the presence of high lactose. Positive control is exerted by cyclic adenosine monophosphate (cAMP) and the cAMP receptor protein (CAP protein). cAMP is produced when there is little glucose, allowing CAP to bind to the activator site and recruit RNA polymerase to initiate transcription. (a) Under conditions of high glucose and low lactose, the *lac I* repressor is bound, preventing transcription. High levels of glucose results in no production of cAMP and, therefore, CAP is not bound at the activator site. (b) Under conditions of high glucose and high lactose, the inducer binds the *lac I* repressor causing the repressor to dissociate from the DNA. High levels of glucose result in no production of cAMP and, therefore, no transcription initiation by CAP. (c) Under conditions of low glucose and low lactose CAP promotes transcription initiation, but RNA polymerase is blocked by bound *lac I* repressor. (d) Under conditions of low glucose and high lactose, CAP is bound and recruits RNA polymerase, and the removal of the *lac I* repressor by the inducer allows for transcription.

such regulatory proteins and DNA is their capacity to recognize symmetrical elements in DNA structure. Binding of the CAP protein involves a region of twofold rotational symmetry. The interaction of the CAP protein with DNA is believed to facilitate the entry of RNA polymerase, allowing transcription of the genes required for lactose utilization.

*Control of RNA transcription elongation.* In organisms ranging from yeast to humans, RNA polymerase II has a C-terminal domain (CTD) that is important for the processivity of the polymerase. Upon initiation, the CTD is hypophosphorylated, but in order for elongation of the transcript to occur the CTD must be phosphorylated. If RNA polymerase II is not phosphorylated, it will not yield full-length transcripts, providing another means of regulation. Thus for very long RNA transcripts, such as viral RNA genomes, it is necessary for the virus to make sure that the entire genome is replicated.

A well-studied example of this regulatory mechanism is the transcriptional control of the Tat-TAR system for elongation in human immunodeficiency virus 1 (HIV-1). Tat is a small protein that contains a basic region consisting of nine positively charged amino acids in close proximity to one another. This region has three important functions: (1) it acts as a nuclear localization signal, (2) it binds to receptors, and (3) it interacts with the transactivation-response region (TAR) in the HIV-1 RNA transcript and enhances RNA elongation. The TAR RNA has a small stem-loop in its structure that forms a docking site for Tat. Once docked, Tat recruits a Tat-associated kinase (TAK, such as P-TEFb) that can phosphorylate the CTD, allowing the full transcription of the HIV-1 genomic RNA.

*Control of RNA transcription termination.* RNA synthesis ends at precise regions of the chromosome, producing mRNAs of precise lengths. Termination of RNA transcription can be controlled in at least two ways: (1) intrinsic terminators can stop RNA polymerase without other proteins or (2) DNA-binding proteins can stop the polymerase. Both systems have been characterized in *E. coli*.

Intrinsic terminators are hairpin loops in the RNA which cause RNA polymerase to slow, and a series of uracil bases after the hairpin that disrupts the DNA-RNA base pairing, causing polymerase release. Terminator proteins act by unwinding the DNA-RNA duplex produced during transcription. In bacteria, one such protein, called rho factor, has six subunits that interact to form a planar annulus around an empty core. There is a built-in gap between the subunit particles, which permits the complex to encircle a nucleic acid strand and subsequently follow the strand to find the transcribing RNA polymerase. Termination itself can be affected by antitermination proteins, such as the *N* gene product of bacteriophage lambda, which can override or attenuate the termination signal.

**Reverse transcriptase.** A number of RNA viruses replicate by a mechanism that involves a DNA-containing replicative form. These viruses contain or produce an enzyme, reverse transcriptase, which

copies RNA sequences to form a complementary DNA (cDNA) strand. This enzyme has many important applications in molecular biology. For example, it can be used to make a DNA copy of mRNA; the copy can be made highly radioactive and used for hybridization studies to locate the corresponding genes in chromosomes, or it can be used to measure how much of the corresponding mRNA is synthesized in a cell. Moreover, the cDNA copy of a message can be incorporated into bacterial plasmids as recombinant DNA, thus permitting amplification of the DNA sequences which code for the mRNA. See REVERSE TRANSCRIPTASE.

### Ribonucleoproteins

As in the case of the DNA-protein complexes, ribonucleoproteins can be classified according to their functions and localization within the cell. Ribonucleoproteins are present in all cells and constitute the genetic material of many viruses; poliovirus, for example, is a ribonucleoprotein.

**mRNA.** The mRNAs themselves occur in association with proteins. When mRNA is synthesized in the nucleus, it is quickly associated with sets of proteins that organize the nascent RNA chains into a more particulate, and presumably protected, structure. These mRNA-binding proteins add a "cap" to the 5' end of the mRNA to help stabilize the molecule and prevent degradation of the transcript. Another enzyme adds adenylic acid residues to the 3' end of the molecule for similar reasons. The poly-A sequence at the 3' end of the mRNA is specifically associated with poly-A-binding proteins. See NUCLEIC ACID.

The shielding of mRNAs by associated proteins appears to be one basis for the stability of messengers in certain cell types, such as oocytes, in which the mRNAs are nonfunctional but remain intact for long periods. Removal of the masking proteins requires proteolysis; this occurs at the time the egg is fertilized, and the unmasked mRNAs become functional in the synthesis of embryonic proteins.

Many RNA-degrading enzymes are known; some function by cleaving the polynucleotide chain at internal linkages, while others remove terminal nucleotides in a stepwise fashion. Certain ribonuclease activities are localized in the cell nucleus, where they degrade portions of the high-molecular-weight precursors of mRNAs. Other ribonucleases are found in the cytoplasm and play a role in mRNA "turnover." Still others are secreted from the cell. One of the best-characterized ribonucleases is ribonuclease A of the pancreas. It contains 124 amino acids and was the first enzyme to be completely sequenced. It is also the first enzyme to be synthesized in the laboratory, by using amino acids and a technique of solid-phase peptide synthesis. See RIBONUCLEASE.

**mRNA splicing.** Splicing is the process of cleaving parts of the mRNA that are not needed for the mature mRNA to be translated. To accomplish this goal, a large complex (approximately 12 MDa) known as the spliceosome, composed of both noncoding RNA and proteins, assists in the removal of introns by an energy-dependent process. Of specific interest

are the small nuclear ribonucleoprotein (snRNP) complexes which contribute almost half of this mass. These five complexes (U1, U2, U4, U5, and U6) are involved in building the lattice for the spliceosome components, recruiting other complexes, base pairing with the mRNA, and directly acting in the removal of introns.

Which introns and exons are included in a mature mRNA is not always fixed. Alternative splicing, a method by which introns and exons from an immature mRNA can be incorporated or excluded in the mature mRNA, can occur. This leads to different mRNA transcripts from the same DNA sequence and allows much more mRNA diversity than that encoded by the DNA. For example, alternative splicing in the *Drosophila melanogaster* gene *Dscam* can potentially generate more than 38,000 different mRNA transcripts from the same gene. These transcripts, based on which introns and exons are used, can then encode proteins with very different and specialized localization properties and functions.

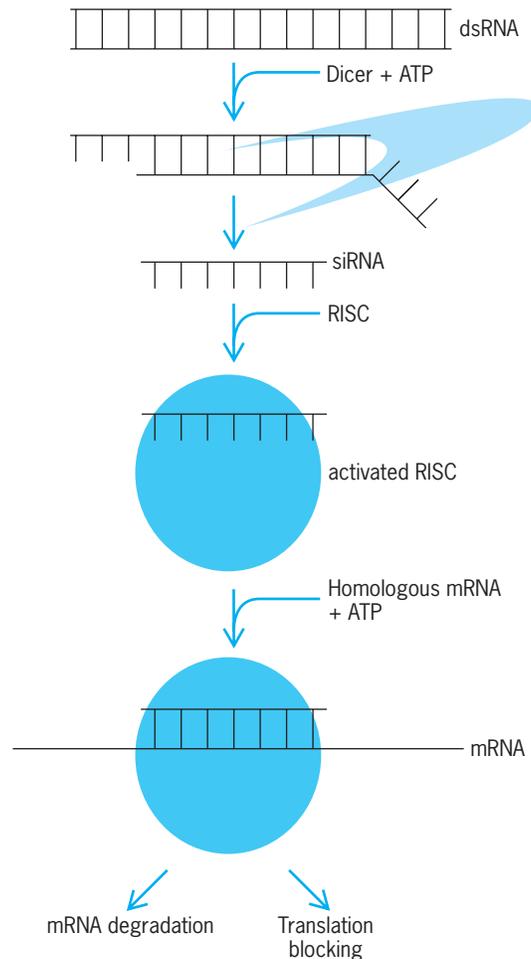
**Ribosome and tRNA.** The role of mRNA is to specify the amino acid sequences of proteins, but translation of the code from RNA to amino acids is conducted by ribonucleoprotein particles called ribosomes. Each ribosome is made up of two subunits which travel along the mRNA chain, taking instructions on how to link the amino acids in the proper sequence. The first step in protein synthesis involves enzymes known as amino acyl transfer RNA (tRNA) synthetases. The role of amino acyl tRNAs is to guide particular amino acids to their proper position in the ribosome, a complex of proteins and RNAs that reads mRNA and translates it into proteins. This positioning is achieved by interactions between nucleotide triplets (anticodons) in the tRNA and complementary triplets (codons) in the mRNA. There are many such enzymes; their job is to activate the amino acid and to attach it to an appropriate tRNA. Each activating enzyme has high specificity both for the amino acid and for the tRNA. More than one tRNA may exist for a particular amino acid, and the specificity of the amino acyl tRNA synthetases must be very high to avoid errors. This is an impressive example of the ability of a protein to recognize specific elements of sequence and structure in an RNA molecule.

The organization of the ribosomes involves many RNA-binding proteins. The large and small ribosomal subunits contain different RNA molecules and different sets of associated proteins. For example, in rabbit reticulocyte ribosomes the small subunit (40S) contains 32 proteins with molecular weights ranging from 8000 to 39,000, while the large subunit contains 39 proteins with molecular weights ranging from 9000 to 58,000. Each ribosomal subunit appears to contain just one copy of each distinct ribosomal protein. The functions of the ribosomal proteins are under intensive investigation. Some appear to be involved in interactions between the ribosomes and amino acyl tRNAs, with factors necessary for protein chain initiation and elongation, in binding the mRNA initiation complex, and in the vectorial transport and release of the nascent polypeptide chains.

**Noncoding RNAs.** Not all RNAs fall neatly into the classes of mRNA, tRNA, or rRNA. Similar to rRNAs or tRNAs, which are not translated into proteins, other noncoding RNAs (ncRNAs) play important roles in biological processes. In fact, 50–75% of all transcription in higher eukaryotes consists of ncRNA. These RNAs function as structural components, catalytic molecules, or regulators in the cell.

*snoRNA.* In a similar class to snRNAs (discussed as splicing components), snoRNAs complex with proteins. They are named for their subcellular localization within the nucleolus. However, while snRNAs are involved in mRNA splicing, snoRNAs help rRNA maturation and are essential for proper rRNA cleavage and modification.

*Xist/Tsix antisense ncRNAs.* ncRNAs are involved in the inactivation of specific genes or even whole chromosomes. An example is the role of ncRNAs in the dosage compensation (X inactivation) mechanism in mammals. X inactivation involves transcriptional silencing of one of the two X chromosomes in



**Fig. 4.** Model of the RNA interference (RNAi) pathway. Long double-stranded RNA (dsRNA), consisting of sense and antisense strands, is cleaved by Dicer (crescent) in an energy (ATP)-dependent manner into short inhibitory RNAs (siRNA). The antisense siRNA is then incorporated into the RNA-induced silencing complex (RISC, circle), activating it and triggering a search for homologous mRNAs. When a match is found, RISC silences the mRNA by either degradation or translation blocking.

female mammals. Inactivation occurs early in development. Through a complex mechanism involving X chromosome counting and epigenetic markings, the X-inactivation center produces the ncRNA *Xist* on the future inactive X chromosome, and the ncRNA *Tsix* on the future active X chromosome (both *Xist* and *Tsix* are polyadenylated and capped ncRNAs). These two ncRNAs are derived from the same DNA region, but are transcribed from opposite DNA strands. Thus, they are antisense RNAs to each other. The *Xist* RNA coats the inactive X chromosome from which it is transcribed and initiates transcriptional silencing of the majority of genes on the inactive X chromosome. The silencing process also involves the recruitment of histone variants and histone modification proteins.

**dsRNA and RNAi.** RNA interference (RNAi) is a mechanism whereby the production of double-stranded RNA (dsRNA) renders homologous mRNA incapable of being translated. The dsRNA is cleaved into short fragments and then unwound so that the antisense RNA can pair with the homologous mRNA. The resulting duplexed regions signal degradation or physically block translation by ribosomes. While the mechanism of RNAi is not fully understood, some components of the pathway have been elucidated (Fig. 4).

Dicer is a family of ATP-dependent RNase III enzymes that cut long dsRNA sequences into small (approximately 20 bases) RNA sequences. Dicer functions in two pathways of RNAi, that of short inhibitory RNAs (siRNA) that lead to mRNA degradation, and microRNAs (miRNA) that inhibit protein translation by binding to complementary sequences in the mRNA. The RNA-induced silencing complex (RISC) is involved in both pathways and mediates the pairing of the antisense siRNA or miRNA to the homologous mRNA. This is an evolutionarily conserved mechanism that is used endogenously by cells to regulate protein production.

Timothy J. Moss, Vincent G. Allfrey, Lori L. Wallrath  
Bibliography. A. M. Denli and G. J. Hannon, RNAi: An ever growing puzzle, *TRENDS Biochem. Sci.*, 28: 196–201, 2003; G. Eckstein and D. M. Lilley, *Nucleic Acids and Molecular Biology*, 1994; B. Lewin, *Genes VIII*, Prentice Hall, 2004; J. A. McCammon and S. C. Harvey, *Dynamics of Proteins and Nucleic Acids*, 1988; D. L. Nelson and M. M. Cox, *Lehninger Principles of Biochemistry*, W. H. Freeman, 2004; H. Tschesche (ed.), *Modern Methods in Protein and Nucleic Acid Research*, 1990.

## Nucleosome

The fundamental histone-containing structural subunit of eukaryotic chromosomes. In most eukaryotic organisms, nuclear deoxyribonucleic acid (DNA) is complexed with an approximately equal mass of histone protein. It was previously thought that the histones formed some sort of coating on the DNA, but in 1974 it was recognized that the DNA-histone complex was in the form of discrete subunits, now

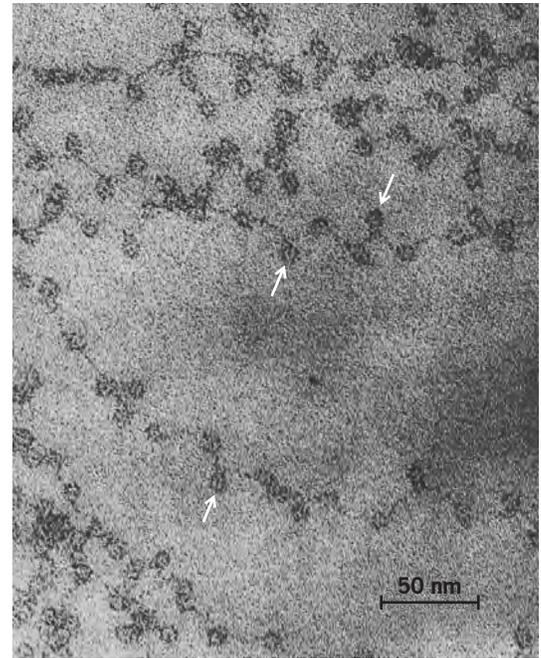


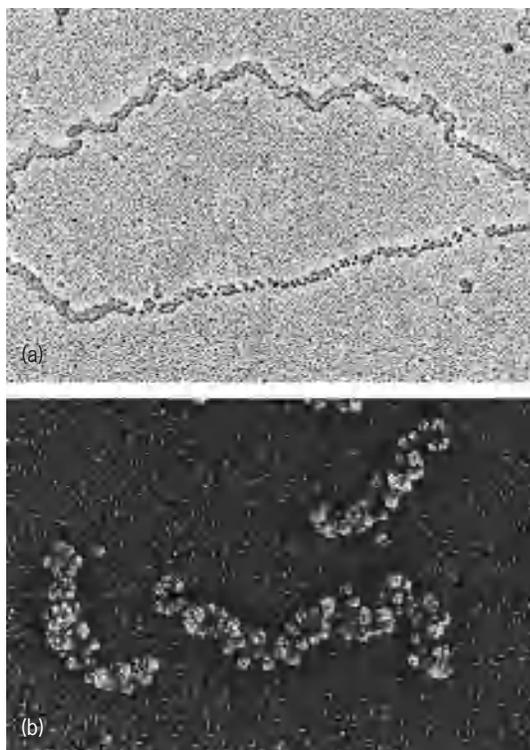
Fig. 1. Electron micrograph of chicken erythrocyte chromatin. The arrows mark individual nucleosomes in ribbonlike zigzag chromatin filaments. (From D. E. Olins and A. L. Olins, *Nucleosomes: The structural quantum in chromosomes*, *Amer. Sci.*, 66:704–711, November–December 1978)

termed nucleosomes. See DEOXYRIBONUCLEIC ACID (DNA).

**Organization.** The nucleosome is organized so that the DNA is exterior and the histones interior. The DNA makes two turns around a core of eight histone molecules, forming a squat cylinder 11 nanometers in diameter and 5.5 nm in height. A short length of linker or spacer DNA connects one nucleosome to the next, forming a nucleosomal chain that has been likened to a beaded string (Fig. 1). This basic structure is found in all forms of chromatin (the DNA-protein complex that forms eukaryotic chromosomes), from the dispersed type in nondividing nuclei to the compact chromosomes visible in mitotic and meiotic nuclei. Nucleosomes have been found in all eukaryotic organisms examined, the only exceptions being some sperm nuclei where protamines replace histones during spermatogenesis, and the dinoflagellate algae, which lack histones as major nuclear components.

A chain of adjacent nucleosomes is approximately sixfold shorter than the DNA it contains. Moreover, chains of nucleosomes can self-assemble into thicker fibers in which the DNA packing ratio approaches 35:1 (Fig. 2). These observations, and the lack of any obvious catalytic activity, led to the assumption that the primary function of the nucleosome consists of organizing and packing DNA. More recently, however, nucleosomes have also been shown to have important gene regulatory functions.

**Nuclease digestion.** Nucleosomes may be prepared for biochemical and biophysical analysis by treating nuclei with certain nuclease enzymes that preferentially attack the linker DNA between nucleosomes.



**Fig. 2.** Chromatin fibers 30 nm in diameter. (a) Fibers from a mouse fibroblast nucleus; the upper fiber shows the typical compact structure, while the bottom fiber has been stretched, revealing individual nucleosomes, some in a zigzag arrangement. (b) Isolated fibers from chicken erythrocyte nuclei; individual nucleosomes are visible in the fibers. (From C. L. F. Woodcock, L.-L. Y. Frado, and J. B. Rattner, *The higher order structure of chromatin: Evidence for a helical ribbon arrangement*, *J. Cell Biol.*, 99:42–52, 1984)

Careful analysis of the digestion products has revealed several important features of nucleosome arrangements. First, the mean spacing between nucleosomes differs between species, and often between tissues of the same organism. The shortest nucleosome repeat length (NRL; average DNA length from one nucleosome to the next) is  $\sim 166$  base pairs for yeast and rat cortical neurons, while the longest is  $\sim 246$  base pairs for sea urchin sperm. Most nuclei examined, however, have spacings in the 180–200 base-pair range. Within one type of nucleus, the spacing is not constant, but varies on either side of the mean value. Even at the level of the DNA base sequence there appears to be no consistently uniform placing of nucleosomes, although there are cases where nucleosomes or groups of nucleosomes are located on specific DNA sequences. In some instances, such “phased” nucleosomes may block access to a DNA regulatory element, and mechanisms for uncovering such elements under appropriate conditions have been demonstrated.

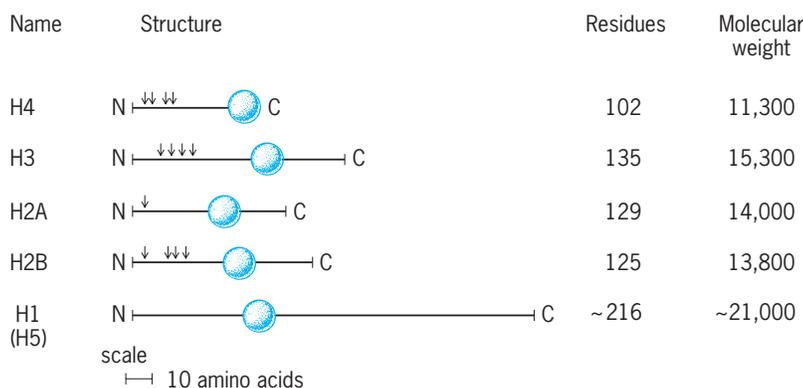
As digestion by micrococcal nuclease (the most commonly used enzyme for preparing nucleosomes) proceeds, first the linker DNA is hydrolyzed, then a brief “pause” in digestion occurs when the mean nucleosomal DNA is  $\sim 166$  base pairs in length. These 166 base pairs complete two turns around the his-

tone core; the particle so produced has been termed a chromatosome. Further digestion reduces the DNA length to  $\sim 145$  base pairs, concomitant with the release of one histone molecule (the H1 class of “linker” histones). The product at this point of digestion is the core nucleosome or core particle, and it is upon this product, with its 1.75 turns of DNA, that most structural studies have been performed. The DNA of the nucleosome core is the most resistant to micrococcal nuclease attack, but eventually it too is cleaved.

Some nucleases do not show preferential affinity for the linker DNA. Deoxyribonuclease I (DNase I) is an example of this type: it produces staggered double-stranded cuts at approximately 10-base-pair intervals within nucleosomal DNA, and it has a preferential affinity for some intranucleosomal sites.

**Histones.** Histones are small basic proteins with common structural features. All have an extended, positively charged amino-terminal region, a globular central domain, and a positively charged carboxy-terminal region, the charges arising mainly from lysine and arginine residues (Fig. 3). Four histone species (H2A, H2B, H3, and H4) are present as two copies each in the nucleosome core, while H1 is present as a single copy in the chromatosome. Histone H1 is thought to reside close to the entry-exit point of the DNA, and to “seal” the two turns of DNA. The core histones may self-associate under appropriate conditions to form a tetramer of H3 and H4 and a dimer of H2A and H2B; further association of one (H3 H4)<sub>2</sub> tetramer with two (H2A H2B) dimers forms the octamer (histone core), which is stable in NaCl concentrations above about 1.0 M. See AMINO ACIDS; PROTEIN.

The core histones are among the most highly conserved proteins known, H4 being the extreme example with only two amino acid changes between pea and calf. However, there are multiple-core histone genes, each coding for a variant protein showing (usually) minor differences in amino acid sequence; and in some cases, crucial roles for these variants have been discovered. For example, a variant of



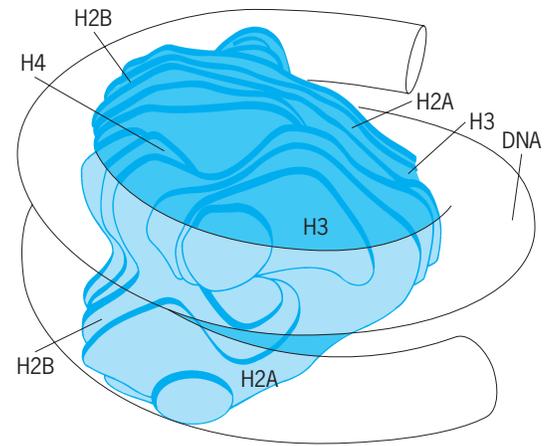
**Fig. 3.** Several characteristic properties of the five major classes of histones. The globular regions of the histones (spheres) correspond to 70–80 amino acid residues and pack into approximate spheres of 2.5 nm diameter. N and C are the amino and carboxyl termini of the peptide chains. Arrows indicate potential sites of acetylation of histone lysine residues. (After D. E. Olins and A. L. Olins, *Nucleosomes: The structural quantum in chromosomes*, *Amer. Sci.*, 66:704–711, November–December 1978)

histone H3 is a required component of centromere nucleosomes, and conserved from yeast to humans. The differential expression of histone variants has also been associated with specific stages in embryonic development.

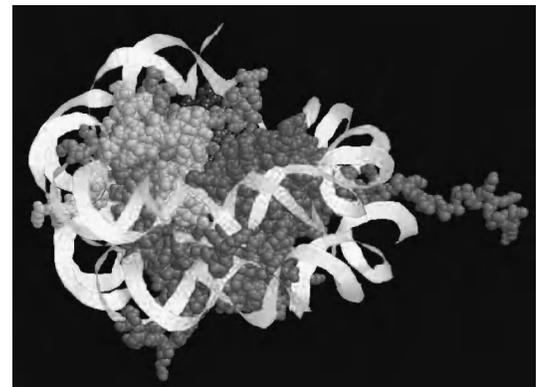
**Histone modifications and the “histone code.”** Another form of histone diversity is amino acid modification through the addition of phosphate, acetate, methyl, and poly (adenosine diphosphate-ribose) groups to specific residues. These modifications are the subject of intensive study, and findings have led to the concept of a “histone code,” which determines the transcriptional availability (as well as other properties) of nucleosomes and groups of nucleosomes. For example, the acetylation of lysines (especially of histone H4) reduces the net positive charge, and is associated with “looser” chromatin and a greater availability to the transcriptional machinery, while methylation of lysines has the opposite effect. Some modifications may recruit enzyme complexes that, for example, include histone acetyl transferases or histone deacetylases. Histone modifications or groups of modifications may be passed to daughter cells through cell division, providing an epigenetic mechanism for the inheritance of transcriptional states (that is, changes that do not affect the genes themselves but alter their expression).

**Forces stabilizing the nucleosome.** If nucleosomes are placed in a high-salt solution (for example, 1.0–2.0 M NaCl), the histone octamer dissociates from the DNA, as would be expected for a complex in which electrostatic interaction between the negatively charged phosphate groups on the DNA and the positive charges on the lysine and arginine residues in the histones predominated. The finding that the highly charged amino-termini project from the histone octamer led to the suggestion that these domains contributed most of the electrostatic “glue” that stabilized nucleosomes. However, experiments showing that nucleosomelike particles could be formed by using octamers that lacked the amino-terminal regions showed clearly that this was not the case, and it now seems likely that these histone domains function in internucleosomal interactions and transcriptional regulation. While the octamer is stable in high salt, it dissociates if the ionic strength is lowered or if it is exposed to urea. Thus, hydrophobic interactions are thought to be largely responsible for maintaining the integrity of the histone core.

**Detailed structure.** The publication in 1997 of a 2.8 Å resolution x-ray structure of the nucleosome constituted a major breakthrough in chromatin research. This technical tour de force was the result of over a decade of work by T. Richmond and colleagues and required producing fully defined nucleosome core particles comprising a palindromic DNA sequence, and bacterially synthesized core histones, together with data collection at liquid nitrogen temperatures using synchrotron radiation. **Figure 4** contrasts the ~20 Å structure and the 2.8 Å structure. The final structure has a twofold axis of symmetry, with the (H3 H4)<sub>2</sub> tetramer forming a U-shaped complex within the two turns of DNA, and the two (H2A



(a)



(b)

**Fig. 4.** Detailed structures of nucleosomes (a) Model of the nucleosome core particle at 20 Å resolution (after R. D. Kornberg and A. Klug, *The nucleosome*, *Sci. Amer.*, 244(2):52–79, 1981). (b) X-ray structure of the nucleosome core particle at 2.8 Å resolution (adapted from MedWeb: Nucleosome structure: <http://medweb.bham.ac.uk/research/chromatin/nucleosome.html>).

H2B) dimers attached to either side of the tetramer. Nucleosomal DNA is not smoothly coiled on the histone octamer, but variously distorted and kinked, and the highly conserved portions of the core histones are folded in a manner that maximizes histone-histone interactions, and also provides a coiled ramp on the outside where positive charges are concentrated. The bulk of the N and C termini of the core histones are not resolved in x-ray structures, suggesting that they are highly mobile even in nucleosome crystals. However, the sites where these histone domains leave the nucleosome is clearly defined. For example, the long N termini of both H3 molecules exit near the linker DNA entry-exit site, suggesting that they may be associated with linker DNA. See X-RAY CRYSTALLOGRAPHY.

**Transcription.** The fate of nucleosomes during the passage of a ribonucleic acid (RNA) polymerase molecule along the DNA has been extensively studied, and appears to be system-dependent. In the case of some ribosomal RNA genes where polymerase molecules are closely packed on the DNA, it is clear that nucleosomes as such cannot exist during

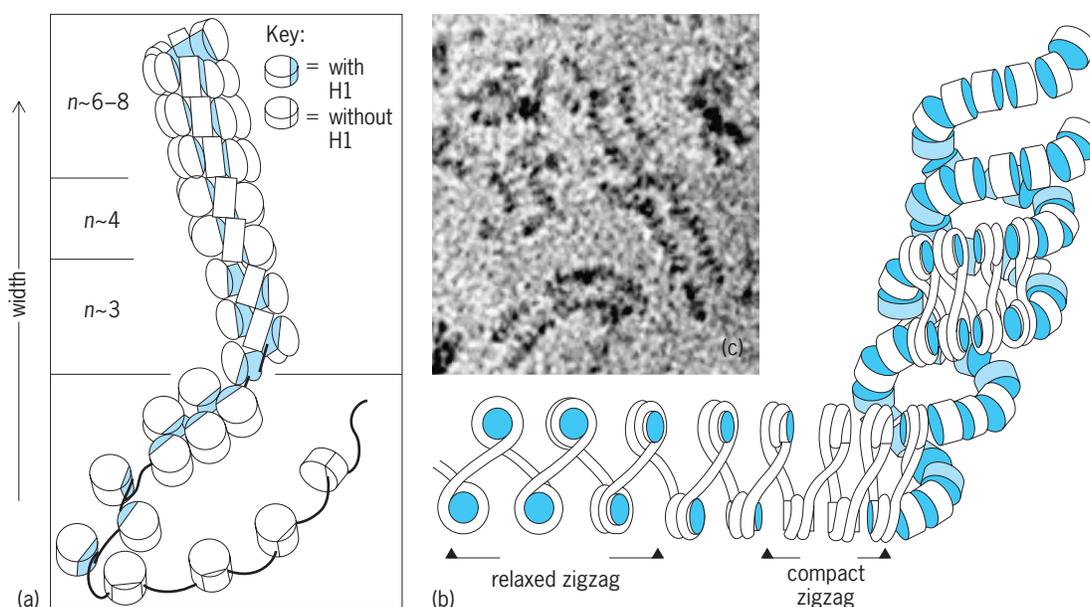
transcription, since electron micrographs show that the genes are extended almost to their full length. Whether the histone octamers are completely dissociated from the DNA during this process or the nucleosomes are merely unfolded is not known. In more typical genes, nucleosomelike particles have been observed between the widely spaced polymerases, suggesting that the perturbations that occur during transcription are of a transitory nature. In some systems, evidence suggests that octamers may be temporarily transferred to “chaperone” complexes during transcription, while other cases indicate a more complex looping out of DNA as the polymerase enters the nucleosome, followed by reassociation with the octamer as the polymerase completes its transit. There is general agreement that actively transcribed genes are depleted in or lack histone H1. See GENE ACTION.

**Nucleosome remodeling.** Nucleosome-remodeling complexes, required for the efficient transcription of some genes, are recently recognized components of the gene regulatory system. These typically contain many proteins (for example, the yeast SWI/SNF complex includes 11 different types of protein and has a mass of over 1 megadalton) including an active ATPase. Remodeling is ATP-dependent and results in nucleosomal DNA becoming more available for transcription. The molecular mechanism of remodeling is not well understood, but is likely to involve a loosening of the DNA-histone interactions, leading to greater mobility and perhaps repositioning of histone octamers.

**Assembly and replication.** Nuclear DNA is replicated semiconservatively; that is, the double helix

splits in two, and new strands are formed on each half in the immediate vicinity of the replication fork. In chromatin, this process is complicated by the presence of nucleosomes, which also must be duplicated along with the DNA. Some principles about the mechanism of chromatin replication are now fairly well established. Unlike the DNA itself, the histone octamers are not replicated semiconservatively. Instead, the old histone octamers remain intact as the replication fork passes along the chromatin strand and are segregated at random to the daughter strands. The mode of formation of new nucleosomes in the wake of the replication fork is complex and still incompletely understood. New nucleosomes, usually bearing modifications that reduce their net charge and complexed with “chaperone” proteins, are formed by the sequential addition of newly synthesized H3 and H4, followed by H2A and H2B, and finally H1. Replication is also accompanied by brief changes in nucleosome spacing: new nucleosomes tend to be closely packed and then gradually change to the mature spacing.

Systems designed for the rapid assembly of nucleosomes (for example, oocytes) contain a large pool of stored histone, and injected DNA is rapidly converted into nucleosomal chromatin. This process can be mimicked in vitro by the incubation of DNA, histones, and oocyte extracts. Effective artificial assembly of nucleosomes can be mediated by factors such as polyglutamic acid, which, like DNA, partially neutralizes the electrostatic charges of the histones. Indeed, by using this polyanion it is possible to produce nucleosomal chromatin from DNA and the four core histones; the subsequent addition of histone H1



**Fig. 5.** Two proposed structures for the arrangement of nucleosomes in the 30-nm fiber. (a) Superhelical or solenoidal (after F. Thoma, T. Koller, and A. Klug, *Involvement of H1 in the organization of the nucleosome and of the salt-dependent superstructures of chromatin*, *J. Cell Biol.*, 83:403–427, 1979). (b) Helical ribbon (after C. L. F. Woodcock, L.-L. Y. Frado, and J. B. Rattner, *The higher order structure of chromatin: Evidence for a helical ribbon arrangement*, *J. Cell Biol.*, 99:42–52, 1984). (c) Micrograph of crosslinked compact chromatin supports a zigzag arrangement. The paired structures are about 30 nm in diameter (reprinted with permission from B. Dorigo et al., *Nucleosome arrays reveal the two-start organization of the chromatin fiber*, *Science*, 306:1571–1573, 2004. © 2004 American Association for the Advancement of Science).

under appropriate conditions then allows the native nucleosomal spacing to be generated.

**Higher-order packing.** The normal state of chromatin in nuclei is not the extended nucleosomal fiber but a much more compact form. Isolated chromatin adopts a fiberlike conformation, about 30 nm in diameter at the appropriate ionic strength (Fig. 2). In the laboratory, the transition from extended to compact fiber can be reversibly induced by altering the ionic composition of the solvent, maximal compaction being observed in about 150 mM monovalent ions, and maximal extension below 5 mM. Above  $\sim 100$  mM monovalent ions, or lower concentrations of divalent or polyvalent cations, chromatin fibers tend to self-associate. Compaction to the 30-nm fiber conformation requires the presence of histone H1, which is also required to establish a basic zigzag architecture (Fig. 1). Despite the ease with which chromatin fibers can be prepared at any desired state of compaction, the way in which the nucleosomes are arranged in the condensed fibers is still controversial due to the difficulty of determining the locations of nucleosomes and linker DNA in compact chromatin fibers.

The most widely discussed proposal for the structure of the 30-nm chromatin fiber is the superhelix or solenoid arrangement, in which the nucleosomal chain is coiled into a continuous shallow helix with 6–8 nucleosomes per turn (Fig. 5a). A second suggestion based on evidence that an intermediate state of compaction is a ribbonlike zigzag structure (for example, regions indicated by arrows in Fig. 1) is that the ribbon arrangement is conserved and simply coils or compresses to form the fiber (Fig. 5b). A major difficulty in studying the compact fiber is its relatively disorganized state, which appears to be an intrinsic property of chromatin with variable linker lengths. Force-extension data from experiments in which isolated fibers are pulled using molecular tweezers support a zigzag over a helical organization. An elegant experiment reported in 2004 analyzed arrays of nucleosomes reconstituted on DNA with regularly spaced nucleosome positioning sequences. The nucleosomes were engineered so that where the arrays were in a compact state, close neighbors would be chemically crosslinked. When viewed in the electron microscope, the crosslinked arrays appeared as ladderlike parallel rows (Fig. 5c). This finding strongly supports the zigzag arrangement, in which alternate nucleosomes are adjacent in compacted chromatin [see region labeled “compact zigzag” in Fig. 5b], over the solenoidal arrangement, in which nearest neighbor nucleosomes along DNA are also nearest neighbors in compact chromatin [see upper region in Fig. 5a]. See CHROMOSOME.

C. L. F. Woodcock

**Bibliography.** K. Luger et al., X-ray structure of the nucleosome core particle at 2.8 Å resolution, *Nature*, 389:251–260, 1997; B. D. Strahl and C. D. Allis, The language of covalent histone modifications. *Nature*, 403:41–43, 2000; A. Wolffe, *Chromatin: Structure and Function*, 3d ed., 1998.

## Nucleosynthesis

Theories of the origin of the elements involve synthesis with charged and neutral elementary particles (neutrons, protons, neutrinos, photons) and other nuclear building blocks of matter, such as alpha particles. The theory of nucleosynthesis comprises a dozen distinct processes, including big bang nucleosynthesis, cosmic-ray spallation in the interstellar medium, and static or explosive burning in various stellar environments (hydrogen-, helium-, carbon-, oxygen-, and silicon-burning, and the *s*-, *r*-, *p*-,  $\gamma$ -, and *v*-processes). Acceptable theories must lead to an understanding of the cosmic abundances observed in the solar system, stars, and the interstellar medium. The curve of these abundances is shown in Fig. 1. Hydrogen and helium constitute about 98% of the total element content by mass and more than 99.8% by number of atoms. There is a rapid decrease with increasing nuclear mass number *A*, although the abundance of iron-group elements like iron and nickel are remarkably large. The processes of nucleosynthesis described in this article attempt to explain the observed pattern. See ELEMENTS, COSMIC ABUNDANCE OF.

Observations of the expanding universe and of the 3-K background radiation indicate that the universe originated in a primordial event known as the big bang about  $15 \times 10^9$  years ago. Absence of stable mass 5 and mass 8 nuclei preclude the possibility of synthesizing the major portion of the nuclei of masses greater than 4 in those first few minutes of the universe, when the density and temperature were sufficiently high to support the necessary nuclear reactions to synthesize elements. See BIG BANG THEORY; COSMIC BACKGROUND RADIATION; COSMOLOGY.

The principal source of energy in stars is certainly nuclear reactions which release energy by fusion of lighter nuclei to form more massive ones. Since the masses of the products are less than those of the constituent parts, the energy released is  $E = mc^2$ , where *m* is the mass difference and *c* is the velocity of light. In 1957 the collaboration of E. M. Burbidge, G. R. Burbidge, W. A. Fowler, and F. Hoyle, and the work of A. G. W. Cameron, laid the foundations for research into the synthesis of the elements, stellar energy generation, and stellar evolution.

**Evidence of nucleosynthesis.** There is considerable evidence that nucleosynthesis has been going on in stars for billions of years. Observations show that the abundance ratio of iron and heavier elements to hydrogen decreases with increasing stellar age. The oldest known stars in the disk of the Milky Way Galaxy exhibit a ratio 10,000 times smaller than in the Sun. This low ratio is understood on the basis of element synthesis occurring in previous-generation stars that evolved to the point of exploding as supernovae, thus enriching the interstellar medium with the nuclei that were synthesized during their lifetimes and by the explosive synthesis that accompanies the ejection of the stellar envelope.

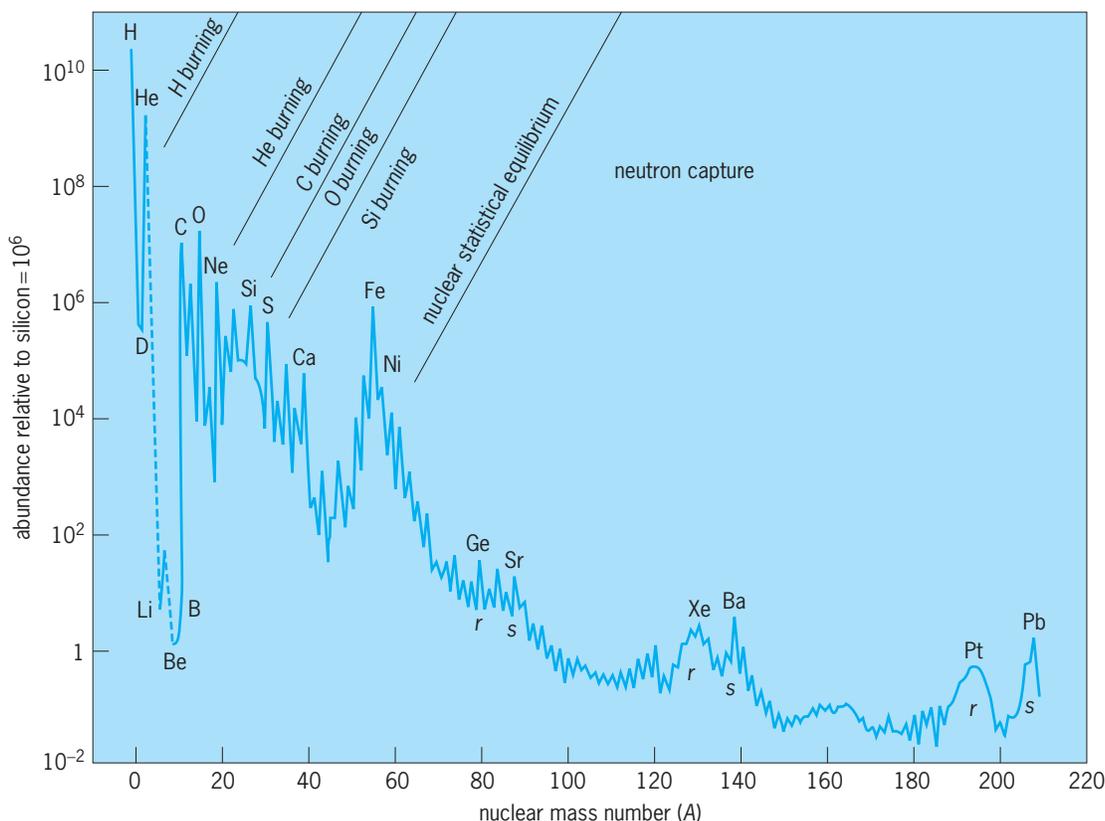


Fig. 1. Schematic diagram of cosmic abundance as a function of nuclear mass number ( $A$ ), based on data of H. E. Suess and H. C. Urey. Predominant nucleosynthetic processes are indicated. (After E. M. Burbidge et al., *Synthesis of elements in stars*, *Rev. Mod. Phys.*, 29:547–650, 1957)

Later-generation stars were then formed from enriched gas and dust. See MILKY WAY GALAXY; STELLAR POPULATION.

The most direct evidence that nucleosynthesis is an ongoing process in stars is provided by the spectra of N-type carbon stars in which convection has dredged the  $s$ -process element technetium to the surface. The technetium is observed through specific absorption lines in the stellar spectrum, and must have been freshly produced, because there is no stable isotope of this element. No technetium is found naturally on Earth, and the stars showing technetium are obviously much older than  $2.6 \times 10^6$  years, the half-life of its most stable isotope. See CARBON STAR; TECHNETIUM.

Another piece of evidence supporting ongoing element synthesis in the Galaxy comes from gamma-ray astronomy. Winds from massive Wolf-Rayet stars, and the ejected envelopes of supernovae, release small amounts of radioactive  $^{26}\text{Al}$  into the interstellar medium. The decay of this isotope results in the emission of gamma-ray line photons with energy  $E_\gamma = 1.809$  MeV. At any given time, the Milky Way is believed to contain about one solar mass of this isotope, distributed throughout the disk. The decay of  $^{26}\text{Al}$  thus leads to a diffuse glow that is confined to the galactic plane. This glow has been detected with gamma-ray telescopes (Fig. 2). See GAMMA-RAY ASTRONOMY; WOLF-RAYET STAR.

A third piece of evidence for ongoing element formation in stars lies in refractory dust grains recovered from meteorites. These micrometer-sized grains (about the size of a human blood cell or smaller) have isotopic patterns that greatly differ from the abundance pattern shown in Fig. 1, which strongly indicates that the grains predate the solar system and in fact condensed in the outflow from stars before the stellar debris fully mixed with the interstellar medium. Certain grains of silicon carbide and graphite have large anomalies in their abundance of  $^{44}\text{Ca}$ . This anomaly resulted from in situ decay of  $^{44}\text{Ti}$  (half-life of about 60 years), which, in turn, means that these grains condensed live  $^{44}\text{Ti}$  no more than several years after that isotope was produced. Similar isotopic anomalies in other presolar meteoritic grains further demonstrate that nucleosynthesis occurs in stars. See COSMOCHEMISTRY; METEORITE.

Finally, certain primitive meteorites show evidence that radioactivities such as  $^{26}\text{Al}$  were alive in the early solar system. This indicates that nucleosynthesis was occurring shortly before the Sun's birth. Indeed, the abundance inferred for these radioactivities is so high that it is likely that they were injected into the forming solar system from a nearby, exploding star.

**Hydrogen burning.** The first static burning process of nucleosynthesis converts hydrogen to helium. In stars of 1.2 or less solar masses, this process

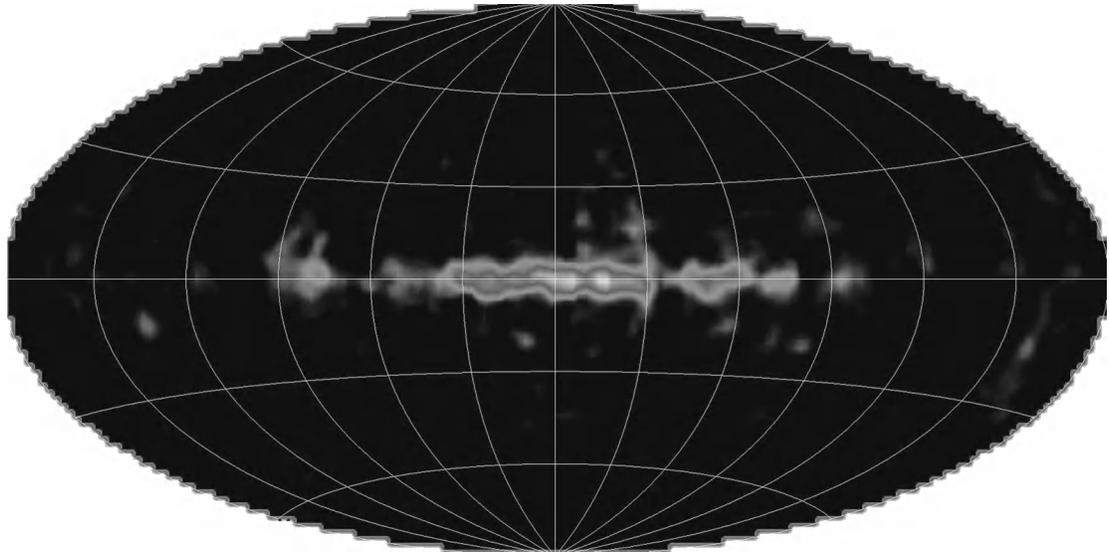


Fig. 2. Map of the Milky Way Galaxy, showing intensity of 1.809-MeV gamma-ray line emission from the decay of the radioactive isotope  $^{26}\text{Al}$ , predominantly made in massive stars. Detection of this emission by gamma-ray telescopes is evidence of ongoing element synthesis in the Galaxy. (COMPTEL/MPE map)

occurs via the proton-proton chain. This chain was also responsible for much of the element synthesis during the big bang, producing the bulk of deuterium and helium, and some of the  $^7\text{Li}$ , observed today. In more massive stars where the central temperatures exceed  $2 \times 10^7$  K, hydrogen burning is accomplished through proton captures by carbon, nitrogen, and oxygen nuclei, in the carbon-nitrogen-oxygen (CNO) cycles, to form  $^4\text{He}$ . The product (ash) of hydrogen burning is helium, but much of the helium produced is consumed in later stages of stellar evolution or is locked up forever in stars of lower mass that never reach the temperatures required to ignite helium burning. The observed abundances of some carbon, nitrogen, oxygen, and fluorine nuclei are attributed to hydrogen burning in the CNO cycles. See CARBON-NITROGEN-OXYGEN CYCLES; NUCLEAR FUSION; PROTON-PROTON CHAIN.

**Helium burning.** When the hydrogen fuel is exhausted in the central region of the star, the core contracts and its temperature and density increase. Helium, the ash of hydrogen burning, cannot be burned immediately due to the larger nuclear charge of helium ( $Z = 2$ ) producing a much higher Coulomb barrier against fusion. When the temperature eventually exceeds about  $10^8$  K, helium becomes the fuel for further energy generation and nucleosynthesis. The basic reaction in this thermonuclear phase is the triple-alpha process in which three  $^4\text{He}$  nuclei (three alpha particles) fuse to form  $^{12}\text{C}$ , a carbon nucleus of mass 12 (atomic mass units). Capture of an alpha particle by  $^{12}\text{C}$  then forms  $^{16}\text{O}$ , symbolically written as  $^{12}\text{C} + ^4\text{He} \rightarrow ^{16}\text{O} + \gamma$ , or simply  $^{12}\text{C}(\alpha, \gamma)^{16}\text{O}$ , where  $\gamma$  represents energy released in the form of electromagnetic radiation. Other reactions that are included in helium burning are  $^{16}\text{O}(\alpha, \gamma)^{20}\text{Ne}$ ,  $^{20}\text{Ne}(\alpha, \gamma)^{24}\text{Mg}$ ,  $^{14}\text{N}(\alpha, \gamma)^{18}\text{F}$  and  $^{18}\text{O}(\alpha, \gamma)^{22}\text{Ne}$ . Fluorine-18, produced when  $^{14}\text{N}$  captures an alpha particle, is unstable and decays by emitting a

positron ( $e^+$ ) and a neutrino ( $\nu$ ) to form  $^{18}\text{O}$  [in short,  $^{14}\text{N}(\alpha, \gamma)^{18}\text{F}(e^+, \nu)^{18}\text{O}$ ]. Because there is likely to be  $^{13}\text{C}$  in the stellar core if hydrogen burning proceeded by the carbon-nitrogen-oxygen cycles, the neutron-producing reaction  $^{13}\text{C}(\alpha, n)^{16}\text{O}$  should also be included with the helium-burning reactions. The neutrons produced by this reaction are probably responsible for the bulk of s-process nucleosynthesis (discussed below). Helium burning is probably responsible for much of the  $^{12}\text{C}$  observed in the cosmic abundances, although in more massive stars the later burning stages will consume the  $^{12}\text{C}$  produced earlier by helium burning. See NUCLEAR REACTION.

**Carbon burning.** Upon exhaustion of the helium supply, if the star has an initial mass of at least 8 solar masses, gravitational contraction of the stellar core can lead to a temperature exceeding  $5 \times 10^8$  K, where it becomes possible for two  $^{12}\text{C}$  nuclei to overcome their high mutual Coulomb-repulsion barrier and fuse to form  $^{20}\text{Ne}$ ,  $^{23}\text{Na}$ , and  $^{24}\text{Mg}$  through reactions such as  $^{12}\text{C}(^{12}\text{C}, \alpha)^{20}\text{Ne}$ ,  $^{12}\text{C}(^{12}\text{C}, p)^{23}\text{Na}$ , and  $^{12}\text{C}(^{12}\text{C}, \gamma)^{24}\text{Mg}$ . Carbon burning can produce a number of nuclei with masses less than or equal to 28 through further proton and alpha-particle captures.

**Oxygen burning.** Carbon burning is followed by a short-duration stage, sometimes referred to as neon burning, in which  $^{20}\text{Ne}$  disintegrates by the reaction  $^{20}\text{Ne}(\gamma, \alpha)^{16}\text{O}$ . The alpha particle released is then captured by a remaining  $^{20}\text{Ne}$  nucleus to produce  $^{24}\text{Mg}$ ; thus, the effective neon burning reaction is  $^{20}\text{Ne} + ^{20}\text{Ne} \rightarrow ^{16}\text{O} + ^{24}\text{Mg}$ . The eventual result is that much of the carbon from helium burning becomes oxygen, which supplements the original oxygen formed in helium burning. This stage is followed by the fusion of oxygen nuclei at much higher temperatures. (Temperatures greater than  $10^9$  K are required for  $^{16}\text{O}$  nuclei to overcome their mutual Coulomb barrier.) Some relevant reactions for oxygen burning are

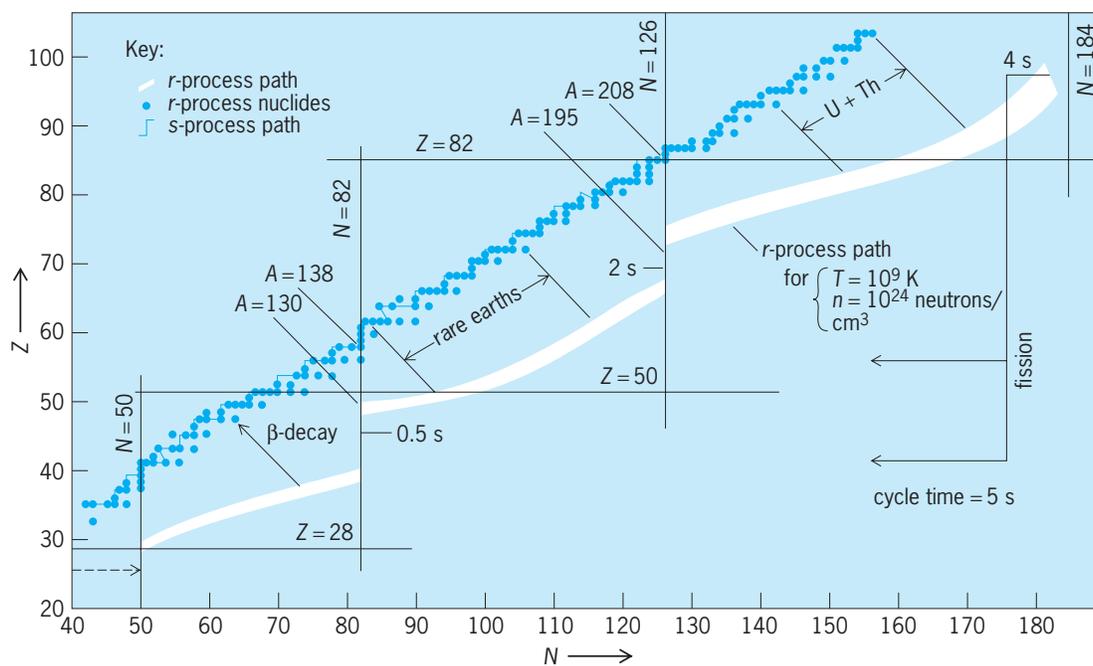


Fig. 3. Neutron-capture paths among the nuclei plotted in the  $Z$ - $N$  plane, showing nuclei identified by the number of protons ( $Z$ ) and number of neutrons ( $N$ ) in the nucleus. (After P. A. Seeger, W. A. Fowler, and D. D. Clayton, *Nucleosynthesis of heavy elements by neutron capture*, *Astrophys. J. Suppl.*, 11(97):121–166, 1965)

$^{16}\text{O}(\alpha, n)^{28}\text{Si}$ ,  $^{16}\text{O}(\alpha, p)^{31}\text{P}$ , and  $^{16}\text{O}(\alpha, \gamma)^{32}\text{S}$ . Nuclei of masses up to  $A = 40$  may be produced in this phase through proton, neutron, and alpha-particle captures.

**Silicon burning.** This process commences when the temperature exceeds about  $3 \times 10^9$  K. In this phase, photodisintegration of  $^{28}\text{Si}$  and other intermediate-mass nuclei around  $A = 28$  produces copious supplies of protons, neutrons, and alpha particles. These particles capture on the seed nuclei left from previous burning stages and thus produce new isotopes up to mass 60, resulting in the buildup of the abundance peak near  $A = 56$  (Fig. 1). Because the binding energy per nucleon is a peak for the iron-group ( $A$  near 60) nuclei, further fusion reactions no longer release energy. Silicon burning is therefore the end stage of stellar burning. Production of nuclei with mass higher than  $A$  near 60 therefore occurs by special processes that are ancillary to the mainline stellar burning stages.

**The s-process.** Because neutrons are neutral particles, their capture is not affected by the Coulomb barrier that inhibits charged-particle reactions. If the number of neutrons per seed nucleus is small, so that time intervals between neutron captures are long compared to the beta-decay lifetimes of unstable nuclei that are formed, the *s*-process (slow process) takes place. The seed nuclei are predominantly in the iron peak (Fig. 1), but the abundances of low-mass nuclei are also affected by neutron-capture processing. In the *s*-process, if neutron capture produces a nucleus that is unstable (due to an excess number of neutrons), the newly formed nucleus undergoes beta decay to a stable isobar by emitting an electron and an antineutrino. The resulting nucleus even-

tually captures another neutron, and the capture-decay step repeats. The presence of free neutrons leads to a chain of capture-decay events that drives the abundances along a unique *s*-process path that zigzags along a single line in the nuclear  $Z$ - $N$  diagram (Fig. 3) near the beta-stable nuclei (valley of beta stability). Given enough neutrons, the *s*-process synthesizes nuclei of masses up to 209, when alpha decay becomes a deterrent to further buildup by neutron capture. See RADIOACTIVITY.

The predominant site of the *s*-process is believed to be located in asymptotic giant branch stars. As these stars evolve on the asymptotic giant branch in the Hertzsprung-Russell diagram, thermonuclear energy is provided by multiple shells in which hydrogen and helium burning takes place. At some point in the evolution, the stellar interior becomes thermally unstable and a sequence of pulses with convective mixing takes place. Based on the observational results provided by heavy-element *s*-process rich stars, there is strong evidence that the reaction  $^{13}\text{C}(\alpha, n)^{16}\text{O}$  is the dominant source of neutrons. The  $^{13}\text{C}$  synthesized at the hydrogen-helium interface on the asymptotic giant branch in between thermal pulses is burned via the  $(\alpha, n)$  reaction during this radiative phase at temperatures of order  $10^8$  K. The neutrons released by this process then drive the *s*-process. An additional source of neutrons is the reaction  $^{22}\text{Ne}(\alpha, n)^{25}\text{Mg}$ , although the neutron-capture reactions induced reset the abundances of key isotopes at the end of thermal pulses.

Convective mixing repeatedly brings *s*-process material to the stellar surface. The technetium observed in some stars provides direct evidence for the *s*-process. Since cross sections for neutron

capture by nuclei with magic neutron numbers ( $N = 50, 82, 126$ ) are small, the  $s$ -process builds a high abundance of these nuclei and produces peaks near  $A = 87, 138, \text{ and } 208$ , labeled  $s$  in the cosmic elemental abundance curve (Fig. 1). See MAGIC NUMBERS; NUCLEAR STRUCTURE.

**The  $r$ -process.** This rapid process occurs when a large neutron flux allows rapid neutron capture, so that seed nuclei capture many neutrons before undergoing beta decay. The rapid neutron capture takes the nuclei far away from the valley of beta stability, into the regime of extremely neutron-rich nuclei. The time scale for the  $r$ -process is very short, 1–100 s, and the abundance of free neutrons is very large. These conditions are found deep inside the interiors of exploding massive stars, supernovae. The  $r$ -process can synthesize nuclei all the way into the transuranic elements. The  $r$ -process path in the chart of nuclei is shown in Fig. 3, where representative conditions for temperature ( $10^9$  K) and neutron density ( $n_n \sim 10^{24} \text{ cm}^{-3}$ ) were chosen. Once these conditions cease to exist, nuclei along this path very quickly beta-decay (along the diagonal lines in Fig. 3) toward the valley of stability. The circles in Fig. 2 indicate the stable end products reached by these decays. The times shown are intervals to reach locations along the  $r$ -process path for these conditions. See SUPERNOVA.

At  $Z$  near 94, the transuranic elements are unstable to neutron-induced fission, which terminates the  $r$ -process path. The resulting fission leads to the production of intermediate-mass nuclei which act as fresh seed nuclei, introducing a cycling effect. The peaks at  $A = 80, 130, \text{ and } 195$ , labeled  $r$  in Fig. 1, are due to the magic neutron numbers  $N = 50, 82, \text{ and } 126$ , and the subsequent decays of unstable nuclei. See NUCLEAR FISSION.

**The  $p$ -process.** This process produces heavier elements on the proton-rich side of the beta valley. The  $p$ -nuclei are blocked from formation by stable nuclei produced by either the  $r$ - or  $s$ -process. The major task for  $p$ -process theory is thus to find ways, other than beta decay, to process the more abundant  $r$ - and  $s$ -nuclei into the less abundant (Fig. 1)  $p$ -nuclei. Burbidge, Burbidge, Fowler, and Hoyle suggested two possible mechanisms that are still under consideration: radiative proton capture, and gamma-induced neutron, proton, or alpha-particle removal reactions. In both cases the temperature should be in excess of  $2\text{--}3 \times 10^9$  K.

The sites of the  $p$ -process have not yet been determined with confidence, and the models do not yet produce the  $p$ -process abundance pattern satisfactorily. The  $p$ -process was thought to take place in high-temperature, proton-rich zones in supernovae, but detailed studies have shown that the right conditions are very hard to find in these stellar environments. Also, several  $p$ -nuclei can be produced by other distinct processes, such as the alpha-rich freeze-out and the  $\nu$ -process. This suggests that  $p$ -nuclei owe their origin to a variety of different synthesis processes and environments; no single site is likely to produce all the  $p$ -nuclei.

Current theories of the  $p$ -process consider the premature termination of  $(\gamma, n)$ -,  $(\gamma, p)$ - and  $(\gamma, \alpha)$ -induced “melting” of nuclei in the shock-heated oxygen-neon layers in type II supernovae, or in exploding carbon-oxygen white dwarfs (type Ia supernovae). When only photodisintegration is important, “ $\gamma$ -process” is used as an alternate process name.

**The  $\nu$ -process.** The neutrino ( $\nu$ ) flux emitted from a cooling proto neutron star alters the yields of explosive nucleosynthesis from type II supernovae. Inelastic scattering of neutrinos (all flavors) off abundant nuclei excites states that can decay via single or multiple nucleon emission. The  $\nu$ -process is probably responsible for significant contributions to the synthesis in nature of about a dozen isotopes. While the neutrino interaction cross section with matter is extremely small (about  $10^{-44} \text{ cm}^2$ ), the high neutrino energies and the large number flux close to the collapsing iron core of a massive star lead to significant synthesis of nuclei, either directly or indirectly.

An example of a direct  $\nu$ -process reaction is  $^{12}\text{C}(\nu, \nu'p)^{11}\text{B}$ , where  $\nu'$  indicates the inelastically scattered neutrino emerging with a reduced energy. Even a very small change in the abundance of  $^{12}\text{C}$  (in the carbon layer of the supernova) implies a large contribution to  $^{11}\text{B}$  synthesis, because the cosmic abundance of boron is small compared to that of carbon (Fig. 1). The  $\nu$ -process thus contributes greatly to lithium, beryllium, and boron synthesis. A similar example is  $^{19}\text{F}$ , which can be efficiently made by  $^{20}\text{Ne}(\nu, p)^{19}\text{F}$  in the neon shell of the star. An example of an indirect (two-step) contribution of neutrino spallation is  $^4\text{He}(\nu, \nu'p)^3\text{H}$  in the helium shell. The  $^3\text{H}$  liberated by the neutrino interaction subsequently interact with the abundant helium to form  $^7\text{Li}$  via  $^4\text{He}(^3\text{H}, \gamma)^7\text{Li}$ .

Two of the rarest stable nuclei in nature,  $^{138}\text{La}$  and  $^{180}\text{Ta}$ , could owe their abundance to neutral-current neutrino interactions with their abundant neighbors in the chart of nuclei via  $^{181}\text{Ta}(\nu, \nu'n)^{180}\text{Ta}$ , and  $^{139}\text{La}(\nu, \nu'n)^{138}\text{La}$ . In addition, the charged-current reaction  $^{138}\text{Ba}(\nu, e^-)^{138}\text{La}$  can contribute significantly. While some contribution to element synthesis by neutrino-induced reactions must occur in nature, the estimated yields are somewhat uncertain due to the difficulties in the calculation of neutrino transport during the supernovae and also during the longer initial cooling phase (approximately 10 s) of the proto neutron star. The rates are very sensitive to the high-energy tail of the nonthermal neutrino spectrum. See NEUTRINO.

**The LiBeB process.** The bulk of the light elements lithium, beryllium, and boron found in the cosmic abundance curve (Fig. 1) cannot have survived processing in stellar interiors because they are readily destroyed by proton capture. Although some  $^7\text{Li}$  originated in the big bang, primordial nucleosynthesis cannot be responsible for the bulk of the  $^7\text{Li}$  in existence. Spallation of more abundant nuclei such as carbon, nitrogen, and oxygen by high-energy protons and alpha particles can account for the

low-abundance nuclides  ${}^6\text{Li}$ ,  ${}^9\text{Be}$ ,  ${}^{10}\text{B}$ ,  ${}^{11}\text{B}$ , and for some  ${}^7\text{Li}$ . The canonical process is spallation of carbon, nitrogen, and oxygen nuclei in the interstellar medium by fast light particles, such as alpha particles and protons, which are abundant in the gas between the stars. These high-energy particles are referred to as cosmic rays, leading to the term “cosmic-ray spallation (CRS) process.”

Supernova-induced shock acceleration in the interstellar medium provides the energetic cosmic-ray particles, which then interact with the carbon, nitrogen, and oxygen nuclei that are also present in the interstellar medium. Observations of lithium, beryllium, and boron abundances in metal-poor (population II) stars indicate that the (number) abundance ratio  ${}^7\text{Li}/\text{H}$  in these stars is about  $1.5 \times 10^{-10}$ , regardless of stellar metallicity, measured by the logarithmic iron-to-hydrogen ratio relative to the Sun:  $[\text{Fe}/\text{H}] = \log(\text{Fe}/\text{H}) - \log(\text{Fe}/\text{H})_{\odot}$ , where the subscript indicates that the last term refers to the solar ratio. The main source of this lithium is the big bang, but some of it could also be made in the  $\nu$ -process. On the other hand,  ${}^6\text{Li}$  is proportional to  $[\text{Fe}/\text{H}]$  and so are boron ( ${}^{10}\text{B}$  and  ${}^{11}\text{B}$ ) and beryllium ( ${}^9\text{Be}$ ). While  ${}^{11}\text{B}$  could be made by the  $\nu$ -process,  ${}^{10}\text{B}$  and  ${}^9\text{Be}$  are “orphans” of nucleosynthesis, made neither in stars nor in the big bang—instead their origin lies 100% in the cosmic rays. The observed nearly linear scaling with  $[\text{Fe}/\text{H}]$  of both beryllium and boron implies that both elements are of primary origin, which suggests a modified CRS process. One proposal involves a reverse CRS process, in which accelerated carbon, nitrogen, and oxygen nuclei interact with interstellar hydrogen and helium. In one version of this proposal, the cosmic-ray particles are assumed to be accelerated supernova ejecta, instead of being particles accelerated (by supernova shocks) out of the interstellar medium. These issues are not yet settled, but it is clear that lithium, beryllium, and boron nucleosynthesis will provide powerful diagnostic tools to study the important dynamic interaction of stellar explosions with the interstellar medium. See COSMIC RAYS.

**Conclusions.** Although stars spend the majority of their lifetimes in hydrogen-burning phases and most of the remaining time in helium burning, the vast majority of the elements are synthesized in the relatively brief time spans of later stages of evolution and the processes described above. Figure 1 serves as a summary of those processes and the mass ranges of elements for which they are responsible. However, there are many unanswered questions regarding which processes and which sites are responsible for certain elements. Shocks propagating through the supernova ejecta and the exposure to neutrinos from the newborn neutron star modify stellar synthesis at the very end of a massive star’s life in ways that are not fully understood.

While many of the scenarios presented in 1957 proved correct, new facets of nucleosynthesis theory were discovered and new processes and sites identified. Astrophysicists are increasing the knowledge of the thermodynamic conditions prevailing in

stars. As a result, further modifications in the theory of the processes of nucleosynthesis are quite likely. See STELLAR EVOLUTION.

Dieter H. Hartmann; Bradley S. Meyer

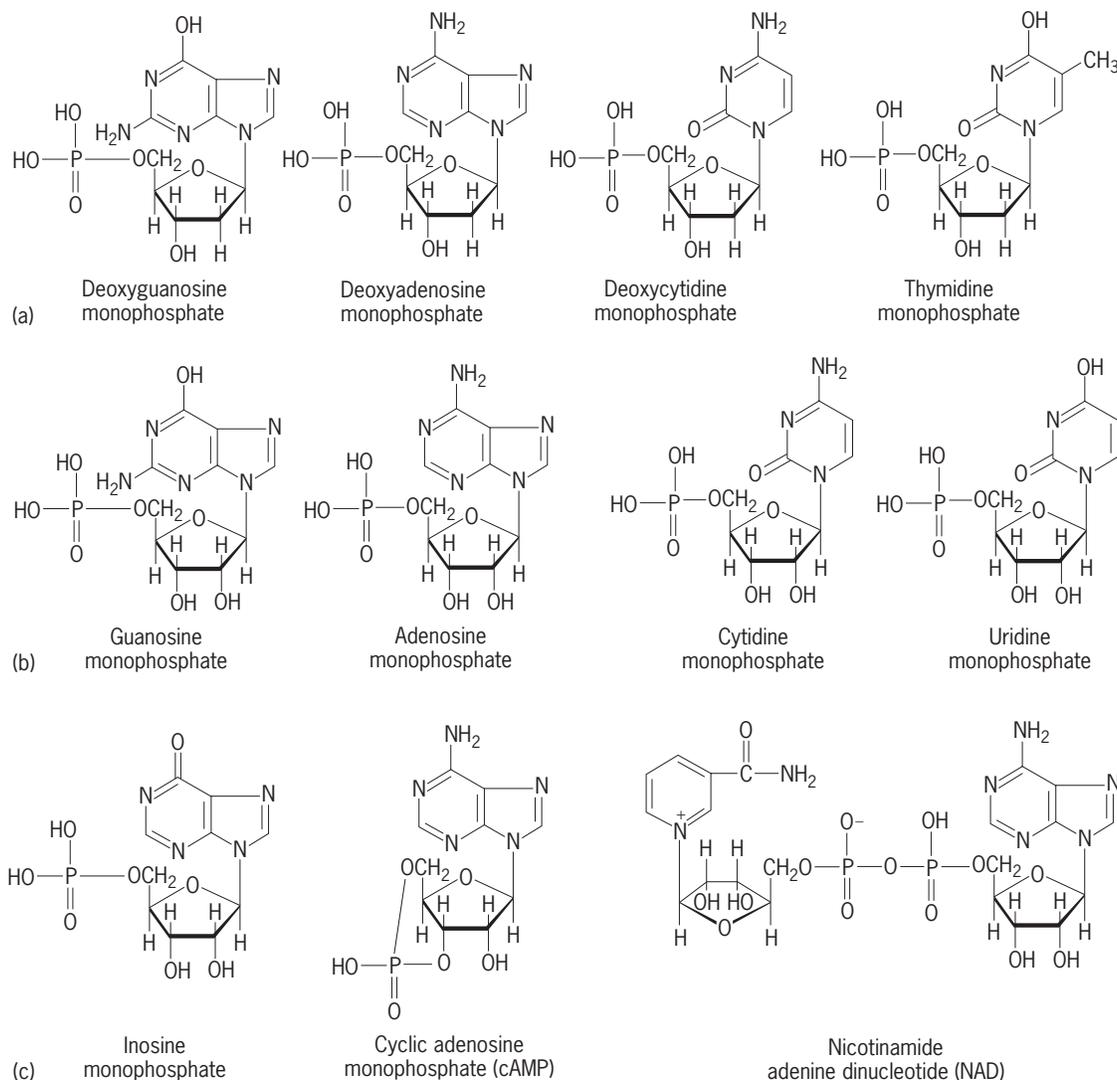
**Bibliography.** D. Arnett, *Supernovae and Nucleosynthesis*, Princeton University Press, 1996; D. D. Clayton, *Principles of Stellar Evolution and Nucleosynthesis*, University of Chicago Press, 1983; B. E. J. Pagel, *Nucleosynthesis and Chemical Evolution of Galaxies*, Cambridge University Press, 1997; C. E. Rolfs and W. S. Rodney, *Cauldrons in the Cosmos*, University of Chicago Press, 1988; G. Wallerstein et al., Synthesis of the elements in stars: Forty years of progress, *Rev. Mod. Phys.*, 69:995–1084, 1997.

## Nucleotide

A cellular constituent that is one of the building blocks of ribonucleic acids (RNA) and deoxyribonucleic acid (DNA). In biological systems, nucleotides are linked by enzymes in order to make long, chain-like polynucleotides of defined sequence. The order or sequence of the nucleotide units along a polynucleotide chain plays an important role in the storage and transfer of genetic information. Many nucleotides also perform other important functions in biological systems. Some, such as adenosine triphosphate (ATP), serve as energy sources that are used to fuel important biological reactions. Others, such as nicotinamide adenine dinucleotide (NAD) and coenzyme A (CoA), are important cofactors that are needed to complete a variety of enzymatic reactions. Cyclic nucleotides such as cyclic adenosine monophosphate (cAMP) are often used to regulate complex metabolic systems. Chemically modified nucleotides such as fluoro-deoxyridine monophosphate (FdUMP) contain special chemical groups that are useful for inactivating the normal function of important enzymes. These and other such compounds are widely used as drugs and therapeutic agents to treat cancer and a variety of other serious illnesses. See ADENOSINE TRIPHOSPHATE (ATP); COENZYME; CYCLIC NUCLEOTIDES; NICOTINAMIDE ADENINE DINUCLEOTIDE (NAD).

**Classification.** Nucleotides are generally classified as either ribonucleotides or deoxyribonucleotides (Fig. 1). Both classes consist of a phosphorylated pentose sugar that is linked via an *N*-glycosidic bond to a purine or pyrimidine base (Fig. 2). The combination of the pentose sugar and the purine or pyrimidine base without the phosphate moiety is called a nucleoside. See PURINE; PYRIMIDINE.

Ribonucleosides contain the sugar D-ribose, whereas deoxyribonucleosides contain the sugar 2-deoxyribose. The four most common ribonucleosides are adenosine, guanosine, cytidine, and uridine. The purine ribonucleosides, adenosine and guanosine, contain the nitrogenous bases adenine and guanine, respectively. The pyrimidine ribonucleosides, cytidine and uridine, contain the bases cytosine and uracil, respectively. Similarly, the most common deoxyribonucleosides include

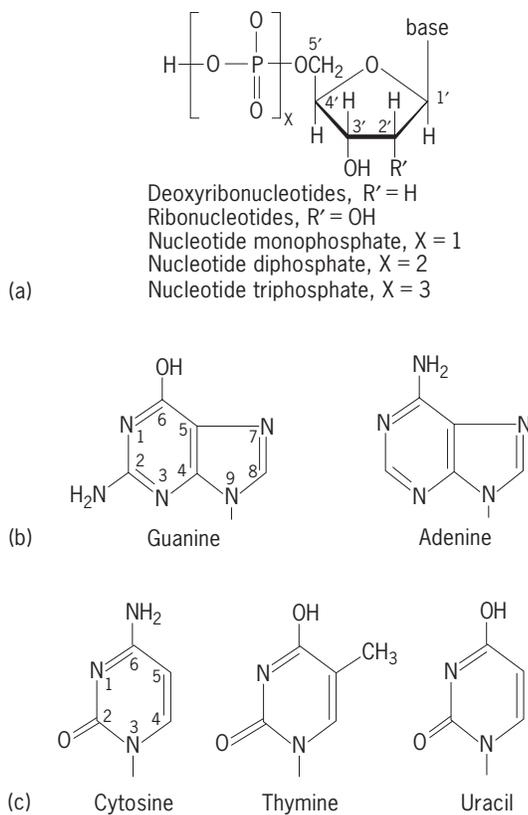


**Fig. 1.** Molecular structures of some common nucleotides. (a) Deoxyribonucleotides. (b) Ribonucleotides. (c) Other important nucleotides.

deoxyadenosine, deoxyguanosine, deoxycytidine, and thymidine, which contains the pyrimidine base thymine. Phosphorylation of the ribonucleosides or deoxyribonucleosides yields the corresponding ribonucleotide or deoxyribonucleotide.

The phosphate moiety of the nucleotide may be attached to any of the hydroxyl groups (5', 3', or 2') that are part of the pentose sugar (Fig. 2a). For example, phosphorylation of the 5' hydroxyl of adenosine yields the nucleotide adenosine 5'-monophosphate (5'-AMP); phosphorylation of thymidine at the 3' position yields thymidine 3'-monophosphate (3'-TMP). Some nucleotides contain one or two additional phosphate molecules that are linked together by pyrophosphate bonds. The phosphate moieties are added to the appropriate nucleotide precursors by enzymes known as kinases. The resulting nucleotide diphosphates and triphosphates play important roles in the synthesis of polynucleotides and in providing the energy for important biochemical reactions.

**Synthesis.** The biological synthesis of nucleotides involves a long series of biochemical reactions. These reactions are carefully controlled by complex feedback mechanisms that regulate the quantities of nucleotides produced. Purine ribonucleotides are synthesized by sequentially building components of the purine ring structure into ribose 5'-phosphate. The initial purine ribonucleotide, known as 5'-inosine monophosphate (5'-IMP), undergoes subsequent chemical modification in order to make the more common ribo-nucleotides 5'-adenosine monophosphate (5'-AMP) and 5'-guanosine monophosphate (5'-GMP). The synthesis of pyrimidine nucleotides involves the direct attachment of a completed pyrimidine molecule, orotic acid, onto 5'-phosphoribosyl-1-pyrophosphate (PRPP). Enzymatic decarboxylation of the resulting pyrimidine ribonucleotide yields 5'-uridine monophosphate (5'-UMP), and amination of 5'-UMP yields the other pyrimidine ribonucleotide, 5'-cytidine monophosphate (5'-CMP). Deoxyribonucleotides are produced by direct



**Fig. 2.** Chemical structure of (a) nucleotides, (b) purine bases, and (c) pyrimidine bases.

reduction of the appropriate ribonucleotide precursors using the enzyme ribonucleotide diphosphate reductase.

Nucleotides can also be obtained by the chemical or enzymatic degradation of polynucleotides. Ribonucleic acid can be readily hydrolyzed to nucleotides under alkaline conditions. Deoxyribonucleic acid is somewhat resistant to this alkaline treatment, however, because of the absence of a hydroxyl group at the 2' position. Enzymatic digestion of polynucleotides can also be used to make both nucleotide 3' and 5' phosphates. Phosphodiesterase, an enzyme in snake venom, degrades polynucleotides from the 3' end to yield nucleoside 5' phosphates. Conversely, phosphodiesterase from the spleen degrades polynucleotides from the 5' end to yield nucleoside 3' phosphates. Other enzymes, which are called restriction enzymes, split the polynucleotides into a series of smaller fragments by hydrolysis of the internucleotide bonds at sequence-specific positions. *See* RESTRICTION ENZYME.

Most nucleic acids contain nucleotides in which the 3' hydroxyl group of the first nucleotide is linked to the 5' phosphate of the second nucleotide. The polymerization of the nucleotide units is most often accomplished by the use of enzymes known as polymerases and transferases. These enzymes utilize nucleotide triphosphates in order to make long-chain polynucleotides. Some of these enzymes, such as Klenow polymerase, use a single-stranded template nucleic acid in order to control the sequence of nu-

cleotide addition. Other enzymes such as terminal deoxynucleotidyl transferase randomly add nucleotides to the 3' end of a growing polynucleotide without the need for a template strand. *See* DEOXYRIBONUCLEIC ACID (DNA); ENZYME; NUCLEIC ACID; RESTRICTION ENZYME; RIBONUCLEIC ACID (RNA).

E. Patrick Groody

**Bibliography.** A. Lesk (ed.), *Computational Molecular Biology: Sources and Methods for Sequence Analysis*, 1989; J. W. Phillis, *Adenosine and Adenine Nucleotides as Regulators of Cellular Function*, 1991; L. Stryer, *Biochemistry*, 4th ed., 1995; L. B. Townsend (ed.), *Chemistry of Nucleosides and Nucleotides*, 1994.

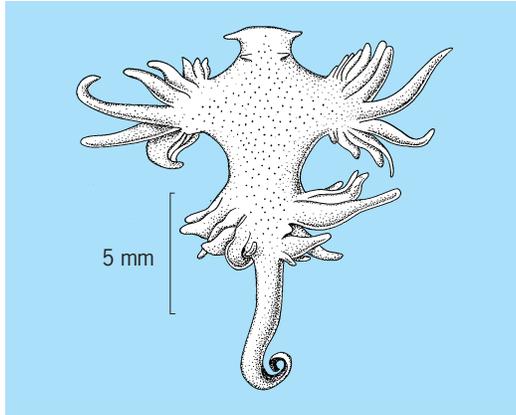
## Nuclide

A species of atom that is characterized by the constitution of its nucleus, in particular by its atomic number  $Z$  and its neutron number  $A-Z$ , where  $A$  is the mass number. Whereas the terms isotope, isotone, and isobar refer to families of atomic species possessing common atomic number, neutron number, and mass number, respectively, the term nuclide refers to a particular atomic species. The total number of stable nuclides is approximately 275. About a dozen radioactive nuclides are found in nature, and in addition, hundreds of others have been created artificially. *See* NUCLEOSYNTHESIS. Henry E. Duckworth

## Nudibranchia

An order of the gastropod subclass Opisthobranchia containing about 2500 living species of carnivorous sea slugs. They occur in all the oceans, at all depths, but reach their greatest size and diversity in warm shallow seas. In this, the largest order of the Opisthobranchia, there is much evidence indicating polyphyletic descent from a number of long-extinct opisthobranch stocks. In all those lines which persist to the present day, the shell and operculum have been discarded in the adult form. In many of them the body has quite independently become dorsally papillate. In at least two suborders these dorsal papillae have acquired the power to nurture nematocysts derived from their coelenterate prey so as to use them for the nudibranch's own defense. In other cases such papillae (usually called cerata) have independently become penetrated by lobules of the adult digestive gland, or they may contain virulent defensive glands or prickly bundles of daggerlike calcareous spicules. So pervasive are the tendrils of the converging lines of evolution that the precise interrelationships between the dendronotacean, arminacean, aeolidacean, and doridacean suborders may never be fully understood.

Nudibranchs feed on every kind of epifaunal animal material, as well as playing a significant part in some planktonic communities. Some nudibranchs



*Glaucus*, a specialized nudibranch known to attack planktonic coelenterates.

(*Cerberilla*) burrow into soft sediments in search of their actinian (sea anemone) prey. Each nudibranch family contains species which feed on broadly similar types of prey. For instance, members of the Tritoniidae all feed upon alcyonarian (soft) corals. The Antiopellidae and Onchidorididae contain species which attack bryozoans. The Coryphellidae, Dendronotidae, Eubbranchidae, Facelinidae, Heroidae, and Lomanotidae contain species that feed principally upon Hydrozoa. An unusual feeding type has evolved independently in members of several families; this is the egg-consumer, for example, the aeolids *Favorinus* and *Calma*. The former takes the eggs of other gastropod mollusks, but the latter feeds on the eggs of cephalopods and teleost fishes. Another highly successful specialization has evolved independently in the Fionidae (*Fiona*) and the Glaucidae (*Glaucus*, see **illus.**; *Glaucilla*); this involves attacking the planktonic coelenterates *Velella*, *Porpita*, and *Physalia* (the Portuguese man-o'-war). *Fiona* is said to attack also stalked barnacles. Glaucids and fionid nudibranchs are circumtropical in distribution.

There has been much argument about the significance of coloration in nudibranchs. The cerata of the aeolids and the dorsal papillae of many dorids or dendronotaceans are often patterned in such flamboyant ways as to foster the theory that these serve to take the attention of a sighted predator away from the fragile head and visceral mass. In many other cryptically marked nudibranchs, the resemblance in color pattern to the normal surroundings is so perfect that it would appear to disguise these mollusks. But against this it can be argued that many nudibranchs live in situations where there is little light (for instance, in deep waters or under boulders) and their color patterns cannot therefore be perceived by potential predators. The truth is that there is very little accurate information about the method of selection of food by predatory fish in different marine habitats. See OPISTHOBRANCHIA. T. E. Thompson

Bibliography. S. P. Parker (ed.), *Synopsis and Classification of Living Organisms*, 2 vols., 1982; T. E. Thompson, *Nudibranchs*, 1976.

## Number theory

The study of the properties and relationships between integers and other special types of numbers. There are many sets of positive integers of particular interest, such as the primes and the perfect numbers. Number theory, of ancient and continuing interest for its intrinsic beauty, also plays a crucial role in computer science, particularly in the area of cryptography.

**Elementary number theory.** This part of number theory does not rely on advanced mathematics, such as complex analysis and ring theory. The basic notion of elementary number theory is divisibility. An integer  $d$  is a divisor of  $n$ , written  $d \mid n$ , if there is an integer  $t$  such that  $n = dt$ . A prime number is a positive integer that has exactly two positive divisors, 1 and itself. The ten smallest primes are 2, 3, 5, 7, 11, 13, 17, 19, 23, and 29. Euclid (around 300 BC) proved that there are infinitely many primes by showing that if the only primes were 2, 3, 5, ...,  $p$ , a prime not in this list could be found by taking a prime factor of the number shown in Eq. (1).

$$N = (2 \cdot 3 \cdot 5 \cdots p) + 1 \quad (1)$$

Primes are the building blocks of the positive integers. The fundamental theorem of arithmetic, established by K. F. Gauss in 1801, states that every positive integer can be written as the product of prime factors in exactly one way when the order of the primes is disregarded. The fundamental ingredient in the proof is a lemma proved by Euclid: if  $a$  is a divisor of  $bc$  and  $a$  and  $b$  have no common factors, then  $a$  divides  $c$ .

The greatest common divisor of the positive integers  $a$  and  $b$ , written  $\gcd(a, b)$ , is the largest integer that divides both  $a$  and  $b$ . Two integers with greatest common divisor equal to 1 are called coprime. Euclid devised a method, now known as the euclidean algorithm, for finding the greatest common divisor of two positive integers. This method works by replacing  $a$  and  $b$ , where  $a > b$ , by  $b$  and the remainder when  $a$  is divided by  $b$ . This step is repeated until a remainder of 0 is reached. See ALGORITHM.

**Perfect numbers and Mersenne primes.** A perfect number is a positive integer equal to the sum of its positive divisors other than itself. L. Euler showed that  $2^{n-1}(2^n - 1)$  is perfect if and only if  $2^n - 1$  is prime. If  $2^n - 1$  is prime, then  $n$  itself must be prime. Primes of the form  $2^p - 1$  are known as Mersenne primes after M. Mersenne, who studied them in the seventeenth century. As of June 2005, 42 Mersenne primes were known, the largest being  $2^{25,964,951} - 1$ , a number with 7,816,230 decimal digits. For most of the time since 1876, the largest known prime has been a Mersenne prime; this is the case because there is a special method, known as the Lucas-Lehmer test, to determine whether  $2^p - 1$  is prime. Only 12 Mersenne primes were known prior to the invention of computers. Since the early 1950s, with the help of computers, Mersenne primes have been found at a

rate averaging approximately one every 2 years. The eight largest Mersenne primes were found as part of the Great Internet Mersenne Prime Search (GIMPS). GIMPS provides optimized software for running the Lucas-Lehmer test. Thousands of people participate in GIMPS, running a total of almost 1 teraflops in the search for Mersenne primes.

All even perfect numbers are given by Euler's formula. Whether there are odd perfect numbers is still an unsolved problem. If an odd perfect number existed, it would have at least eight distinct prime factors and be larger than  $10^{300}$ .

*Congruences.* If  $a - b$  is divisible by  $m$ , then  $a$  is called congruent to  $b$  modulo  $m$ , and this relation is written  $a \equiv b \pmod{m}$ . This relation between integers is an equivalence relation and defines equivalence classes of numbers congruent to each other, called residue classes. Congruences to the same modulus can be added, subtracted, and multiplied in the same manner as equations. However, when both sides of a congruence are divided by the same integer  $d$ , the modulus  $m$  must be divided by  $\gcd(d, m)$ . There are  $m$  residue classes modulo  $m$ . The number of classes containing only numbers coprime to  $m$  is denoted by  $\phi(m)$ , where  $\phi(m)$  is called the Euler phi function. The value of  $\phi(m)$  is given by Eq. (2).

$$\phi(m) = m \prod_{p|m} \left(1 - \frac{1}{p}\right) \quad (2)$$

Fermat's little theorem states that if  $p$  is prime and  $a$  is coprime to  $p$ , then formula (3)

$$a^{p-1} \equiv 1 \pmod{p} \quad (3)$$

is true. Euler generalized this congruence by showing that formula (4)

$$a^{\phi(m)} \equiv 1 \pmod{m} \quad (4)$$

is valid whenever  $a$  and  $m$  are coprime.

The linear congruence  $ax \equiv b \pmod{m}$  is solvable for  $x$  if and only if  $d \mid b$  where  $d = \gcd(a, m)$ . Under this condition, it has exactly  $d$  incongruent solutions modulo  $m$ . If  $p$  is prime, then the congruence  $ax \equiv 1 \pmod{p}$  has exactly one solution for each  $a$  not divisible by  $p$ . This solution is the inverse of  $a$  modulo  $p$ . This implies that the residue classes modulo  $p$  form a finite field of  $p$  elements. See FIELD THEORY (MATHEMATICS).

The simultaneous system of congruences  $x \equiv a_i \pmod{m_i}$ ,  $i = 1, 2, \dots, r$ , where the moduli  $m_i$ ,  $i = 1, 2, 3, \dots, r$ , are pairwise coprime positive integers, has a unique solution modulo  $M$ , where  $M$  is the product of these moduli. This result, called the Chinese remainder theorem, was known by ancient Chinese and Hindu mathematicians.

*Primality testing and factorization.* Since every composite integer has a prime factor not exceeding its square root, to determine whether  $n$  is prime, it is necessary only to show that no prime between 2 and  $\sqrt{n}$  divides  $n$ . However, this simple test is extremely inefficient when  $\sqrt{n}$  is large. Since generating large

primes is important for cryptographic applications, better methods are needed.

By Fermat's little theorem, if  $2^n$  is not congruent to 2 modulo  $n$ , then  $n$  is not prime; this was known to the ancient Chinese, who may have thought that the converse was true. However, it is possible that  $2^n \equiv 2 \pmod{n}$  without  $n$  being prime; such integers  $n$  are called pseudoprimes to the base 2. More generally, an integer  $n$  is called a pseudoprime to the base  $b$  if  $n$  is not prime but  $b^n \equiv b \pmod{n}$ . A Carmichael number is a number  $n$  that is not prime but that is a pseudoprime to all bases  $b$ , where  $b$  is coprime to  $n$  and less than  $n$ . The smallest Carmichael number is 561. In 1992 it was shown that there are infinitely many Carmichael numbers.

A positive integer  $n = 2^s t$ , where  $t$  is odd, is called a strong pseudoprime to the base  $b$  if either  $b^t \equiv 1 \pmod{n}$  or  $b^{2^j t} \equiv -1 \pmod{n}$  for some integer  $j$  with  $0 \leq j \leq s-1$ . The fact that  $n$  is a strong pseudoprime to the base  $b$  for at most  $(n-1)/4$  bases with  $1 \leq b \leq n-1$  when  $n$  is a composite integer is the basis for a probabilistic primality test which can be used to find extremely large numbers almost certainly prime in just seconds: Pick at random  $k$  different positive integers less than  $n$ . If  $n$  is composite, the probability that  $n$  is a strong pseudoprime to all  $k$  bases is less than  $(1/4)^k$ .

The naive way to factor an integer  $n$  is to divide  $n$  by successive primes not exceeding  $\sqrt{n}$ . If a prime factor  $p$  is found, this process is repeated for  $n/p$ , and so on. Factoring large integers  $n$  in this way is totally infeasible, requiring prohibitively large times. This shortcoming has led to the development of improved factorization techniques. Techniques including the quadratic sieve, the number field, and the elliptic curve factoring methods have been developed, making it possible to factor numbers with as many as 150 digits, although a large number of computers may be required. However, the factorization of integers with 200 digits is considered far beyond current capabilities, with the best-known methods requiring an astronomically larger time.

Interest in primality testing and factorization has increased since the mid-1970s because of their importance in cryptography. The Rivest-Shamir-Adleman (RSA) public-key encryption system uses the product of two large primes, say with 100 digits each, as its key. The security of the system rests on the difficulty of factoring the product of these primes. The product of the two primes held by each person is made public, but the separate primes are kept secret. See CRYPTOGRAPHY.

Although probabilistic primality tests can be used to generate primes rapidly, no practical test has been found that determines with certainty whether a particular positive integer is prime. An important and surprising theoretical breakthrough in this direction was made in 2002 by M. Agrawal, N. Kayal, and N. Saxena. They were able to find a polynomial-time algorithm that proves that a positive integer is prime, if it indeed is. So far, the best algorithms developed using their ideas are not of practical value in cryptographic applications.

*Sieves.* In the third century B.C., Eratosthenes showed how all primes up to an integer  $n$  can be found when only the primes  $p$  up to  $\sqrt{n}$  are known. It is sufficient to delete from the list of integers, starting with 2, the multiples of all primes up to  $\sqrt{n}$ . The remaining integers are all the primes not exceeding  $n$ .

Many problems, including the relative density of the set of twin primes (primes that differ by 2 such as 17 and 19), or the Goldbach conjecture that every even integer greater than 2 is the sum of two primes, have been attacked by sieve methods with partial success. Nevertheless, in spite of improvements of the method, it is not known whether the number of twin primes is finite or infinite or whether Goldbach's conjecture is true or false. (Goldbach's conjecture has been verified by computers up to  $2 \times 10^{17}$ .) In 1966 J. Chen showed that all even integers other than 2 are sums of a prime and another integer that is either prime or a product of only two primes. (The evidence that there are infinitely many twin primes continues to mount. As of 2005, the largest known twin primes were  $33,218,925 \times 2^{169,690} \pm 1$ .)

*Primitive roots and discrete logarithms.* The order of  $a$  modulo  $m$ , where  $a$  and  $m$  are coprime, is the least positive integer  $x$  such that formula (5)

$$a^x \equiv 1 \pmod{m} \tag{5}$$

is satisfied. Euler's theorem shows such an integer exists since  $a^{\phi(m)} \equiv 1 \pmod{m}$ . An integer  $a$  is called primitive root modulo  $m$  if the order of  $a$  modulo  $m$  equals  $\phi(m)$ , the largest possible order. The integer  $m$  has a primitive root if and only if  $m = 2, 4, p^k$ , or  $2p^k$  where  $p$  is an odd prime.

A solution of the congruence  $b^x \equiv c \pmod{n}$  for  $x$  is known as a discrete logarithm to the base  $b$  of  $c$  modulo  $n$ . The security of many cryptographic systems is based on the difficulty of finding discrete logarithms. The computational complexity of finding discrete logarithms modulo a prime  $p$  is similar to that of factoring an integer  $n$  of similar size to  $p$ . Several important methods for finding discrete logarithms modulo a prime are based on sieve techniques. The problem of finding discrete logarithms modulo  $p$  is much easier when  $p - 1$  has only small prime factors.

*Quadratic residues and reciprocity.* Quadratic residues modulo  $m$  are integers that are perfect squares modulo  $m$ . More precisely,  $a$  is a quadratic residue modulo  $m$  if the congruence shown in formula (6)

$$x^2 \equiv a \pmod{m} \tag{6}$$

has a solution for  $x$ , where  $a$  and  $m$  are coprime. If  $p$  is an odd prime, the Legendre symbol  $(a/p)$  is defined to be +1 if  $a$  is a quadratic residue modulo  $p$ , and to be -1 if  $a$  is a quadratic nonresidue modulo  $p$ .

The law of quadratic reciprocity, first proved by Gauss, states that if  $r$  and  $q$  are odd primes, then Eq. (7)

$$\left(\frac{p}{q}\right) \left(\frac{q}{p}\right) = (-1)^{[(p-1)/2][(q-1)/2]} \tag{7}$$

is valid. Gauss found eight different proofs of the law quadratic reciprocity himself, and new proofs are found on a regular basis with more than 200 different proofs known. Moreover, Eq. (8)

$$\left(\frac{-1}{p}\right) = (-1)^{(p-1)/2} \tag{8}$$

is true. The Legendre symbol can be generalized to odd composite moduli (Jacobi symbol) and to even integers (Kronecker symbol); reciprocity laws hold for these more general symbols.

*Quadratic forms.* The expression  $F(x, y) = Ax^2 + Bxy + Cy^2$ , where  $A, B$ , and  $C$  are integers and  $x$  and  $y$  are variables that take on integer values, is called a binary quadratic form with discriminant  $B^2 - 4AC$ . Two forms  $F(x, y) = Ax^2 + Bxy + Cy^2$  and  $F_1(x', y') = A_1x'^2 + B_1x'y' + C_1y'^2$  are equivalent if there are integers  $a, b, c$ , and  $d$  such that  $ad - bc = \pm 1$ ,  $x' = ax + by$ , and  $y' = cx + dy$ . Equivalent binary quadratic forms have the same discriminant and represent the same integers. The number of classes of equivalent forms with a given discriminant is finite. There are only finitely many negative discriminants with a given class number, with exactly nine with class number one. In the theory of binary quadratic forms with positive discriminant  $D$ , Pell's equation,  $x^2 - Dy^2 = \pm 1$ , plays a fundamental role.

*Sums of squares and similar representations.* Pierre de Fermat observed, and Euler proved, that every prime number congruent to 1 modulo 4 is the sum of two squares. J. L. Lagrange showed that every positive integer is the sum of at most four squares, and for some integers four squares are indeed needed. E. Waring conjectured in 1782 that to every positive integer  $k$  there is a number  $g(k)$  such that every natural number is the sum of at most  $g(k)$   $k$ th powers. This was first proved in 1909 by David Hilbert. Further work has been devoted to finding  $g(k)$ , the least integer such that every positive integer is the sum of at most  $g(k)$   $k$ th powers and  $G(k)$ , the least integer such that every sufficiently large integer is the sum of at most  $G(k)$   $k$ th powers. It is known that  $G(2) = 4$  and  $G(4) = 16$ , but the value of  $G(3)$  is unknown. Although there is strong numerical evidence that  $G(3) = 4$ , all that is currently known is that  $4 \leq G(3) \leq 7$ . New and improved bounds for  $G(k)$  for small integers  $k$  are established with some regularity such as the bounds  $G(19) \leq 134$  and  $G(20) \leq 142$  established by R. C. Vaughn and T. D. Wooley in 1999.

*Diophantine equations.* Single equations or systems of equations in more unknowns than equations, with restrictions on solutions such as that they must all be integral, are called diophantine equations, after Diophantus who studied such equations in ancient times. A wide range of diophantine equations have been studied. For example, the diophantine equation (9)

$$x^2 + y^2 = z^2 \tag{9}$$

has infinitely many solutions in integers. These solutions are known as pythagorean triples, since they correspond to the lengths of the sides of right

triangles where these sides have integral lengths. All solutions of this equation are given by  $x = t(u^2 - v^2)$ ,  $y = 2tuv$ , and  $z(u^2 + v^2)$ , where  $t$ ,  $u$ , and  $v$  are positive integers.

Perhaps the most notorious diophantine equation is Eq. (10).

$$x^n + y^n = z^n \tag{10}$$

Fermat's last theorem states that this equation has no solutions in integers when  $n$  is an integer greater than 2 where  $xyz \neq 0$ . Establishing Fermat's last theorem was the quest of many mathematicians over 200 years. In the 1980s, connections were made between the solutions of this equations and points on certain elliptic curves. Using the theory of elliptic curves, A. Wiles completed a proof of Fermat's last theorem based on these connections in 1995.

The discovery of the proof of Fermat's last theorem has led to the formulation of a more powerful conjecture by Andrew Beal, an amateur mathematician. Beal's conjecture asserts that the Diophantine equation (11)

$$x^a + y^b = z^c \tag{11}$$

has no solution in positive integers  $x, y, z, a, b$ , and  $c$ , where  $a \geq 3, b \geq 3, c \geq 3$ , and the integers  $x, y$ , and  $z$  are pairwise coprime.

**Algebraic number theory.** Attempts to prove Fermat's last theorem led to the development of algebraic number theory, a part of number theory based on techniques from such areas as group theory, ring theory, and field theory. Gauss extended the concepts of number theory to the ring  $R[i]$  of complex numbers of the form  $a + bi$ , where  $a$  and  $b$  are integers. Ordinary primes  $p \equiv 3 \pmod{4}$  are also prime in  $R[i]$ , but  $2 = -i(1 + i)^2$  is not prime, nor are primes  $p \equiv 1 \pmod{4}$  since such primes split as  $p = (a + bi)(a - bi)$ . More generally, an algebraic number field  $R(\theta)$  of degree  $n$  is generated by the root  $\theta$  of a polynomial equation  $f(x) = 0$  of degree  $n$  with rational coefficients. A number  $\alpha$  in this field is called an algebraic integer if it satisfies an algebraic equation with integer coefficients with initial coefficient 1. The algebraic integers in an algebraic number field form an integral domain. But, prime factorization may not be unique; for example, in  $R[\sqrt{-5}]$ ,  $21 = 3 \cdot 7 = (1 + 2\sqrt{-5}) \cdot (1 - 2\sqrt{-5})$  where each of the four factors in the two products is prime. To restore unique factorization, the concept of ideals is needed, as shown by E. E. Kummer and J. W. R. Dedekind. See RING THEORY.

**Analytic number theory.** There are many important results in number theory that can be established by using methods from analysis. For example, analytic methods developed by G. F. B. Riemann in 1859 were used by J. Hadamard and C. J. de la Vallée Poussin in 1896 to prove the famous prime number theorem. This theorem, first conjectured by Gauss about 1793, states that  $\pi(x)$ , the number of primes not exceeding  $x$ , behaves as shown in Eq. (12).

$$\lim_{x \rightarrow \infty} \frac{\pi(x)}{x/\log x} = 1 \tag{12}$$

These methods of Riemann are based on  $\zeta(s)$ , the function defined by Eq. (13),

$$\zeta(s) = \sum_{n=1}^{\infty} \frac{1}{n^s} = \prod_p \frac{1}{1-p^{-s}} \tag{13}$$

where  $s = \sigma + it$  is a complex variable; the series in this equation is convergent for  $\sigma > 1$ . Via an analytic continuation, this function can be defined in the whole complex plane. It is a meromorphic function with only a simple pole of residue 1 at  $s = 1$ . However, the fundamental theorem of arithmetic can be used to show that this series equals the product over all primes  $p$  shown in Eq. (13). It can be shown that  $\zeta(s)$  has no zeros for  $\sigma = 1$ ; this result and the existence of a pole at  $s = 1$  suffice to prove the prime number theorem. Many additional statements about  $\pi(x)$  have been proved. Riemann's work contains the still unproved so-called Riemann hypothesis: all zeros of  $\zeta(s)$  have a real part not exceeding  $1/2$ . See COMPLEX NUMBERS AND COMPLEX VARIABLES.

Another important result that can be proved by analytic methods is Dirichlet's theorem, which states that there are infinitely many prime numbers in every arithmetic progression  $am + b$ , where  $a$  and  $b$  are coprime positive integers.

**Diophantine approximation.** A real number  $x$  is called rational if there are integers  $p$  and  $q$  such that  $x = p/q$ ; otherwise  $x$  is called irrational. The number  $b^{1/m}$  is irrational if  $b$  is an integer which is not the  $m$ th power of an integer (for example,  $\sqrt{2}$  is irrational). A real number  $x$  is called algebraic if it is the root of a monic polynomial with integer coefficients; otherwise  $x$  is called transcendental. The numbers  $e$  and  $\pi$  are transcendental. That  $\pi$  is transcendental implies that it is impossible to square the circle. See CIRCLE; E (MATHEMATICS).

The part of number theory called diophantine approximation is devoted to approximating numbers of a particular kind by numbers from a particular set, such as approximating irrational numbers by rational numbers with small denominators. A basic result is that, given an irrational number  $x$ , there exist infinitely many fractions  $b/k$  that satisfy the inequality (14),

$$\left| x - \frac{b}{k} \right| < \frac{1}{ck^2} \tag{14}$$

where  $c$  is any positive number not exceeding  $\sqrt{5}$ . However, when  $c$  is greater than  $\sqrt{5}$ , there are irrational numbers  $x$  for which there are only finitely many such  $b/k$ . The exponent 2 in inequality (14) cannot be increased, since when  $x$  is algebraic and not rational the Thue-Siegel-Roth theorem implies that the inequality (15),

$$\left| x - \frac{b}{k} \right| < \frac{1}{ck^{2+\epsilon}} \tag{15}$$

where  $\epsilon$  is any number greater than zero (however small), can have at most only finitely many solutions  $b/k$ .

In 1851, J. Liouville showed that transcendental numbers exist; he did so by demonstrating that the

number  $x$  given by Eq. (16)

$$x = \sum_{j=1}^{\infty} 10^{-j} \tag{16}$$

has the property that, given any positive real number  $m$ , there is a rational number  $b/k$  that satisfies Eq. (17).

$$\left| x - \frac{b}{k} \right| < \frac{1}{k^m} \tag{17}$$

See ALGEBRA; NUMBERING SYSTEMS; ZERO.

**Additive number theory.** Problems in additive number theory can be studied using power series. Suppose that  $m_1, m_2, \dots, m_k, \dots$ , is a strictly increasing sequence of positive integers, such as the sequence of perfect squares, the sequence of prime numbers, and so on. If the power series of Eq. (18) is raised to the  $q$ th power to yield Eq. (19),

$$f(x) = \sum_{k=1}^{\infty} x^{m_k} \tag{18}$$

$$f(x)^q = \sum_{j=1}^{\infty} A_j x^j \tag{19}$$

then the coefficients  $A_j$  represent the number of times that  $j$  can be written as the sum of  $q$  elements from the set  $\{m_j\}$ . If it can be shown that  $A_k$  is positive for all positive integers  $k$ , then it has been shown that every positive integer  $k$  is the sum of  $q$  numbers from this set. Function-theoretic methods can be applied, as they have been by G. H. Hardy and J. E. Littlewood and by I. M. Vinogradov, to determine the coefficients  $A_k$ . For example, if  $m_k$  is the  $k$ th power of a positive integer and  $q$  is sufficiently large, Waring's theorem can be proved. Similarly, if  $m_k$  is the  $k$ th odd prime, this technique has been used to show that every sufficiently large odd number is the sum of at most three primes.

A partition of the positive integer  $n$  is a decomposition of  $n$  into a sum of positive integers, with order disregarded. Euler showed that the generating function for the number of partitions  $p(n)$  is given by Eq. (20).

$$\sum_{n=0}^{\infty} p(n)x^n = \prod_{m=1}^{\infty} (1 - x^m)^{-1} \tag{20}$$

A recurrence formula for  $p(n)$  can be derived from this formula, and in 1917 Hardy and S. Ramanujan derived an asymptotic formula for  $p(n)$  from it using function-theoretic methods. The first term is formula (21).

$$p(n) \approx \frac{\exp(\pi\sqrt{2n/3})}{4n\sqrt{3}} \tag{21}$$

Ramanujan discovered that the function  $p(n)$  satisfies many congruences, such as Eqs. (22) and (23).

$$p(5n + 4) \equiv 0 \pmod{5} \tag{22}$$

$$p(7n + 5) \equiv 0 \pmod{7} \tag{23}$$

Euler showed that the number of partitions into odd parts equals the number of partitions into distinct parts. For instance, 7 has five partitions into

odd parts (7, 1 + 1 + 5, 1 + 3 + 3, 1 + 1 + 1 + 1 + 3, and 1 + 1 + 1 + 1 + 1 + 1) and five partitions into distinct parts (7, 1 + 6, 2 + 5, 3 + 4, 1 + 2 + 4). Many additional results have been proved about partitions. See COMBINATORIAL THEORY.

Kenneth H. Rosen

**Bibliography.** T. M. Apostol, *Introduction to Analytic Number Theory*, Springer, 1986; J. R. Goldman, *The Queen of Mathematics: An Historically Motivated Guide to Number Theory*, A K Peters, 1997; Grosswald, *Topics from the Theory of Numbers*, 2d ed., Birkhauser, 1982; G. H. Hardy and E. M. Wright, *An Introduction to the Theory of Numbers*, 5th ed., Oxford, 1979; R. A. Mollin, *Algebraic Number Theory*, CRC Press, 1999; I. Niven, H. S. Zuckerman, and H. L. Montgomery, *An Introduction to the Theory of Numbers*, 5th ed., Wiley, 1991; K. H. Rosen, *Elementary Number Theory and Its Applications*, 5th ed., Addison-Wesley, 2005; W. Sierpinski and A. Schinzel, *Elementary Theory of Numbers*, 2d ed., North-Holland, 1988; J. Steuding, *Diophantine Analysis*, CRC Press/Chapman & Hall, 2005; J. Stopple, *A Primer of Analytic Number Theory*, Cambridge University Press, 2003.

## Numbering systems

A numbering system is a systematic method for representing numbers using a particular set of symbols. The most commonly used numbering system is the decimal system, based on the number 10, which is called the basis or radix of the system. The basis tells how many different individual symbols there are in the system to represent numbers. In the decimal system these symbols are the digits 0, 1, 2, 3, 4, 5, 6, 7, 8, 9. The range of these numbers varies from 0 to  $(10 - 1)$ . This is a particular case of a more general rule: Given any positive basis or radix  $N$ , there are  $N$  different individual symbols that can be used to write numbers in that system. The range of these numbers varies from 0 to  $N - 1$ .

In the computer and telecommunication fields, three of the most frequently used numbering systems are the binary (base 2), the octal (base 8), and the hexadecimal (base 16). The binary system has only two symbols: 0 and 1. Either of these symbols can be called a binary digit or a bit. The octal system has eight symbols: 0, 1, 2, 3, 4, 5, 6, 7. The hexadecimal system has 16 symbols: 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, A, B, C, D, E, F. In the hexadecimal system A stands for 10, B for 11, C for 12, D for 13, E for 14, and F for 15. The reason for choosing single letters to represent numbers higher than 9 is to keep all individual symbols single characters. See BIT.

Although "digits" generally refers to the individual symbols of the decimal system, it is common to call the individual symbols of the other numbering systems by that name. To explicitly state the basis on which a number is written, a subscript will be used to the lower right of the number. In this way  $11001_2$  is identified as a binary number, and  $113_{16}$  is identified

as a hexadecimal number. Subscripts are not used to represent decimal numbers.

All the numbering systems mentioned so far are positional systems. That is, the value of any symbol depends on its position in the number. For example, the value of 2 in the decimal number 132 is that of two units, whereas its value in decimal 245 is that of two hundreds. In the decimal system, the rightmost position of a number is called the ones ( $10^0$ ) place, the next position from the right is called the tens ( $10^1$ ) place, the next position the hundreds ( $10^2$ ) place, and so on. Observe that the powers increase from the right. The power of the rightmost digit is zero, the power of the next digit is one, the power of the next digit is two, and so on. These powers are sometimes called the weight of the digit.

**Conversion between bases.** In any positional system, the decimal equivalent of a digit in the representation of the number is the digit's own value, in decimal, multiplied by a power of the basis in which the number is represented. The sum of all these powers is the decimal equivalent of the number. The corresponding powers of each of the digits can be better visualized by writing superscripts beginning with 0 at the rightmost digit, and increasing the powers by 1 as we move toward the left digits of the number. For example, the decimal equivalent of the hexadecimal number 1A2C can be calculated as follows: (1) number the digits from right to left using superscripts; (2) use these superscripts as the powers of the basis and apply the general rule indicated before. After step 1 the number looks like this:  $1^3A^22^1C^0$ . According to step 2, the value of each digit and the decimal equivalent of the number are:

$$\begin{aligned} 1A2C_{16} &= (1 \cdot 16^3) + (10 \cdot 16^2) + (2 \cdot 16^1) + (12 \cdot 16^0) \\ &= (1 \cdot 4096) + (10 \cdot 256) + (2 \cdot 16) + (12 \cdot 1) \\ &= (4096) + (2560) + (32) + (12) \\ &= 6700 \end{aligned}$$

In this example, A and C have been replaced by their decimal equivalents of 10 and 12 respectively. This result indicates that, in the number 1A2C, the digit 1 has a value of 4096 units, the A has a value of 2560 units, the 2 has a value of 32, and the C has a value of 12 units. The decimal equivalent of the entire number is 6700. A similar procedure can be used to find the equivalents of  $175_8$  and  $11001_2$ .

$$\begin{aligned} 175_8 &= (1 \cdot 8^2) + (7 \cdot 8^1) + (5 \cdot 8^0) = 125 \\ 11001_2 &= (1 \cdot 2^4) + (1 \cdot 2^3) + (0 \cdot 2^2) + (0 \cdot 2^1) \\ &\quad + (1 \cdot 2^0) = 25 \end{aligned}$$

*Decimal numbers to other bases.* The conversion of a given decimal number to another basis  $r$  ( $r > 0$ ) is carried out by initially dividing the given decimal number by  $r$ , and then successively dividing the resulting quotients by  $r$  until a zero quotient is obtained. The decimal equivalent is obtained by writing the remainders of the successive divisions in the opposite

order in which they were obtained. For example, to convert decimal 41 to binary using this method, the number 41 is initially divided by the binary basis, that is, by  $r = 2$ , to obtain a first quotient and a remainder. This and all successive quotients are then divided by 2, until a quotient of zero is obtained. The remainders in the opposite order in which they were obtained give the representation of the equivalent binary number. This process is illustrated below.

Number	Quotient when dividing by 2	Remainder
41	20	1
20	10	0
10	5	0
5	2	1
2	1	0
1	0	1



Thus, the binary equivalent of decimal 41 is 101001. That is,  $41_{10} = 101001_2$ . This result can be verified by converting 101101 to its decimal equivalent according to the procedure described earlier. In fact,  $101001 = (1 \cdot 2^5) + (0 \cdot 2^4) + (1 \cdot 2^3) + (0 \cdot 2^2) + (0 \cdot 2^1) + (1 \cdot 2^0) = 41$ . Following a similar procedure and dividing by 16 and 8 respectively, it can be verified that  $5460 = 1554_{16}$  and  $876 = 1554_8$ .

*Binary to hexadecimal or octal and vice versa.* The table shows the decimal numbers 1 through 15 written in binary, octal, and hexadecimal. Since each four-bit binary number corresponds to one and only one hexadecimal digit and vice versa, the hexadecimal system can be viewed as a shorthand notation of the binary system. Similar reasoning can be applied to the octal system. This one-to-one correspondence between the symbols of the binary system and the symbols of the octal and hexadecimal system provides a method for converting numbers between these bases.

To convert binary numbers to hexadecimal, the following procedure may be used: (1) Form four-bit groups beginning from the rightmost bit of the number. If the last group (at the leftmost

Binary	Octal	Hexadecimal	Decimal
0001	1	1	1
0010	2	2	2
0011	3	3	3
0100	4	4	4
0101	5	5	5
0110	6	6	6
0111	7	7	7
1000	10	8	8
1001	11	9	9
1010	13	A	10
1011	14	B	11
1100	14	C	12
1101	15	D	13
1110	16	E	14
1111	17	F	15

position) has fewer than four bits, add extra zeros to the left of the bits in this group to make it a four-bit group. (2) Replace each four-bit group by its hexadecimal equivalent. Using this procedure, it can be verified that the hexadecimal equivalent of binary 111110110110101111 is 3EDAF. A process that is almost the reverse of the previous procedure can be used to convert from hexadecimal to binary. However, there is no need to add extra zeros to any group since each hexadecimal number will always convert to a group with four binary bits.

A similar process can be followed to convert a binary number to octal, except that in this case three-bit groups must be formed. Individual octal numbers will always convert to groups with three binary bits.

**Computer representation of integers.** Numerical data in a computer are generally written in basic units of storage made up of a fixed number of consecutive bits. The most commonly used units in the computer and communication industries are the byte (8 consecutive bits), the word (16 consecutive bits), and the double word (32 consecutive bits).

*Nonnegative integers.* A number is represented in each of these units by setting the bits according to the binary representation of the number. By convention, bits in any of these units are numbered from right to left, beginning with zero. The rightmost bit is called the least significant bit. The leftmost bit, numbered ( $n - 1$ ), where  $n$  is the number of bits in the unit, is called the most significant bit.

Since a bit can have only two values, given  $n$  bits, there are  $2^n$  different numbers that can be represented. The range of these numbers varies from 0 to  $2^n - 1$ .

*Negative integers.* To represent negative numbers, one of the bits is chosen to represent the sign. By convention, the leftmost bit is considered the sign bit. A value of 0 in the sign bit indicates a positive number, whereas a value of 1 indicates a negative number. This convention applies to bytes, words, and double words.

In addition to this sign convention, computer manufacturers use another convention, called the two's complement notation, to represent negative numbers. Two's complement is a three-step process: (1) Obtain the positive binary representation of the number. (2) Complement all bits of the number. That is, change all 0's to 1's and vice versa. (3) Add 1 to the binary configuration of the number obtained in the previous step. For example, the two's complement of decimal  $-45$  (using a word as the storage unit) can be obtained as follows: (1) Obtain the binary representation of  $+45$ . The procedure indicated earlier yields the result that  $45 = 000000000101101_2$ . (2) Complement all bits to obtain 1111111110010010. (3) Add 1 to the result of the previous step:

$$\begin{array}{r} 111111111010010 + \\ \underline{\phantom{111111111010010} 1} \\ 111111111010011 \leftarrow \text{Two's complement} \\ \phantom{111111111010011} \phantom{\leftarrow} \text{representation of } -45 \end{array}$$

To find out the positive counterpart of a given negative two's complement number, follow steps 2 and 3 indicated before. In this case, given the two's complement representation of  $-45$ , the representation of 45 should be 000000000101101. See NUMERICAL REPRESENTATION (COMPUTERS).

**Arithmetic operations.** Addition of binary numbers in a computer is carried out using a procedure similar to that of the addition of decimal numbers. The rules for binary addition are shown below. The sign bit is treated as any other bit.

+	0	1
0	0	1
1	1	0 with a carry of 1

The addition of the binary numbers 01010 and 00111 is shown below. Carries are indicated in italics.

$$\begin{array}{r} \phantom{0000}111 \\ \text{Augend} \quad 00001010 + \\ \text{Addend} \quad \underline{00000111} \\ \text{Sum} \quad \phantom{0000}00010001 \end{array}$$

In this example, when the two rightmost bits are added, the result is 1. When the next two binary 1's of the second column (from the right) are added, the result is zero with a carry of 1. This carry is then added to the zero of the third column of the augend. The result is 1. This 1 is then added to the 1 of the third column of the addend to produce another zero at the sum and another carry. The remaining bits of the sum are obtained in similar manner.

Subtracting binary numbers is very similar to the subtraction of decimal numbers; however, in borrowing from a higher unit, 2 is borrowed instead of 10. The similarity between the binary and decimal operations may be noted in the examples below. In the decimal system, the 0's to the left of the rightmost 0 became 9 (the basis minus 1). Likewise in the binary operation the 0's to the left of the rightmost 0 became 1 (the basis minus 1).

Decimal	Binary
7000 -	10000 -
15	111
<hr/> 6985	<hr/> 01001

To add two's complement numbers, the usual rules for adding binary numbers are followed and the carry out of the leftmost bit is ignored, if there is one. The binary number B can be subtracted from A simply by adding to A the two's complement of B. That is,  $A - B = A + (-B)$ , where  $-B$  is the two's complement of B. These processes are illustrated below. The decimal equivalents of the two's complement

numbers are indicated in parentheses. Here, A = 00011011 (27) and B = 00000100 (4).

Column A	Column B
00011011 + (27)	00011011 + (27)
00000100 (4)	11111100 (-4) ← two's complement notation
00011111 (31)	00010111 (23)

*Overflows.* Since all arithmetic operations are carried out using storage units of fixed size, there is the possibility that the result of an operation exceeds the capacity of the storage units. In this case, an overflow condition is said to have occurred. When the result is too small to be represented in the storage unit, an underflow is said to have occurred. Overflows will always occur if either of the following conditions is satisfied: (1) There is a carry into the sign bit and no carry out of it. (2) There is no carry into the sign bit and there is a carry out of the sign bit. In the addition of column B, for example, there is a carry into the most significant bit and a carry out of the most significant bit. Therefore, there is no overflow. Subtractions of two's complement numbers are carried out in the usual way; however, if a borrow is needed in the leftmost place, the borrow should be carried out as if there were another bit to the left. The following example illustrates this:

Minuend	00100011 -
Subtrahend	00111001
	<hr/>
	11101010

*Binary multiplication and division.* Multiplication and division of binary numbers is remarkably simple since the user can carry out these two operations by treating the numbers as if they were decimal numbers. The examples below illustrate these operations, using a byte as the basic storage unit.

To multiply 6 (=0110<sub>2</sub>) times 2 (=0010) we can proceed as follows:

Multiplicand	00110 ×
Multiplier	0010
	<hr/>
	0000 +
	0110
	0000
	0000
	<hr/>
	0001100 (=12)

*Hexadecimal and octal operations.* Operating with long strings of 0's and 1's is an error-prone task. For this reason it is preferable to carry out arithmetic operations in the hexadecimal system. Adding hexadecimal numbers is a process similar to that of adding decimal and binary numbers. However, it is necessary to remember that there is a carry whenever the sum exceeds F. In the subtraction operations, "16" is borrowed from the higher units to the left. To simplify this process, it is convenient to think in decimal notation. Therefore, in our heads we may translate letters into their decimal equivalent, do the operations in decimal, and translate the results back to hex-

adecimal. The following example illustrates the similarities between the arithmetic operations of both systems.

Column A (hexadecimal)	Column B (decimal)
I	
EDF4 +	7955 +
263	3162
<hr/>	<hr/>
F057	11117

The addition of the numbers of the second column (from the right) in column A is carried out as follows: Since 15 + 6 = 21 and we can write 21 = 1\*16 + 5, then we write 5 and carry 1. Similarly since 16 = 1\*16 + 0, we write 0 and carry 1. A similar situation is encountered in the addition of the decimal numbers of the second column of column B.

To multiply or divide numbers in hexadecimal, it is preferable to transform the numbers to binary, do the operations in binary, and then translate back the results to hexadecimal.

Operations in octal follow a similar pattern for both addition and subtraction. However, there is always a carry whenever the result of an addition exceeds 7. When borrowing, we always borrow 8 from the higher units to the left. The following examples illustrate this. The decimal equivalents of the numbers are shown in parentheses.

I	
623 + (403)	751 -
126 (86)	126
<hr/>	<hr/>
751 (489)	623

Octal quantities can be multiplied or divided by transforming the numbers to binary, carrying out the operations in binary, and transforming the results back to octal. See DIGITAL COMPUTER.

Ramon A. Mata-Toledo

*Bibliography.* H. Kruglak et al., *Basic Mathematics with Applications to Science and Technology*, 2d ed., Schaum's Outline Series, McGraw-Hill, 1998; R. Mata-Toledo and P. Cushman, *Introduction to Computer Science with Examples in Visual Basic, C, C++, and Java*, Schaum's Outline Series, McGraw-Hill, 2000; C. Maxfield and A. Brown, *Bebop Bytes Back. An Unconventional Guide to Computers*, Doone Publications, 1997.

## Numerical analysis

The development and analysis of computational methods (and ultimately of program packages) for the minimization and the approximation of functions, and for the approximate solution of equations, such as linear or nonlinear (systems of) equations and differential or integral equations. Originally part of every mathematician's work, the subject is now often taught in computer science departments because of the tremendous impact which computers have had on its development. Research focuses mainly on the numerical solution of (nonlinear) partial differential equations and the minimization of functions. See COMPUTER.

Numerical analysis is needed because answers provided by mathematical analysis are usually symbolic and not numeric; they are often given implicitly only, as the solution of some equation, or they are given by some limit process. A further complication is provided by the rounding error which usually contaminates every step in a calculation (because of the fixed finite number of digits carried).

Even in the absence of rounding error, few numerical answers can be obtained exactly. Among these are (1) the value of a piecewise rational function at a point and (2) the solution of a (solvable) linear system of equations, both of which can be produced in a finite number of arithmetic steps. Approximate answers to all other problems are obtained by solving the first few in a sequence of such finitely solvable problems. A typical example is provided by Newton's method: A solution  $c$  to a nonlinear equation  $f(c) = 0$  is found as the limit  $c = \lim_{n \rightarrow \infty} x_n$ , with  $x_{n+1}$  a solution to the linear equation  $f(x_n) + f'(x_n)(x_{n+1} - x_n) = 0$ , that is,  $x_{n+1} = x_n - f(x_n)/f'(x_n)$ ,  $n = 0, 1, 2, \dots$ . Of course, only the first few terms in this sequence  $x_0, x_1, x_2, \dots$  can ever be calculated, and thus one must face the question of when to break off such a solution process and how to gauge the accuracy of the current approximation. The difficulty in the mathematical treatment of these questions is amplified by the fact that the limit of a sequence is completely independent of its first few terms.

In the presence of rounding error, an otherwise satisfactory computational process may become useless, because of the amplification of rounding errors. A computational process is called stable to the extent that its results are not spoiled by rounding errors. The extended calculations involving millions of arithmetic steps now possible on computers have made the stability of a computational process a prime consideration.

**Interpolation and approximation.** Polynomial interpolation provides a polynomial  $p$  of degree  $n$  or less that uniquely matches given function values  $f(x_0), \dots, f(x_n)$  at corresponding distinct points  $x_0, \dots, x_n$ . The interpolating polynomial  $p$  is used in place of  $f$ , for example in evaluation, integration, differentiation, and zero finding. Accuracy of the interpolating polynomial depends strongly on the placement of the interpolation points, and usually degrades drastically as one moves away from the interval containing these points (that is, in case of extrapolation). See EXTRAPOLATION.

When many interpolation points (more than 5 or 10) are to be used, it is often much more efficient to use instead a piecewise polynomial interpolant or spline. Suppose the interpolation points above are ordered,  $x_0 < x_1 < \dots < x_n$ . Then the cubic spline interpolant to the above data, for example, consists of cubic polynomial pieces, with the  $i$ th piece defining the interpolant on the interval  $[x_{i-1}, x_i]$  and so matched with its neighboring piece or pieces that the resulting function not only matches the given function values (hence is continuous) but also has a continuous first and second derivative.

Interpolation is but one way to determine an approximant. In full generality, approximation involves several choices: (1) a set  $P$  of possible approximants, (2) a criterion for selecting from  $P$  a particular approximant, and (3) a way to measure the approximation error, that is, the difference between the function  $f$  to be approximated and the approximant  $p$ , in order to judge the quality of approximation. Much studied examples for  $P$  are the polynomials of degree  $n$  or less, piecewise polynomials of a given degree with prescribed breakpoints, and rational functions of given numerator and denominator degrees. The distance between  $f$  and  $p$  is usually measured by a norm, such as the  $L_2$  norm,  $(\int |f(x) - p(x)|^2 dx)^{1/2}$  or the uniform norm  $\sup_x |f(x) - p(x)|$ . Once choices 1 and 3 are made, one often settles 2 by asking for a best approximation to  $f$  from  $P$ , that is, for an element of  $P$  whose distance from  $f$  is as small as possible. Questions of existence, uniqueness, characterization, and numerical construction of such best approximations have been studied extensively for various choices of  $P$  and the distance measure. If  $P$  is linear, that is, if  $P$  consists of all linear combinations

$$\sum_{i=1}^n a_i p_i$$

of certain fixed functions  $p_1, \dots, p_n$ , then determination of a best approximation in the  $L_2$  norm is particularly easy, since it involves nothing more than the solution of  $n$  simultaneous linear equations.

**Solution of linear systems.** Solving a linear system of equations is probably the most frequently confronted computational task. It is handled either by a direct method, that is, a method that obtains the exact answer in a finite number of steps, or by an iterative method, or by a judicious combination of both. Analysis of the effectiveness of possible methods has led to a workable basis for selecting the one that best fits a particular situation.

*Direct methods.* Cramer's rule is a well-known direct method for solving a system of  $n$  linear equations in  $n$  unknowns, but it is much less efficient than the method of choice, elimination. In this procedure the first unknown is eliminated from each equation but the first by subtracting from that equation an appropriate multiple of the first equation. The resulting system of  $n - 1$  equations in the remaining  $n - 1$  unknowns is similarly reduced, and the process is repeated until one equation in one unknown remains. The solution for the entire system is then found by back-substitution, that is, by solving for that one unknown in that last equation, then returning to the next-to-last equation which at the next-to-last step of the elimination involved the final unknown (now known) and one other, and solving for that second-to-last unknown, and so on.

This process may break down for two reasons: (1) when it comes time to eliminate the  $k$ th unknown, its coefficient in the  $k$ th equation may be zero, and hence the equation cannot be used to eliminate the  $k$ th unknown from equations  $k + 1, \dots, n$ ; and (2) the process may be very unstable. Both

difficulties can be overcome by pivoting, in which one elects, at the beginning of the  $k$ th step, a suitable equation from among equations  $k, \dots, n$ , interchanges it with the  $k$ th equation, and then proceeds as before. In this way the first difficulty may be avoided provided that the system has one and only one solution. Further, with an appropriate pivoting strategy, the second difficulty may be avoided provided that the linear system is stable. Explicitly, it can be shown that, with the appropriate pivoting strategy, the solution computed in the presence of rounding errors is the exact solution of a linear system whose coefficients usually differ by not much more than roundoff from the given ones. The computed solution is therefore close to the exact solution provided that such small changes in the given system do not change its solution by much. A rough but common measure of the stability of a linear system is the condition of its coefficient matrix. This number is computed as the product of the norm of the matrix and of its inverse. The reciprocal of the condition therefore provides an indication of how close the matrix is to being noninvertible or singular.

*Iterative methods.* The direct methods described above require a number of operations which increases with the cube of the number of unknowns. Some types of problems arise wherein the matrix of coefficients is sparse, but the unknowns may number several thousand; for these, direct methods are prohibitive in computer time required. One frequent source of such problems is the finite difference treatment of partial differential equations (discussed below). A significant literature of iterative methods exploiting the special properties of such equations is available. For certain restricted classes of difference equations, the error in an initial iterate can be guaranteed to be reduced by a fixed factor, using a number of computations that is proportional to  $n \log n$ , where  $n$  is the number of unknowns. Since direct methods require work proportional to  $n^3$ , it is not surprising that as  $n$  becomes large, iterative methods are studied rather closely as practical alternatives.

The most straightforward iterative procedure is the method of substitution, sometimes called the method of simultaneous displacements. If the equations for  $i = 1, \dots, n$  are shown in Eq. (1), then

$$\sum_{j=1}^n a_{ij}x_j = b_i \quad (1)$$

the  $r$ th iterate is computed from the  $r - 1$ st by solving the trivial equations for  $x_i^{(r)}$  shown in Eq. (2) for

$$\sum_{j \neq i} a_{ij}x_j^{(r-1)} + a_{ii}x_i^{(r)} = b_i \quad (2)$$

$i = 1, \dots, n$ , where the elements  $x_i^{(0)}$  are chosen arbitrarily. If for  $i = 1, \dots, n$ , the inequality

$$\sum_{j \neq i} |a_{ij}| \leq |a_{ii}|$$

holds for some  $i$ , and the matrix is irreducible, then  $x_i^{(r)} \rightarrow x_i$  is the solution. For a matrix to be irreducible, the underlying simultaneous system must

not have any subset of unknowns that can be solved for independently of the others. For practical problems for which convergence occurs, analysis shows the expected number of iterations required to guarantee a fixed error reduction to be proportional to the number of unknowns. Thus the total work is proportional to  $n^2$ .

The foregoing procedure may be improved several ways. The Gauss-Seidel method, sometimes called the method of successive displacements, represents the same idea but uses the latest available values. Equation (3) is solved for  $x_i^{(r)}$ ,  $i = 1, \dots, n$ . The

$$\sum_{j < i} a_{ij}x_j^{(r)} + a_{ii}x_i^{(r)} + \sum_{j > i} a_{ij}x_j^{(r-1)} = b_i \quad (3)$$

Gauss-Seidel method converges for the conditions given above for the substitution method and is readily shown to converge more rapidly.

Further improvements in this idea lead to the method of successive overrelaxation. This can be thought of as calculating the correction associated with the Gauss-Seidel method and overcorrecting by a factor  $\omega$ . Equation (4) is first solved for  $y$ . Then

$$\sum_{j > i} a_{ij}x_j^{(r)} + a_{ii}y + \sum_{j < i} a_{ij}x_j^{(r-1)} = b_i \quad (4)$$

$x_i^{(r)} = x_i^{(r-1)} + \omega(y - x_i^{(r-1)})$ . Clearly, choosing  $\omega = 1$  yields the Gauss-Seidel method. For problems of interest arising from elliptic difference equations, there exists an optimum  $\omega$  that guarantees a fixed error reduction in a number of iterations proportional to  $n^{1/2}$ , and thus in total work proportional to  $n^{3/2}$ .

A number of other iterative techniques for systems with sparse matrices have been studied. Primarily they depend upon approximating the given matrix with one such that the resulting equations can be solved directly with an amount of work proportional to  $n$ . For a quite large class of finite difference equations of interest, the computing work to guarantee a fixed error reduction is proportional to  $n^{5/4}$ . The work requirement proportional only to  $n \log n$  quoted earlier applies to a moderately restricted subset.

*Overdetermined linear systems.* Often an overdetermined linear system has to be solved. This happens, for example, if one wishes to fit the model [Eq. (5)]

$$p(x) = \sum_{j=1}^n a_j p_j \quad (5)$$

to observations  $(x_i, y_i)_{i=1}^m$  with  $n < m$ . Here one would like to determine the coefficient vector  $\mathbf{a} = (a_1, \dots, a_n)^T$  so that  $p(x_i) = y_i$ ,  $i = 1, \dots, m$ . In matrix notation, one wants  $\mathbf{Aa} = \mathbf{y}$ , with  $\mathbf{A}$  the  $m$ -by- $n$  matrix  $[p_j(x_i)]$ . If  $n < m$ , one cannot expect a solution, and it is then quite common to determine  $\mathbf{a}$  instead by least squares, that is, so as to minimize the "distance"  $(\mathbf{y} - \mathbf{Aa})^T(\mathbf{y} - \mathbf{Aa})$  between the vectors  $\mathbf{y}$  and  $\mathbf{Aa}$ . This leads to the so-called normal equations  $\mathbf{A}^T\mathbf{Aa} = \mathbf{A}^T\mathbf{y}$  for the coefficient vector  $\mathbf{a}$ . But unless the "basis functions"  $p_1, \dots, p_n$  are chosen very carefully, the condition of the matrix  $\mathbf{A}^T\mathbf{A}$  may be very

bad, making the elimination process outlined above overly sensitive to rounding errors. It is much better to make use of a so-called orthogonal decomposition for  $A$ , as follows. See LEAST-SQUARES METHOD.

Assume first that  $A$  has full rank (which is the same thing as assuming that the only linear combination  $p$  of the functions  $p_1, \dots, p_n$  that vanishes at all the points  $x_1, \dots, x_m$  is the trivial one, the one with all coefficients zero). Then  $A$  has a  $QR$  decomposition, that is,  $A = QR$ , with  $Q$  an orthogonal matrix (that is,  $Q^T = Q^{-1}$ ), and  $R$  an  $m$ -by- $n$  matrix whose first  $n$  rows contain an invertible upper triangular matrix  $R_1$ , while its remaining  $m-n$  rows are identically zero. Then  $(\mathbf{y} - A\mathbf{a})^T \cdot (\mathbf{y} - A\mathbf{a}) = (Q^T\mathbf{y} - \mathbf{R}\mathbf{a})^T(Q^T\mathbf{y} - \mathbf{R}\mathbf{a})$  and, since the last  $m-n$  entries of  $\mathbf{R}\mathbf{a}$  are zero, this is minimized when the first  $n$  entries of  $\mathbf{R}\mathbf{a}$  agree with those of  $Q^T\mathbf{y}$ , that is  $R_1\mathbf{a} = [(Q^T\mathbf{y})(i)]^n_1$ . Since  $R_1$  is upper triangular, this system is easily solved by back-substitution, as outlined above. The  $QR$  decomposition for  $A$  can be obtained stably with the aid of Householder transformations, that is, matrices of the simple form  $H = I - (2/\mathbf{u}^T\mathbf{u})\mathbf{u}\mathbf{u}^T$ , which are easily seen to be orthogonal and even self-inverse, that is,  $H^{-1} = H$ . In the first step of the process,  $A$  is premultiplied by a Householder matrix  $H_1$  with  $\mathbf{u}$  so chosen that the first column of  $H_1A$  has zeros in rows 2, ...,  $m$ . In the next step, one premultiplies  $H_1A$  by  $H_2$  with  $\mathbf{u}$  so chosen that  $H_2H_1A$  retains its zeros in column 1 and has also zeroes in column 2 in rows 3, ...,  $m$ . After  $n - 1$  such steps, the matrix  $R := H_{n-1} \dots H_1A$  is reached with zeros below its main diagonal, and so  $A = QR$  with  $Q := H_1 \dots H_{n-1}$ .

The situation is somewhat more complicated when  $A$  fails to have full rank or when its rank cannot be easily determined. In that case, one may want to make use of a singular value decomposition for  $A$ , which means that one writes  $A$  as the product  $USV$ , where both  $U$  and  $V$  are orthogonal matrices and  $S = (s_{ij})$  is an  $m$ -by- $n$  matrix that may be loosely termed "diagonal," that is,  $s_{ij} = 0$  for  $i \neq j$ . Calculation of such a decomposition is more expensive than that of a  $QR$  decomposition, but the singular value decomposition provides much more information about  $A$ . For example, the diagonal elements of  $S$ , the so-called singular values of  $A$ , give precise information about how close  $A$  is to a matrix of given rank, and hence make it possible to gauge the effect of errors in the entries of  $A$  on the rank of  $A$ . See CURVE FITTING; INTERPOLATION; LINEAR SYSTEMS OF EQUATIONS; MATRIX THEORY.

**Differential equations.** Classical methods yield practical results only for a moderately restricted class of ordinary differential equations, a somewhat more restricted class of systems of ordinary differential equations, and a very small number of partial differential equations. The power of numerical methods is enormous here, for in quite broad classes of practical problems relatively straightforward procedures are guaranteed to yield numerical results, whose quality is predictable.

*Ordinary differential equations.* The simplest system is the initial value problem in a single unknown,  $y' = f(x, y)$ , and  $y(a) = \eta$ , where  $y'$  means  $dy/dx$ , and

$f$  is continuous in  $x$  and satisfies a Lipschitz condition in  $y$ ; that is, there exists a constant  $K$  such that for all  $x$  and  $y$  of interest,  $|f(x, y) - f(x, z)| \ll K|y - z|$ . The problem is well posed and has a unique solution.

The Euler method is as follows:  $y_0 = \eta$ , and Eq. (6)

$$y_{i+1} = y_i + hf(x_i, y_i) \tag{6}$$

holds for  $i = 0, 1, \dots, [(b - a)/b] - 1$ . Here  $b$  is a small positive constant, and  $x_i = a + ib$ . Analysis shows that as  $b \rightarrow 0$ , there exists a constant  $C$  such that  $|y_k - y(x_k)| \leq Cb$ , where  $y(x_k)$  is the value of the unique solution at  $x_k$ , and  $a \leq x_k \leq b$ . This almost trivial formulation of a numerical procedure thus guarantees an approximation to the exact solution to the problem that is arbitrarily good if  $b$  is sufficiently small, and it is easy to implement. A trivial extension of this idea is given by the method of Heun, Eq. (7).

$$y_{i+1} = y_i + \frac{1}{2}b[f(x_i, y_i) + f(x_i + b, y_i + hf(x_i, y_i))] \tag{7}$$

The method of Heun similarly is guaranteed to approximate the desired solution arbitrarily well since there exists another constant  $C_1$  such that the  $\{y_i\}$  satisfy relation (8). This is clearly asymptotically bet-

$$|y_k - y(x_k)| \leq C_1b^2 \tag{8}$$

ter than the method of Euler. It is readily found to be practically superior for most problems. Further improvement of this type is offered by the classical Runge-Kutta method, Eq. (9),

$$y_{i+1} = y_i + b\phi(x_i, y_i, b) \tag{9}$$

where  $\phi(x, y, b) = \frac{1}{6}[k_1 + 2k_2 + 2k_3 + k_4]$ , and  $k_1 = f(x, y)$ ,  $k_2 = f(x + b/2, y + bk_1/2)$ ,  $k_3 = f(x + b/2, y + bk_2/2)$ , and  $k_4 = f(x + b, y + bk_3)$ . For this method there exists a constant  $C_2$  such that  $|y_k - y(x_k)| \leq C_2b^4$ .

The foregoing methods are called single-step since only  $y_i$  is involved in the computation of  $y_{i+1}$ . The single-step methods yielding the better results typically require several evaluations of the function  $f$  per step. By contrast, multistep methods typically achieve high exponents on  $b$  in the error bounds without more than one evaluation of  $f$  per step. Multistep methods require use of  $y_{i-\alpha}$  or  $f(x_{i-\alpha}, y_{i-\alpha})$  or both for  $\alpha = 0, 1, \dots, j$  to compute  $y_{i+1}$ . Typical is the Adams-Bashforth method for  $j = 5$ , Eq. (10).

$$y_{i+1} = y_i + \frac{b}{1440}[4277f(x_i, y_i) - 7923f(x_{i-1}, y_{i-1}) + 9982f(x_{i-2}, y_{i-2}) - 7298f(x_{i-3}, y_{i-3}) + 2277f(x_{i-4}, y_{i-4}) - 475f(x_{i-5}, y_{i-5})] \tag{10}$$

Analysis shows the solution of this Adams-Bashforth procedure to satisfy relation (11) for some

$$|y_k - y(x_k)| \leq C_3b^6 \tag{11}$$

constant  $C_3$ . Many valuable multistep methods have been studied. If  $f$  is nontrivial to evaluate, the multistep methods are less work to compute than the

single-step methods for comparable accuracy. The chief difficulty of multistep methods is that they require  $j$  starting values and, therefore, cannot be used from the outset in a computation.

*Partial differential equations.* Methods used for partial differential equations differ significantly, depending on the type of equation. Typically, parabolic equations are considered for which the prototype is the heat flow equation, (12), with  $u(x,0)$  given on  $x \in$

$$\frac{\partial^2 u}{\partial x^2} = \frac{\partial u}{\partial t} \tag{12}$$

$[0,1]$ , say, and  $u(0,t)$  and  $u(1,t)$  given for  $t > 0$ . A typical finite difference scheme is shown in Eq. (13),

$$(w_{i,n})_{x\bar{x}} = \frac{w_{i,n+1} - w_{i,n}}{k} \tag{13}$$

where  $i = 1, \dots, 1/b - 1$ ,  $w_{0,n} = u(0,t_n)$ , and  $w_{1/b,n} = u(1,t_n)$ , with  $(w_i)_{x\bar{x}} = (w_{i+1} - w_i)/b$ ,  $(w_i)_{\bar{x}} = (w_{i-1})_x$ , and  $w_{i,n}$  is the function defined at  $x_i = ib$ ,  $t_n = nk$ . Analysis shows that as  $b, k \rightarrow 0$ , the solution  $w_{i,n}$  satisfies  $|w_{i,n} - u(x_i, t_n)| < C(b^2 + k)$  for some constant  $C$  if  $k/b^2 \leq 1/2$ , but for  $k/b^2$  somewhat larger than  $1/2$ ,  $w_{i,n}$  bears no relation at all to  $u(x_i, t_n)$ .

The restriction  $k/b^2 \leq 1/2$  can be removed by using the implicit difference equation, (14), but

$$(w_{i,n+1})_{x\bar{x}} = \frac{w_{i,n+1} - w_{i,n}}{k} \tag{14}$$

now simultaneous equations must be solved for  $w_{i,n+1}$  each step. The inequality,  $|w_{i,n} - u(x_i, t_n)| < C(b^2 + k)$ , still holds for some constant  $C$ . An improved implicit formulation is the Crank-Nicolson equation (15).

$$1/2(w_{i,n} + w_{i,n+1})_{x\bar{x}} = \frac{w_{i,n+1} - w_{i,n}}{k} \tag{15}$$

As  $b, k \rightarrow 0$ , solutions satisfy relation (16) for some

$$|w_{i,n} - u(x_i, t_n)| < C(b^2 + k^2) \tag{16}$$

constant  $C$ ; again  $b/k$  is unrestricted. Such techniques can extend to several space variables and to much more general equations. The work estimates given above for iterative solution of simultaneous equations, such as Eq. (15), apply for two-space variables.

Work using variational techniques to approximate the solution of parabolic and elliptic equations has been most fruitful for broad classes of nonlinear problems. The technique reduces partial differential equations to systems of ordinary differential equations. Analysis shows that solutions obtained approximate the desired solution, as within a constant multiple of the best that can be achieved within the subspace of the basis functions used. Practical utilization suggests this to be the direction for most likely future developments in the treatment of partial differential systems. See DIFFERENTIAL EQUATION.

Carl de Boor

*Bibliography.* W. Ames, *Numerical Methods for Partial Differential Equations*, 3d ed., 1992; K. E. Atkinson, *Elementary Numerical Analysis*, 2d ed., 1993;

J. C. Butcher, *The Numerical Analysis of Ordinary Differential Equations*, 1987; M. Friedman and A. Kandel, *Fundamentals of Computer Numerical Analysis*, 1993; C. F. Gerald and P. O. Wheatley, *Applied Numerical Analysis*, 6th ed., 1998; V. A. Patel, *Numerical Analysis*, 1993; H. R. Schwarz and J. Waldvogel, *Numerical Analysis: A Comprehensive Introduction*, 1989.

## Numerical representation (computers)

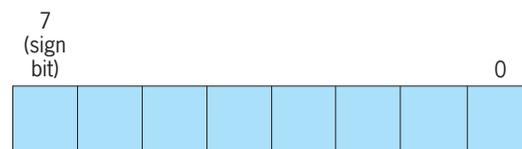
Numerical data in a computer are written in basic units of storage made up of a fixed number of consecutive bits. The most commonly used units in the computer and communication industries are the byte (8 consecutive bits), the word (16 consecutive bits), and the double word (32 consecutive bits). A number is represented in each of these units by setting the bits according to the binary representation of the number. By convention the bits in a byte are numbered, from right to left, beginning with zero. Thus, the rightmost bit is bit number 0 and the leftmost bit is number 7. The rightmost bit is called the least significant bit, and the leftmost bit is called the most significant bit. Higher units are numbered also from right to left. In general, the rightmost bit is labeled 0 and the leftmost bit is labeled  $(n - 1)$ , where  $n$  is the number of bits available. See BIT.

Since each bit may have one of two values, 0 or 1,  $n$  bits can represent  $2^n$  different unsigned numbers. The range of these nonnegative integers varies from 0 to  $2^n - 1$ .

**Sign bit.** To represent positive or negative numbers, one of the bits is chosen as the sign bit. By convention, the leftmost bit (or most significant bit) is considered the sign bit. This convention is shown in the **illustration** for a byte. A value of 0 in the sign bit indicates a positive number, whereas a value of 1 indicates a negative one. A similar convention is followed for higher storage units, including words and double words.

**Representation of integers.** Traditionally, the computer representation of integer numbers has been referred to as fixed-point representation. This usage assumes that there is a "binary" point located at the right end of the bit-string representation of the number. Several conventions have been proposed to represent integer values in a computer.

*Sign-magnitude convention.* To represent negative numbers using  $n$  bits, in this convention, the sign bit is set to 1 and the remaining  $n - 1$  bits are used to represent the absolute value of the number (its magnitude). Positive numbers are represented in a similar way except that the sign bit is set to 0. The range



Representation of a byte in computer storage.

of integer values that can be represented with this convention using  $n$  bits varies from  $-2^{n-1} + 1$  to  $2^{n-1} - 1$ . For example, the representation of +75 using this convention is 01001011<sub>2</sub>. The sign bit is 0 since the number is positive. The remaining 7 bits are used to represent the magnitude. Similarly, the representation of -75 is 11001011<sub>2</sub>. The sign bit is set to 1 since the number is negative. The subscript (2 at the right of the least significant digit indicates that the number is binary).

Although this convention is very simple to implement, computer manufacturers have not favored it due to the following drawbacks: First, there are two different representations for the number zero (10000000 and 00000000). Second, to carry out arithmetic operations with numbers of different sign requires that the magnitudes of the individual numbers involved be determined in advance to determine the sign of the result.

*One's complement notation.* This convention has the advantage of making the addition of two numbers with different signs the same as for two numbers with the same sign. Positive numbers are represented in the usual way. To represent a negative number in this convention, find the binary representation of the absolute value of the number and then complement all the bits, including the sign bit. To complement the bits of a binary number, change all 0's to 1's and all 1's to 0's. For example, using a word, the one's complement binary representations of decimal 15 and -15 are 0000000000001111<sub>2</sub> and 1111111111110000<sub>2</sub> respectively.

One's complement notation also suffers from the "negative and positive" zero malady. When a computer is performing computations, the hardware needs to check for both representations of zero. Arithmetic operations in one's complement notation sometimes require some additional checks at the hardware level. This and the double representation of zero have made this convention an unpopular choice.

*Two's complement notation.* The most widely used convention for representing negative numbers is the two's complement notation. This variation of the one's complement method corrects the problems of the negative zero and the additional checks at the hardware level required by the one's complement. The representation of positive numbers in this new convention is similar to that of the one's complement. To represent negative numbers in this convention, the following algorithm can be used: Scan the positive binary representation of the number from right to left until the first 1-bit is encountered. Leave this bit and all the 0-bits to its right, if any, unchanged. Complement all bits to the left of this first 1-bit. For example, according to this algorithm, the two's complement representation of 00010100 is 11101100.

The range of numbers that can be represented using this method with  $n$  bits is  $-2^{n-1}$  to  $2^{n-1} - 1$ . The range of negative numbers is one unit larger than the range of positive values. See NUMBERING SYSTEMS.

**Representation of real numbers.** The convention for representing real numbers, also called floating-point

numbers, is similar to the scientific notation for representing very small or very large real numbers. The representation of floating-point has the format  $\pm f \times r^p$ . Here,  $f$  is called the mantissa and is restricted to values in a small range, usually,  $0 \leq f < 1$ . The value  $r$ , called the radix, is generally restricted to the values 2, 8, or 16. The power  $p$  can be either positive or negative. On a computer, the values of both  $f$  and  $p$  are restricted by the number of bits used to store them.

Computer manufacturers sometimes define their own standards for representing floating-point numbers. The range of the numbers that can be represented depends upon the number of bits used to represent  $f$  and  $p$ . Some typical range values are  $0.29 \times 10^{-38}$  to  $1.7 \times 10^{38}$  (with  $f$  and  $p$  represented by 23 and 8 bits respectively), and  $0.84 \times 10^{-4932}$  to  $0.59 \times 10^{4932}$  (with  $f$  and  $p$  represented by 112 and 15 bits respectively). See DIGITAL COMPUTER.

Ramon A. Mata-Toledo

Bibliography. S. Baase, *VAX Assembly Language*, 2d ed., 1992; R. P. McArthur, *Logic to Computing*, 1991.

## Numerical taxonomy

As defined by P. H. A. Sneath and R. R. Sokal, the grouping by numerical methods of taxonomic units based on their character states. According to G. G. Simpson, taxonomy is the theoretical study of classification, including its bases, principles, procedures, and rules. The application of numerical methods to taxonomy, dating back to the rise of biometrics in the late nineteenth century, has received a great deal of attention with the development of the computer and computer technology. Numerical taxonomy provides methods that are objective, explicit, and repeatable, and is based on the ideas first put forward by M. Adanson in 1963. These ideas, or principles, are that the ideal taxonomy is composed of information-rich taxa ("taxon," plural "taxa," is the abbreviation for taxonomic group of any nature or rank as suggested by H. Lam) based on as many features as possible, that a priori every character is of equal weight, that overall similarity between any two entities is a function of the similarity of the many characters on which the comparison is based, and that taxa are constructed on the basis of diverse character correlations in the groups studied.

In the early stages of development of numerical taxonomy, phylogenetic relationships were not considered. However, numerical methods have made possible exact measurement of evolutionary rates and phylogenetic analysis. Furthermore, rapid developments in the techniques of direct measurement of the homologies of deoxyribonucleic acid (DNA), and ribonucleic acid (RNA), between different organisms now provide an estimation of "hybridization" between the DNAs of different taxa and, therefore, possible evolutionary relationships. Thus, research in numerical taxonomy often includes analyses of the chemical and physical properties of the nucleic acids

of the organisms; the data from these analyses are correlated with phenetic groupings established by numerical techniques. See PHYLOGENY; TAXONOMIC CATEGORIES.

**Mathematical approach.** Classification is recognized by philosophers as one of the main elements in human language. Two broad types of classification have been distinguished, natural and artificial. The essential difference between the two types in the extreme forms is that the objects in a natural classification show a likeness in a large number of attributes, whereas those in an artificial classification show a likeness in only a small number of attributes, possibly only one.

Two methods of classification, at the extreme, are fairly distinct: the typological method, linked with a natural classification, and the definitional method, linked with an artificial classification. The definitional classification consists of selecting one or perhaps two attributes and defining the groups as all the objects which show these attributes. The typological method is not a deliberate process but one by which the classes gradually emerge around a notional type. The Adansonian principle, based on an empirical approach, estimates overall similarity between organisms, based on the examination of a large number of equally weighted features. Several formulas have been suggested for calculating similarity between organisms. Resemblance is expressed by coefficients of similarity usually ranging between unity and zero, the former for perfect agreement and the latter for no agreement. Coefficients of dissimilarity (distance) can also be employed, in which case the range is between zero and an undefined positive value, the former for identity and the latter for maximal distance or disparity. Resemblance coefficients are usually tabulated in the form of a matrix, with one coefficient for every pair of taxonomic entities. The similarity index is the simplest of the coefficients employed to calculate similarity:  $S = n_s / (n_s + n_d)$ , where  $n_s$  represents the number of positive features shared by two strains being compared and  $n_d$  represents the number of features positive for one strain and negative for the other (and vice versa). An example of an  $S$ -value matrix is shown in Fig. 1.

A variety of clustering methods has been employed in numerical taxonomy, including sequential, agglomerative, hierarchic, and nonoverlapping clustering. Dendrograms, an example of which is given in Fig. 2, are most commonly employed to represent taxonomic structure resulting from cluster analysis. These have the advantage of being readily interpretable as conventional taxonomic hierarchies. Dendrograms are, in general, distinguished into phenograms representing phenetic relationships and cladograms representing evolutionary branching sequences.

Mathematics, in one form or another, whether a calculation of similarity indices, factor analysis to summarize complex relationships in terms of a few factors, or discriminatory analysis (the identification of an entity with an existing group by using attributes in such a way that the chance of correct identifica-

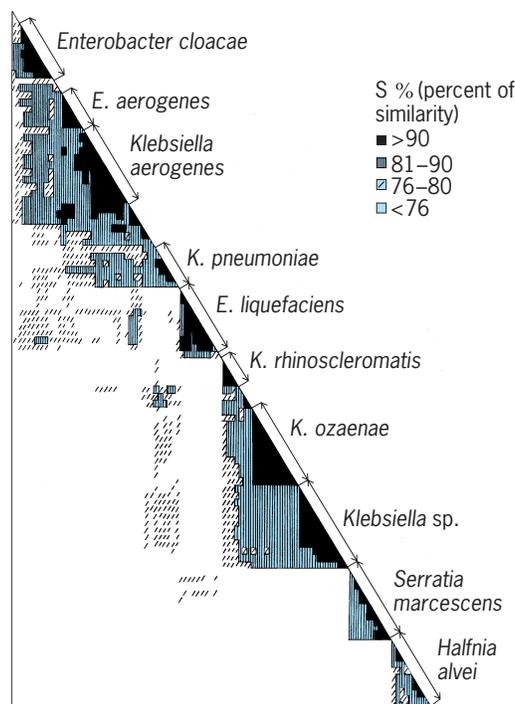


Fig. 1. Similarity matrix showing relationships between clusters of strains of enteric bacteria. (After R. Johnson et al., *Numerical taxonomy study of the Enterobacteriaceae*, *Int. J. System. Bacteriol.*, 25(1):12-37 1975)

tion is as high as possible), is successfully employed in classification. The availability of electronic computers renewed interest and sparked new developments in the mathematical approach to classification. A large number of attributes scored for each individual in a large sample set are easily managed by computer, the tool of the numerical taxonomist. Scoring taxonomic data for handling by the computer poses two questions: what is the actual scoring process and what number of features (unit characters) must be scored for an operational taxonomic unit (OTU) in numerical analysis? Simply stated, the data are scored in format suitable for transfer into the computer, via punched tape, punched cards, or magnetic tape, and for handling by programmed instructions stored in the computer. The data are usually scored as + or 1 for positive features, - or 0 for features negative or absent, and NC or any alphabetical or numerical term for features which are not applicable or not included in the calculation. Characters are broken down into unit characters, with the aim that each unit, as far as possible, should contribute one new item of information that is relevant to taxonomy; that is, the unit will permit distinguishing one kind of organism from another, or the unit varies from one kind of organism to another. The unit character has been defined by Sneath and Sokal as "a taxonomic character of two or more states, which within the study at hand cannot be subdivided logically, except for subdivision brought about by the method of coding." The kinds of information which are relevant are decided by the taxonomist.

The general parameter sought in the comparison of samples of numbers of characters is the estimate

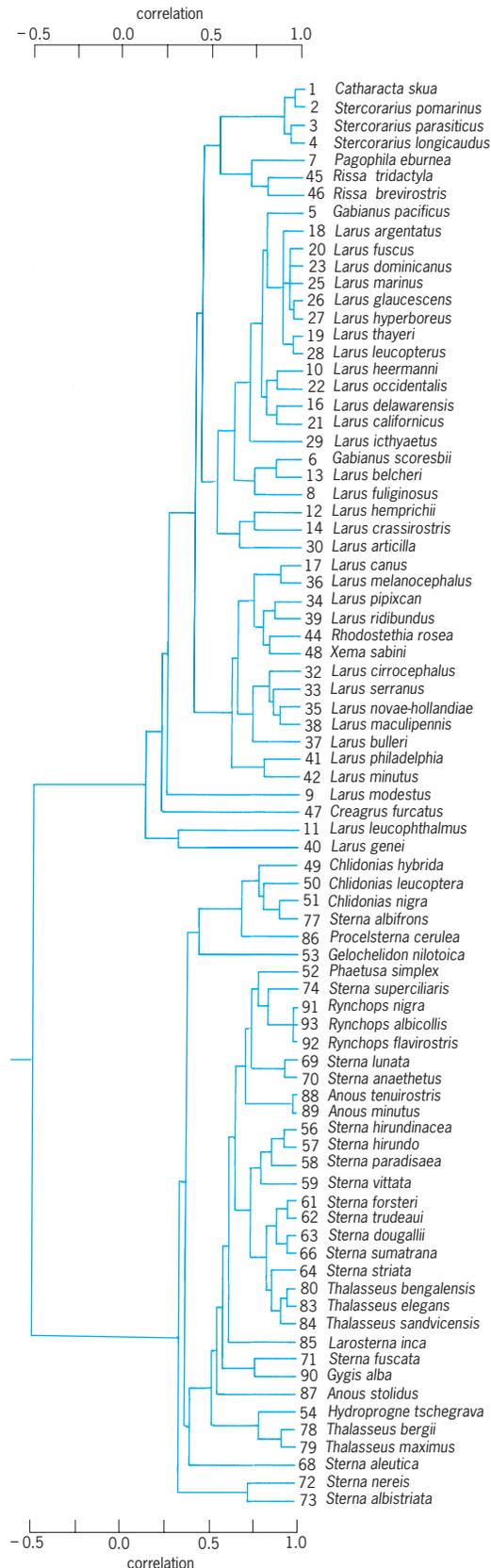


Fig. 2. Phenogram of 81 species of gulls and terns based on UPGMA cluster analysis of correlation coefficients on 51 skeletal measurements. (After G. O. Schnell, *A phenetic study of the suborder Lari (Aves), II. Phenograms, discussions, and conclusions*, *System. Zool.*, 19:264–302, 1970)

of similarity or total phenetic resemblance; one assumes that the phenetic information of each OTU is finite. Thus, the second question can be answered by applying the matches asymptote hypothesis of Sokal and Sneath, the best approach to the question of “how many characters should be sampled for a general taxonomy?” since it utilizes a testable empirical standard. The hypothesis simply assumes that as the number of characters sampled increases, the value of the similarity coefficient becomes more stable; eventually a further increase in the number of characters is not warranted by the corresponding mild decrease in the width of the confidence band of the coefficient. C. D. Michener and Sokal ventured a suggestion of the number of characters—not less than 60. The most practical advice to follow is that as many characters as is feasible should be distributed as widely as possible over the various body regions, life history stages, tissues, and levels of organization of the organisms. The capacities of computers which are available for numerical taxonomy computations rarely impose data storage limitations on the investigator.

Computer-assisted identification and classification of organisms have been made possible by establishment of data matrices that are stored in the computer, using keys or discriminant functions to achieve identification of unknown specimens. Online computer identification is now used to a great extent in microbiology and pathology.

The applications of numerical taxonomy cover many scientific disciplines since mathematical methods of classification can be applied to any material and in any discipline. Zoologists and botanists with a tradition of Linnaean taxonomy and a more or less standardized procedure find that where numerical methods have been used, the method has proved valuable for predictive purposes. Fields, besides botany and zoology, in which numerical methods of classification have been successful are linguistics, archeology, social anthropology, psychiatry, sociology, medical diagnosis, microbiology, ecology, biogeography, and earth sciences. Numerical taxonomy studies have spanned a variety of organisms, including vertebrates, arthropods, dicotyledons, monocotyledons, bacteria, and viruses. *See ANIMAL SYSTEMATICS; PLANT TAXONOMY.*

**Nucleic acid data.** Microbiologists have been ready and willing to apply the methods of numerical taxonomy, chiefly because extremely scanty paleontological (fossil) records of microorganisms exist; hence microbiologists have not been evolutionarily oriented. Developments in molecular biology which provide DNA-DNA and DNA-RNA pairings between organisms, that is, essentially detailed comparisons of the DNA base sequence of two organisms, have made evolutionary conclusions at least theoretically possible. *See MOLECULAR BIOLOGY.*

Numerical taxonomy analyses of several bacterial genera, namely *Escherichia*, *Serratia*, *Pseudomonas*, and *Staphylococcus*, show strong correlation between clusters obtained by numerical taxonomy and DNA base composition data for these

strains. Bacterial strains clustered by mathematical methods and shown to share high similarity indices share identical overall DNA base composition. The identical or very nearly identical overall DNA base composition thus is considered to be a highly correlative feature of genetically interacting microbes. In addition, phenetic groupings, obtained by numerical analysis, have been shown to be homologous, with respect to DNA/DNA reassociation analyses. Thus, DNA base sequence studies have proven useful in elucidating the genotype-phenotype relationships revealed by numerical taxonomy. See BACTERIAL TAXONOMY.

R. R. Colwell

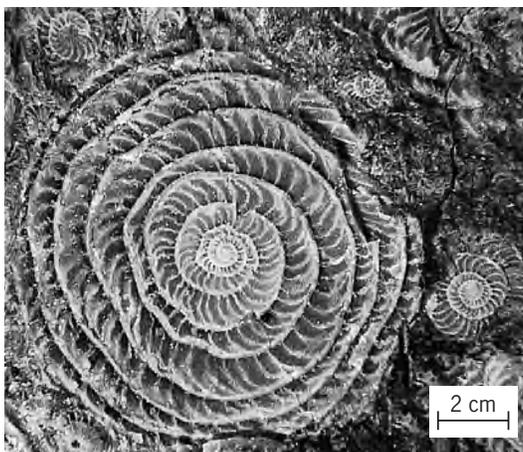
**Bibliography.** D. L. Quicke, *Principles and Techniques of Contemporary Taxonomy*, 1993; B. Rosner, *Fundamentals of Biostatistics*, 5th ed., 1999; S. Selvin, *Practical Biostatistical Methods*, 1995; G. G. Simpson, *Principles of Animal Taxonomy*, 1990; P. H. A. Sneath and R. R. Sokal, *Numerical Taxonomy: The Principles and Practice of Numerical Classification*, 1973; R. R. Sokal and J. F. Rohlf, *Biometry: The Principles and Practice of Statistics in Biological Research*, 3d ed., 1994.

## Nummulites

A genus of unicellular shelled protocist (protozoa) of the class Foraminifera (order Rotaliida, family Nummulitidae). The shells of *Nummulites* and other large, internally complex foraminifers commonly occurred in rock-forming abundances on continental shelves and around oceanic islands throughout subtropical and tropical regions during the Cenozoic. Resulting limestones are important hydrocarbon reservoirs from the Philippines and Indonesia westward to Trinidad and Venezuela. Nummulitic limestones have been used as building materials around the Mediterranean; the pyramids of Egypt were constructed of blocks of Eocene age. *Nummulites* shells are useful stratigraphic and paleoenvironmental indicators; the Paleogene was formerly known as the Nummulitic Period. See LIMESTONE.

**Shell morphology.** The shells, also known as tests, are composed of finely perforate calcite; may be discoidal, lenticular, or globular in shape; and can be as much as 6 in. (15 cm) in diameter, though sizes of 1–20 mm are more common. Shells are planispirally enrolled whorls of tiny undivided chambers (see **illustration**). The first-formed chamber (proloculus) and second chamber (deuteroconch) are connected by a central pore. In subsequent chambers, the test opening (aperture) consists of rows of pores at the base of the apertural face. Within the wall, a system of canals includes a distinct marginal cord at the outer periphery, branching sutural canals, and radial passages connecting canals of successive whorls.

**Ecology.** Modern representatives of the Nummulitidae are benthic (bottom-dwelling) foraminifers that harbor diatom (single-celled alga) symbionts. The symbionts live within the host cytoplasm and assist the host in metabolism and calcification by providing photosynthetically produced carbon. Paleonto-



**Nummulites** in a borehole core through the early Eocene El Garia Formation of the Hasdrubal Gas Field, offshore Tunisia. The large individual developed from gametes; the smaller specimens were likely produced asexually. (Photo courtesy of Simon Beavington-Penney)

logical studies indicate that fossil *Nummulites* lived in similar symbioses. Because the diatoms require sunlight to photosynthesize, the foraminifers must live at depths to which some sunlight can penetrate. The shapes of the shells indicate the amount of both light and water motion that the foraminifers were exposed to in life; globular shells represent abundance of both, while very flat, delicate tests indicate minimal light and water motion. Since both light and wave motion decrease exponentially with depth, geologists use the shell shapes to estimate water depth at which the foraminifers lived. Very flat, delicate forms can be found living today as deep as 500 ft (~150 m).

**Reproduction.** Reproduction is through alternation of sexual and asexual generations. Haploid asexual generations are produced by multiple fission; the parent extrudes its cytoplasm, in which hundreds of embryos (proloculus plus deuteroconch) develop. The embryos incorporate both the cytoplasm and symbionts of the parent before dispersing away from the empty parent shell. The offspring grow for several months, then may produce either another asexual generation or millions of tiny, identical gametes. Embryos resulting from the fusion of two gametes are much smaller than asexually produced embryos, yet individuals grow to much larger sizes before reproducing. Thus, dimorphism is the norm, with the diploid generation characterized by larger adult shells (typically less than 5 mm) with tiny embryos (less than 0.02 mm) while haploid asexual generations have smaller adult shells (usually less than 5 mm) with larger embryos (0.1 mm or more). See FORAMINIFERIDA. Alfred R. Loelich, Jr.; Pamela Hallock

**Bibliography.** S. J. Beavington-Penney and A. Racey, Ecology of extant nummulitids and other larger benthic foraminifera: Applications in palaeoenvironmental analysis, *Earth-Sci. Rev.*, 67(3–4):219–265, 2004; J. Hohenegger, E. Yordanova, and A. Hatta, Remarks on West Pacific Nummulitidae (foraminifera), *J. Foram. Res.*, 30(1):3–28, 2000; M. Holzmann,

J. Hohenegger, and J. Pawlowski, Molecular data reveal parallel evolution in nummulitid, *J. Foram. Res.*, 33(4):277-284, 2003; L. M. A. Purton and M. D. Brasier, Giant protist *Nummulites* and its Eocene environment: Life span and habitat insights from delta O-18 and delta C-13 data from Nummulites and *Venericardia*, Hampshire basin, UK, *Geology*, 27(8): 711-714, 1999; A. Racey, A review of Eocene nummulite accumulations: Structure, formation and reservoir potential, *J. Petrol. Geol.*, 24(1):79-100, 2001; B. K. Sen Gupta (ed.), *Modern Foraminifera*, Kluwer Academic, Dordrecht, The Netherlands, 1999.

## Nursing

The application of principles from the basic sciences, social sciences, and humanities to assist healthy and sick individuals and their families or other caring persons in performing those activities that contribute to the individuals' physical and mental well-being and that they would perform unaided if able to do so. Nursing includes providing physical and emotional care, promoting comfort, serving as patient advocates, assisting in rehabilitative efforts, teaching self-care and health promotion activities, and administering treatments prescribed by a licensed physician or dentist. Patient-care activities are conceived and coordinated so as to help individuals gain independence as rapidly as possible or maintain an optimal level of function. When multiple health-care providers are involved, nursing coordinates patient-care efforts to improve the quality of care.

Nursing practice is conducted in a variety of settings, including hospitals, community facilities, private homes, nursing homes, schools, industry, physician's offices, the military, and civil service arenas. Standards for nursing practice and licensure are governed by state nurse practice acts and are directed by professional nursing organizations.

Two types of nurses are legally recognized in the United States: the registered professional nurse and the licensed practical nurse. Licensed practical or vocational nurses (LPNs or LVNs) are trained to perform uncomplicated patient-care tasks in hospitals or other health-care facilities under the aegis of registered nurses or physicians. Having successfully completed approximately 1 year of training in a hospital, community college, or other vocational setting, they are licensed by state boards of nursing after passing the licensing exam.

Registered professional nurses may be educated in numerous ways. Associate degree programs (2 years), hospital-based diploma schools (3 years), and baccalaureate degree programs (4 years) are the most commonly chosen routes. Several generic master's programs have also been established for those who enter nursing with a college degree. While all of these programs prepare nurses to assume beginning clinical roles, the baccalaureate degree is the first professional degree and allows the recipient greater flexibility and advancement. After completing an

approved program, nurses are required to successfully complete their state licensure examination before they are recognized as registered professional nurses.

Master's and doctoral education is directed toward preparing nursing leaders. Advanced study in nursing may be pursued by qualified individuals who are seeking expanded nursing roles. These roles include, but are not limited to, clinical nurse specialist (who has been trained for work in a specific field of medicine), nurse practitioner (who has additional training for performing primary care procedures), certified nurse midwife, nurse anesthetist, nurse educator, nurse administrator, and nurse researcher.

Professional nursing is guided by numerous professional organizations. The American Nurses' Association, founded in 1896, serves as the professional organization for registered nurses. Its role is to foster high standards of nursing care, promote nurses' economic welfare, and improve the general working conditions of nursing.

The National League for Nurses resulted from a merger (1952) of the National League of Nursing Education, the National Organization for Public Health Nurses, and the Association of Collegiate League for Nursing. Supporting both licensed practical and registered nurses, it provides numerous educational services and is the recognized voluntary national accrediting body for schools of nursing.

A third organization, the International Council of Nurses, was established in 1899. It facilitates communication for nurses of varying nations on issues of patient welfare, health promotion, and the advancement of the nursing profession. In addition, more than 40 specialty organizations provide support for nursing's various constituencies. See MEDICINE.

Judith A. Vessey

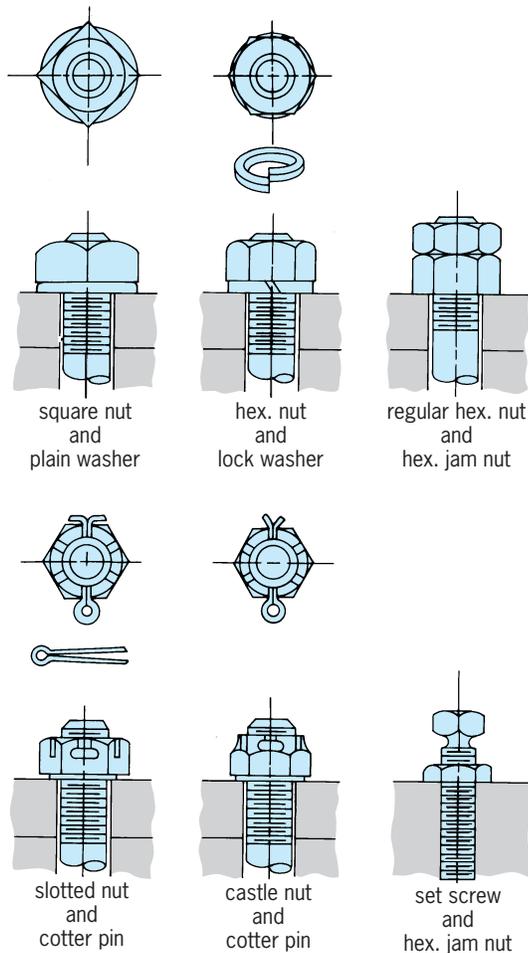
Bibliography. L. H. Aiken, *Nursing in the 1980's: Crises, Opportunities, Challenges*, 1982; American Nurses' Association, *Facts about Nursing*, 86-87, 1987; P. Benner, *From Novice to Expert: Excellence and Power in Clinical Practice*, 1984; M. P. Donahue, *Nursing, the Finest Art: An Illustrated History*, 1985; F. Nightingale, *Notes on Nursing: What It Is and What It Is Not*, 1860.

## Nut (engineering)

In mechanical structures, an internally threaded fastener. Plain square and hexagon nuts for bolts and screws are available in three degrees of finish: unfinished, semifinished, and finished. There are two standard weights: regular and heavy. For specific applications, there are other standard forms such as jam nut, castellated nut, slotted nut, cap nut, wing nut, and knurled nut (see **illus.**).

Hexagon jam nuts are used as locking devices to keep regular nuts from loosening and for holding set screws in position. They are not as thick as plain nuts. Jam nuts are available in semifinished form in both regular and heavy weight.

Castellated and slotted nuts have slots so that a cotter pin or safety wire can hold them in place.



Examples of standard forms of nuts.

They are commonly used in the automotive and allied fields on operating machinery where nuts tend to loosen. The slotted nut is a regular hexagon nut with slots cut across the flats of the hexagon. Slotted nuts are standardized in regular and heavy weights in semifinished hexagon form. Finished thick slotted nuts are available. Castle nuts are hexagonal with a cylindrical portion above, through which slots are cut.

Machine screw and stove-bolt nuts may be either square or hexagonal. Hexagon machine-screw nuts may have the top chamfered at 30° with a plain bearing surface, or both top and bearing surfaces may be chamfered. Square nuts have flat surfaces without chamfer. Square nuts are available with a coarse thread; hexagon nuts may be supplied with either coarse or fine thread.

Wing and knurled nuts are designed for applications where a nut is to be tightened or loosened by using finger pressure only. *See* SCREW FASTENER.

Warren J. Luzadder

## Nut crop culture

The cultivation of plants, primarily trees, that produce nuts. The term nut is used loosely. Generally, a nut is defined as any edible fruit or seed enclosed in a hard shell. Botanically, a nut is a hard, inde-

hiscent, one-seeded (nut), pericarp (shell) generally resulting from a compound ovary of a flower. Indehiscent means the shell does not split open spontaneously when ripe. Examples are chestnuts, filberts, and acorns. Technically, a nut is a dry edible fruit consisting of a kernel or seed enclosed in a woody shell. Only a fraction of the 80 fruits and seeds designated as nuts fit this description. The peanut is notable because the plant is a herbaceous annual legume in which the nuts are analogous to peapods that mature underground. *See* FLOWER.

Although most nuts are spherical, they can vary considerably in size, shape, shell, and hull. Popular edible nuts range in size from the small sunflower seed to the large coconut. Nuts can be round (macadamias), elongated (pistachios and almonds), kidney shaped (cashews), or triangular (brazil nuts), or have the distinctive double-lobed kernel of the walnut and pecan.

The protective shell surface is called a hull (or husk) and serves a variety of purposes. Chestnut hulls are spiny. The coconut husk is thick and fibrous. The cashew hull contains a caustic mix of cardol and anacardic acid that causes a serious allergic reaction worse than that of its relative poison ivy. Extracts of walnut hulls at different stages of maturity have been used as yellow and brown dyes.

**Crop plants.** Plants producing nut crops are very diverse in their botanical classification, type, and climatic and cultural requirements. Pine nuts are the seeds of conifers from cool temperate climates, coconuts and brazil nuts are from a tropical palm and evergreen respectively, and pecans and walnuts are from large deciduous trees of warm temperate zones. Many nut crops were originally produced from wild trees that received little care (for example, pine nuts, brazil nuts, pistachios, and black walnuts). However, as demand increased and harvesting and processing technology improved, nut crops began to be cultivated as intensively as other fruits.

**Cultivation.** Cultural methods differ according to the nature of the plant, its soil requirements, the climate, available labor, and other factors. As with most orchard crops, commercially cultivated nut trees are now grafted onto rootstocks that allow better adaptation to the soil or climate, and disease or pest resistance. Relative to other tree fruits nut crop culture offers some distinct advantages in the ability to use machinery rather than hand labor for harvesting. Pruning of mature nut trees is not as critical as that of stone fruits because pruning does not play as strong a role in nut size, color development, or internal quality, as it does in fleshy fruits. Because most nut kernels are enclosed in a hard shell (often with a hull) and are too numerous to pick by hand, harvesting is frequently done by shaking-and-sweeping machines. Further, the time and length of harvest, processing, and storage is not as critical as with fleshy fruits. Thus, where the technology is available, nut crop culture is unique among the tree crops in being mechanically pruned and harvested. *See* AGRICULTURAL MACHINERY.

**Disease.** As with other orchard crops, pest and disease control is important in commercial plantings.

Effective control measures are related to the life history of the particular pest or disease, since each may require a different approach in cultural practice or spray program. Generally, nut crops are more amenable to organic production or integrated pest management because they have protective shells. Nut crops also have an advantage over fleshy fruits in that if a pathogen or pest does not cause economic damage by decreasing production or harm shell appearance or kernel quality, it need not be controlled. See PLANT PATHOLOGY.

**Economic value.** The edible part of most nuts, the kernel, is high in fat (30–70%) and low in carbohydrates (generally under 25%). Although nuts contain high-quality protein, they are considered primarily a source of fat and are rarely a dietary staple in modern cultures. The most common nuts, such as walnuts, almonds, pecans, and cashews, are consumed primarily as snack food or as a confectionery additive, less frequently in savory foods. Other nuts, such as the betel nuts, are masticatories, that is, nuts that are chewed but not consumed.

Not all nuts are edible or chewable. The tung nut, the source of tung oil—the active drying agent in varnish and paint with superior preservative and waterproofing qualities—is toxic. The tropical candlenut

oil tree (*Aleurites triloba*) produces a poisonous nut that is a source of oil for illuminants, soaps, paints, and wood preservatives. The nutlike seeds of *Nuxvomica* contain strychnine, which in appropriate concentrations has been used as a heart stimulant or rat poison. The crushed nuts of the Nigerian Calabar vine (*Physostigma venenosum*) were fed to persons suspected of crimes; if the suspect vomited and survived, he was judged innocent. Physostigmine, the alkaloid used to contract pupils, is extracted from the dried nuts.

Nut-bearing plants were particularly important in supplying food for early populations and wildlife. Archeological records of primitive humans are scarce, but sites from Mesolithic to recent times contain nut and shell fragments indicating nut consumption. Changes in nut-producing trees caused changes in animal populations. The disappearance of the native chestnut in the eastern United States caused a drop in the wild turkey population, and the encroachment of civilization on the supply of beechnuts and acorns contributed to the extinction of the passenger pigeon.

With increasing world population and diversity in diets, commercially produced nut crops have become economically important (see table). Political

Nuts important in world commerce

Common name (scientific name)	Plant type	Origin	Climatic zone adaptation	Principal uses and producing areas
Almond ( <i>Prunus amygdalus</i> )	Small deciduous tree	Asia Minor	Temperate, subtropical; planted	Food; Mediterranean
Brazil nut ( <i>Bertholletia excelsa</i> )	Very large evergreen tree	Amazon basin	Tropical; wild	Food; Brazil, Bolivia
Black walnut ( <i>Juglans nigra</i> )	Large deciduous tree	Central and eastern United States	Temperate; mostly wild	Food, lumber, ground-up shells; United States
Cashew ( <i>Anacardium occidentale</i> )	Evergreen tree to 40 ft (12 m)	Tropical America	Tropical; wild and planted	Food, shell oil; India, East Africa
Chestnut ( <i>Castanea</i> sp.)	Large, spreading, deciduous tree	Eastern United States, Europe, Asia	Temperate; wild and planted	Food, lumber; Europe, China, Japan
Coconut ( <i>Castanea</i> sp.)	Large palm	Probably Polynesia	Tropical; wild and planted	Food, edible and industrial oil, fiber; Philippines, tropics of Pacific
Cola nut ( <i>Cola acuminata</i> )	Evergreen tree to 40 ft (12 m)	West Africa	Tropical; wild and planted	Stimulant; West Africa
English or Persian walnut ( <i>Juglans regia</i> )	Large deciduous tree	Eastern Europe, Asia	Temperate; wild and planted	Food, oil, lumber; Mediterranean basin, California, China
Filbert ( <i>Corylus avellana</i> )	Deciduous shrub or small tree	Eastern Europe, Asia Minor	Temperate; wild and planted	Food; Turkey, Italy, Spain, northwestern United States
Macadamia nut ( <i>Macadamia ternifolia</i> )	Evergreen tree to 60 ft (18 m)	Queensland, New South Wales	Tropical, subtropical; planted	Food; Australia, Hawaii
Palm nut ( <i>Elaeis guineensis</i> )	Large palm	West and Central Africa	Tropical; wild and planted	Edible and industrial oil; West Africa, Indonesia, Brazil
Peanut ( <i>Arachis hypogaea</i> )	Annual crop plant	Brazil	Tropical, subtropical, warm temperate; planted	Food, edible oil; India, Africa, United States
Pecan ( <i>Carya pecan</i> )	Large, deciduous tree	Southern and central United States	Warm temperate; wild and planted	Food; central and southern United States
Pine nuts ( <i>Pinus</i> sp.)	Evergreen conifers	Europe, Asia, southwestern North America	Temperate; wild	Food; southern Europe, Asia, North America
Pistachio nut ( <i>Pistacia vera</i> )	Small, spreading, deciduous tree	Asia Minor	Warm temperate; wild and planted	Food; Iran, Turkey, Syria, Italy
Tung nut ( <i>Aleurites fordii</i> )	Small tree to 20 ft (6 m)	Central Asia	Warm temperate, subtropical; planted	Industrial oil, paints; China, southern United States

events have also played a role in the economic development of some nut crops. The pistachio, for example, is the most successful plant introduction to the United States in the twentieth century. California would not have developed the successful pistachio industry it now has had the world's primary pistachio producer, Iran, not been disrupted by war in the late 1970s. At that point, California had a small annual production but responded so strongly to the sharply increased world demand that pistachios have become the third largest nut crop in California. See ALMOND; BRAZIL NUT; CASHEW; CHESTNUT; COCONUT; COLA; FILBERT; MACADAMIA NUT; PEANUT; PECAN; PINE NUT; PISTACHIO; WALNUT.

Louise Ferguson

Bibliography. E. A. Meninger, *Edible Nuts of the World*, 1977; F. Rosengarten, Jr., *The Book of Edible Nuts*, 1984.

### Nutation (astronomy and mechanics)

In mechanics, nutation is a bobbing motion that accompanies the precession of a spinning rigid body, such as a top. Astronomical nutation refers to irregularities in the precessional motion of rotating bodies. A well-studied example is the irregularities in the precessional motion of the equinoxes caused by the varying torque applied to the Earth by the Sun and Moon. Astronomical nutation should not be confused with nutation as defined in mechanics; the latter is present even if the source of the torques is unvarying.

**Nutation of tops.** In simple precession, the axis of a top with a fixed point of contact sweeps out a cone, whose axis is the vertical direction. In the general motion, the angle between the axis of the top and the vertical varies with time (Fig. 1). This motion of the top's axis, bobbing up and down as it precesses, is known as nutation.

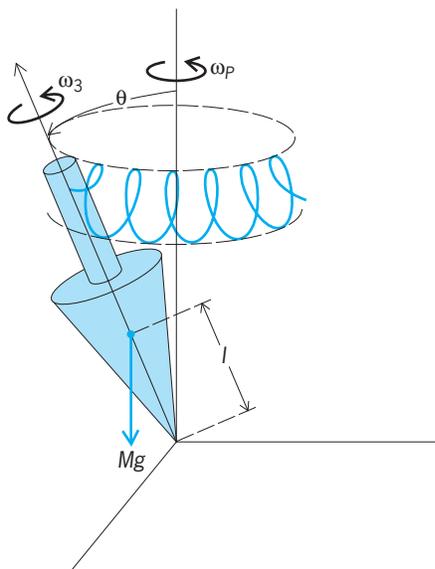


Fig. 1. Nutation and precession of a spinning top with a fixed point of contact.

Slow precession and fast nutation are commonly observed in the motion of a rapidly spinning top. For this case, the equations of motion have a simple approximate solution, which will be presented below. The angular frequency of the precessional motion is given by Eq. (1), where  $\omega_3$  is the angular velocity

$$\omega_p \simeq \frac{Mgl}{I_3\omega_3} \quad (1)$$

about the spin axis of the top,  $I_3$  is the moment of inertia about this axis,  $M$  is the mass of the top,  $l$  is the distance of the center of mass from the point of contact, and  $g$  is the gravitational acceleration. The nutation is governed by the angular frequency given in Eq. (2), where  $I$  is the moment of inertia about an

$$\omega_N = \frac{I_3\omega_3}{I} \quad (2)$$

axis that is perpendicular to the spin axis of the top.

The location of the spin axis can be specified by an azimuthal angle  $\phi$  and a polar angle  $\theta$ . Suppose that the top is started at a polar angle  $\theta(t=0) = \theta_0$  with an initial azimuthal angular velocity  $\phi(t=0) = \omega_0$ . Then the approximate solutions to the equations of motion are given by Eqs. (3), where  $C$  is given by Eq. (4).

$$\theta(t) = \theta_0 + C \sin \theta_0 (1 - \cos \omega_N t) \quad (3a)$$

$$\phi(t) = \omega_p t - C \sin \omega_N t \quad (3b)$$

$$C = \frac{\omega_p - \omega_0}{\omega_N} \quad (4)$$

This solution for  $\theta(t)$  oscillates between limits of  $\theta_0$  and  $\theta_0 + 2C \sin \theta_0$ ; the top's spin axis bobs up and down between cones of these angles. The maximal excursion in  $\theta$  depends on  $C$ , and hence can be changed by the initial conditions on  $\omega_0$  and  $\omega_3$ . The precession angle  $\phi$  has a sinusoidal time dependence determined by  $\omega_N$  in addition to the steady precession term. When the top is set in motion with  $\omega_0 = \omega_p$ , it will undergo precession with no nutation. Typical curves traced out by the end of the top's symmetry axis are shown in Fig. 2.

The nutation frequency  $\omega_N$  in Eq. (2) is proportional to the spin  $\omega_3$  of the top while the precession frequency  $\omega_p$  in Eq. (1) is inversely proportional to  $\omega_3$ . Also, the nutation amplitude  $C$  is inversely proportional to  $\omega_N$ . Thus it is difficult to observe the nutation of a very fast top. However, a buzzing tone with the nutation frequency can be heard when a fast top is spun on a resonant surface. See PRECESSION; RIGID-BODY DYNAMICS. Vernon D. Barger

**Astronomical nutation.** The rotating Earth can be regarded as a spinning symmetrical top with small angular speed but large angular momentum, the latter due to its large mass. The gravitational attractions of the Sun and Moon cause the Earth's axis to describe a cone about the normal to the plane of its orbit. However, the magnitude of these gravitational attractions is continually varying, due to the changing positions in space of Sun, Moon, and Earth. The

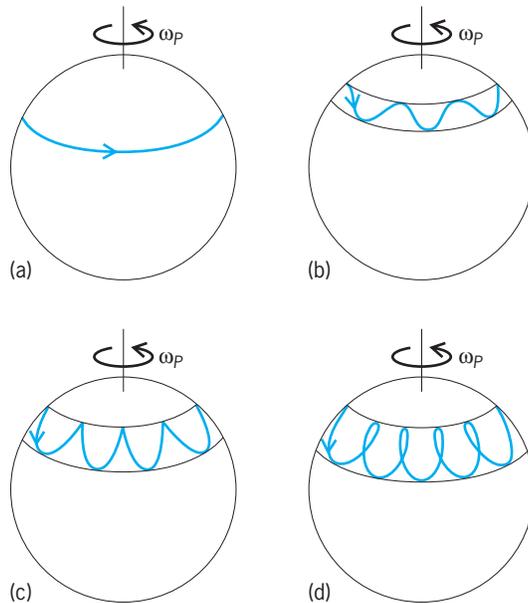


Fig. 2. Path traced out by the symmetry axis of a top with a fixed point of contact for different initial angular velocities  $\omega_0$  of the  $\phi$  motion. (a)  $\omega_0 = \omega_p$ . (b)  $\omega_0 = \frac{1}{2}\omega_p$ . (c)  $\omega_0 = 0$ . (d)  $\omega_0 = -\omega_p$ .

Moon's orbit is continually changing its position in such a way that the celestial pole undergoes a nodding (nutations) as well as a periodic variation in the rate of advance. The largest nutation is about  $9''/2$ , and occurs in a period of a little less than 19 years; that is, the celestial pole completes a small ellipse of semimajor axis  $9''/2$  in about 19 years. See PRECESSION OF EQUINOXES.

There are lesser nutation effects in the Earth-Moon system which are due to the motion of the Moon's nodes, the changing declination of the Sun, the changing declination of the Moon, and the effects of the dynamics due to the nonrigid Earth. See CELESTIAL MECHANICS.

Ray E. Bolz

The gas giants Jupiter, Saturn, Uranus, and Neptune are much more massive than their satellites and they also have no fixed surface features, so astronomical nutations of these planets cannot be observed. Pluto should show nutation caused by its satellite Charon; however, not enough data on the rotation of Pluto have been collected to detect this nutation. Although the Earth-Moon system is the only measured example of astronomical nutation, the discovery of planets circling other stars provides the theoretical possibility of other examples. Again, such effects are far beyond current or foreseeable observational abilities. See EXTRASOLAR PLANETS; PLANET; PLUTO.

John L. Safko

Bibliography. V. Barger and M. Olsson, *Classical Mechanics: A Modern Perspective*, 2d ed., McGraw-Hill, 1995; H. Goldstein, C. P. Poole, Jr., and J. L. Safko, *Classical Mechanics*, 3d ed., Addison Wesley, 2002; D. Halliday, R. Resnick, and K. Krane, *Physics*, 5th ed., Wiley, 2002; D. Hestenes, *New Foundations for Classical Mechanics*, Kluwer Academic, 1999; P. K. Seidelman (ed.), *Explanatory Supplement to the As-*

*tronomical Almanac*, rev. ed., University Science Books, 1992; K. Symon, *Mechanics*, 3d ed., Addison Wesley, 1971.

## Nutmeg

A delicately flavored spice obtained from the nutmeg tree (*Myristica fragrans*), a native of the Moluccas, or Spice Islands. The tree is a dark-leaved evergreen 30–60 ft (9–18 m) high, and is a member of the nutmeg family (Myristicaceae). The golden-yellow, mature fruits resemble apricots (see *illus.*).



Nutmeg (*Myristica fragrans*), mature fruits. (USDA)

They gradually lose moisture and when completely ripe, the husk (pericarp) splits open, exposing the shiny brown seed covered with a red, fibrous, aromatic aril which is the mace. The kernel inside the seed coat is the nutmeg of commerce. Fruits are produced throughout the year and are picked when the husks split open. The mace is removed from the husks, flattened, and dried. It is used in making pickles, ketchup, and sauces. When the seeds are thoroughly dried, the shells are cracked off, the kernels are removed and sorted, and often treated with lime to prevent damage by insects. Grated nutmeg is used in custards, puddings, and other sweet dishes, also in various beverages. Nutmeg oil is used in medicine, perfumery, and dentifrices, and in the tobacco industry. See MAGNOLIALES; SPICE AND FLAVORING.

Perry D. Strausbaugh; Earl L. Core

## Nutrition

The science of nourishment, including the study of the nutrients that each organism must obtain from its environment to maintain life and health and to reproduce. Although each kind of organism has its distinctive needs, which can be studied separately, a far-reaching biochemical unity in nature has been discovered which gives vastly more coherence to the whole subject. Many nutrients, such as amino acids, minerals, and vitamins, needed by higher organisms may also be needed by the simplest forms of life—single-celled bacteria and protozoa. The recognition of this fact has made possible highly important developments in biochemistry.

Mammals need for their nutrition (aside from water and oxygen) a highly complex mixture of more than 40 chemical substances, including amino acids; carbohydrates; certain lipids; fibers (for preventing constipation and diverticular disease and for slowing the absorption of carbohydrates); a great variety of minerals, including several that are required only in minute amounts, commonly referred to as trace minerals; and vitamins—organic substances of diverse structure that are treated as a group only because as nutrients they are required in relatively small amounts. There are also a few known conditional nutrients—essential components of cellular biochemistry that are not normally nutrients, because they are made internally in adequate amounts, but that become nutrients under stressful conditions of increased need or diminished supply. Vitamin D would not be considered a nutrient if adequate exposure of the skin to sunlight were consistently assured. Conditional nutrients are known in infancy and with injury, disease, genetic defects, and the use of some pharmaceuticals. *See* AMINO ACIDS; CARBOHYDRATE METABOLISM; LIPID METABOLISM; PROTEIN METABOLISM; VITAMIN; VITAMIN D.

Most nutrients were recognized in the nineteenth century, but the vitamins and some trace minerals did not become known as fundamental parts in the machinery of all living things until the early twentieth century. The discovery of vitamins, and some of the trace minerals, originally came about through the recognition of deficiency diseases, such as beriberi, scurvy, pellagra, and rickets, which arise because of specific nutritional lacks and can be cured or prevented by supplying the needed nutrients.

Different species of mammals have some distinctive nutritional needs. Guinea pigs, monkeys, and humans, for example, require an exogenous supply of ascorbic acid (vitamin C) to maintain life and health, whereas most animals, including rats, do not. Ascorbic acid, however, is an essential part of the metabolic machinery of animals that do not need an exogenous supply. Rat tissues, for example, are relatively rich in ascorbic acid; unlike guinea pigs, these animals are genetically endowed with biochemical mechanisms for producing ascorbic acid from carbohydrate. *See* ASCORBIC ACID.

The amounts of nutrients adequate for health are not well known for several reasons, including un-

certainities in how to define adequacy, substantial individual variations due to genetic differences, and environmental influences. Even less is known about alterations in nutritional needs caused by various disease conditions, use of pharmaceuticals, and injury. More is known about many aspects of the nutrition of livestock, laboratory animals, and pets than is known about human nutrition, and considerably greater care and expertise are exercised with prized animals than with humans, in part because animal diets can be more easily controlled.

Early workers in human nutrition focused on the minimum amounts needed to prevent or cure acute deficiency diseases, such as scurvy and beriberi. Since that time, the Recommended Dietary Allowances and Adequate Intakes (RDAs and AIs, collectively called Dietary Reference Intakes) in the United States and similar recommendations in other countries include consideration of biochemical criteria of adequacy. They also include approximate adjustments for age, sex, and pregnancy and lactation, along with rough estimates for some other sources of individual variation. However, statistical data needed to adequately assess individual variations are not yet available for any nutrient.

Interests have shifted toward what may be more nearly optimal nutritional intakes based on the amounts needed to promote health (not merely to avoid disease or biochemical deficiency), longevity, and resistance to chronic disorders, including cardiovascular disease, cancer, hypertension, and diabetes. Increasing evidence indicates that nutrients also protect against environmental pollutants and some human birth defects that formerly were not believed to be nutrition-related. Animal studies have long shown that diets adequate for youths and adults may not be adequate for good reproduction, and that deficiencies or imbalances of virtually all nutrients can cause myriad physical and mental defects.

Besides nutrients, which play essential roles in the biochemical machinery of all cells, there are thousands of additional substances in plant foods that are not essential parts of cellular biochemistry but nevertheless are beneficial to mammals. These optional dietary substances, called phytochemicals, were first recognized in the late twentieth century, and their significance is poorly understood. There are also a few known phytochemical-like substances in meat and dairy products. Many of these nonessential substances function as antioxidants. Others stimulate the immune system or production of detoxification enzymes; modulate hormones or gene expression; or act as antibacterial, antifungal, or antiviral agents. Some are pigments that plants produce to limit damage by sunlight. In humans, phytochemicals seem to help protect against heart disease, cancer, hypertension, and many other major and minor disorders. They are widely distributed in probably all plants, including fruits, vegetables, melons, berries, whole grains, beans, nuts and seeds, as well as in plant extracts such as tea, wine, and coffee.

Nutrients, conditional nutrients, and phytochemicals collectively have been called nutraceuticals,

which may be defined as substances considered to be a food, or part of a food, that provide health benefits, including the prevention and treatment of disease. Some organizations promote research and development of nutraceutical supplements, with a potentially separate legal status intermediate between nutrients and pharmaceuticals.

Because of the biochemical unity of living organisms, all whole foods (those that have not undergone processing that greatly alters their nutritional value) contain a wide assortment of the nutrients needed by humans and other mammals. Thus diets derived from a variety of whole foods tend to supply adequate amounts of all nutrients. Guided in part by some innate abilities to select proper kinds and amounts of various whole foods, all existing animal species have at least survived without scientific nutritional knowledge. However, experience shows that suboptimal nutrition prevails in all of nature, and nutrition can always be improved.

Probably all foods contain small amounts of naturally occurring toxins and antinutrients, some of which are deactivated by cooking. The possible roles of these minor components in nutrition are generally only dimly perceived. Their amounts seem mostly insignificant, and some may play useful roles. For example, dietary fiber is known to reduce the availability of some minerals, yet its presence in all plant foods is considered valuable.

For modern humans, the problems of suboptimal nutrition have increased with the advent and extensive consumption of technologically derived, refined foods. In advanced western nations, more than half of the dry weight and energy content of food supplies derives from purified sugars, separated fats, alcohol, and milled grains. These nonwhole "foods" have lost most or all of the nutrients present in the whole foods from which they derive. Although they are manufactured in a wide variety of appealing products that satisfy hunger, their excessive use facilitates various kinds of malnutrition and overconsumption that do not occur readily with whole foods. Modern dietary guidelines and nutrition education focus substantially on partially replacing nonwhole foods with whole grains, legumes, low-fat meats and dairy products, fish, vegetables, fruits, and nuts that retain their natural biochemical unity.

One of the bases for interest in nutrition is the fact that individuals who have differing genetic backgrounds have differing nutritional needs; for this reason, various human ills may arise because the individuals concerned do not get all of the nutrients in amounts compatible with their own distinctive requirements.

Every cell and tissue in the entire body requires continued adequate nutrition in order to perform its functions properly. A multitude of functions, involving the production of specific chemical substances (hormones and prostaglandins, for example) and the regulation of numerous processes, are performed by cells and tissues. Therefore, it is clear that improper nutrition may produce or contribute to almost every

type of illness. Nutritional and medical research are yielding important advances in using improved nutrition to prevent, cure, and ameliorate disease and illness. For example, diet composition has been linked to obesity, heart disease, cancer, diabetes, hypertension, osteoporosis, kidney stones, and other prominent diseases of advanced nations. Obesity long has been considered a simple result of excessive consumption of energy and fat, compounded by insufficient exercise. However, this concept is evolving. While Americans have decreased fat consumption in recent decades, and low-fat cookies and other nonwhole foods have proliferated, rates of obesity and diabetes have surged, especially in children. A new concept of obesity focuses more on the proper functioning of appetite and less on the conscious control of energy consumption. Refined carbohydrates, especially in the form of soft drinks, seem to stimulate abnormal appetite and other obesity-related ills. Whole-food sources of fat, such as nuts, avocados, and fish, appear beneficial not only for controlling obesity but also for heart disease, diabetes, and other modern disorders. The benefits of whole-food sources of carbohydrate and fat are likely highly complex, involving the chemical natures of the carbohydrate, fat, and rate of absorption, broad nutrient contents, and possibly flavors and other phytochemicals that act as appetite suppressants. The 2005 Dietary Guidelines for Americans emphasize increased consumption of whole grains, fruits, vegetables, and nuts, and decreased consumption of refined carbohydrates, solid fats, and partially hydrogenated fats. Other advances include the therapeutic use of parenteral nutrition (partial or total intravenous feeding). See DISEASE; MALNUTRITION; METABOLIC DISORDERS.

Roger J. Williams; Donald R. Davis

Bibliography. Y. Bao and R. Fenwick (eds.), *Phytochemicals in Health and Disease*, CRC Press, Boca Raton, FL, 2004; Institute of Medicine, *Dietary Reference Intakes*, National Academies Press, Washington, D.C., 6 volumes on nutrients, 1997 to 2004; M. Rechcigl, *Handbook of Naturally Occurring Food Toxicants*, CRC Press, Boca Raton, FL, 1983; M. E. Shils et al., (eds.), *Modern Nutrition in Health and Disease*, 9th ed., Lippincott Williams & Wilkins, Baltimore, 1998; M. R. Werbach, *Nutritional Influences on Illness*, 2d ed., Third Line Press, Tarzana, CA, 1996; W. C. Willett and P. J. Skerrett, *Eat, Drink, and Be Healthy: The Harvard Medical School Guide to Healthy Eating*, Simon & Schuster Source, New York, 2001; R. J. Williams, *Physician's Handbook of Nutritional Science*, C.C. Thomas, 1978.

## Nymphaeales

An order of flowering plants that have previously been included in the same larger grouping as Magnoliales, the Magnoliidae. Deoxyribonucleic acid (DNA) sequence studies have demonstrated that Nymphaeales as previously defined contain two

families, Ceratophyllaceae (water hornwort) and Nelumbonaceae (lotus), that are not closely related to the others. Remaining in the order is the family Nymphaeaceae, the waterlilies, from which a small group of tropical plants, Cabombaceae, are split by some scientists. Nymphaeaceae contain nearly 100 species of fresh-water aquatics that are typically found in river and lake systems throughout the world. The ovaries of these plants are filled with mucilage, which mediates pollen tube growth from the

stigmas to the ovules, and they have either inaperturate or monosulcate pollen. A spectacular plant is the Amazonian water lily (*Victoria*), which has leaves up to 15 ft (5 m) in diameter. The water lily family has been shown by deoxyribonucleic acid analyses to be one of the oldest lineages of flowering plants and distantly related to all others as well. The family members are relics of the early diversification of the flowering plants. See EUMAGNOLIIDS; PLANT KINGDOM; POLLEN.

Mark Chase





## Oak — Ozonolysis

### Oak

A genus (*Quercus*) of trees, some of which are shrubby, with about 200 species, mainly in the Northern Hemisphere. About 50 species are native in the United States. All oaks have scaly winter buds, usually clustered at the ends of the twigs, and single at the nodes. The fruit is a nut (acorn) surrounded at the base by an involucre, the acorn cup. The pith is star-shaped. The leaves are simple and usually lobed. See FAGALES.

Oaks furnish the most important hardwood lumber in the United States. Principal uses are for charcoal, barrels, building construction, flooring, railroad

ties, mine timbers, boxes, crates, vehicle parts, ships, agricultural implements, caskets, woodenware, fence posts, piling, and veneer. Oak is also used for pulp and paper products. See PAPER.

**Eastern oaks.** The oaks of the eastern United States are divided into two main categories, the black oak group and the white oak group.

*Black oak group.* In this group the leaf lobes are bristle-tipped (**Fig. 1**). Acorns ripen in 2 years, and winter buds are pointed.

The northern red oak (*Q. rubra*), which may attain a height of 75 ft (23 m), grows in the eastern half of the United States, except the extreme South and Southeast. It can be recognized by the large acorns

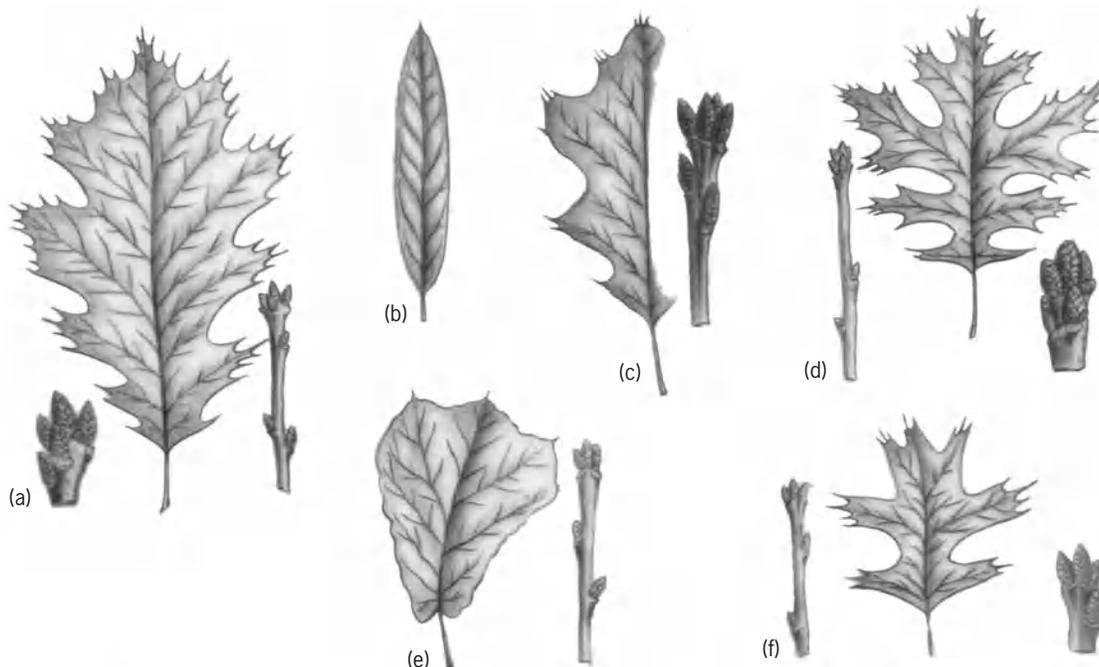


Fig. 1. Eastern oaks (black group), with bristle-tipped leaf lobes. (a) Red oak (*Quercus rubra*), terminal buds, leaf, and twig. (b) Willow oak (*Q. phellos*), leaf. (c) Black oak (*Q. velutina*), half leaf and terminal buds. (d) Scarlet oak (*Q. coccinea*), twig, leaf, and terminal buds, (e) Blackjack oak (*Q. marilandica*), leaf and twig. (f) Pin oak (*Q. palustris*), twig, leaf, and terminal buds.

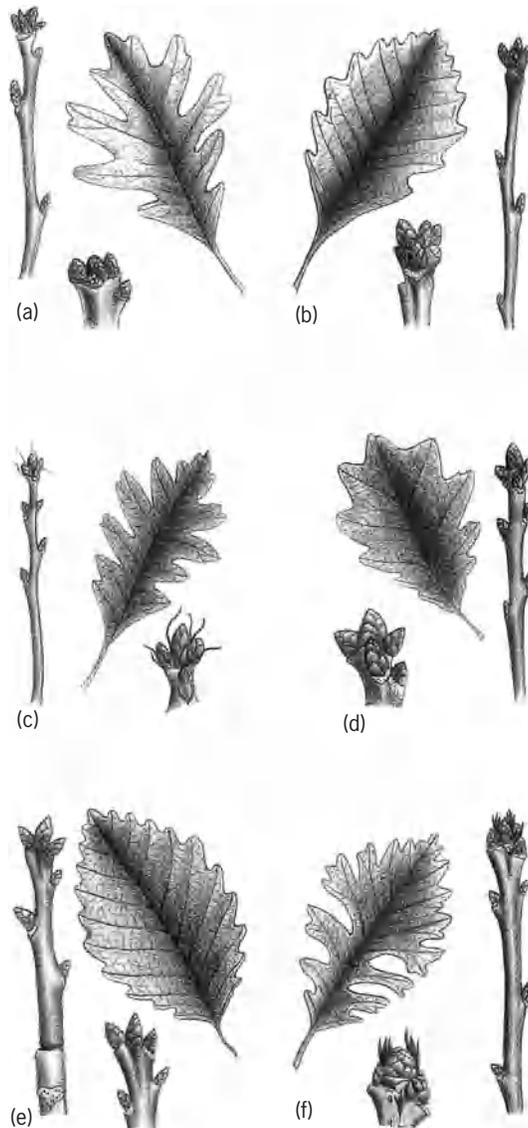


Fig. 2. Eastern white and western oaks showing leaf, twig, and terminal buds for each species. (a) White oak (*Quercus alba*). (b) Swamp white oak (*Q. bicolor*). (c) Turkey oak (*Q. cerris*). (d) English oak (*Q. robur*). (e) Chestnut oak (*Q. prinus*). (f) Bur oak (*Q. macrocarpa*).

and flattish cups and by the red, usually shiny, winter buds. In old trees the bark is comparatively smooth with shallow vertical grooves. It is the most important timber tree of the black oak group and is also a popular shade tree.

Other commercially valuable species in the eastern United States include scarlet oak (*Q. coccinea*), a highly prized ornamental tree with deeply cut leaves which turn a brilliant crimson in the fall; pin oak (*Q. palustris*), with tiny acorns, and small, deeply cleft leaves; black oak (*Q. velutina*), one of the most common species, with large five-sided, pubescent (hairy) winter buds and rough, black bark; southern red oak (*Q. falcata*), with long, often sickle-shaped leaf lobes; and blackjack oak (*Q. marilandica*), with triangular leaves.

Eastern oaks with entire or almost entire leaves are the water oak (*Q. nigra*), laurel oak (*Q. laurifolia*), and willow oak (*Q. phellos*; Fig. 1b). The lumber of

all these species is similar and usually is classed as red oak lumber.

The live oak (*Q. virginiana*), a medium-sized tree of the South Atlantic and Gulf Coast regions, has evergreen leaves.

**White oak group.** This group has rounded leaf lobes, acorns which ripen in 1 year, and winter buds which are usually rounded (Fig. 2).

White oak (*Q. alba*) furnishes the most valuable hardwood lumber of all Eastern trees. In this same group is the bur oak (*Q. macrocarpa*), a large tree of the eastern United States and adjacent Canada. Its acorns are large and edible.

The chestnut oak subgroup is characterized by leaves with numerous small rounded teeth. This subgroup is represented chiefly by the chestnut oak (*Q. prinus*), of the Appalachian Mountain and Ohio Valley regions; the swamp chestnut oak (*Q. michauxii*); and the swamp white oak (*Q. bicolor*).

**Western oaks.** In the western United States the native oaks do not have the same high commercial value as timber trees, their place being taken by the valuable western conifers. However, species of importance are the California black oak (*Q. kelloggii*), which grows in Oregon and California; the California live oak (*Q. agrifolia*), with persistent leaves, found in California and Baja California; and the California white oak (*Q. lobata*).

The English oak (*Q. robur*) and its varieties are cultivated in the United States, and have leaves of the white oak type, as well as the turkey oak (*Q. cerris*), which has shallowly lobed leaves (Fig. 2c and d). See FOREST AND FORESTRY; TREE.

Arthur H. Graves; Kenneth P. Davis

## Oasis

An isolated fertile area, usually limited in extent and surrounded by desert. The term was initially applied to small areas in Africa and Asia typically supporting trees and cultivated crops with a water supply from springs and from seepage of water originating at some distance. However, the term has been expanded to include areas receiving moisture from intermittent streams or artificial irrigation systems. Thus the floodplains of the Nile and Colorado rivers can be considered vast oases, as can arid areas irrigated by humans. See DESERT.

**River valleys.** Other well-known desert rivers include the Tigris-Euphrates in the Middle East, Hwang-Ho (Yellow River) in China, and the Amu Darya and Syr Darya in Central Asia. All of these support luxuriant, although sometimes intermittent, vegetation along their courses. Even the most extreme deserts, such as the Namib along the western coast of southern Africa and the Atacama-Peruvian Desert of western South America, have through-flowing streams rising in adjacent mountain areas. In a broad sense all of these are oases.

The very highly developed Salt River Valley in Arizona, in which the city of Phoenix is located, could

be considered an immense oasis. Its cultivated crops, including palms, fruit trees, and vegetables, have much in common with those of the oases of the Sahara. The area around Indio, California, might be considered even more like African oases because a wide variety of date palms that were imported from the Sahara produce the highest-quality dates.

On a much more modest scale are the palm oases of the Colorado and Sonoran deserts of the southwestern United States. These usually small areas, located in canyons and at the foot of mountains, are characterized by the native palm *Washington filifera*. One such oasis has been developed into Palm Springs, the well-known California winter resort area.

**Desert gardens.** The term oasis usually brings to mind the widely scattered desert gardens of the Sahara and Arabian deserts. It is commonly associated with palm trees, lush vegetation, Arabs, and camels. These oases have a character distinct from that of irrigated river valleys. They may be large and support a population of more than 500,000 such as Damascus in Syria, or they may be small, only an acre or less in extent. Formerly, most of these oases were reached only by camel caravan, but now many are accessible by air, railroad, or paved highway. Some receive their moisture from artesian sources, others from intermittent streams with either surface or underground flow. Some are watered by deep wells or by horizontal wells called foggaras by the Arabs, kánats or ganats by the people of southeastern Asia, and karez by the Mongolians.

**Water-collection systems.** A foggara is made by digging a ditch into the desert on a gentle slope starting at the edge of a basin or dry stream bed. When the ditch becomes too deep to be maintained as an open course, it becomes a tunnel which is extended outward, with wells constructed to the surface at intervals. These chains of wells connected by the tunnel reach into moist sand. Water collects in the tunnel and flows to the basin, where it may be used for irrigation or for human or animal consumption.

Various methods are used to take water from wells. In the shallow wells of the Gobi Desert, water can be reached with a short rope to which a stick and a bag are attached. The well man pulls up the rope until he can reach the stick, which he uses to lift the bag, and then pours water into a trough for animals or into wooden casks for human use.

Well sweeps are common in the Sahara. These consist of a long pole swung between upright posts near the well. The butt of the pole is weighted with rocks or other heavy material. At the well end of the pole is a rope and bucket with which water can be lifted from the well.

The use of pulleys is common at most oases. The rope and pulley may be operated by donkey, ox, or camel power. The water bucket or bag has a long spout that folds back when the bag is lowered and is straightened out by a short rope at the well curb so that water flows into ditches or storage ponds. Modern pumps operated by windmills or engines are less common.

**Climate and soil.** Oases are restricted to climatic regions where precipitation is insufficient to support crop production. Such regions may be classified as extremely arid (annual rainfall less than 2 in. or 50 mm), arid (annual rainfall less than 10 in. or 250 mm), and semiarid (rainfall less than 20 in. or 500 mm). Many African and Asian oases are in extremely arid areas. Most oases are found in warm climates. Some authorities restrict the use of the term to regions where date palms grow and produce. This means long frost-free growing seasons.

Oasis soils are weakly developed, high in organic matter but often saline, and have been strongly affected by human occupation.

**Economics and agricultural practices.** The date palm (*Phoenix dactylifera*) is the tree of greatest economic importance. The living tree provides shade, the fruit is eaten by the Arabs, and the cracked seeds are fed to the camels. In winter dead leaves are cut out and the dead palm leaves used for fuel for cooking fires. Trunks of old trees are used for houses and serve as simple bridges across irrigation ditches. The fiber at the base of the leaves is used to make rope and a course cloth.

Date palms are planted about a rod apart, and the spaces between are planted with cereals or garden vegetables. The plots are usually small, often not more than a few square feet. Each property may be surrounded by mud walls, which often have broken glass or cacti on top to discourage intruders. Under the date palm canopy may be found apricot, fig, mulberry, peach, pear, and orange trees together with the pomegranate and table grape. Vegetables include artichokes, beans, carrots, melons, peas, potatoes, squashes, and radishes. Cereals include wheat, barley, and rye, and in some oases alfalfa is produced. Various ornamental plants are also grown. Occasionally the vine-covered tree canopy is so dense that little sunlight reaches the ground.

**Habitation.** Oasis habitations are located at the edge of the gardens, on slopes above the flat plain where topography permits. The flat-roofed houses are made from dirt and may be whitewashed. Most have walled courtyards. In the Sahara there is usually a central square with a mosque. In the larger oases there are shops around the square, but many of the smaller oases depend for their needs upon itinerant traders, who congregate on weekly market days.

**Natural vegetation.** Among the native plants restricted to waste areas and abandoned plots are a wide variety of phreatophytes, including tamarisk (*Tamarix* sp.), cattail (*Typha latifolia*), sedges, and rushes; farther away many halophytes are found where drainage is poor; and beyond this fringe, desert vegetation takes over.

**Geographic distribution.** Some of the better-known oases are listed below.

Talfilalet, Morocco, has an area of 200 mi<sup>2</sup> (518 km<sup>2</sup>) and is reputed to be the largest oasis in the Sahara.

Algeria has a number of well-known oases. Biskra in northeastern Algeria is a dateshipping center and health resort; El Goléa in the central region is noted

for its gardens; Ghardai; a has a mosque with a minaret 300 ft (91 m) high; Laghouat is noted for its luxuriant vegetation, and the vine-covered palms make a nearly sunproof canopy; the Ouargla Oasis in east central Algeria is an ancient rambling town reputed to have had 500,000 date palms in earlier times; the Touggourt area, including the Souf and Oued Rhir oases, is noted for its Deglet Nur dates; the Tidikelt and Touat oasis region in central Algeria includes the Aoulef, In-Salah, and Tit oases, all of which are important date producers.

Tozeur Oasis, in Tunisia, is 110 mi (177 km) west of Gabès and is a producer of dates, olives, and vegetables.

Kufra Oasis, in Libya, is a camel-breeding center.

Dakhla is the most populous (21,000) oasis in Egypt, with 40 mi<sup>2</sup> (104 km<sup>2</sup>) in cultivation; the Kharga Oasis in south-central Egypt is known as the Great Oasis; the Siwa Oasis in the Libyan Desert of western Egypt is 100 ft (30 m) below sea level.

In Saudi Arabia, Hasa Oasis (Hofuf is the principal city), located along the Persian Gulf and extending some distance inland, is the largest oasis in the Arabian Peninsula; the Qatif Oasis, which is also on the Persian Gulf, is smaller but is an important agricultural center.

Damascus, Syria, and the surrounding area with over 600,000 people may be the most populous oasis in the Afro-Asian area.

William G. McGinnies

Bibliography. A. D. Abrahams and A. J. Parsons (eds.), *Geomorphology of Desert Environments*, 1993; S. Aritt, *The Living Earth Book of Deserts*, 1993; J. Flegg, *The Deserts*, 1993; W. G. McGinnies, B. J. Goldman, and P. Paylore, *Deserts of the World: An Appraisal of Research into Their Physical and Biological Environments*, 1968.

## Oats

An agricultural crop grown for its grain and straw in most countries of the temperate zones of the world. In the major oat-growing states of the midwestern United States (Iowa, North Dakota, South Dakota, Minnesota, and Wisconsin) the crop is raised for grain, whereas in the Southern states (Texas, Oklahoma, and Georgia) it is used for pasture or a combination of pasture and grain. About 90% of the annual oat grain production is used for animal feeds, and about 10% is processed into food for humans, for example, oatmeal and other cereal products. See ANIMAL FEEDS; CEREAL.

Among the cereal grains grown in the United States, oats are exceeded in importance only by corn, wheat, and sorghum. See CORN; SORGHUM; WHEAT.

In general, oats are a cool-season crop which requires a moist climate. They grow well on both light and heavy soils if sufficient moisture and fertility nutrients are available. In the Central and Northern states, oats are spring-sown; but in the Southern states they are fall-sown. In the Corn Belt they are grown in crop rotation with corn, soybeans, and forages.

**Origin and description.** Oats probably became an agricultural crop about 2000 years ago, most likely starting in the Mediterranean area. The fifteen species of oats in the genus *Avena* are divided into three groups on the basis of chromosome number: 14, 28, or 42. Of the seven species that make up the 14-chromosome group, only *A. strigosa* is grown commercially on small agricultural acreage in Portugal and Brazil. The two 42-chromosome species, *A. fatua* and *A. sterilis*, grow in natural stands in the eastern Mediterranean area and provide wild pastures for sheep and goats. Seeds of both wild species shatter at maturity to ensure annual reseeding.

Crosses among the 28-chromosome species and among the 42-chromosome species are easy to make and the hybrids are fully fertile, but only certain 14-chromosome species can be intercrossed. Crosses among species with different chromosome numbers can be made also, but usually the hybrids are sterile or only partly fertile. See BREEDING (PLANT); GENETICS; REPRODUCTION (PLANT).

Within the 42-chromosome cultivated species there is wide variation among varieties for all plant traits. Oats belong to the Graminae (grass) family; thus the oat plant forms a crown at the soil surface from which a fibrous root system penetrates the soil. Most of the roots are concentrated in the upper foot of soil, but some grow to a depth of 5 ft (1.5 m). Under thick seeding only one or two culms develop, but when plants are spaced, 10–30 culms may develop. Culms usually grow 2–5 ft (0.6–1.5 m) tall, and they are terminated with inflorescences called panicles. Each panicle usually bears 10–75 spikelets on its numerous branches (Fig. 1). A spikelet is enclosed by two papery glumes and bears two or three florets, each with an ovary, two stigmas, and three anthers enclosed in a lemma and palea (Fig. 2). The flowers are normally self-pollinated but 1–2% outcrossing may occur. The stem has 7–9 nodes, and a leaf grows at each node. Internodes of the stems are hollow. See CYPERALES; FLOWER; GRASS CROPS; INFLORESCENCE.

In most varieties the lemma and palea adhere to the oat seed after threshing. A trait used to determine market grade of oats is the color of the lemma, which may be white, yellow, gray, brown, red, or black. The



Fig. 1. Oat panicles with many branches and spikelets.

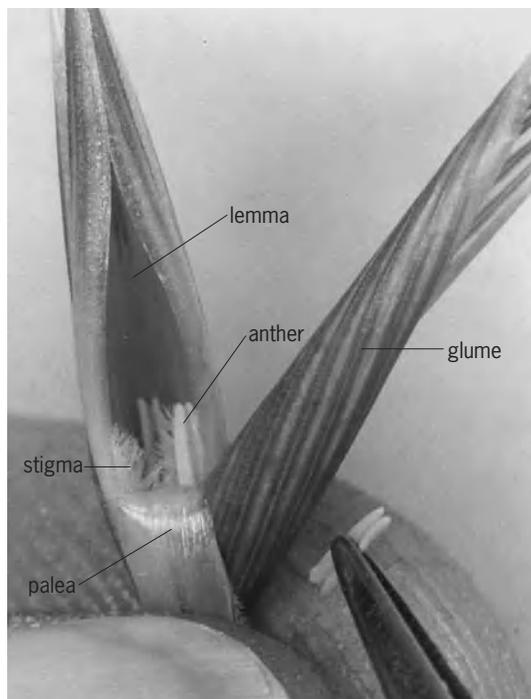


Fig. 2. An oat flower; two oat grains are shown at the lower right.

major trait that distinguishes wild from cultivated oats is seed shattering. In cultivated species the seed attachment is persistent, and it can be separated from the panicle only by threshing.

**Varieties.** The world collection of oats, maintained by the U.S. Department of Agriculture, contains more than 14,000 lines of 42-chromosome types. These represent lines from wild species and from varieties produced at breeding stations. The collection represents a vast range of genetic types that can be used for varietal improvement. After 1960, breeders put special emphasis on improving lodging and disease resistance, grain quality, and yield of new oat varieties. These varieties have been developed by crossing strains from the world collection to obtain better combinations of genetic traits. A variety called "multiline" was developed to control rusts. While each multiline variety is uniform for all agronomic traits, it contains several genes for resistance to a single disease.

**Cultural practices.** In the Corn Belt, oats usually follow corn or soybeans in crop rotation. They are used as a companion crop for forage seedings; that is, alfalfa and oats are planted simultaneously, and the oats are harvested for grain in the first year and alfalfa for hay in subsequent years. Spring oats are seeded in March or April at a rate of 2–3 bu/acre (0.2–0.3 m<sup>3</sup>/hectare) on disked or plowed land. Winter oats are sown in October or November. Mature oats are harvested with a combine, either direct or after being windrowed to dry. For safe storage oat grain should contain no more than 13% moisture. Straw of harvested oats is either worked into the soil for humus or baled and stored for use as bedding for

livestock. See AGRICULTURAL SOIL AND CROP PRACTICES; HUMUS; LEGUME FORAGES.

**Uses.** Oat grain usually contains 10–16% protein, which makes it especially good for rations for young livestock and for human food. Some strains of *A. sterilis* have up to 30% protein. Dehulled oat seeds are used extensively for making breakfast cereals, such as porridge made from rolled oats. In general, oat grain contains adequate quantities of minerals and B vitamins for normal diet. The fat content is low. Oat hulls are used to make furfural, an important chemical in nylon manufacturing.

Kenneth J. Frey

**Diseases.** Under certain conditions, any major oat disease can reduce yield and quality of oat grain over large areas. Individual oat fields may be very severely damaged or, in some cases, destroyed.

**Rusts.** Crown rust (caused by *Puccinia coronata*) is probably the most destructive oat disease. Losses vary greatly from year to year, ranging from negligible to more than 30% of the potential crop over large areas. The uredial or repeating stage (the one that does the damage; Fig. 3) first appears on the oat leaves several weeks before ripening of the grain. The bright orange, powdery uredial pustules produce spores that are blown to other plants, in which they produce a new generation every 8–10 days. In warmer areas such as the Gulf Coast of the United States, the pathogen maintains itself in the uredial stage throughout the winter on winter-grown oats. Spores then blow north as spring-sown oats begin to grow in the North. An alternative, "complete" life cycle occurs in climates having cold winters where buckthorn (*Rhamnus cathartica*) is present. As the oats mature, small black lesions called telia appear. Spores in the telia survive cold northern winters and germinate in spring to form another spore type that infects young buckthorn leaves, but not oats. Still another spore type, produced on the buckthorn (Fig. 4), can infect oats but not buckthorn. Blown to oats, they produce uredia to complete the life cycle. Stem rust (caused by *P. graminis*), which also can cause severe losses, is distinguished from crown rust by its brick-red uredia, which are more common on the stem and leaf sheaths (Fig. 5), and by the fact that its alternate host is the common barberry (*Berberis*).



Fig. 3. Susceptible oats (foreground) killed by crown rust, and oats resistant (background) to crown rust.

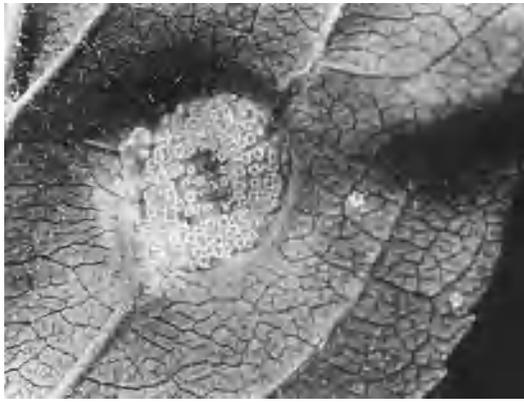


Fig. 4. Crown rust on leaf of alternate host buckthorn (*Rhamnus cathartica*).



Fig. 5. Stem rust on oat stems. (Photograph by J. A. Browning)

Both the crown- and stem-rust fungi consist of many pathogenic strains or races that differ only in their ability to attack different oat varieties. Starting in the 1920s, single genes conferring high resistance were used to develop resistant oat varieties; however, new races of fungus capable of parasitizing the new varieties then developed. The use of other resistance genes in newer varieties was followed by the appearance of still more new rust races. Several approaches are used to break this cycle: (1) Oat varieties that carry not one but several genes for resistance have been developed. (2) Varieties have been developed which consist of a mixture of visibly similar lines, each of which contains a different gene for resistance; this approaches the diversity that protects natural populations. (3) General resistance controlled by large numbers of minor-effect genes has been used; this is difficult in breeding programs but may be longer-lasting than other types of resistance. (4) Tolerance has also been used, tolerance being

the ability of a variety that appears susceptible to withstand infection without serious loss of yield or quality. (5) Geographical deployment of genes has been used; this calls for the use of resistance genes in the South different from those used in oat varieties grown in the North—the rust fungi moving north in the spring would confront oat varieties having different effective resistance genes than the varieties on which the spores had been produced.

Eradication of the alternate hosts (buckthorn and barberry) will reduce early spring inoculum and slow the appearance of new races, because genetic recombination of the fungus occurs on these hosts. Fungicides are effective but not economical.

*Smut.* Unlike the rusts, smut (*Ustilago* sp.) can cause serious losses in drier areas. In certain years, over half the panicles in individual fields have been smutted, and losses over larger areas have averaged as much as 20%. Smut appears as dark brown to black spore masses in place of the seed (Fig. 6). The spores are blown to young seeds on healthy plants and germinate, and the fungus remains dormant until the seed is planted. The fungus infects the plant as it emerges from the seed and eventually produces spore masses in place of the seed to complete the cycle. The smut fungus, like the rust fungi, consists of pathogenic strains. Smut can be controlled by use of resistant varieties or chemical seed treatment.

*Other fungi.* A number of other fungi, including species of *Septoria*, *Helminthosporium*, and *Fusarium*, occasionally cause economic losses. Some varieties have a degree of genetic resistance, and good cultural practices are recommended.

Oat roots and crowns also are attacked by a number of fungi, which may destroy very young plants and damage roots of older plants. The extent of such losses is not clear, but plants grown under optimum conditions of temperature, water supply, and fertility are usually not seriously affected. Seed treatment and use of moderately resistant varieties may also be beneficial.

*Bacteria and viruses.* Bacterial blight (caused by *Pseudomonas* sp.) usually appears early in spring as halo-like lesions on the leaves. It often looks serious, but



Fig. 6. Smutted (left) and healthy oats. (Photograph by J. A. Browning)

significant reduction in yield is rare. Yellow dwarf, however, caused by the aphid-transmitted barley yellow dwarf virus, rivals the rusts in economic importance. Infected plants are dwarfed, turn yellow or reddish, mature early, and are partially sterile. Controlling the disease by controlling its aphid vectors is generally not practical. A high degree of resistance to yellow dwarf is not known, but the use of varieties having moderate resistance will usually give satisfactory control.

*Gray speck.* Several nonparasitic conditions are regarded as diseases of oats. The most common is gray speck, caused by manganese deficiency. It can be controlled by the application of manganese salts to the soil or to the foliage. The use of more tolerant varieties will generally control gray speck. *See* PLANT PATHOLOGY.

M. D. Simons

**Processing.** The milling of oats is less complex than wheat milling and has many similarities to rice or barley milling operations because there is limited fractionation of the kernel. The oat grain is covered with a coarse, adhering hull (about one-quarter of the kernel by weight) which must be removed prior to production of ingredients or consumer foods. Oats as received at the mill house are termed green oats and must be cleaned to remove foreign seeds and trash. Cleaning requires the use of air aspiration to remove low-density materials, reciprocating sizing sieves to segregate oat grains by varying slot width, and continuous disk separators. Disk separators are a series of specialized horizontal rotating disks, each with indentations which are designed to pick up seeds smaller than the oat kernel, leaving clean oats in stream. Clean, sound oats are heated slowly prior to hull removal. The green oats have active lipolytic enzymes (lipases) which will catalyze hydrolysis of triglycerides and yield free fatty acids. This hydrolytic process results in the development of rancid off-flavors which greatly reduces acceptability and the shelf-life stability of processed oat products. Conventional milling of green oats will result in complete hydrolysis of the lipid by lipases within 3 days.

**Preheating and dehulling.** The heating, drying, or roasting procedures inactivate lipases, facilitate hull removal, and impart a distinctive roasted flavor to the oat product. Oat drying is commonly accomplished by heating the grain to approximately 200°F (93°C) for about 1 h. Steam-jacketed open-pan dryers are used in series in which the grain is surface heated, mixed, and conveyed across a series of pans to provide a uniform controlled heat process. Such heating reduces grain moisture content by 3–4%, inactivates enzymes, improves hull separation, and develops a roasted flavor. Alternately, direct steam heating (atmospheric pressure) for 2–3 min will reduce enzyme activity. Excessive steaming can, however, promote undesirable oxidative rancidity.

Roasted oats are air-cooled and size-graded prior to dehulling. Classical dehullers used for oats consist of precisely adjusted rollers; however, modern dehulling systems utilize an impact dehuller which imparts centrifugal acceleration to the sized grain and hull shattering on a rubber impact ring. In the

dehuller, grain enters the top center of the chamber, is accelerated on a high-speed finned rotor, and is discharged to impact on the peripheral rubber ring. Higher yields of whole groats are obtained with less input energy with these modern impact dehullers. Dehulling conditions may be precisely controlled to yield optimum hull separation. Under the most sophisticated system, impact dehullers can be adjusted to produce good-quality kernels without prior heat treatment.

The products of the dehuller are primarily the whole kernel or groat and the fiber hull which are readily separated by air aspiration. The low-density oat hulls possess particularly high levels of fiber and pentosans which are suitable feedstock for industrial production of furfural through high-temperature acid hydrolysis and dehydration. Whole, cleaned oat groats are not frequently available in the commercial food market. Selected large-sized groats are excellent for puffing into ready-to-eat cereals.

**Groat flaking.** Oat cutting and flaking procedures provide more extensive utilization of groats in the form of rolled oats. Steel cutting of the groat is accomplished by passage between a rotating pick-up drum and knives to yield two to four uniform pieces and a small amount of flour. These sieved pieces are then steamed and flaked with heavy rollers while moist, hot, and pliable to provide thin rolled oat flakes. The flakes are subsequently air-dried to assure shelf stability. Cooking time of rolled oats used as a hot breakfast cereal is dependent on the extent of presteaming and the final thickness of the flake. Thin flaked oats acquire an increased surface area (relative to weight) which facilitates hydration and reduces cooking time. Steamed whole groats may be flaked to a greater thickness to produce a so-called old-fashioned rolled oat requiring about 5 min cooking time. Quick-cooking rolled oats are prepared from steamed, steel-cut groats and are rolled to a thin flake (50% thinner than whole groat flakes), and require about 1 min boiling to prepare a palatable product. Instant oat cereals, requiring only the addition of boiling water and mix-in-the-bowl stirring to dissolve, are prepared with very thin rolled flakes and formulated with guar gums. The thin rolled cooked oat flakes rapidly hydrate, and the cereal obtains its viscous cooked consistency from guar gum, a highly branched galactomannan. *See* GUM.

Oat flour is obtained from further reduction and sieving or hammer milling of the whole groat or flaked product. This high-protein flour is frequently used in the formulations of ready-to-eat cereals and many prepared baby foods. Milled oat flour has appreciable antioxidant activity, similar to the phenolic antioxidants propyl gallate and butylated hydroxytoluene (BHT) and can be used as an ingredient to extend shelf stability. Composite flours blended from oat flour and other cereals providing high protein content and extended shelf life have been proposed as suitable for world feeding programs. *See* FOOD MANUFACTURING.

Mark A. Uebersax

**Bibliography.** H. J. Brounlee and F. L. Genderson, Oats and oat products: Culture, botany, seed

structure, milling, composition, and uses, *Cereal Chem.*, 15:257–272, 1938; R. C. Hosney, *Principles of Cereal Science and Technology*, 2d ed., 1994; N. L. Kent and A. D. Evers, *Technology of Cereals: An Introduction for Students of Food and Science Agriculture*, 4th ed., 1994; K. Lorenz and K. Kulp (eds.), *Handbook of Cereal Science and Technology*, 1990; J. H. Spitz, *Crop Physiology and Cereal Breeding*, 1991; P. J. Wood (ed.), *Oat Bran*, 1993.

## Obesity

The presence of excess body fat. The great prevalence of this condition, its severe consequences for physical and mental health, and the difficulty of treating it make the prevention of obesity a major public health priority.

**Definition.** The concept of excess body fat implies comparison to an ideal level of body fat associated with optimal health and longevity. Such an ideal has not been defined, however, primarily because accurate and convenient measures of body fat content are unavailable. Instead, obesity is most often defined in terms of body weight relative to height, since both height and weight are easily measured. Obesity is considered to begin at a weight-for-height that is 20–30% above desirable weight, with this desirable weight taken as the midpoint of ranges of weight associated with the greatest longevity in studies of life-insured individuals. See ADIPOSE TISSUE.

In population surveys, obesity is defined as a body weight that meets or exceeds the 85th percentile of the Body Mass Index (BMI), an index of weight-for-height that correlates well with body fat content. The BMI is calculated as weight in kilograms divided by height in meters squared for men and to the power 1.5 for women. By using the range of BMIs for adults aged 20–29 as a standard, 34 million (about 25%) of all adults in the United States aged 20 to 74 are obese, and severe obesity—at or above the 95th percentile of the BMI—affects nearly 10% of the adult population.

The prevalence of obesity increases with age, is higher in women than men, and is highest among the poor and minority groups. Severe obesity has become more common in the United States since the 1970s, but trends in overall prevalence are uncertain; inconsistencies in definitions of overweight used in national surveys have made valid comparisons difficult. See AGING.

**Risks.** Obesity increases the likelihood of high blood cholesterol, high blood pressure, and diabetes, and therefore of the diseases for which such conditions are risk factors—coronary heart disease, stroke, and kidney disease. It also increases the likelihood of gallbladder disease and cancers of the breast and uterus. Thus, obesity increases overall mortality rates, and it does so in proportion to the degree and duration of overweight. Individuals who become obese at the earliest ages are at highest risk of premature mortality. Distribution of excess fat to the upper

body rather than the lower body may also increase risk.

The precise levels of overweight at which health risks begin to increase have not been established. Studies designed to define such ranges have been limited by inadequate numbers of subjects, study periods that are too short, or failure to control for the consequences of cigarette smoking, chronic disease risk factors, weight loss due to disease, or duration of overweight. Despite these limitations, it appears that weight-for-height ranges associated with the longest survivals are significantly below average weights in the population.

Psychosocial consequences of obesity are also of concern. Obese individuals are subjected to negative social attitudes and to discrimination in employment, housing, education, and health care. This is perhaps because of mistaken beliefs that people become and remain overweight because of poor self-control.

**Causes.** The causes of most cases of obesity are poorly understood. At the simplest level, obesity results from an excess of energy (caloric) intake over expenditure, but this statement does not explain why some individuals can eat as much as they like without gaining weight while others remain overweight despite constant dieting. Studies of genetically obese animals and those with damage to the part of the brain called the hypothalamus suggest that individuals may balance body weight around a “setpoint” that is maintained—without conscious control—by variations in metabolic rate in response to caloric intake. When too many calories are consumed, the metabolic rate increases and burns off the excess energy as heat; when calories are restricted (as in dieting), the metabolic rate decreases to compensate. Although the existence of setpoints is unproven, this theory offers a testable hypothesis to explain why most obese persons find weight loss so difficult to maintain. The factors that might establish setpoint levels are undefined but are most likely to include inherited metabolic characteristics, level of caloric intake, and level of physical activity, alone or in combination. See METABOLIC DISORDERS.

Variations in the prevalence of obesity among population groups suggest a genetic basis for the condition. To eliminate the possibility that the differences are due to cultural or other environmental factors, investigators have compared body weights in biological and adopted children and in identical twins. Although the results of these studies have not always been consistent, they generally indicate that the body weights of children resemble those of the biological parents more than the adoptive parents and that the weights of identical twins are more similar than those of other siblings regardless of who raised them.

The complexity of body-weight regulatory mechanisms suggests that obesity is not due to a single cause but, like other chronic diseases, is multifactorial in origin. Specific inherited differences that might influence setpoints include differences in nearly every anatomic, neurologic, and biochemical factor known to affect food intake and utilization,

energy metabolism, and energy expenditure. See ENDOCRINE MECHANISMS; ENERGY METABOLISM.

Although excessive caloric intake might seem to be an obvious cause of obesity, it has not been possible to demonstrate that overweight individuals consume more calories (relative to body weight) than people of normal weight. Most surveys indicate that obese individuals consume fewer calories for their weight than lean individuals, an observation at least partially explained by the fact that it takes fewer calories to maintain fatty tissue than lean muscle tissue. Suggestions that the obese display impaired control of hunger and satiety, respond inappropriately to external hunger signals (for example, taste, time of day), or demonstrate abnormal eating patterns do not seem to be generalizable.

Obese individuals from earliest infancy to old age display reduced energy expenditures. Population studies also reveal an association between sedentary life style and obesity, but it is uncertain from these observations whether reduced activity is a cause or a consequence of being overweight. Evidence that moderate exercise exerts a regulatory action on body weight beyond its immediate effect on caloric output is incomplete. Exercise raises the metabolic rate and increases the proportion of lean body tissue, and clinical studies demonstrate that highly active people are leaner even if they consume more calories. Spontaneous physical activity—fidgeting—has been identified as a potentially significant factor in obesity prevention. Although more information is needed about the relationship between caloric intake, caloric expenditure, and body weight, it seems evident that inadequate caloric expenditure deserves at least as much attention as excessive caloric intake in obesity causation.

**Treatment.** Because the causes of obesity are incompletely understood, it is difficult to formulate effective treatment strategies. Weight loss requires a decrease in caloric intake, an increase in caloric expenditure, or both. Typically, treatments have emphasized reductions in caloric intake by diet, induction of more restrictive eating patterns by behavior modification, control of hunger by drugs, or limitations on food intake by surgery. None of these methods is free of cost or risk, and none has been outstandingly successful in inducing long-term weight loss; the great majority of obese people who attempt to lose weight eventually gain it back. The costs, risks, and ineffectiveness of traditional weight-loss methods explains why increased physical activity is becoming a major focus of obesity prevention and treatment. Preliminary studies suggest that programs combining diet and exercise help obese individuals lose more weight and maintain losses longer than either program does separately. A rational approach to prevention is the adoption of a life style that includes considerable physical activity along with a diet that contains adequate but not excessive calories. See FOOD; NUTRITION. Marion Nestle

Bibliography. K. D. Brownell and J. P. Foreyt (eds.), *Handbook of Eating Disorders: Physiology, Psychology, and Treatment of Obesity, Anorexia, and*

*Bulimia*, 1986; J. E. Manson et al., Body weight and longevity: A reassessment, *J.A.M.A.*, 257:353–358, 1987; National Institutes of Health Consensus Development Panel, Consensus conference statement: Health implications of obesity, *Ann. Int. Med.*, 103:1073–1077, 1985; E. Ravussin et al., Reduced rate of energy expenditure as a risk factor for body-weight gain, *N. Engl. J. Med.*, 318:467–472, 1988; A. J. Stunkard and E. Stellar, *Eating and Its Disorders*, Association for Research in Nervous and Mental Disease, vol. 62, 1984; T. A. Wadden and A. J. Stunkard, Social and psychological consequences of obesity, *Ann. Int. Med.*, 103:1062–1067, 1985.

## Object-oriented programming

A computer-programming methodology that focuses on data items rather than processes. Traditional software development models assume a top-down approach. A functional description of a system is produced and then refined until a running implementation is achieved. The refinement process breaks each system function down into subfunctions which combine to implement the higher-level function. The emphasis of the approach is on the transformation of inputs to outputs. Data structures (and file structures) are proposed and evaluated based on how well they support the functional models.

The object-oriented approach focuses first on the data items (entities, objects) that are being manipulated. The emphasis is on characterizing the data items as active entities which can perform operations on and for themselves. It then describes how system behavior is implemented through the interaction of the data items.

**Basic concepts.** The essence of the object-oriented approach is the use of abstract data types, polymorphism, and reuse through inheritance.

Abstract data types define the active data items described above. A traditional data type in a programming language describes only the structure of a data item. An abstract data type also describes operations that may be requested of the data item. It is the ability to associate operations with data items that makes them active. The abstract data type makes operations available without revealing the details of how the operations are implemented, preventing programmers from becoming dependent on implementation details. The definition of an operation is considered a contract between the implementor of the abstract data type and the user of the abstract data type. The implementor is free to perform the operation in any appropriate manner as long as the operation fulfills its contract. Object-oriented programming languages give abstract data types the name class. See ABSTRACT DATA TYPE.

Polymorphism in the object-oriented approach refers to the ability of a programmer to treat many different types of objects in a uniform manner by invoking the same operation on each object. Because the objects are instances of abstract data types, they may implement the operation differently as long as

they fulfill the agreement in their common contract. The process of finding the implementation of an operation for an object is referred to as dynamic binding (various languages implement dynamic binding in different ways).

A new abstract data type (class) can be created in object-oriented programming simply by stating how the new type differs from some existing type. A feature that is not described as different will be shared by the two types, constituting reuse through inheritance. Inheritance is useful because it replaces the practice of copying an entire abstract data type in order to change a single feature. Copying abstract data types to make small changes causes problems because it unnecessarily duplicates program code and causes an increase in maintenance effort. Reuse through inheritance changes only the features that are different and remembers the commonality between types.

The elements of object-oriented programming have much in common with automobiles. Drivers operate automobiles without detailed knowledge of their internal implementation. They do not need to know the details because they have an agreement with the automobile manufacturers on how cars should behave. A second commonality is that most drivers can operate most cars because all cars have essentially the same interface, even though different kinds of cars implement the operations differently. This characteristic is polymorphism. The third commonality is found by noticing that different kinds of automobiles often are not greatly different from one another. An upscale car may simply be a version of a basic car with more expensive options and a different body style. Because there is so much in common between the them, the upscale model may be described simply by stating how it is different from the basic model. A feature that is not described as different can be assumed to be the same. This characteristic is reuse through inheritance.

**Classes as abstract data types.** In the object-oriented approach, a class is used to define an abstract data type, and the operations of the type are referred to as methods. An instance of a class is termed an object instance or simply an object. To invoke an operation on an object instance, the programmer sends a message to the object.

Example (1) illustrates an Employee class with a simple set of operations.

```
CLASS Employee =
  METHODS
    FUNCTION name : string;
    FUNCTION Position : string;
    ...
END; (1)
```

This declaration indicates that a programmer who holds an employee object can ask the employee for its name or its position.

**Inheritance.** As noted above, inheritance allows a programmer to define a new class simply by stating

how it differs from an existing class. The new class is said to inherit from the original class and is referred to as a subclass of the original class. The original class is referred to as the superclass of the new class.

Example (2) illustrates two classes which inherit

```
CLASS Hourly_Employee INHERITS FROM
  Employee =
  METHODS
    FUNCTION hourly_salary : monetary;
    FUNCTION hours_this_month : integer;
    FUNCTION wages : monetary;
END; (2)

CLASS Salaried_Employee INHERITS FROM
  Employee =
  METHODS
    FUNCTION salary : monetary;
    FUNCTION wages : monetary;
END;
```

from the Employee class described above. The classes extend the Employee class by providing different kinds of salary information and a way to calculate wages from that information. An instance of the Salaried-Employee class can respond to the salary or wages messages and can also respond to the name or position messages because Salaried-Employee inherits from the Employee class.

**Polymorphism.** As noted above, polymorphism allows different kinds of objects to be treated in a uniform manner. For example, both an Hourly-Employee object instance and a Salaried-Employee object instance can respond to the name message (inherited), the position message (inherited), or the wages message. Someone who wants to know wages does not need to know which kind of employee object is involved because either will respond to these messages.

**Programming languages.** The object-oriented languages in most widespread use are Smalltalk and C++. Smalltalk is an interpreted language which presents a very uniform object model in which even elementary data types are modeled as classes. The two basic operations of the language are assignment and sending a message to an object. Variables have no declared type and may refer to an instance of any class at run time. This approach provides for great flexibility in programming but can also result in run-time errors if an object is found not to respond to a message that has been received. Smalltalk supports only single inheritance. C++ is a compiled, statically typed language that was created as an extension to the C language. Basic types such as int and char are modeled as in C. Variables must be declared with types, and the compiler performs static type checking. Static type checking ensures that there is always a response for every message that is sent. C++ has lower run-time overhead than Smalltalk but at a cost of less flexibility. See COMPUTER PROGRAMMING; PROGRAMMING LANGUAGES. John J. Shilling

Bibliography. B. J. Cox, *Object-Oriented System Building: A Revolutionary Approach*, 1995;

W. R. LaLonde and J. R. Pugh, *Inside Smalltalk*, 1990; S. B. Lippman, *C++ Primer*, 1989; B. Meyer, *Object-Oriented Software Construction*, 1988; J. Rumbaugh et al., *Object-Oriented Modeling and Design*, 1991.

## Obolellida

A small extinct order of articulated brachiopods that ranges in age from Early to Middle Cambrian and includes the earliest known calcitic brachiopods.

**Classification.** The order Obolellida is included within the class Obolellata, subphylum Rhynchonelliformea, phylum Brachiopoda.

Phylum Brachiopoda

Subphylum Rhynchonelliformea

Class Obolellata

Order Obolellida

Superfamily Obolelloidea

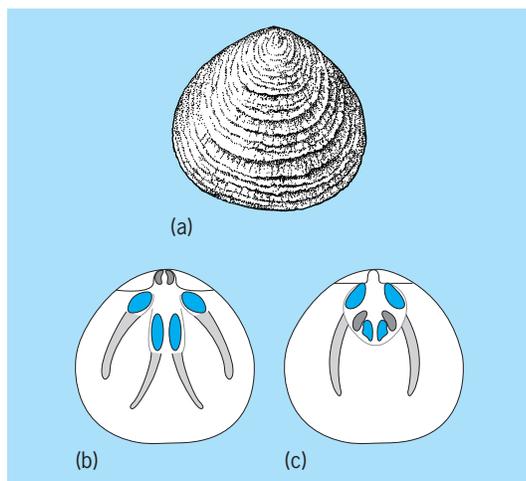
Family Obolellidae

5 genera (Lower Cambrian)

Family Trematobolidae

3 genera (Early-Middle Cambrian)

**Morphology.** Obolellids have a biconvex, calcite, impunctate (lacking holes) shell with an elongated oval shape and a laminar secondary layer. The order includes forms that have primitive articulation of the ventral and dorsal valves—consisting of paired ventral denticles (hinge teeth), where the pedicle (fleshy stalk) attaches to the shell, and dorsal sockets along the internal posterior margin (hinge line)—and forms that lack denticles, such as the genus *Obolella*. The ventral valve (formerly named pedicle valve) has a well-defined, low and relatively short flat shelf (interarea) at the posterior. The pedicle opening



*Obolella* (a) Shell exterior (after J. Hall and J. M. Clarke, *An introductory to the study of the Brachiopoda intended as a handbook for the use of students, Report of the N.Y. State Geologist for 1891, 1894*). (b) Dorsal valve interior, and (c) Ventral valve interior, with schematic illustration of musculature and mantle canals (b and c reprinted with permission from R. L. Kaesler, ed., *Treatise on Invertebrate Paleontology*, courtesy of and © 2000, Geological Society of America and University of Kansas).

(delthyrium) is located between the valves, and is either uncovered, as in *Obolella*, or closed off by convex plates (pseudodeltidium).

The mantle canal system lacks bifurcations (baculate type) in both valves. The ventral and dorsal valves contained main mantle canals known as *vascula lateralia*. The dorsal valve (formerly named brachial valve) also contained a secondary mantle canal, called the *vascula media*.

The muscle arrangement is very similar to that of other articulated brachiopods: in the dorsal valve the pairs of anterior and posterior adductor muscles form a muscle field, while the single pair of oblique diductor muscles is attached at the bottom of the pedicle opening (notothyrial cavity). In the ventral valve, the pair of diductor muscles are located between the pairs of anterior adductor and posterior adductor muscles. Members of this group were presumably epifaunal and sessile. See BRACHIOPODA; RHYNCHONELLIFORMEA. Christian Emig

Bibliography. V. Iu. Gorjansky and L. E. Popov, Morphology, systematic position and origin of the inarticulate brachiopods with calcareous shells [in Russian], *Paleontologicheskii Zh.*, 1985(3):3–14, 1985; V. Iu. Gorjansky and L. E. Popov, On the origin and systematic position of the calcareous-shelled inarticulate brachiopods, *Lethaia*, 19:233–240, 1986; R. L. Kaesler (ed.), *Treatise on Invertebrate Paleontology, Part H, Brachiopoda*, Geological Society of America, Boulder, Co, and University of Kansas, Lawrence, vol. 2, 2000; Yu. L. Pelman, Early and Middle Cambrian inarticulate brachiopods of the Siberian Platform [in Russian], *Trudy Inst. Geologii i Geofiziki, Sibirskoe otdelenie*, vol. 316, 1977; L. E. Popov, L. E. Holmer, and M. G. Bassett, Radiation of the earliest calcareous brachiopods, in P. Copper and J. Jin (eds.), *Brachiopods, Proc. 3d Int. Brachiopod Congr.*, Sudbury, Ontario, Canada, Sept. 2–5, 1995, pp. 209–213, A. A. Balkema, Rotterdam, Brookfield, 1996; A. J. Rowell, The genera of the brachiopod superfamilies Obolellacea and Siphonotretacea, *J. Paleontol.*, 36:136–152, 1962; A. J. Rowell, Inarticulate, in R. C. Moore (ed.), *Treatise on Invertebrate Paleontology, Part H, Brachiopoda*, pp. 260–296, Geological Society of America, Boulder, Co, and University of Kansas Press, Lawrence, 1965; A. Williams et al., A supraordinal classification of the Brachiopoda, *Phil. Trans. Roy. Soc. Lond. B*, 351:1171–1193, 1996.

## Observatory, astronomical

A telescope or telescopes, their protective enclosures (if any), support and headquarters buildings, and the staff of astronomers, engineers, technicians, and other support personnel. The telescopes can be optical or infrared (reflecting or refracting) inside a corotating dome, or radio dishes without enclosures.

The on-site support building contains the control room with the computers and control electronics to operate and point the telescope, as well as the data acquisition computers and electronics for detector

instruments. It houses laboratories for testing and calibrations, a machine shop for emergency repair and fabrication of parts, and storage for inactive instruments or telescope optics. These areas are often incorporated into a single building along with the telescope and dome. Since telescopes are generally located at remote sites to maximize their usefulness, the administration and support offices are often located in a separate headquarters building miles away in a town or university.

The *2005 Astronomical Almanac* lists over 500 observatories in 61 countries, and these are only some of the active ones. That list omits thousands of amateur telescopes, space-borne telescopes, and nontraditional observatories (such as neutrino and gravitational).

**Mauna Kea Observatories.** The 13,800-ft-high (4200-m) volcanic mountain Mauna Kea, in Hawai'i, offers many examples of the various configurations possible. The 3.6-m (142-in.) Canada-France-Hawaii Telescope has everything enclosed in the traditional cylindrical building with the telescope and dome on top. The W. M. Keck Observatory operates the world's two largest telescopes, the 9.8-m (386-in.) multisegmented Keck I and Keck II, with an attached support building between the domes. The 8.2-m (323-in.) Japanese Subaru Telescope has a large support building downslope from the telescope dome connected by a tunnel. The 8.0-m (315-in.) Gemini

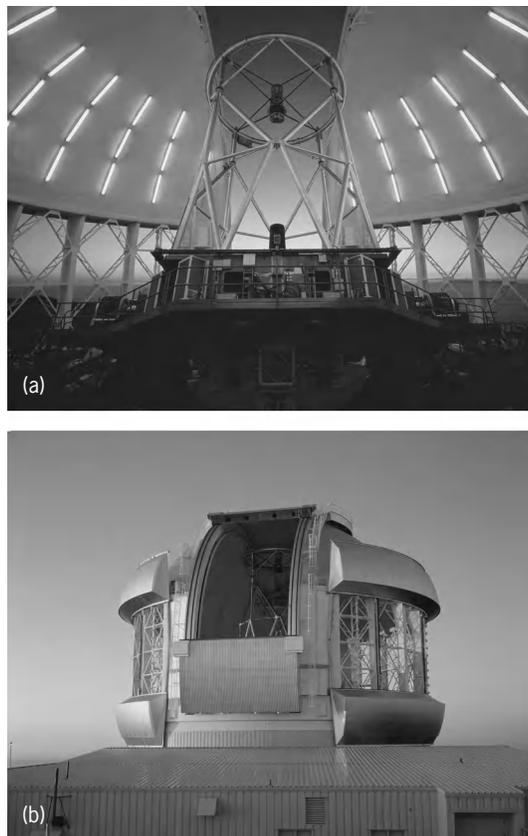
North Telescope (**Fig. 1**) has the support building on the side of the dome, but the same observatory also operates a twin 8.1-m telescope, Gemini South, a hemisphere away on Cerro Pachon in Chile. The University of Hawai'i-Mānoa operates a 2.2-m (87-in.) telescope and the NASA 3.0-m (118-in.) Infrared Telescope Facility. All these telescopes have headquarters in towns miles away. Just below the summit, the Sub-Millimeter Array consists of a control building and numerous small radio telescope dishes that can be moved to different positions via a transporter. Collectively these telescopes (and the others on the mountain) are referred to as the Mauna Kea Observatories (**Fig. 2**), yet each telescope is an observatory in its own right.

**Sizes and types of telescopes.** Telescopes are tools to collect the electromagnetic radiation commonly known as light. This light can vary in wavelength from the kilometer length of very long radio waves, through microwaves and submillimeter waves, infrared or heat radiation, half-micrometer optical or visual wavelengths, and ultraviolet radiation, to the very short nanometer-length x-rays and gamma-rays. Astronomical objects emit radiation at all these wavelengths, and telescopes have been designed to collect that light. The type of light collected determines the type of telescope and detector technology. Thus there are radio telescopes that are metallic steerable dishes, optical telescopes that use lenses, optical-infrared telescopes that use glass mirrors with a reflective coating, and x-ray telescopes that use grazing incident plates to focus the high-energy radiation. *See* ELECTROMAGNETIC RADIATION; LIGHT; TELESCOPE.

The resolution (diffraction limit) of an ideal telescope depends on only two factors, the wavelength of light observed and the size (diameter) of the collecting area:  $\text{resolution} = \text{wavelength}/\text{size}$ . To get the sharpest images, it is necessary to either observe at very small wavelengths or use larger telescopes. In practice, the Earth's atmosphere either completely absorbs some types of light or blurs the images so that the ideal is not met. Astronomers circumvent these phenomena by placing telescopes on high mountain summits, or even in space completely above the atmosphere. Adaptive optics techniques can also mitigate the blurring scintillation seen as twinkling caused by the motions of the air. The drive toward large telescopes is simply to gain the best resolution possible and to collect the most light. Larger telescopes are not built to magnify the image. *See* ADAPTIVE OPTICS; RESOLVING POWER (OPTICS).

The **table** lists the major active professional observatory telescopes and their sizes. **Figure 3** is a graph of the total collecting area of the entire world's telescopes throughout history.

**Ancient observatories.** The human eyes can be considered an astronomical observatory that consists of a pair of 5-mm (0.2-in.) telescopes. Naked-eye astronomy developed for thousands of years, and many early civilizations built structures to aid them in measuring recurring celestial events central for



**Fig. 1.** Gemini North Observatory. (a) Telescope and interior of dome. (b) Exterior of dome. Vent gates surrounding the dome allow air to flow throughout the dome, thereby minimizing distortion of images due to swirling. (Neelon Crawford, *Polar Fine Arts, and Gemini Observatory*)

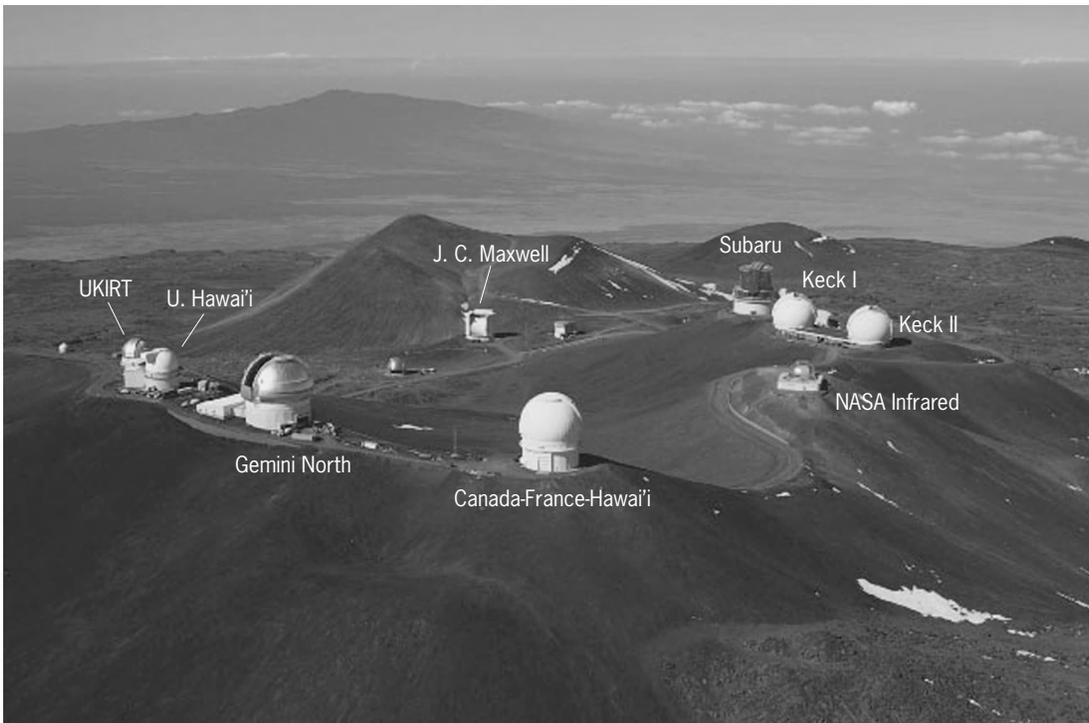


Fig. 2. Mauna Kea Observatories. (University of Hawai'i Institute for Astronomy)

development of their calendars. These archeological structures, which ranged from modest circles of stones to the circle of monoliths at Stonehenge in England, can rightfully be called astronomical observatories. See ARCHEOASTRONOMY; EYE (VERTEBRATE).

**Optical and infrared observatories.** These are the most common types of observatory. They are often located in remote mountains to minimize the effects of contaminating artificial lights and atmospheric blurring. For national observatories this can mean the best site in the country [such as that of the 6-m (236-in.) Bolshoi Telescope in Russia]. Some countries now place their largest telescopes at the best sites in the world [such as the United Kingdom's 4.2-m (165-in.) Herschel and 2.5-m (98-in.) Isaac Newton telescopes at La Palma in the Canary Islands, and the Japanese Subaru Telescope on Mauna Kea]. Groups of countries sometimes pool their resources to jointly operate these larger telescopes. Most observatories are affiliated with or managed by a university or research organization.

Optical observatories study planets, stars, nebulae, and galaxies. A large aperture is needed to collect the faint light of these sources. Sizes typically range from 0.5 to 10 m (20–400 in.). The smaller telescopes can use glass lenses, the largest refractor being the University of Chicago 1.0-m (40-in.) Yerkes. For the largest telescopes, engineering and manufacturing constraints dictate the reflecting design. The 9.8-m (386-in.) Keck Telescope is the world's largest reflector. The Very Large Telescope Project seeks to combine the light from four 8.2-m (323-in.) telescopes to effectively have the light-gathering capability of a 16-m (630-in.) telescope. Studies have been undertaken of a 30–50-m (1200–2000-in.) telescope.

The next generation of large ground-based telescopes truly will be enormous. A 30-m (100-ft) telescope, originally named CELT (California Extremely Large Telescope) but now called the TMT (Thirty Meter Telescope), is being proposed for location in Chile or Hawai'i as a testbed for the European Southern Observatory's 100-m (330-ft) telescope OWL (Overwhelmingly Large Telescope). Also proposed are a 50-m (160-ft) telescope for the Canary Islands called Euro50, a United States effort named MaxAT (30–50 m or 100–160 ft), the Giant Segmented Mirror Telescope (30-m or 100-ft) based on the Keck design, and a Canadian effort to upgrade the 3.6-m (142-in.) Canada-France-Hawai'i Telescope on Mauna Kea to a 20-m (65-ft) telescope. The reason

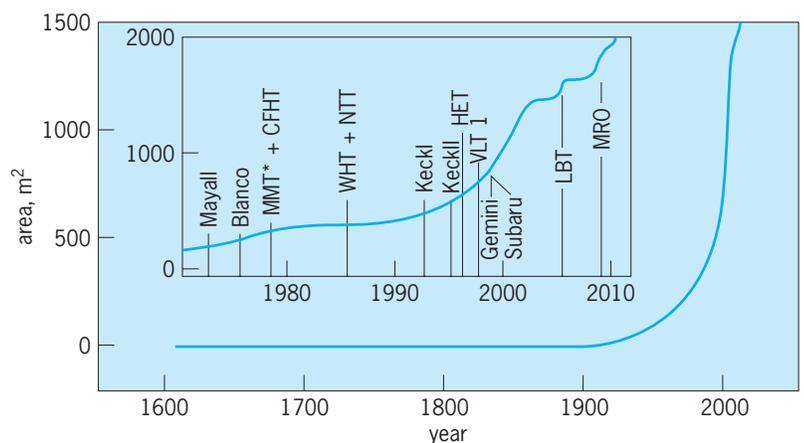


Fig. 3. Total collecting area of the world's telescopes, 1610–2010. Inset details the years 1971–2010 (data after 2005 are projected); names and dates of major observatories are included.  $1 \text{ m}^2 = 10.76 \text{ ft}^2 = 1550 \text{ in.}^2$  (Marianne Takamiya, Institute for Astronomy, University of Hawai'i-Mānoa; updated in 2005 by John Hamilton)

Major astronomical observatories*						
Observatory	Location	Type	Size		Date	Name
			meters	inches		
Human eye	Earth	Refractor	8 mm	0.32	Prehistory	
Galileo	Italy	Refractor	0.016	0.63	1609	
Galileo	Italy	Refractor	0.026	1.02	1610	
Bath, England	United Kingdom	Reflector	1.22	40	1789	William Herschel's 12.2-m (40-ft) focal length telescope
Birr, Ireland	United Kingdom	Reflector	1.83	60	1845	Lord Rosse's Leviathan
Athens	Greece	Refractor	0.6	24	1862	Newall
Treptow	Germany	Refractor	0.7	28	1869	
U.S. Naval Observatory	United States	Refractor	0.66	26.0	1873	Washington
Vienna	Austria	Reflector	0.66	26.0	1880	Grosse
Nice	France	Refractor	0.7	28	1886	
Lick	United States	Refractor	0.9	35	1888	
Paris	France	Refractor	0.8	31	1889	
Paris	France	Reflector	1.0	39	1893	
Herstmonceux	England	Refractor	0.66	26.0	1897	Thompson
Yerkes	United States	Refractor	1.0	39	1897	
Potsdam	Germany	Reflector	0.8	31	1899	
Mount Wilson	United States	Reflector	1.5	59	1908	
Lowell	United States	Reflector	1.1	43	1910	
Hamburg	Germany	Reflector	1.0	39	1911	
Mount Wilson	United States	Reflector	2.5	98	1917	Hooker
Dominion Astrophysical	Canada	Reflector	1.9	75	1918	
Johannesburg	South Africa	Refractor	0.66	26.0	1926	Innes
Geneva	Switzerland	Reflector	1.0	39	1927	
Milan	Italy	Reflector	1.0	39	1929	
Boyden	South Africa	Reflector	1.5	59	1930	
Stockholm	Sweden	Reflector	1.0	39	1931	
Haute Provence	France	Reflector	0.8	31	1932	
Harvard	United States	Reflector	1.5	59	1934	Wyeth
David Dudley Observatory	Canada	Reflector	1.9	75	1935	
McDonald	United States	Reflector	2.1	83	1939	Struve
Padua	Italy	Reflector	1.2	47	1942	
National	Argentina	Reflector	1.5	59	1942	
Haute Provence	France	Reflector	1.2	47	1943	
Palomar	United States	Reflector	5.1	200	1948	Hale
Palomar	United States	Solar	1.2	47	1948	Oschin
Crimean	Russia	Reflector	1.2	47	1952	
Mount Stromlo Observatory	Australia	Reflector	1.3	51	1954	
Hamburg	Germany	Reflector	1.2	47	1954	
Mount Observatory	Australia	Reflector	1.9	75	1955	
U.S. Naval Observatory	United States	Reflector	1.0	39	1955	
Haute Provence	France	Reflector	1.93	76	1958	
Lick	United States	Reflector	3.0	118	1959	Shane
Helwan	Egypt	Reflector	1.9	75	1960	
Okayama	Japan	Reflector	1.9	75	1960	
Schwarzschild	Germany	Solar	1.3	51	1960	
Byurakan	Russia	Solar	1.0	39	1960	
Sternberg	Russia	Reflector	1.3	51	1960	
Kitt Peak National Observatory	United States	Reflector	0.9	35	1960	
Tonanzintla	Mexico	Reflector	1.0	39	1961	
Crimean	Russia	Reflector	2.6	102	1961	Shajn
Dominion Astrophysical	Canada	Reflector	1.2	47	1961	
Royal	South Africa	Reflector	1.0	39	1961	
Uppsala	Sweden	Solar	1.0	39	1962	
Nizamiah	India	Reflector	1.2	47	1962	
Kitt Peak National Observatory	United States	Solar	1.5	59	1962	McMath
Lowell	United States	Reflector	1.8	71	1962	Perkins
U.S. Naval Observatory	United States	Reflector	1.5	59	1963	
Pic du Midi	France	Reflector	1.1	43	1964	
Siding Spring	Australia	Reflector	1.0	39	1964	
Kitt Peak National Observatory	United States	Reflector	2.1	83	1964	
Catalina	United States	Reflector	1.5	59	1965	University of Arizona
La Silla	Chile	Reflector	1.0	39	1966	
Kitt Peak National Observatory	United States	Reflector	0.9	35	1966	
Cerro Tololo InterAmerican	Chile	Reflector	1.5	59	1967	
Shemakha	Russia	Reflector	2.0	79	1967	
Northwestern	United States	Reflector	1.0	39	1967	
Ondrejov	Czechoslovakia	Reflector	2.0	79	1967	
Toledo	United States	Reflector	1.0	39	1967	
McDonald	United States	Reflector	2.7	106	1968	Harlan Smith
La Silla	Chile	Reflector	1.5	59	1968	

Major astronomical observatories (cont.)						
Observatory	Location	Type	Size		Date	Name
			meters	inches		
Yerkes	United States	Reflector	1.0	39	1968	
Kitt Peak, McGraw	United States	Reflector	2.4	94	1968	
Milan	Italy	Reflector	1.4	55	1968	
Ontario	Canada	Reflector	1.2	47	1968	
Figl	Austria	Reflector	1.5	59	1969	
Haute Provence	France	Reflector	1.52	59.8	1969	
La Silla	Chile	Solar	1.0	39	1969	
Kitt Peak, Steward	United States	Reflector	2.3	91	1969	Bok
Sacramento Peak	United States	Solar	1.6	63	1969	
Mount Hopkins	United States	Reflector	1.5	59	1970	
Baja	Mexico	Reflector	1.5	59	1970	
Mount Lemmon	United States	Infrared	1.0	39	1970	University of Arizona
Palomar	United States	Reflector	1.5	59	1970	
Mauna Kea	Hawai'i	Reflector	2.2	87	1970	University of Hawai'i
Los Campanas	Chile	Reflector	1.0	39	1971	Swope
Wise	Israel	Reflector	1.0	39	1971	
Mount Lemmon	United States	Infrared	1.5	59	1972	Minnesota, California
Tenerife	Canary Islands	Infrared	1.5	59	1972	
Siding Spring	Australia	Solar	1.2	47	1973	UK Schmidt
South African Astronomical Observatory	South Africa	Reflector	1.0	39	1973	
Padua	Italy	Reflector	1.8	71	1973	Copernicus
Kitt Peak National Observatory	United States	Reflector	4.0	157	1973	Mayall
Cerro Tololo InterAmerican	Chile	Reflector	1.0	39	1973	
South African Astronomical Observatory	South Africa	Reflector	1.9	75	1974	
Turin	Italy	Reflector	1.0	39	1974	
Konkoly	Hungary	Reflector	1.0	39	1974	
Kiso	Japan	Solar	1.1	43	1974	Kiso Schmidt
Mount Lemmon	United States	Infrared	1.5	59	1974	NASA
Center for Astronomical Investigations	Venezuela	Reflector	1.0	39	1975	
Calar Alto	Spain	Reflector	1.2	47	1975	
Struve	Russia	Reflector	1.5	59	1975	
Kitt Peak, McGraw	United States	Reflector	1.3	51	1975	
Pennsylvania State	United States	Reflector	1.5	59	1975	
National	Greece	Reflector	1.2	47	1975	
Siding Spring	Australia	Reflector	3.9	154	1975	Anglo-Australian
Cerro Tololo InterAmerican	Chile	Reflector	4.0	157	1976	Blanco
Special Astrophysical Observatory	Russia	Reflector	6.0	236	1976	Bolshoi Teleskop Azimutal'nyi (BTA)
Byurakan	Armenia	Reflector	2.6	102	1976	
University of Virginia	United States	Reflector	1.0	39	1976	
La Silla	Chile	Reflector	3.6	142	1976	
Los Campanas	Chile	Reflector	2.5	98	1976	DuPont
University of Wyoming	United States	Infrared	2.3	91	1977	
Quebec	Canada	Reflector	1.6	63	1978	
Mauna Kea	Hawai'i	Infrared	3.8	150	1978	United Kingdom Infrared Telescope (UKIRT)
Lick	United States	Reflector	1.0	39	1979	
Mount Hopkins	United States	Reflector	4.5	177	1979	Multi-Mirror Telescope (MMT; original)
Rio de Janeiro	Brazil	Reflector	1.6	63	1979	
Calar Alto	Spain	Reflector	1.5	59	1979	
Mauna Kea	Hawai'i	Infrared	3.0	118	1979	Infrared Telescope Facility (IRTF)
Mauna Kea	Hawai'i	Reflector	3.6	142	1979	Canada-France-Hawai'i Telescope (CFHT)
Calar Alto	Spain	Reflector	3.5	138	1979	
Calar Alto	Spain	Reflector	2.2	87	1979	
Pic du Midi	France	Reflector	2.0	79	1979	
Baja	Mexico	Reflector	2.1	83	1979	
La Silla	Chile	Reflector	1.4	55	1979	
Rozhen	Bulgaria	Reflector	2.0	79	1980	
Gornegrat	Switzerland	Infrared	1.5	59	1980	
La Palma	Canary Islands	Reflector	2.5	98	1982	Isaac Newton
La Palma	Canary Islands	Reflector	1.0	39	1983	Kapteyn
Siding Spring	Australia	Reflector	2.3	91	1984	
La Silla	Chile	Reflector	2.2	87	1985	
Tuorla	Finland	Reflector	1.0	39	1985	
La Palma	Spain	Reflector	4.2	165	1986	William Herschel Telescope (WHT)
La Silla	Chile	Reflector	3.6	142	1986	New Technology Telescope (NTT)
La Palma	Canary Islands	Reflector	2.5	98	1989	Nordic
Mauna Kea	Hawai'i	Reflector	9.82	387	1991	Keck I
University of British Columbia (UBC)-Laval	Canada	Reflector <sup>†</sup>	2.7	106	1992	

Major astronomical observatories ( <i>cont.</i> )						
Observatory	Location	Type	Size		Date	Name
			meters	inches		
Mount Graham	United States	Reflector	1.8	71	1994	Vatican Advanced Technology Telescope
Kitt Peak National Observatory	United States	Reflector	3.5	138	1994	Wynn
Apache Point	United States	Reflector	3.5	138	1995	Astrophysical Research Consortium (ARC)
Mauna Kea	Hawai'i	Reflector	9.82	387	1996	Keck II
Mauna Kea	Hawai'i	Reflector	8.2	323	1998	Subaru
Apache Point	United States	Reflector	2.5	98	1998	Sloan
La Palma	Canary Islands	Reflector	3.5	138	1998	Telescopio Nazionale Galileo (TNG)
Mauna Kea	Hawai'i	Infrared	8.0	315	1998	Gemini North Fred Gillette Telescope
McDonald (Mt. Fowlkes)	United States	Reflector <sup>†</sup>	9.2	362	1998	Hobby-Eberly Telescope (HET)
Mt. Hopkins	United States	Reflector	6.5	256	1999	Multi-Mirror Telescope (MMT; new)
Cerro Paranal	Chile	Reflector	8.2 <sup>‡</sup>	323	1999	Very Large Telescope (VLT) 1, Antu
Cerro Paranal	Chile	Reflector	8.2 <sup>‡</sup>	323	2000	VLT 2, Kueyen
Las Compañas	Chile	Reflector	6.5	256	2000	Magellan 1, Walter Baade Telescope
Cerro Pachon	Chile	Infrared	8.0	315	2000	Gemini South
Las Compañas	Chile	Reflector	6.5	256	2001	Magellan 2, Landon Clay Telescope
Cerro Paranal	Chile	Reflector	8.2 <sup>‡</sup>	323	2001	VLT 3, Melipal
Paranal	Chile	Reflector	8.2 <sup>‡</sup>	323	2002	VLT 4, Yepun
Vancouver, British Columbia	Canada	Reflector <sup>†</sup>	6.0	236	2004	Large Zenith Telescope
Haleakalā, Maui	Hawai'i	Robotic reflector	2.0	79	2004	Faulkes Telescope North
La Palma	Canary Islands	Reflector	10.4	409	2005	Gran Telescopio Canarias
Cerro Pachon	Chile	Reflector	4.1	161	2005	Southern Astrophysical Research (SOAR)
Sutherland	South Africa	Reflector <sup>†</sup>	10.0	394	2005	South African Large Telescope (SALT)
Chelmos	Greece	Reflector	2.3	91	2005	Aristarchos
Cerro Paranal	Chile	Reflector	2.6	102	2005	Very Large Telescope Survey Telescope (VLTST)
Mount Graham	United States	Reflector	2 × 8.4 <sup>§</sup>	331	2006	Large Binocular Telescope (LBT) 1 and 2
Cerro Paranal ("NTT peak")	Chile	Reflector	4.0	157	2006	Visible and Infrared Survey Telescope for Astronomy (VISTA)
Siding Spring	Australia	Robotic reflector	2.0	79	2007	Faulkes Telescope South
Magdalena, New Mexico	United States	Reflector, interferometer	1 × 2.4, 10 × 1.4	1 × 94, 10 × 55	2008	Magdalena Ridge Observatory (MRO)
Airborne; Ames, California	United States, Germany	Reflector mounted on Boeing 747SP	2.5	98	2008	Stratospheric Observatory for Infrared Astronomy (SOFIA)**
Mauna Kea	Hawai'i	Robotic reflector	0.9	35	2008	University of Hawai'i at Hilo
Flagstaff, Arizona	United States	Reflector	4.2	165	2009	Discovery Channel Telescope (DCT)
Haleakalā, Maui	Hawai'i	Reflector, solar only	4.0	157	2010	Advanced Technology Solar Telescope
Mauna Kea	Hawai'i	Reflector, interferometer	6 × 2.0, 2 × 9.82	6 × 79, 2 × 387,	2010?	Keck Interferometer with outrigger telescopes
Mauna Kea	Hawai'i	All-sky reflector	4 × 1.8	4 × 71	2010?	Pan STARRS
	Chile, Mexico, or Canary Islands	Reflector	8.4 (6.9 clear aperture)	331 (272 clear aperture)	2012	Large Synoptic Survey Telescope (LSST)
Xinglong	China	Reflector, collector for optical fibers	6.67 × 6.05	263 × 238	?	Large Sky Area Multi-Object Fiber Spectroscopic Telescope (LAMOST)

<sup>\*</sup>Optical, infrared, and solar telescopes only; radio and space telescopes omitted. Compiled by Marianne Takamiya, Institute for Astronomy, University of Hawai'i-Mānoa; updated in 2005 by John Hamilton.

<sup>†</sup>Not fully steerable.

<sup>‡</sup>The Very Large Telescope Project seeks to combine the light from the four 8.2-m (323-in.) telescopes, VLT 1–4, to effectively have the light-gathering capability of a 16-m (630-in.) telescope.

<sup>§</sup>Light from the two Large Binocular Telescopes will be combined, so that the light-gathering area as seen by the detector will be 11.8 m (465 in.).

\*\*Replacement for Kuiper Airborne Observatory (KAO), a Lockheed C-141 with 0.9-m (35-in.) infrared-optimized reflector. Flew 1971–1995.

that there are no plans for 12-, 15-, or 18-m telescopes is related to the increase in the collecting area of a telescope, whose light-gathering power increases as the square of the radius. Thus, a 100-m telescope will collect 100 times the amount of light collected by a 10-m telescope, even though it is "only" ten times bigger. It is currently felt that such an increase in light is necessary for answering many of the significant questions in modern astronomy.

Most optical observatories can also observe in

the near-infrared region. Infrared observatories have been optimized to work further into the longer (thermal) infrared. Since glass absorbs infrared light, all infrared telescopes are of the reflecting type. Special design techniques must be used, as the telescope itself glows in the infrared. The secondary mirror can be rapidly tilted (chopped) to sequentially view the (glowing) sky and the sky plus source. When averaged, the thermal effects of the sky will cancel, as they are random. Infrared observatories must be

placed in dry climates, as the water vapor in the air absorbs the infrared radiation. High mountain sites also tend to be dry, as water vapor content decreases with altitude. *See* INFRARED ASTRONOMY.

**Radio observatories.** Radio observatories have become an essential complement to optical observatories. With their longer wavelengths, the radio telescopes can see through cosmic clouds and measure the properties (temperature, pressure, chemical composition, and velocities) of gases which pervade the universe.

Since the wavelengths are so long, the telescopes must be large to achieve a high resolution (the ability to discriminate between two nearby sources). Radio dishes are typically tens to hundreds of meters wide. The largest movable radio telescope (100 m or 328 ft in diameter) is operated by the Max Planck Institute for Radio Astronomy at Effelsberg, Germany. The 305-m (1000-ft) telescope at Arecibo, Puerto Rico, uses a natural valley for its dish and a movable secondary feed to point about the sky.

The longer radio wavelengths allow the use of interferometers, where the signal from several small radio telescopes can be mathematically combined to yield results as if they had been collected from a very large dish. The Very Large Array radio observatory in Socorro, New Mexico, has twenty-seven 25-m (82-ft) antennas arranged in a Y pattern, which can yield the resolution of a single telescope 36 km (22 mi) wide and a sensitivity of a dish 130 m (426 ft) wide. Even larger synthetic telescopes can be created with baselines of intercontinental lengths. The Very Long Baseline Array uses ten 25-m (82-ft) antennas spread from Hawai'i to the Virgin Islands, allowing the resolution of a telescope dish that distance across. *See* RADIO TELESCOPE.

**Solar observatories.** Solar observatories have telescopes optimized for an objective opposite to that of the optical telescopes. Here the problem is too much light (and heat). Solar telescopes overcome this problem by greatly reducing the amount of light, by using smaller telescopes, and by passing the light through very narrow filters. Daytime observations also present the problem of more severe air convection around the telescope due to the heat of the sunlight. The Big Bear Solar Observatory is built in the center of a lake to mitigate this convection and achieve sharp images. When high resolution is needed (which calls for larger telescopes), the design must include means to reject the solar heat and reduce the convection turbulence in the telescope itself. The National Solar Observatory operates the Vacuum Telescope at Kitt Peak, Arizona, where the sunlight is reflected off a coelostat mirror into a vacuum tube on its way to the detector; the absence of air means that there is no disturbing convection. Larger solar telescopes are now being built, such as the 4-m (157-in.) Advanced Technology Solar Telescope. *See* SUN.

**Airborne and space observatories.** Crewless balloons have been used to hoist telescopes above the absorbing atmosphere for gamma-ray, x-ray, and ultraviolet observations. Their lower cost (compared to space missions) is offset by the short duration

of the observations. The National Aeronautics and Space Administration (NASA) successfully operated the Kuiper Airborne Observatory for 20 years. This facility had a 0.9-m (36-in.) telescope aboard a converted C-141 military cargo plane flying at an altitude of 12.5 km (41,000 ft), which is above 85% of the atmosphere and 99% of the water vapor. The successor is the Stratospheric Observatory for Infrared Astronomy (SOFIA), a Boeing 747-SP aircraft carrying a 2.5-m (98-in.) telescope. In addition, NASA has used high-flying U2 spy planes to measure variations in the cosmic microwave radio background.

Space observatories have amply proven their worth despite their initial high cost and limited mission lifetimes. Unimpeded high-energy gamma-ray and x-ray observations from observatories such as *Einstein* and *Compton* have redefined the universe. The *Cosmic Background Explorer (COBE)* microwave satellite has observed the earliest viewable portions of the universe, the cosmic microwave background. A followup mission, the *Wilkinson Microwave Anisotropy Probe (WMAP)*, has enabled cosmologists to accurately measure many of the physical parameters of the early universe. The Hubble Space Telescope, an orbiting 2.4-m (94-in.) optical-infrared telescope, has achieved unparalleled images of the sky, from exploding supernovae, to star-forming regions, to the deepest (farthest) images of the universe. An infrared-optimized counterpart, the 0.85-m (33-in.) *Spitzer Space Telescope* (formerly *SIRTF, Space Infrared Telescope Facility*), was launched into orbit in 2003. The success of these space observatories has led to the planning of a 6-m (236-in.) successor, the *James Webb Space Telescope*. *See* COSMIC BACKGROUND RADIATION; GAMMA-RAY ASTRONOMY; SATELLITE (ASTRONOMY); X-RAY TELESCOPE.

**Virtual observatories.** With the advent of digital format in astronomical images and spectra, large data-storage facilities, data-mining software, and the World Wide Web, the concept of a virtual observatory has been established. It will allow researchers global electronic access to combined archived data from many ground and space telescopes. Today, at most major telescopes, all new data are archived (and released publicly after the usual proprietary period). Efforts are underway to digitally convert old photographic plates for computer storage and retrieval. Leaders in this field are the United States, with the National Virtual Observatory; the European Southern Observatory, with the Astrophysical Virtual Observatory; and the United Kingdom's AstroGrid project. This effort is seen as so important that these groups have formed the International Virtual Observatory Alliance to ensure commonality and compatibility with data transfers, protocols, and formats. Soon professional astronomers can observe while never visiting a telescope. *See* DATA MINING; WORLD WIDE WEB.

**Detectors and instruments.** An astronomical instrument is the working heart of an observatory. The telescope exists solely to collect and funnel light into the instrument. Astronomical instruments range from a simple eyepiece for direct viewing (with different magnifications), to a camera for imaging on

film or electronic detectors, to a spectrograph which records the wavelength distribution of light energy (analogous to a rainbow). For radio telescopes the instrument is an amplifier that isolates and boosts the weak celestial electric signal. See ASTRONOMICAL SPECTROSCOPY; CAMERA; SPECTROGRAPH.

Before the invention of photography, an eyepiece was used as the instrument with the human eye as the detector. Observations consisted of written notes and sketches. When photography was invented, one of the first uses was to image the Moon through a telescope. Photography soon became the mainstay for astronomical observations as it presented a permanent unbiased record of an event. Most observatories had well-equipped darkrooms, and techniques for extracting as much information as possible from photographic emulsions, such as hypersensitizing and plate baking in a nitrogen atmosphere. The electronic revolution introduced the photomultiplier tube and photon counters, which were very useful for photometry (the precise measurement of stellar brightness). See ASTRONOMICAL IMAGING; PHOTOMULTIPLIER.

The charge-coupled device (CCD) has now virtually supplanted the photographic plate. Its advantages include high quantum efficiency, large linear response range, and the digitized nature of the data, allowing immediate display on computer terminals and ease of computer data reduction. To support night-vision technology, charge-coupled devices have also been developed as useful detectors in the infrared region. Photography is still useful for wide-field applications, but arrays of large-format charge-coupled devices are closing the gap there also. See CHARGE-COUPLED DEVICES; INFRARED IMAGING DEVICES.

**Unconventional observatories.** Besides observing light (electromagnetic radiation), astronomers are mapping the universe via neutrinos, particles that can transverse matter with a minimum of interaction. Neutrinos are as numerous as photons of light. Neutrinos have been observed from the core of the Sun. Interpretations of observed solar neutrino fluxes have led to the discovery of oscillations between the three types of neutrinos, the electron,  $\mu$  (mu), and  $\tau$  (tau) neutrinos. This has led to the discovery that neutrinos are not massless after all. Neutrino bursts have been detected from an exploding supernova (SN 1987A). With the discoveries of dark matter and dark energy as majority components of the universe, astroparticle physics will come to play an ever-increasing role in astronomy. The Super-Kamiokande Detector in Japan and the Sudbury Neutrino Observatory are examples. Proposed new projects such as AUGER in Argentina, ICECUBE in Antarctica, and Ashra in Hawai'i seek to observe the very highest energy cosmic rays and neutrinos. See COSMOLOGY; DARK ENERGY; DARK MATTER; NEUTRINO; NEUTRINO ASTRONOMY; SOLAR NEUTRINOS.

Gravitational wave observatories attempt to detect ripples (gravitational waves or gravitons) in the fabric of space-time itself. It is thought that these can be generated by collapses of massive black holes. There

are two main approaches to detection, laser interferometry and resonant mass detectors. Examples of each are the Laser Interferometer Gravitational-Wave Observatory (LIGO) and the TIGA (Truncated Icosahedral Gravitational Wave Antenna) project. See GRAVITATION RADIATION; TELESCOPE.

John Hamilton

**Bibliography.** S. J. Dick, *Sky and Ocean Joined: The U.S. Naval Observatory 1830–2000*, Cambridge University Press, 2002; R. Kerrod, *Hubble: The Mirror on the Universe*, Firefly Books, 2003; H. T. Kirby-Smith, *U.S. Observatories: A Directory and Travel Guide*, Van Nostrand Reinhold, 1976; K. Krisciunas, *Astronomical Centers of the World*, Cambridge University Press, 1988; S. Laustsen, C. Madsen, and R. M. West, *Exploring the Southern Sky: A Pictorial Atlas from the European Southern Observatory (ESO)*, Springer, 1987; W. P. McCray, *Giant Telescopes: Astronomical Ambition and the Promise of Technology*, Harvard University Press, 2004; R. Naeye, *Signals from Space: The Chandra X-Ray Observatory*, Raintree Steck-Vaughn, 2000; T. Orr, *The Telescope (Inventions That Shaped the World)*, Scholastic, 2004; C. C. Petersen and J. C. Brandt, *Vision of the Cosmos*, Cambridge University Press, 2003; A. Sandage, *Centennial History of the Carnegie Institution of Washington*, vol. 1: *The Mount Wilson Observatory: Breaking the Code of Cosmic Evolution*, Cambridge University Press, 2005; F. Watson, *The Stargazer: The Life and Times of the Telescope*, Da Capo Press, 2005.

## Obsessive-compulsive disorder

A type of anxiety disorder (commonly referred to as OCD) characterized by recurrent, persistent, unwanted, and unpleasant thoughts (obsessions) or repetitive, purposeful ritualistic behaviors that the person feels driven to perform (compulsions). A cardinal feature of this disorder is an awareness of the irrationality or excess of the obsessions and compulsions accompanied by an inability to control them.

**Characteristics.** Typical compulsions include an irresistible urge to wash (particularly the hands) or clean, to check doors to confirm that they are locked, to return repeatedly to appliances to make sure they are turned off, to touch, to repeat, to count, to arrange, or to save. Typical obsessions include overconcern about dirt and contamination, fear of acting on violent or aggressive impulses, feeling overly responsible for the safety of others, abhorrent religious (blasphemous) and sexual intrusions, and inordinate concern with arrangement or symmetry. Obsessions may accompany compulsions, or compulsions may occur alone.

The most common subtype of the disorder is washing. The washers are driven to perform repeated handwashing and are obsessed with a fear of dirt, germs, and contamination. They may spend several hours each day washing their hands or showering. Typically they try to avoid what they perceive to be

sources of contamination, like door knobs, electric switches, or dust.

A second presentation for obsessive-compulsive disorder involves pathologic doubt coupled with compulsive checking. Some patients have an incessant need for symmetry, but typically the checker is concerned for the safety of others. Checking, which is enough to resolve normal uncertainty, often contributes to doubt and leads to even more checking.

The difference between obsessive-compulsive disorder and milder forms of obsession or compulsion seen in otherwise healthy people is that for the sufferer the obsessions or compulsions cause marked distress, are time-consuming, and significantly interfere with the person's normal routine, occupational functioning, usual social activities, and relationships with others.

Although the obsessions and compulsions may seem "crazy," a typical patient with obsessive-compulsive disorder is acutely aware of the irrationality or excessiveness of his or her fears or behaviors, yet is unable to control them. Such self-awareness is associated with a tendency to hide the symptoms, and usually persons with obsessive-compulsive disorder are successful in concealing the problem, because in areas other than that of the obsession or compulsion they are perfectly reasonable. This secretiveness may explain why the prevalence of obsessive-compulsive disorder in the general population, which has been reported to be more than 2.0% in several recent epidemiological studies in the United States, was once thought to be 0.05%.

Onset in adolescence occurs in about a third of cases. In another third symptoms appear in early adulthood, and in the last third they start later in life. If not treated appropriately, the disorder is often chronic, with waxing and waning of symptoms.

**Treatment.** Obsessive-compulsive disorder is generally resistant to traditional psychotherapy, which has tried to trace the condition to conflicts of early childhood. An effective mode of psychotherapy is behavioral therapy, in which the patients are gradually exposed to their feared or triggering situation but are prevented from performing accompanying compulsions. For example, patients with a fear of contamination from germs are instructed to touch a very dirty-looking, "contaminated" cloth, but they are not allowed to perform handwashing afterward. This approach, which focuses on treating the symptoms rather than trying to understand their origin, seems to be more effective in treating the ritualistic behavior (compulsions) than the pervasive thoughts (obsessions).

Obsessive-compulsive disorder is also refractory to most drugs used to treat anxiety, depression, and psychosis. However, it often eases with medications that affect the brain's serotonergic system, such as clorimipramine, fluvoxamine, and fluoxetine.

Serotonin, like other neurotransmitters, is released into the gap between two nerve cells (synapse) and later must be removed, through a process called reuptake, before the presynaptic cell can be fired again. Clorimipramine, fluvoxamine, and flu-

oxetine are unique among the psychoactive drugs as they specifically block the reuptake of serotonin in the synapse. The observation that the disorder responds only to drugs that alter the serotonin function in the brain suggests a pivotal role for serotonin in this disorder. Indeed, compared with normals, persons with obsessive-compulsive disorder appear particularly sensitive to the behavioral effect of methachlorophenylpiperazine (mCPP), a compound that activates the serotonergic system and may create new obsessions that usually last as long as the serotonergic system is activated, that is, a couple of hours. *See* SEROTONIN; SYNAPTIC TRANSMISSION.

Another approach in studying the psychobiology of psychiatric disorders is examination of the functional anatomy of the brain of those with the condition using such techniques as positron emission tomographic (PET) and magnetic resonance imaging (MRI), which have revealed abnormalities in the cortex (frontal lobes) and basal ganglia along with decreased caudate volume. Obsessive-compulsive disorder thus may reflect a discordance between the brain's most advanced region (the cortex) and one of its more primitive areas (the basal ganglia). *See* MEDICAL IMAGING.

The specific response of patients with obsessive-compulsive disorder to serotonergic drugs, their hypersensitivity to activation of the serotonergic system, and the distinct functional anatomy differences found in those patients suggest a biological cause for this disorder. In this regard, obsessive-compulsive disorder represents a shift from a psychological to a neurobiological approach in the study of anxiety disorders. *See* ANXIETY DISORDERS; NEUROTIC DISORDERS.

Joseph Zohar

Bibliography. E. Hollander (ed.), *Current Concepts in Obsessive-Compulsive Disorder*, 1994; J. L. Rapoport, The biology of obsessions and compulsions, *Sci. Amer.*, pp. 83-89, March 1989; J. Zohar, E. B. Foa, and T. R. Insel, Obsessive-compulsive disorder, in *Treatments of Psychiatric Disorders*, American Psychiatric Association, pp. 2095-2105, 1989; J. Zohar and T. R. Insel, Obsessive-compulsive disorder: Psychobiological approaches to diagnosis, treatment and pathophysiology, *Biol. Psychiat.*, 22: 667-687, 1987.

## Obsidian

A volcanic glass, usually of rhyolitic composition, formed by rapid cooling of viscous lava. The color is jet-black because of abundant microscopic, embryonic crystal growths (crystallites) which make the glass opaque except on thin edges. Iron oxide dust may produce red or brown obsidian. *See* LAVA.

Obsidian usually forms the upper parts of lava flows. Well-known occurrences are Obsidian Cliffs in Yellowstone Park, Wyoming; Mount Hekla, Iceland; and the Lipari Islands off the coast of Italy. Less commonly, obsidian forms selvages of dikes and sills. *See* IGNEOUS ROCKS; VOLCANIC GLASS.

Carleton A. Chapman

## Occultation

The temporary blocking from view of one celestial body by another. The occulting body is the one closer to the observer, and can be a planet, moon, ring system, or other body, usually in the solar system. The occulted body is smaller in apparent, projected size and is usually a distant star, although it can also be a spacecraft radio signal, as in the case of the *Voyager* spacecraft radio occultations at the outer planets, or another solar system body. Examples of occultations are that of a star by a planet, a lunar occultation of a star, and an occultation by Pluto of its satellite Charon. Although a solar eclipse is not usually thought of in these terms, this event is actually an occultation of the Sun by the Moon. See ECLIPSE; PLAUET; SATELLITE (ASTRONOMY); SOLAR SYSTEM.

Observations of occultations can reveal information about the physical size of the blocking body, the structure of its atmosphere (bound, in the case of planetary atmospheres, or unbound in the case of comets), or the structure of its rings. Occultations of stars by the Moon are used primarily to study the structure of the occulted stars, quantities such as their projected size, and limb darkening, though the figure of the lunar limb used to make predictions of Bailey's beads at eclipses is based on lunar occultations.

**Observations.** Although occultations provide a wealth of information about solar system objects, the observation of an occultation is primarily a monitoring of stellar brightness over time. In fact, the best data are obtained when the occulting solar system body is as close to invisible as is possible. Observations of stellar occultations typically require fast sampling rates, monitoring the stellar brightness several times each second. Such observations are possible with photoelectric photometers and charge-coupled devices (CCDs) designed for such rapid readout. The quality of the resulting light curve depends on the stellar brightness and the brightness of the background. The resulting light curve, or plot of stellar brightness against time, shows the star at full level,

then a drop in brightness when star is occulted, and then a rise of the signal back to the full level (see **illustration**). See ASTRONOMICAL IMAGING; CHARGE-COUPLED DEVICES.

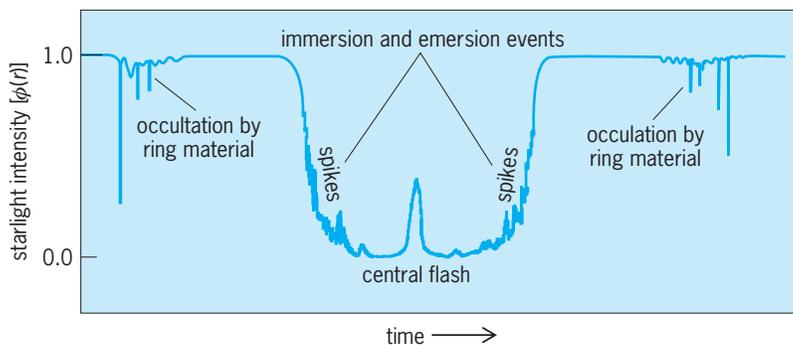
The spatial resolution of a stellar occultation is limited by Fresnel diffraction, not by the seeing disk as for imaging observations. This limit depends on both the wavelength at which the observations are made and the distance between the observer and the occulting body. At Pluto and at visible wavelengths (500 nm), the limiting occultation spatial resolution is approximately 1 mi or 1.7 km. When this is compared to the imaging spatial resolution at a superb observing site (seeing  $\sim 0.5$  arcsec) of 9000 mi or 14,500 km on Pluto, the advantage of occultations is obvious. See DIFFRACTION; TELESCOPE; TWINKLING STARS.

Analysis of stellar occultation data involves converting the time series light curve into a spatial scan, using knowledge of the geometry of the occulting body and the observer. These calculations rely heavily on solar system ephemerides; for occultations by small bodies in the outer solar system, such as Pluto, Triton, comets, and Kuiper Belt objects, the accuracy of event predictions are constrained by the accuracy of their ephemerides. To compensate for these uncertainties in the data analysis, typically two or more light curves from different observing sites are required, since the circular silhouette has different-length chords at different parts of the surface. These are then combined and the ephemeris uncertainties are removed. See EPHEMERIS.

**Applications.** Occultations by asteroids have been observed extensively to determine the shapes of these bodies. Several observers with fixed-location or portable telescopes can map the shape of an asteroid, including any departures from a purely spherical shape. Attempts are being made to extend the utility of occultations for determining sizes of solid-surface bodies to cometary nuclei and Kuiper Belt objects. See ASTEROID; COMET; KUIPER BELT.

For bodies with atmospheres, the stellar signal disappears gradually because of refraction of the stellar light ray by the body's atmosphere. Analysis of the occultation light curves when combined with knowledge of the atmospheric composition (usually obtained through spectral observations) can reveal the temperature in the atmosphere at a particular altitude or pressure level, and how the temperature varies with altitude. If any extinction is present in the form of haze or clouds, it can be detected as well. This method has been used on all planets with even minimal atmospheres, from Venus out to Pluto (including Earth). In stellar occultation observations yielded the first direct detection of Pluto's atmosphere, although its presence had been suspected earlier from spectral analysis of surface ices. See PLUTO; REFRACTION OF WAVES.

Occultations can be used to study the response of an atmosphere to seasonal variations, such as Pluto's large changes in solar distance and heating due to its substantial orbital eccentricity, and Triton's large swings in subsolar latitude or orientation toward the Sun. In fact, some have predicted a complete



**Schematic light curve of a central occultation by a ringed planet with substantial atmosphere.** The relative brightness of the star,  $\phi(r)$ , is plotted as a function of time or distance from the center of the planet  $r$ . The decrease in stellar brightness occurs when ring material obscures the star and when the atmosphere refracts the starlight. Immersion and emersion events provide information about the temperature, composition, and structure of the upper atmosphere. When the observer, occulting body, and star are precisely aligned, an additional feature is visible, the central flash formed by focusing the starlight around the entire atmosphere. (After J. L. Elliot, *Stellar occultation studies of the solar system*, reprinted with permission from *Annual Review of Astronomy and Astrophysics*, volume 17 © 1979 by Annual Reviews, www.annualreviews.org)

freeze-out of Pluto's atmosphere as it recedes from the Sun. (Pluto's closest passage to the Sun was in 1989, and its farthest distance will be in 2114.) Occultation observations have revealed surprising pressure increases in both Pluto's and Triton's atmospheres since about 1990, perhaps in response to small surface temperature changes carried up to the atmosphere via vapor-pressure equilibrium. In addition, occultations by both bodies have allowed study of the nonsphericity of both atmospheres, possibly attributed to zonal winds. See NEPTUNE.

Understanding of planetary rings has been revolutionized by stellar occultation observations, both by spacecraft and by ground-based observers. The ring system of Uranus was discovered during a stellar occultation in 1977; prior to this, only the bright rings of Saturn were known to astronomers. Since the discovery of Uranus' rings, many occultations by the rings of Saturn, Uranus, and Neptune have been observed. Multiple occultation observations are combined to study the shapes of rings and many have been found to be noncircular, driven by resonances with nearby satellites. Dynamical studies of ring lifetimes suggest relatively short-lived systems of perhaps a few million years (much shorter than the age of the solar system at  $4.5 \times 10^9$  years). Ring occultation studies suggest one reason why such varied ring systems can be viewed today: The satellite resonances prevent the spread and dissipation of ring material, in effect freezing the rings in place. See SATURN; URANUS.

Amanda S. Bosh

**Bibliography.** J. K. Beatty and A. Killian, Discovering Pluto's atmosphere, *Sky Telesc.*, 76:624-627, 1988; D. P. Cruikshank, Triton, Pluto, and Charon, in J. K. Beatty et al. (eds.), *The New Solar System*, 4th ed., pp. 285-296, Cambridge University Press, Sky Publishing Corp., 1999; J. L. Elliot, E. Dunham, and R. L. Millis, Discovering the rings of Uranus, *Sky Telesc.*, 53:412-416, 1977; J. L. Elliot et al., The recent expansion of Pluto's atmosphere, *Nature*, 424:165-168, 2003; R. G. French et al., Dynamics and structure of the Uranian rings, in J. T. Bergstralh et al. (eds.), *Uranus*, pp. 327-409, University of Arizona Press, 1991; B. Sicardy et al., Large changes in Pluto's atmosphere as revealed by recent stellar occultations, *Nature*, 424:168-170, 2003.

## Ocean

One of the major subdivisions of the interconnected body of salt water that occupies almost three-quarters of the Earth's surface. Earth is the only planet in the solar system whose surface is covered with significant quantities of water. Of the nearly 1.4 billion cubic kilometers of water found either on the surface or in relatively accessible underground supplies, more than 97% is in the oceans; most of the rest is in glacier-covered Greenland and Antarctica (Table 1). During much of Earth's history the oceans have been a difficult, if not quite impassable, barrier for the movement of land-based plants and animals from one continent or island to another. This article describes and compares some of the

	Volume	
	km <sup>3</sup>	%
Oceans	1,348,000,000	97.39
Polar ice caps and glaciers	27,820,000	2.01
Ground water	8,062,000	0.58
Lakes and rivers	225,000	0.02
Atmosphere	13,000	0.001

major oceans and their features. See GLACIOLOGY; OCEANOGRAPHY.

Oceans and the seas that connect them cover some 73% of the surface of the Earth, with a mean depth of 3729 m (Table 2). More than 70% of the oceans have a depth between 3000 and 6000 m. Less than 0.2% of the oceans have depths as great as 7000 m (Table 3).

The oceans are cold and salty. Some 50% have a temperature between 0 and 2°C and a salinity between 34.0 and 35.0‰. To a high degree of approximation, a salinity of 34‰ is the equivalent of 34 grams of salt in a kilogram of seawater. Water with a temperature above a few degrees Celsius is confined to a relatively thin surface layer of the ocean. Figure 1 shows the distribution of both temperature and salinity in the world's oceans. See SEAWATER.

Nearly all elements known to humankind have been found dissolved in seawater, and those that have not are assumed to be present. However, all but a few are found in very small amounts. Sodium chloride accounts for some 85% of the dissolved salts, and an additional four ions (sulfate, magnesium, calcium, and potassium) bring the total to more than 99.3%.

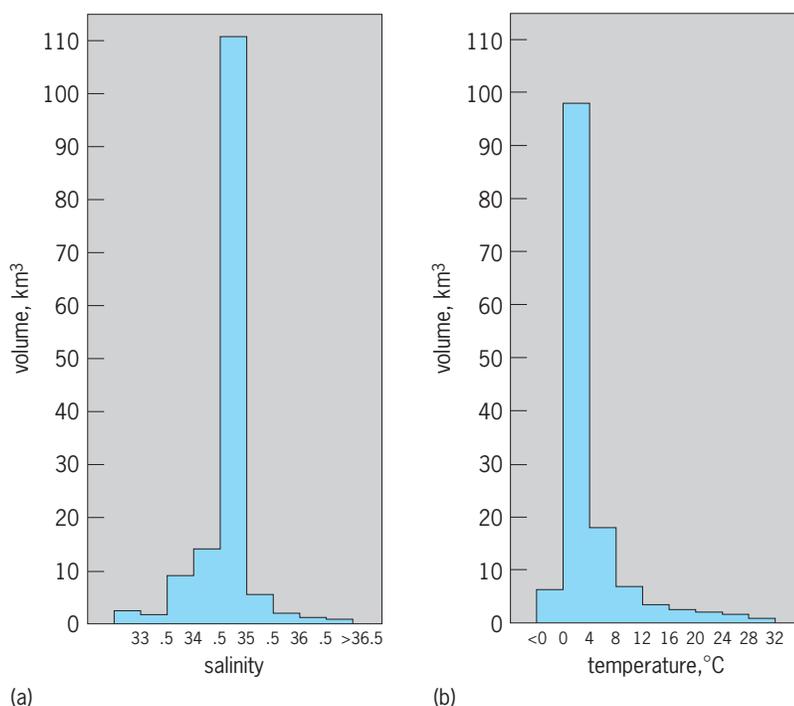


Fig. 1. Range of (a) temperature and (b) salinity in the world's oceans.

**TABLE 2. Ocean basin characteristics**

	Area, km <sup>2</sup>	Volume, km <sup>3</sup>	Mean depth, m
Pacific	181,344,000	714,410,000	3940
Atlantic	94,314,000	337,210,000	3575
Indian	74,118,000	284,608,000	3840
Arctic	12,257,000	13,702,000	1117
Total	362,033,000	1,349,929,000	3729

**TABLE 3. Distribution of oceanic depth**

Depth, m	Area, km <sup>2</sup>	Volume, %	Cumulative %
0–200	27,123,000	7.49	7.49
200–1000	16,012,000	4.42	11.91
1000–2000	15,844,000	4.38	16.29
2000–3000	30,762,000	8.50	24.79
3000–4000	75,824,000	20.94	45.73
4000–5000	114,725,000	31.69	77.42
5000–6000	76,753,000	21.20	98.62
6000–7000	4,461,000	1.23	99.85
7000–8000	380,000	0.10	99.96
8000–9000	115,000	0.03	99.99
9000–10,000	32,000	0.01	100.00
10,000–11,000	2,000	0.00	100.00

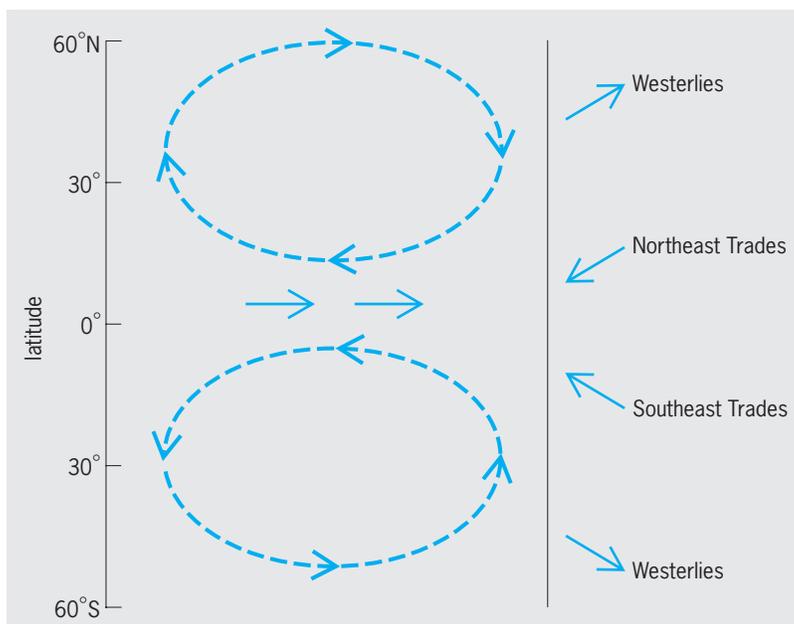
The ratio of ions is remarkably constant from one ocean to another and from top to bottom of each. Until about 1960, oceanographers calculated the total salinity by a titration process that was essentially equivalent to measuring only the chloride ion and multiplying by a constant. They could be confident that the uncertainty in their calculation was no more than  $\pm .02$  because of variations in the distribution of different ions within their samples. Calculation of the

salinity equivalent is now based on measurement of the electrical conductivity of the seawater, and is an order of magnitude more accurate.

Ocean salinity is primarily controlled by the balance of precipitation, river runoff, and evaporation of water at the sea surface (Table 4). The highest salinities are found in major evaporation basins with little rainfall or river runoff, such as the Red Sea. The lowest salinities are found near the mouths of major rivers such as the Amazon. Some of the lowest open-ocean salinities are found in the relatively small and isolated Arctic Ocean, which contains only 1% of the total volume of the oceans but drains several of the world's largest river systems, the Mackenzie from Canada and the Lena, Yenisei, and Ob from Russia. See ARCTIC OCEAN; RED SEA.

The tropics receive much more heat from the Sun than do the polar regions. However, the average temperature gradient between high and low latitudes is significantly less than might be expected, because the oceans are continually transporting excess heat (warm water) from the tropics toward the Poles and returning colder water toward the tropics. This process of moving excess heat from lower (south of 40°) to higher (north of 40°) latitudes is shared approximately equally by the oceans and the atmosphere. A significant part of the ocean heat exchange process is carried out by the major ocean currents, the "named" currents such as the Gulf Stream, Brazil Current, California Current, and Kuroshio. These currents are primarily driven by the winds, and there is considerable similarity in their pattern from one ocean basin to another. See GULF STREAM; KUROSHIO.

The average winds over the North and South Atlantic as well as the North and South Pacific oceans come out of the west (westerlies) at the middle



**Fig. 2. Role of the winds in establishing major current gyres (closed circulatory systems that are larger than whirlpools or eddys).** (From J. A. Knauss, *Introduction to Physical Oceanography*, 2d ed., Prentice Hall, 1997)

**TABLE 4. Water balance of the world's oceans\***

	<i>P</i>	<i>E</i>	<i>P</i> − <i>E</i>	River runoff	Runoff + <i>P</i> − <i>E</i>
Arctic	97	53	+44	307	+351
Pacific	1292	1202	+90	69	+159
Indian	1043	1294	−251	72	−179
Atlantic	761	1133	−372	197	−175

\* Precipitation (*P*) and evaporation (*E*) are in millimeter per year. River runoff is in millimeter per year equivalent for the appropriate ocean basin.

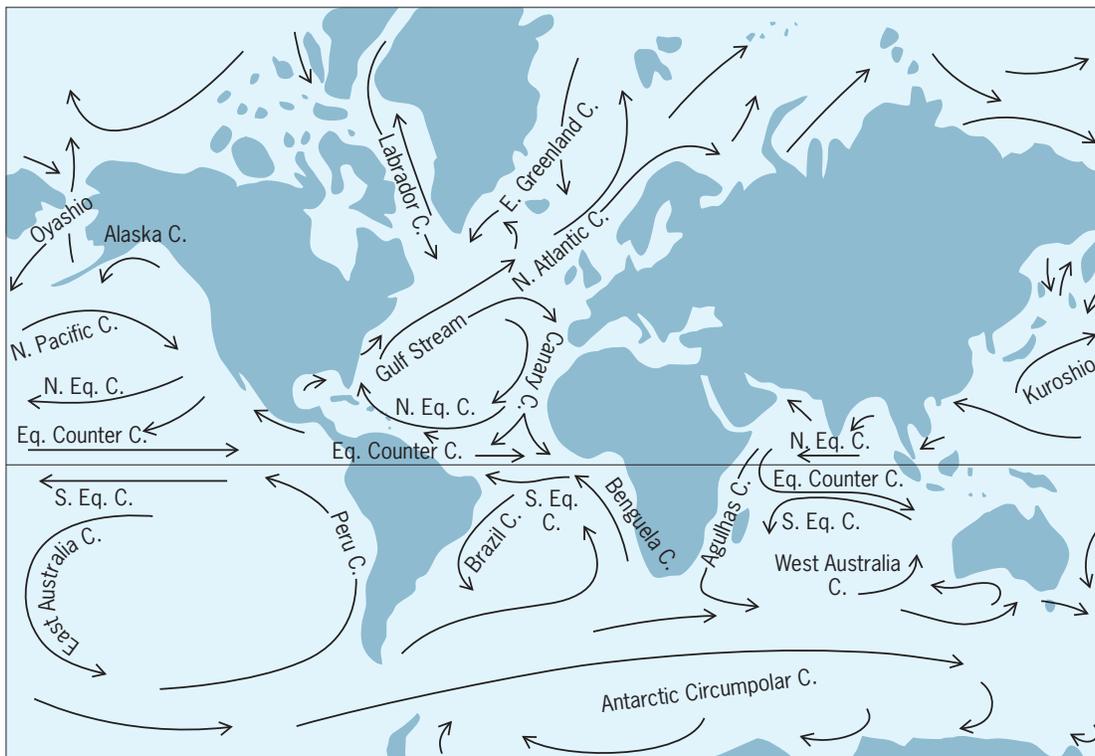


Fig. 3. Major currents of the world's oceans.

latitudes and from the east at the lower latitudes (trade winds). The frictional drag of these winds on the surface of the water imparts a spin or torque to the surface of the ocean, clockwise in the Northern Hemisphere and counterclockwise in the Southern Hemisphere (Fig. 2). The resulting basin-wide current gyres are apparent in the diagram of the major ocean currents in Figure 3, allowance being made for the shapes and boundaries of the different ocean basins. The major exception is the Indian Ocean north of the Equator, where the circulation is strongly influenced by the winds of the seasonal monsoon. See ATLANTIC OCEAN; CORIOLIS ACCELERATION; EQUATORIAL CURRENTS; INDIAN OCEAN; OCEAN CIRCULATION; PACIFIC OCEAN.

John A. Knauss

Bibliography. J. A. Knauss, *Introduction to Physical Oceanography*, 2d ed., Prentice Hall, 1997; M. E. Q. Pilson, *An Introduction to the Chemistry of the Sea*, Prentice Hall, 1998; E. I. Seibold and W. H. Berger, *The Sea Floor: An Introduction to Marine Geology*, 3d ed., Springer, 1996.

### Ocean circulation

The general circulation of the ocean. The term is usually understood to include large-scale, nearly steady features, such as the Gulf Stream, as well as current systems that change seasonally but are persistent from one year to the next, such as the Davidson Current, off the northwestern United States coast and the equatorial currents in the Indian Ocean. A great number of energetic motions have periods of a month or two and horizontal scales of a few hundred kilometers—a very low-frequency turbulence, collectively called eddies. Energetic motions are also concentrated near the local inertial period (24 h, at 30° latitude) and at the periods associated with tides (primarily diurnal and semidiurnal). See TIDE.

#### Surface Circulation

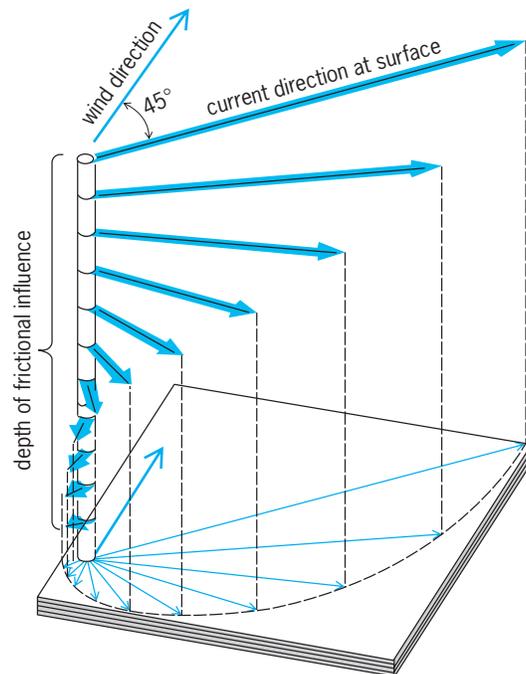
The greatest single driving force for currents, as for waves, is the wind. Furthermore, the ocean absorbs heat at low latitudes and loses it at high latitudes. The resultant effect on the density distribution is

coupled into the large-scale wind-driven circulation. Some subsurface flows are caused by the sinking of surface waters made dense by cooling or high evaporation.

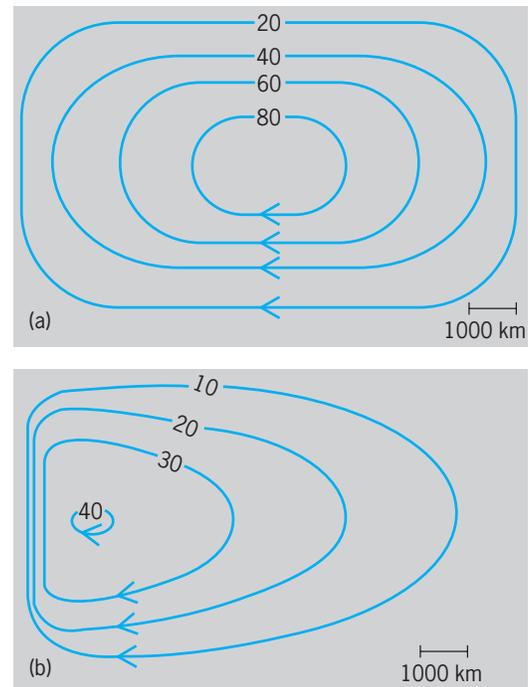
**Wind stress.** Air does not merely glide along the surface of the water, but exerts a frictional effect, or wind stress, which causes the surface water to be carried along with it. The movement of this thin layer on the surface of the water is conveyed by an internal turbulent friction to the deeper levels. The eventual result of such interaction, in a limitless homogeneous sea under the influence of a steady wind, would be a pure drift current. The resulting current distribution is illustrated by the so-called Ekman spiral (Fig. 1). There is a current at the sea surface at an angle to the right of the direction of the wind in the Northern Hemisphere. With increasing depth, the current turns farther toward the right and gradually subsides. When the direction of this current reaches an angle of  $180^\circ$  to the flow on the sea surface, the speed of the current is only  $1/23$  that of the surface water. This depth is called the depth of frictional influence. For example, at a latitude of  $50^\circ$  this depth amounts to about 200 ft (60 m) when the windspeed reaches 13.8 knots (7 m/s).

Observations show that the angle between wind and surface velocity is usually between  $30$  and  $45^\circ$ ; the surface velocity is about 2% of the wind speed. Both these factors depend on how strongly the upper layers are stratified and how effectively the vertical turbulent motions transmit the stresses downward.

The wind-driven currents are in addition to any other currents that are present—from the large-scale



**Fig. 1.** Ekman spiral. The arrows show the direction and magnitude of the purely wind-driven current, as it changes with depth. The angle between the wind and the surface current is a function of the vertical mixing by turbulent motions; if this mixing is uniform with depth, the angle is  $45^\circ$  as shown.



**Fig. 2.** Streamlines showing currents for (a) an ocean on a nonrotating globe and (b) an ocean on a uniformly rotating globe in which the Coriolis forces increase with the geographic latitude.  $1000 \text{ km} = 620 \text{ mi}$ . (After H. Stommel, *The Gulf Stream*, University of California Press, 1965)

density distribution, eddies, tides, and so on. One important feature of the wind-driven currents that is independent of stratification or turbulent intensity is that the total volume transport is at right angles to the wind. It is understandable, therefore, that upwelling of cold deep water occurs at a coast when the wind blows parallel to the coast, with the coast on the left-hand side of the wind. See UPWELLING; WIND STRESS.

An outstanding feature of ocean currents is that intense flows (such as the Gulf Stream) occur on the western sides of the oceans. These flows were shown to result from the variation of Coriolis parameter with latitude (Fig. 2).

**Surface currents.** Except in western boundary currents, and in the Antarctic Circumpolar Current, the system of strong surface currents is restricted mainly to the upper 330–660 ft (100–200 m) of the sea. The mid-latitude anticyclonic gyres, however, are coherent in the mean well below 3300 ft (1000 m). The average speeds of the open-ocean surface currents remain mostly below 0.4 knot (20 cm/s). Exceptions to this are found in the western boundary currents, such as the Gulf Stream, and in the Equatorial Currents of the three oceans, all of which have velocities of 2–4 knots (1–2 m/s). Knowledge of the surface currents is based in part on direct measurements of the current, and more generally on dead reckoning from ships. See DEAD RECKONING.

The primary causes of surface currents are wind stress and internal pressure forces resulting from the density distribution. Frictional forces and the Coriolis acceleration influence the surface currents. The effect of the wind is greatest when the direction

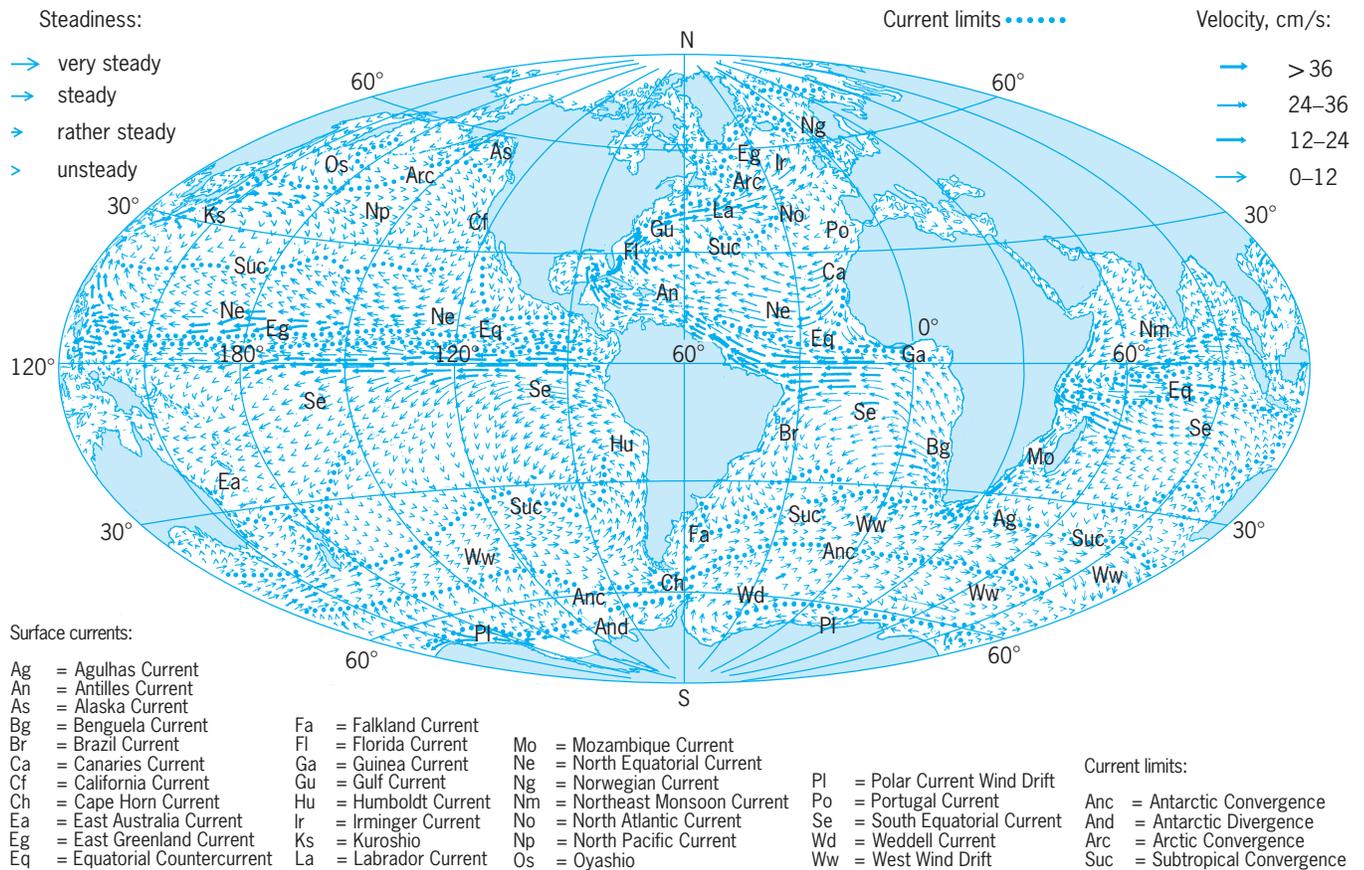


Fig. 3. Surface currents of the oceans in February and March. 1 cm/s = 0.019 knot.

and strength of the wind are steady; this is the case in lower and middle latitudes. In these latitudes an anticyclonic (clockwise, in the Northern Hemisphere) current system corresponds to the anticyclonic wind system (Fig. 3). Surface currents which flow in a westerly direction in the lower latitudes are part of this system (the North and South Equatorial Currents). The continuation of these currents is found along the eastern sides of these continents in narrow and strong surface currents directed toward the poles—the Western Boundary Currents; the Gulf Stream, Brazil Current, and Somali Current (only in the summer of the Northern Hemisphere); and the Agulhas Current, Kuroshio, and East Australia Current. In middle latitudes these currents turn and flow in an easterly direction (North Atlantic Current, North Pacific Current, and West Wind Drift of the Southern Hemisphere). On the eastern sides of the oceans this pattern contains surface currents directed toward the Equator—the Eastern Boundary Currents; the Canary, Benguela, West Australia, California, and Humboldt currents.

Embedded in the system of the North and South Equatorial Currents, which flow to the west, are the Equatorial Countercurrents, flowing to the east. These are found about 5° north of the Equator in all three oceans. An additional easterly flowing countercurrent is sometimes found in the eastern Pacific Ocean, south of the South Equatorial Current.

A large subsurface current is found centered on the Equator. This Equatorial Undercurrent is about 240 mi (400 km) wide but only 660 ft (200 m) thick (centered at a depth of 500-660 ft or 150-200 m), flowing to the east with peak speeds of 2-3 knots (1-1.5 m/s). This current is absent in the Indian Ocean in Northern Hemisphere summer; in the Pacific it is sometimes called the Cromwell Current. See EQUATORIAL CURRENTS.

In higher latitudes the currents tend to follow the coasts and shelf edges; in the Northern Hemisphere the continents lie to the right-hand side of the current when one looks in the downstream direction. Examples are the Norwegian, East Greenland, West Greenland, Labrador, and Alaska currents. Islands in this fashion are, so to speak, surrounded by currents moving in a clockwise direction (for example, Iceland by the Irminger, North Iceland, and East Iceland currents). It is for this reason that the western sides of continents in these latitudes are bordered by comparatively warm waters coming from lower latitudes, whereas off the eastern coasts of the same latitudes there are cold waters from higher latitudes. For example, in the Norwegian Current the surface temperature in summer is 18°F (10°C) higher than in the East Greenland Current; both are at the same latitude. Thus surface currents are of great climatic importance.

All surface currents contain vertical components

that vary from region to region. These vertical current components are influenced by converging or diverging winds, acceleration of the currents, and other factors. The vertical components are certainly very small [for example, a large value would be a speed of 0.0012 in./s (0.003 cm/s) in the upwelling area of the California Current], but in the long run they are of great importance when considering the balance of heat and all sea-contained substances. See MARITIME METEOROLOGY; UPWELLING.

One important idea about the large mid-ocean gyres is that their transport is determined not simply by the wind stress but by the latitudinal variation of the east-west winds.

The maximum transport in the Florida Current occurs in June or July, almost 6 months out of phase. The seasonal variation is from about  $9.5$  to  $12 \times 10^8$  ft<sup>3</sup>/s ( $27$  to  $33 \times 10^6$  m<sup>3</sup>/s) in volume transport in the Straits of Florida; this transport increases to approximately  $35 \times 10^8$  ft<sup>3</sup>/s ( $100 \times 10^6$  m<sup>3</sup>/s) as the Stream passes the longitude of Cape Cod.

The idea has been advanced that a substantial portion of the transport of the Western Boundary Currents does not continue to the northeast but returns southerly in a weak countercurrent just offshore.

**Current variability.** Most of the major ocean currents have been known fairly well for some time. However, advances have been made in trying to understand the variations in ocean currents. The instantaneous path of the Gulf Stream (Fig. 4) has been determined by observations of its sharp surface-temperature gradient. The instantaneous path is much more sinuous than the long-term average path; the wavelike path variations will change entirely in a period of roughly 2 weeks to a month. Some wavelike meanders grow to have very large amplitudes, which become unstable and “pinch off” to form

“Gulf Stream rings,” or eddies (Fig. 4). The water in the center of a ring is cold or warm, depending merely upon the phase of the meander that pinched off. These eddies usually drift to the west or southwest.

South of Cape Hatteras, the Gulf Stream may be found going  $45$  to  $90^\circ$  away from its “normal” path. Downstream from Cape Hatteras, the meanders may grow to hundreds of miles away from the mean path, which becomes convoluted; a ship traveling perpendicular to the Stream, for example, could cross it three times.

### Deep Circulation

The deep circulation results in part from the wind stress and in part from the internal pressure forces which are maintained by the budgets of heat, salt, and water. Both groups of forces are dependent upon atmospheric influences. Apart from Coriolis and frictional forces, the topography of the sea bottom exercises a decisive influence on the course of deep circulation.

**Marginal seas.** The deep circulation in marginal seas depends largely on the climate of the region, whether arid or humid.

*Arid climates.* Under the influence of an arid climate, evaporation is greater than precipitation. The marginal sea (Fig. 5) is therefore filled with relatively salty water of a high density. Its surface lies at a lower level than that of the neighboring ocean. Examples of this type are the Mediterranean Sea, Red Sea, and Persian Gulf. At the connection between the two seas there is water of a slightly lower density from the ocean flowing in at the surface. The water from the marginal sea flows over the sill into the ocean, where it sinks to a level at which it finds water corresponding to its density. Substantial vertical mixing takes place during this initial flow. At the deeper level it then spreads horizontally. The waters from the Mediterranean Sea and the Red Sea, because of their high salinity, can be followed far out into the Atlantic and Indian oceans, respectively. See INDIAN OCEAN; MEDITERRANEAN SEA.

*Humid climates.* The deep circulation of marginal seas in humid climates shows a different pattern, however (Fig. 5b). The level of the sea is higher than in the neighboring ocean. Therefore, the surface water with its lower density and accordingly its lower salinity flows outward, and the relatively salty ocean water of higher density flows over the sill into the marginal sea. Examples of this circulation are the Baltic Sea with the shallow Darsser and Drogden rises, the Norwegian and Greenland fiords, and the Black Sea with its entrance through the Bosphorus.

The Black Sea is an example of a special case. The sill depth always remains in the water of low density—that is, the outflowing upper layer. The renewal of deep water, and with it the deep circulation, comes to a complete halt. The result is that the oxygen is entirely used up, and poisonous hydrogen sulfide is generated. Below depths of 660 ft (200 m), only anaerobic organisms live in the Black Sea.

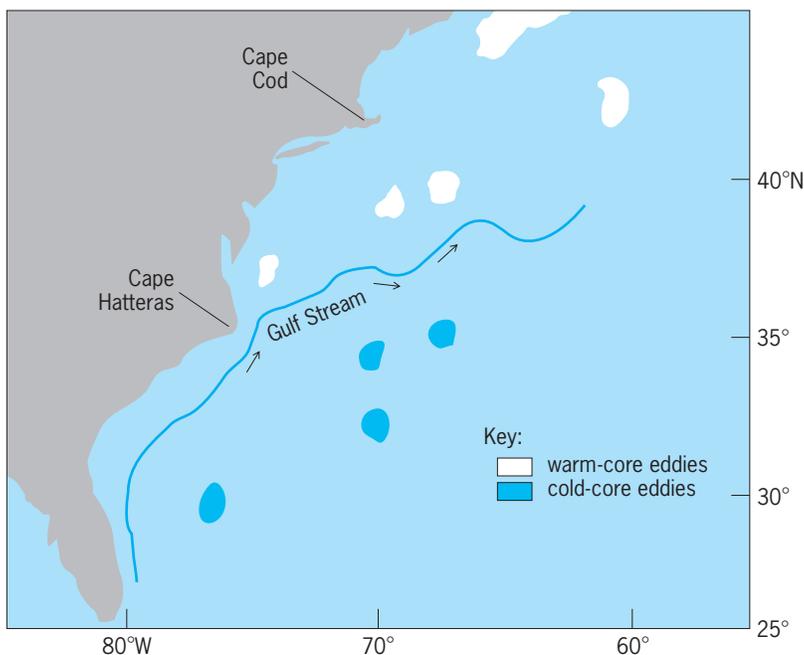


Fig. 4. Position of the Gulf Stream at the beginning of October 1975.

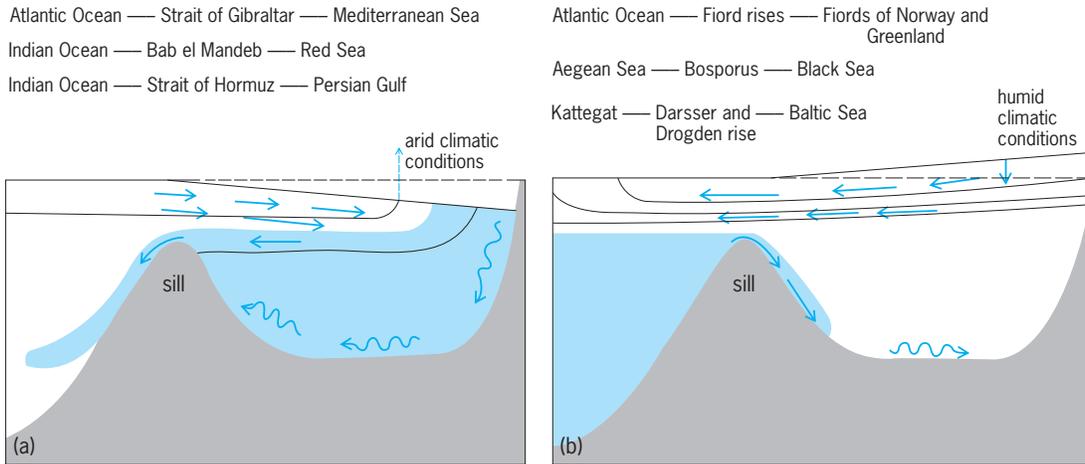


Fig. 5. Schematic representation of the circulation in marginal seas and over their sills (a) in arid climates, for example, Mediterranean Sea, Red Sea, and Persian Gulf; (b) in humid climates, for example, Black Sea, Baltic Sea, and fjords of Norway and Greenland.

If the sill depth interferes only occasionally with the lighter water of the upper levels, as in the entrances to the Baltic Sea, the renewal of deep water is interrupted at intervals. In the Baltic Sea these interruptions sometimes last for several years. See BALTIC SEA; BLACK SEA; FIORD.

**Oceans.** The deep circulation in the oceans is more difficult to perceive than the circulation in the marginal seas. In addition to the internal pressure forces, determined by the distribution of density and the piling up of water by the wind, there are also the influences of Coriolis forces and large-scale turbulence. There are areas in tropical latitudes in which

the surface water, as a result of strong evaporation, has a relatively high density. In thermohaline convection, the water sinks while flowing horizontally until it reaches a density corresponding to its own, and then spreads out horizontally. In this way the colder and deeper levels of the oceans take on a layered structure consisting of the so-called bottom water, deep water, and intermediate water. In the Atlantic Ocean the deep-water circulation is strong on the western side of the ocean, where measurable speeds—roughly 4 in./s (10 cm/s)—are found (Fig. 6). The Bottom Water comes from very cold water masses that have sunk along the edge of

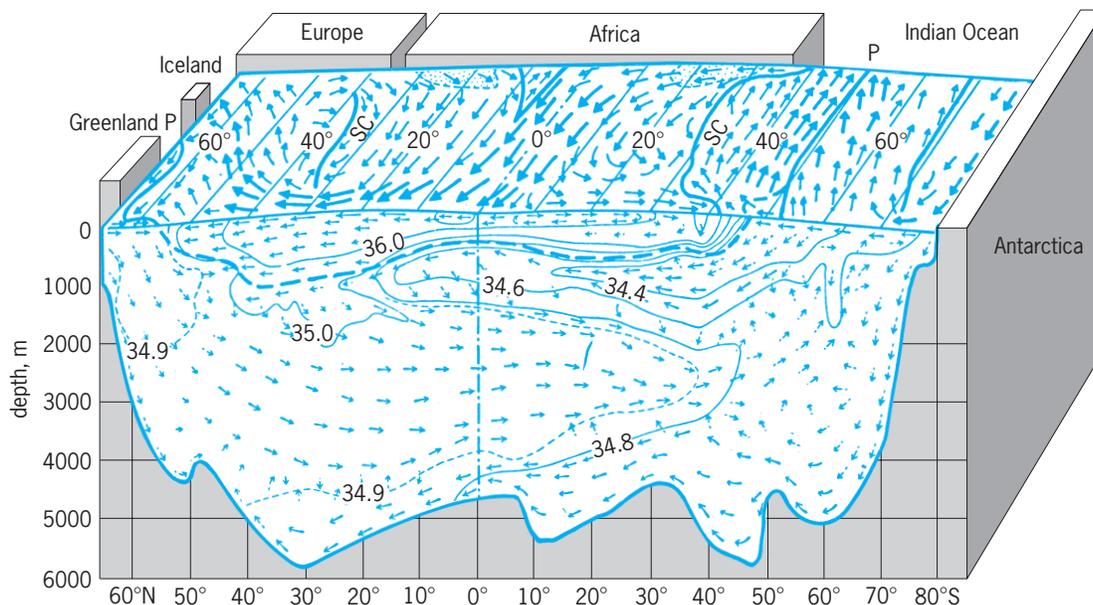


Fig. 6. Schematic representation of the surface and deep circulation in the Atlantic Ocean. All arrows show current directions; on the surface thin arrows indicate speeds of 0.1–0.8 knot (5–40 cm/s), and thick arrows indicate speeds of 0.8–2.9 knots (40–150 cm/s). SC indicates convergence of surface currents in subtropical waters. P indicates oceanic polar front where cold-water masses from polar and subpolar geographical latitudes meet relatively warm waters of temperate zone. In vertical section the heavy broken line shows division between warm- and cold-water spheres, and other lines indicate equal salinities. 1 m = 3.3 ft.

Antarctica. In the layer immediately above, or Deep Water, the water masses have their origin in the far North Atlantic. In the next layer above, or Intermediate Water, the water south of about 30°N comes from the Southern Hemispheric polar front.

There are five major areas where the surface water becomes denser and sinks.

1. The Norwegian Sea, where the water with the highest density of the world oceans is formed. This supplies the North Polar Sea with cold deep water, as well as the North Atlantic Ocean with cold bottom water to a latitude of approximately 50°N.

2. The Antarctic continental slope in the Weddell Sea, where water temperatures beneath the winter pack ice are near freezing (for seawater, -1.9°C or 28.6°F). Favored by the topography of the ocean floor, this water flows northward in the western Atlantic to the foot of the Grand Banks, as well as through the Romanche Deep on the Equator (through the Mid-Atlantic Ridge) into the eastern side of the Atlantic. It also spreads northward into the Indian and Pacific oceans, flowing toward the Equator on the western side of each ocean.

3. In the Labrador and Irminger seas, where the Deep Water (between depths of 3300 and 13,000 ft or 1000 and 4000 m) of the Atlantic Ocean originates. Apart from low temperatures, this water is distinguished by its richness in oxygen. This water has been traced around South America and into the Pacific Ocean.

4. The polar front at latitude of 50°S. Here cool water with low salinity is formed, supplying the Antarctic Intermediate Water of the Atlantic, Indian, and Pacific oceans.

5. The polar front in the North Pacific Ocean, where North Pacific Intermediate Water is formed. The distribution of this water in the North Pacific Ocean has been studied extensively.

The deep-sea circulation of the Atlantic Ocean appears to be the most active in comparison with the Indian and Pacific oceans, because the important sources of thermohaline convection are found in the Atlantic. In addition, the continental barrier of South America seems to force water from the surface currents of the South Atlantic into the North Atlantic Ocean. To compensate for the loss of surface water in the South Atlantic, there is a more active deep-water circulation, in which North Atlantic Deep Water flows southward into the South Atlantic Ocean. *See ATLANTIC OCEAN; PACIFIC OCEAN.* Wilton Sturges

### Mesoscale Eddies

Wherever oceanographers have made long-term current and temperature measurements, they have found energetic fluctuations with periods of several weeks to several months. These low-frequency fluctuations (compared to tides) are caused by oceanic mesoscale eddies which are in many respects analogous to the atmospheric mesoscale pressure systems that form weather. Like the weather, mesoscale eddies often dominate the instantaneous current, and are thought to be an integral part of the ocean's general circulation.

**Time and space scales.** The most exhaustive field studies of mesoscale eddies have been conducted in the Sargasso Sea of the western North Atlantic. A wide variety of eddy types and sizes have been found, and probably others remain to be discovered. A typical Sargasso Sea eddy has a period of about 2 months and a diameter of about 210 mi (350 km). Some eddies are approximately circular in plan view, while others may be quite elongated or irregular. The eddy vertical structure is most commonly found to be baroclinic, with the largest currents near the sea surface, but barotropic eddies that have almost uniform currents from the sea surface to the ocean bottom have also been observed. Currents may be as large as 12 in./s (30 cm/s), and temperature surfaces displaced from their average depth by up to 330 ft (100 m). The sea surface displacement is about 4 in. (10 cm) and can be detected by satellite-borne radar altimeters.

The horizontal momentum balance for mesoscale eddies is well approximated by the geostrophic vertical shear. Eddy currents may thus be calculated from observations of the temperature and salinity field (which may be used to calculate density), and either an observed reference current or sea surface displacement.

**Distribution and sources.** Eddies occur in virtually all oceans and seas, but their amplitude varies greatly from place to place. The largest amplitudes are found on the western sides of the oceans in conjunction with the strongest ocean currents (the Gulf Stream in the North Atlantic, the Kuroshio in the North Pacific) and near the Equator. The typical Sargasso Sea eddy is representative of the powerful eddies found in such regions. Much weaker eddies are found in the ocean interior, distant from major currents.

This consistent pattern of eddy amplitude suggests that instabilities of western boundary currents are an important source of eddy energy. In some cases the instability results in a pinched-off current ring made up of the western boundary current water, but on other occasions it results in eddy motions in the surrounding water. Atmospheric forcing by variable winds can also generate eddies, and is probably most important at low latitudes where the horizontal scales of the oceanic eddies best match the scales of the atmospheric forcing.

**Eddy dynamics.** Eddies propagate and evolve in ways that depend in large part upon the eddy amplitude. If the eddy amplitude is small (currents less than a few centimeters per second) as it is over much of the eastern North Atlantic and Pacific, then the eddies tend to behave like linear planetary waves. Their phase speed is westward at typically 2 in./s (5 cm/s), and their group speed is eastward at a slightly smaller rate. If the eddy current amplitude is larger than the phase speed, for example, in the Sargasso Sea, then the eddies undergo continual scale change and exchange energy with neighboring eddies. This sort of chaotic, random behavior has the characteristics of turbulence. Mesoscale eddies that behave this way are a very effective mechanism for mixing heat, chemicals, and other tracers on an ocean basin-wide

scale. One of the most important research problems for contemporary oceanography is to learn how this eddy mixing affects the general circulation of the oceans.

James F. Price

### Measurement of Ocean Currents

One current-measuring method is to measure the flow of water past a point. Just as winds are measured by attaching a propeller-driven anemometer to a tower or building, current meters are attached to buoys anchored in the ocean, or lowered from anchored or floating ships. Unlike the wind anemometers attached to a fixed structure, floating ships move with the wind and current, and even anchored ships and buoys are not truly stationary. In the absence of any ocean current, these instruments measure the movement of the current meter through the water. Separating the movement of the current meter through the water from the velocity of the water past a fixed point can be a challenge. A variety of techniques have been employed either to minimize the movement of the meter or to accurately determine its movement, but all such measurements leave some uncertainty, and where the currents are weak, the order of a few hundredths of a meter per second, the uncertainty is often of the same order as the current that one is attempting to measure.

A second method measures the drift of a free body such as a drogoue, a current-measuring device consisting of a weighted parachute and attached surface buoy. The earliest reliable current charts were constructed using a variation of this method, and much of the most recent information has been gathered by this technique. Systematic collection of ship drift data began in the midnineteenth century. The navigator would note the ship's noon position based on stars, sun, and dead reckoning, and calculate where the ship should be in 24 h. In the absence of very strong winds, the navigator would assume that the difference in the dead-reckoned position and the actual position revealed by stars and sun was a result of surface currents. By the end of the century, remarkably good charts of the average ocean surface currents were available.

**Surface measurements.** Surface drogues, carefully designed to minimize windage, equipped with Global Positioning System (GPS) receivers are the modern-day descendant of this technique. These surface drogues can report their position several times a day to a passing communication satellite, which in turn relays the drogues' positions to a central shore station. Thousands of these drogues are being deployed around the world, and surface current charts of unprecedented accuracy are now becoming available.

**Subsurface measurements.** Subsurface drifters measure currents at depth. The most common technique is for the container of the subsurface drifter to be less compressible than seawater. Weighted at the surface to be more dense than seawater, the density difference between seawater and float decreases as the float sinks until the densities match, at which pressure the float will remain. These sub-

surface floats can be designed to sink to a predetermined pressure surface with the equivalent accuracy of a few meters in depth. A more elegant technique is to design the container to have the same compressibility as seawater and to weigh the floats to sink to a given density surface. These so-called isopycnal (constant density) floats follow a given density surface rather than a given pressure surface and thus provide a better measure of the path of a given parcel of water at depth. *See* INSTRUMENTED BUOYS; SATELLITE NAVIGATION SYSTEMS.

Tracking subsurface floats is more difficult than tracking surface floats. One method is to design a float that can make itself lighter, return to the surface, and broadcast its position using the GPS signal to locate itself, as do the surface floats. These floats can be made to take on additional ballast, to sink, and to continue to follow along a given pressure surface. Subsurface floats of this type have been designed to float along at depth, come to the surface and signal their location, then sink again, as many as 150 times during a four-year period. By insuring that the submerged time is long compared to the time it takes for the float to move from depth to the surface and back, the aliasing of the subsurface current measurement is minimized. A second way of tracking subsurface floats is by using underwater sound. One method is for the subsurface float to carry a sound source and be tracked in a manner analogous to that used in tracking submarines. Another method is to anchor a number of buoys in the area where the floats are to be deployed, and attach sound sources to the buoys and recording hydrophones to the floats. The buoys have accurate clocks and transmit sound signals on a regular schedule that are recorded by the floats. The floats are programmed to come to the surface months or years later and transmit those recorded signals to a passing satellite, which in turn transmits those recordings to a shore station laboratory. Noting the difference in times of arrival of each of the transmitted signals and using simple triangulation geometry, each float's position can be determined as a function of time with an accuracy of better than a nautical mile.

**Geostrophic current.** Historically, much of the information about currents below the ocean surface comes not from direct measurement, but from the assumption that these currents are in geostrophic balance (where the Coriolis force balances exactly the horizontal pressure gradient). By careful measurement of temperature and salinity as a function of depth at two locations, one can determine seawater density, and in turn calculate the hydrostatic pressure as a function of depth. The difference in density at the two stations results in differences in hydrostatic pressure, which in turn allows one to calculate as a function of depth changes in the geostrophic current normal to an imaginary line connecting the two stations. Recent improvements in satellite altimeters now allow one to measure differences of a few centimeters in sea level surface. Knowing the slope of the sea surface with respect to its equal-potential surface, one can calculate the geostrophic surface

current. See CORIOLIS ACCELERATION; OCEANOGRAPHY; SEAWATER. John A. Knauss

Bibliography. J. R. Apel, *Principles of Ocean Physics*, 2d ed., 2001; R. A. Davis, Jr., *Oceanography: An Introduction to the Marine Environment*, 2d ed., 1996; W. J. Emery (ed.), *Descriptive Physical Oceanography*, 5th ed., 1982; T. S. Garrison, *Oceanography: An Invitation to Marine Science*, 2d ed., 1996; M. N. Hill (ed.), *The Sea*, vol. 1, 1982; F. W. Smith and F. A. Kalber (eds.), *Handbook in Marine Science*, 2 vols., 1974; H. Stommel, *A View of the Sea: A Discussion Between a Chief Engineer and an Oceanographer about the Machinery of the Ocean Circulation*, 1991; K. Stowe, *Essentials of Ocean Science*, 1988; T. Teramoto, *Deep Ocean Circulation: Physical and Chemical Aspects*, 1993.

## Ocean waves

Propagating oscillations in the ocean which carry energy and momentum from one region to another. Most ocean waves are caused directly or indirectly by wind blowing across the sea surface. Many waves can propagate through the ocean thousands of miles from where they are generated.

**Surface waves.** Ocean surface waves are propagating disturbances at the atmosphere-ocean interface. They are the most familiar ocean waves. Surface waves are also seen on other bodies of water, including lakes and rivers. See WAVE MOTION IN LIQUIDS.

A simple sinusoidal wave train is characterized by three attributes: wave height ( $H$ ), the vertical distance from trough to crest; wavelength ( $L$ ), the horizontal crest-to-crest distance; and wave period ( $T$ ), the time between passage of successive crests past a fixed point. The phase velocity ( $C = L/T$ ) is the speed of propagation of a crest. For a given ocean depth ( $b$ ), wavelength increases with increasing period. The restoring force for these surface waves is predominantly gravitational. Therefore, they are known as surface gravity waves, unless their wavelength is shorter than 1.8 cm (0.7 in.), in which case surface tension provides the dominant restoring force.

**Classification.** Surface gravity waves may be classified according to the nature of the forces producing them. Tides are ocean waves induced by the varying gravitational influence of the Moon and Sun. They have long periods, usually 12.42 h for the strongest constituent. Storm surges are individual waves produced by the wind and dropping barometric pressure associated with storms; they characteristically last several hours. Earthquakes or other large, sudden movements of the Earth's crust can cause waves, called tsunamis, which typically have periods of less than an hour. Wakes are waves resulting from relative motion of the water and a solid body, such as the motion of a ship through the sea or the rapid flow of water around a rock. Wind-generated waves (Fig. 1), having periods from a fraction of a second to tens of seconds, are called wind waves. Like tides, they are ubiquitous in the ocean, and continue to travel well



Fig. 1. Open-ocean wind waves driven by gale-force (18 m/s or 40 mi/h) winds.

beyond their area of generation. The ocean is never completely calm. See STORM SURGE; TIDE; TSUNAMI.

**Wind waves.** The growth of wind waves by the transfer of energy from the wind is not fully understood. At wind speeds less than 1.1 m/s (2.5 mi/h), a flat water surface remains unruffled by waves. Once generated, waves gain energy from the wind by wave-coupling of pressure fluctuations in the air just above the waves. For waves traveling slower than the wind, secondary, wave-induced airflows shift the wave-induced pressure disturbance downwind so the lowest pressure is ahead of the crests. This results in energy transfer from the wind to the wave, and hence growth of the wave.

If a constant wind blows over a sufficient length of ocean, called the fetch, for a sufficient length of time, a wave field develops whose statistical characteristics depend only on wind velocity. In particular, the spectrum of sea-surface elevation for such a fully-developed sea has the form of Eq. (1), where  $f$

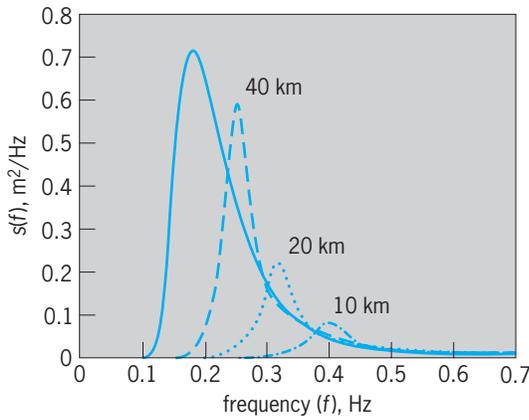
$$S(f) = A \frac{g^2}{f^5} e^{-1.25(f_m/f)^4} \quad (1)$$

is frequency ( $= 1/T$ ),  $g = 9.8 \text{ m/s}^2$  (32 ft/s<sup>2</sup>) is gravitational acceleration,  $f_m = 0.13 g/U$  is the frequency of the spectral peak ( $U =$  wind speed at 10 m or 32.8 ft elevation), and  $A = 5.2 \times 10^{-6}$  is a constant.

The fetch is limited near a coast with the wind blowing offshore, and the waves grow as they propagate toward the open ocean. In such a limited-fetch situation, Eq. (1) is modified:  $A$  and  $f_m$  become dependent on the fetch length, and the peak in the spectrum is enhanced. Figure 2 shows spectral forms of waves generated by a moderate breeze for various fetches and in the open sea. For faster wind speeds, the spectral peaks grow in height and shift to lower frequencies.

Even when the mean wind blows from a single direction, the surface waves that it generates are seen to travel in a variety of directions centered on the downwind direction. The directional spectrum of such a wave field can be approximately represented by a formula, such as Eq. (2), where  $\theta$  is the angle

$$S(f, \theta) = \begin{cases} S(f) \frac{2}{\pi} \cos^2 \theta & \text{if } \theta \leq 90^\circ \\ 0 & \text{if } \theta > 90^\circ \end{cases} \quad (2)$$



**Fig. 2.** Spectra of waves from a 7 m/s (16 mi/h) wind in the open sea (solid line), and for three limited fetches: 40, 20, 10 km (25, 12, 6 mi). (Theoretical relations are from W. J. Pierson and L. Moskowitz, *J. Geophys. Res.*, 69:5181–5190, 1964, for the open sea, and K. Hasselmann et al., *Deutsch. Hydrogr. Z., Reihe A (8<sup>o</sup>), Nr. 12, 1973, for fetch-limited seas*)

of wave propagation relative to the downwind direction.

An observer asked to estimate average wave height typically gives a value that is about the average height of the highest one-third of the waves actually present. This statistic, represented as  $H_{1/3}$ , is called the significant wave height. In the open sea, Eq. (3) applies.

$$H_{1/3} = 0.24 \frac{U^2}{g} \quad (3)$$

If a wind blows steadily over a known fetch for a period of time, the resulting  $H_{1/3}$  may be estimated from Fig. 3.

Because of viscosity, surface waves lose energy as they propagate, short-period waves being dampened more rapidly than long-period waves. Waves with long periods (typically 10 s or more) can travel thousands of kilometers with little energy loss. Such waves, generated by distant storms, are called swell. Equations (1), (2), (3) and Figs. 2 and 3 assume that the waves present were generated by local wind, with no significant swell present. See VISCOSITY.

The highest wind waves are produced by large intense storm systems that last for a day or longer. Such systems of very low atmospheric pressure form in the Gulf of Alaska, the region around Iceland, and the Weddell Sea. Off the west coast of Canada, there have been several measurements of individual waves with heights around 30 m (100 ft). Northwest of Hawaii, on February 7, 1933, the Navy tanker USS *Ramapo* encountered the largest open-ocean wind waves ever reliably observed with heights that were reported to be at least 34 m (112 ft).

When waves propagate into an opposing current, they grow in height. For example, when swell from a Weddell Sea storm propagates northeastward into the southwestward-flowing Agulhas Current off South Africa, high steep waves are formed. Many large ships in this region have been severely damaged by such waves.

Because actual ocean waves consist of many components with different periods, heights, and direc-

tions, occasionally a large number of these components can, by chance, come in phase with one another, creating a freak wave with a height several times the significant wave height of the surrounding sea. According to linear theory, waves with different periods propagate with different speeds in deep water, and hence the wave components remain in phase only briefly. But nonlinear effects are bound to be significant in a large wave. In such a wave, the effects of nonlinearity can compensate for those of dispersion, allowing a solitary wave to propagate almost unchanged. Consequently, a freak wave can have a lifetime of a minute or two. See SOLITON.

**Linear theory.** For waves sufficiently small that linear theory applies, Eq. (4), gives the phase velocity

$$C = \begin{cases} (g/k)^{1/2} & \text{deep water } b > 0.4L \\ (gb)^{1/2} & \text{shallow water } b < 0.04L \end{cases} \quad (4)$$

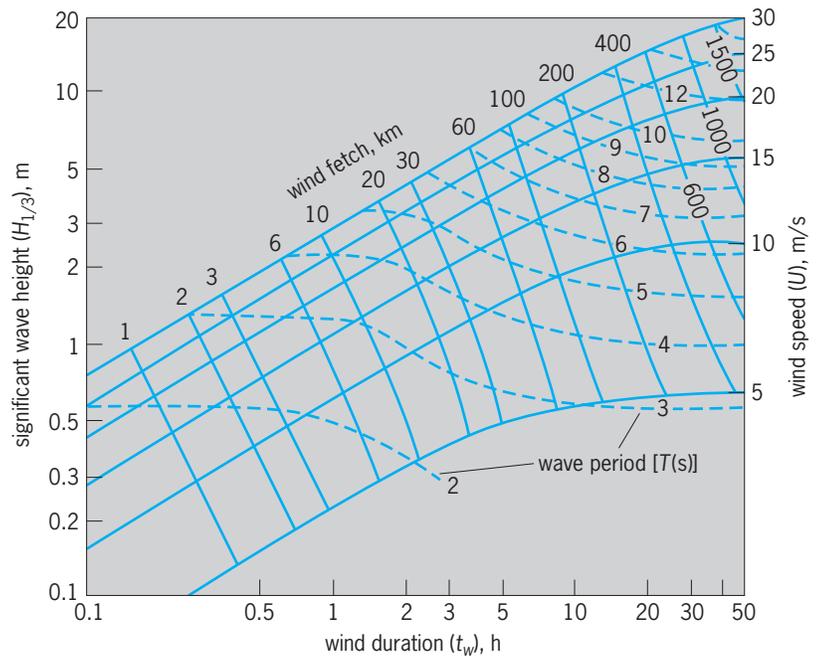
to an accuracy of 1%. Note that wavenumber  $k = 2\pi/L$ . See WAVE MOTION IN LIQUIDS.

Ocean wave energy  $E$  per unit surface area depends only on wave height [Eq. (5)]. This energy propagates at the group velocity [Eq. (6)] and pro-

$$E = \frac{1}{8} \rho g H^2 \quad (5)$$

$$C_g = \begin{cases} \frac{1}{2}C & \text{deep water} \\ C & \text{shallow water} \end{cases} \quad (6)$$

duces a power flux per unit distance along the wave



**Fig. 3.** Relationship of significant wave height  $H_{1/3}$  and characteristic wave period  $T_m$  ( $=1/f_m$ ) to wind speed, duration, and fetch. To estimate the wave characteristics resulting from a wind of velocity  $U$  which has blown steadily for a time  $t_w$ , find one of the six heavy curves labeled on the right with the appropriate  $U$ , follow this curve to the correct  $t_w$ , and, at that point, read the value of  $H_{1/3}$ . The position of this point, relative to the broken curves, indicates the wave period  $T_m$ . If the waves are fetch-limited,  $H_{1/3}$  grows with time only until the heavy curve intersects the appropriate fetch line, at which point  $H_{1/3}$  and  $T_m$  remain fixed. Open-sea conditions exist on the right side of the diagram. (After World Meteorological Organization, *Guide to Wave Analysis and Forecasting*, no. 702, 1988)

front [Eq. (7)]. For example, waves in the deep sea

$$P = C_g E \quad (7)$$

with 1.8-m (6-ft) height and 10-s period carry a power flux of 30 kW/m (13 hp/ft). Available power fluxes of this magnitude are representative of many coastal regions, and represent a substantial renewable energy resource. Various devices with efficiencies of 50% or greater have been developed to convert this wave energy to electrical energy. See HYDROELECTRIC GENERATOR.

Surface-gravity waves cause pressure fluctuations and particle motions that are largest near the surface and decrease with depth. In deep water, this dependence on depth  $d$  below the still-water level is exponential:  $e^{-kd}$ .

**Stokes drift.** To first-order (linear theory), particle orbits resulting from wave motion are closed loops. But since a particle moves in the direction of wave propagation in the loop's upper part and in the reverse direction in its lower part, the forward motion is at a shallower level than the reverse motion. Consequently, the forward motion is slightly stronger than the reverse motion, and the orbital loops do not quite close. As a result, there is a second-order (nonlinear) net particle motion called Stokes drift. In deep water, the velocity associated with this motion is in the direction of wave propagation [Eq. (8)]. In

$$\bar{u} = C(ak)^2 e^{-2kd} \quad (8)$$

shallower water, Stokes drift contributes to sediment transport onto beaches, and along beaches when the waves approach the coast at an angle. See NEARSHORE PROCESSES.

**Shoaling and breaking.** Waves approaching the shore from the open ocean are affected in several ways. As they become shallow-water waves, velocity  $C$  decreases with decreasing water depth  $b$  [Eq. (4)]. Consequently, waves approaching the shore at an angle are refracted so their crests are brought nearly parallel to the shoreline. Wave period  $T$  does not change as  $b$  decreases, so wavelength  $L = CT$  must decrease according to Eq. (4). Shallowing also causes a growth in wave height as in relation (9), where the

$$H \propto b^{-r} \quad (9)$$

value of  $r$  depends on the character of the waves. For example, in the absence of dissipation,  $r = 1/4$  according to linear theory for shallow-water waves [Eqs. (4)–(7)], while  $r = 1$  for solitary waves of moderate amplitude.

As wave height grows, nonlinear terms in the equations of motion become significant. At first, this results in a vertical asymmetry in the wave shape, with crests becoming more peaked and troughs more rounded. Then, as the wave moves into shallow water, a strong horizontal asymmetry develops in which the forward face of the wave becomes progressively steeper than the backward face. This pro-

cess typically continues until the wave breaks. The resulting breakers can have a variety of forms (collapsing, plunging, surging, and so forth), depending on the height of the entering waves and the slope of the seabed. A rough criterion for breaking is shown in relation (10), or, using Eq. (4) for shallow water,  $H > b$ .

$$H > \frac{C^2}{g} \quad (10)$$

In the open ocean also, waves often break (Fig. 1). In this case, the criterion given by relation (10) becomes  $H > 1/k$  [using Eq. (4) for deep water]. Whitecaps from deep-water breaking waves begin to appear at wind speeds of about 0.45 m/s (10 mi/h), while at wind speeds above 27 m/s (60 mi/h) all the high waves are breaking.

**Wave measurement.** There are three classes of instruments for measuring ocean surface waves: those which are at the air-sea interface, those which are below it, and those which are above it. Because so many techniques can be used, only a few typical examples of each class will be described.

*At the surface.* Ocean surface waves can be measured from a dock with a wave staff held vertically in the water. The varying position  $\eta(t)$  of the sea surface on the staff is sensed in a variety of ways, for example, by seawater-shorting of the submerged part of a resistance wire wound along the staff.

An accelerometer-instrumented buoy on a slack mooring line can provide a record of  $\eta(t)$  by double integration of the vertical acceleration. If, in addition, the buoy contains tilt sensors, it is capable of providing the directional spectrum  $S(f, \theta)$  of the waves. Typically, data from a wave-measuring buoy are telemetered to a receiving station on shore. See INSTRUMENTED BUOYS.

*Below the surface.* The most common instrument of this class is the subsurface pressure sensor. The pressure measurement must be made at a level deep enough that it is always submerged; this has the advantage of reducing vulnerability to damage by ships and breaking waves. The main disadvantage of the method is the need to compensate for the frequency-dependent depth attenuation of the measured wave-induced pressure signal  $p(t)$  in converting it to surface elevation  $\eta(t)$ .

A narrow-beam inverted echo sounder can also be used to make subsurface wave measurements. It is placed on the sea floor and directed upward, so the acoustic echo time from the surface is a measure of sea-surface elevation. But variations in temperature and salinity affect the speed of sound in water, and hence affect instrument calibration. Also, bubbles in the water can cause spurious acoustic reflections. See ECHO SOUNDER.

*Above the surface.* Ground-based high-frequency (3–30 MHz) radar systems can provide information on wave height and direction from the backscattered signal. Ranges of 50–500 km (30–300 mi) are feasible, but with over-the-horizon sky-wave systems relying on ionospheric reflection, ranges beyond 3200 km (2000 mi) have been achieved.

TABLE 1. Sea height code\*

Code	Height, ft†	Description of sea surface
0	0	Calm, with mirror-smooth surface
1	0–1	Smooth, with small wavelets or ripples with appearance of scales but without crests
2	1–3	Slight, with short pronounced waves or small rollers; crests have glassy appearance
3	3–5	Moderate, with waves or large rollers; scattered whitecaps on wave crests
4	5–8	Rough, with waves with frequent whitecaps; chance of some spray
5	8–12	Very rough, with waves tending to heap up; continuous whitecapping; foam from whitecaps occasionally blown along by wind
6	12–20	High, with waves showing visible increase in height, with extensive whitecaps from which foam is blown in dense streaks
7	20–40	Very high, with waves heaping up with long frothy crests that are breaking continuously; amount of foam being blown from the crests causes sea surface to take on white appearance and may affect visibility
8	40+	Mountainous, with waves so high that ships close by are lost from view in the wave troughs for a time; wind carries off crests of all waves, and sea is entirely covered with dense streaks of foam; air so filled with foam and spray as to affect visibility seriously
9		Confused, with waves crossing each other from many and unpredictable directions, developing complicated interference pattern that is difficult to describe; applicable to conditions 5–8

\* After *Instruction Manual for Oceanographic Observations*, H. O. Publ. 607, 2d ed., U.S. Navy Hydrographic Office, 1955.  
† 1 ft = 0.3 m.

From aircraft, stereo-photographs can be taken of the sea surface and analyzed photogrammetrically, but this is a laborious process. Also, laser or narrow-beam radar ranging may be used to measure profiles of the sea surface. See LIDAR; PHOTOGRAMMETRY; RADAR.

Two radar techniques are presently used for wave measurements from satellites. The radar altimeter (13.5 GHz) observes the reflection of pulses directed vertically. The deformation of the reflected pulse carries information about significant wave height  $H_{1/3}$  in the irradiated patch of ocean (several kilometers in diameter). The synthetic-aperture radar (SAR; >1 GHz) obliquely irradiates a patch of ocean surface (about 100 km or 60 mi in size), and uses pulse timing and phase information in the backscattered signal to obtain spatial resolution approaching 10 m (30 ft). The backscattering is from sea-surface roughness on scales comparable to the radar wavelength (several centimeters or inches), but longer-wavelength components appear in the SAR image by hydrodynamic interaction, electromagnetic modulation, and effects of wave motion. As a result, it appears that directional wave spectra  $S(f, \theta)$  may be obtained from SAR images, but the procedure is not yet fully understood. See METEOROLOGICAL SATELLITES.

Mark Wimbush

**Sea state.** Sea state is the description of the ocean surface or state of the sea surface with regard to wave action. Wind waves in the sea are of two types: those still growing under the force of the wind are called sea; those no longer under the influence of the wind that produced them are called swell. Differences between the two types are important in forecasting ocean wave conditions.

*Sea.* Those waves which are still growing under the force of the wind have irregular, chaotic, and unpredictable forms. The unconnected wave crests are only two to three times as long as the distance between crests and commonly appear to be traveling in different directions, varying as much as  $20^\circ$  from the dominant direction. As the waves grow, they form

regular series of connected troughs and crests with wave lengths commonly ranging from 12 to 35 times the wave heights. Wave heights only rarely exceed 55 ft (17 m). The appearance of the sea surface is termed state of the sea (Table 1).

The height of a sea is dependent on the strength of the wind, the duration of time the wind has blown, and the fetch (distance of sea surface over which the wind has blown).

*Swell.* As sea waves move out of the generating area into a region of weaker winds, a calm, or opposing winds, their height decreases as they advance, their crests become rounded, and their surface is smoothed. These waves are more regular and more predictable than sea waves and, in a series, tend to show the same form or the same trend in characteristics. Wave lengths generally range from 35 to 200 times wave heights.

The presence of swell indicates that recently there may have been a strong wind, or even a severe storm, hundreds or thousands of miles away. Along the coast of southern California long-period waves are believed to have traveled distances greater than 5000 mi (8000 km) from generating areas in the South Pacific Ocean. Swell can usually be felt by the roll of a ship, and, under certain conditions, extremely long and high swells in a glassy sea may cause a ship to take solid water over its bow regularly.

When swell is obscured by sea waves, or when the components are so poorly defined that it is impossible to separate them, it is reported as confused. (Table 2).

*In-between state.* Often both sea waves and swell waves, or two or more systems of swell, are present in the same area. When waves of one system are superimposed upon those of another, crests may coincide with crests and accentuate wave height, or troughs may coincide with crests and cancel each other to produce flat zones. This phenomenon is known as wave interference, and the wave forms produced are extremely irregular. When wave systems cross each other at a considerable angle, the

TABLE 2. Swell-condition code\*

Code	Description	Height, ft <sup>†</sup>	Length, ft <sup>†</sup>
0	No swell	0	0
	Low swell	1–6	
1	Short or average		0–600
2	Long		600+
	Moderate swell	6–12	
3	Short		0–300
4	Average		300–600
5	Long		600+
	High swell	12+	
6	Short		0–300
7	Average		300–600
8	Long		600+
9	Confused		

\*After *Instruction Manual for Oceanographic Observations*, H. O. Publ. 607, 2d ed., U.S. Navy Hydrographic Office, 1955.

<sup>†</sup> 1 ft = 0.3 m.

apparently unrelated peaks and hollows are known as a cross sea.

**Breaking waves.** The action of strong winds (greater than 12 knots or 6.2 m/s) sometimes causes waves in deeper water to steepen too rapidly. As the height-length ratio becomes too large, the water at the crest moves faster than the crest itself and topples forward to form whitecaps.

As waves travel over a gradually shoaling bottom, the motion of the water is restricted and the wave train is telescoped together. The wave length decreases, and the height first decreases slightly until the water depth is about one-sixth the deep-water wave length and then rapidly increases until the crest curves over and plunges to the water surface below. Swell coming into a beach usually increases in height before breaking, but wind waves are often so steep that there is little if any increase in height before breaking. For this reason, swell that is obscured by wind waves in deeper water often defines the period of the breakers.

The zone of breakers, or surf, includes the region of white water between the outermost breaker and the waterline on the beach. If the sea is rough, it may be impossible to differentiate between the surf inshore and the whitecaps in deep water just beyond. Neil A. Benfer

**Capillary waves.** Capillary waves, or ripples, occur at the interface between two fluids when the principal restoring force is surface tension. Ripples generated by wind on the ocean and lakes are important for the initiation of turbulence in both media, transfer of gases between air to water, and scattering of electromagnetic and sound waves.

Capillary waves have short wave lengths. Both the phase and group velocities increase as wave lengths become shorter, and group velocity is greater than phase velocity. Dissipation of the waves by molecular viscosity is very rapid. The characteristic shape of the water surface, sinusoidal for small amplitudes, becomes distorted with more sharply curved troughs than crests as wave amplitude increases. In all these respects, ripples contrast sharply with gravity waves.

Mathematical formulas have been derived that relate the phase velocity, group velocity, and frequency

to the wave length of low-amplitude waves on still water in the absence of wind. In such waves, gravity and surface tension play equal roles in the restoring force; longer waves are considered to be gravity waves, and shorter are capillary waves. See SURFACE TENSION.

In nature, ripples are observed to grow rapidly when the wind blows, and to die away rapidly when the wind stops. When the water surface is uncontaminated, ripples die away to 37% of their original amplitude in a period of time that is related to the wave length and the kinematic viscosity of water. For example, for a wave with a wave length of 0.67 in. (17 mm) and a kinematic velocity of 10 m/s, the period of time is 3.8 s. When the water surface is contaminated, as by an oil film or other surface-active agent, ripples are damped still more rapidly, because the contaminated surface tends to act as an inextensible film against which the water motions due to the ripples must rub. In such a case for the example given, the period of time becomes 0.86 s.

Under low-wind conditions, this increased damping almost completely inhibits ripple growth: the surface appears glassy smooth and is called a slick. It has been observed that even short gravity waves grow at an inappreciable rate under such conditions; the interpretation here is that the fine scale of roughness presented to the wind by a rippled surface is involved in the formation and growth of gravity waves. However, ripples have been observed to form on clean water in the absence of wind by nonlinear processes occurring at the sharp crests of short, steep gravity waves; consequently the formation of both gravity waves and ripples is an interconnected process. Ripple trains formed under low-wind conditions derive their energy at the expense of gravity waves and are called parasitic capillaries. In this case, the capillary wave train is propagating in a moving stream of water, the orbital current of the underlying gravity wave. The longer wave parts of the capillary train would have a lower phase velocity in still water than either the shorter wave-length parts of the train or of the gravity wave, but they maintain their position relative to the crest of the gravity wave because they ride in a favorable part of the orbital current of the gravity wave. Short capillaries, having high phase velocity, do not need this aid to keep in phase, and their higher group velocity enables them to proceed to a leading position where the gravity-wave orbital velocity vanishes or even opposes their motion. Charles S. Cox

**Internal waves.** Internal waves are wave motions of stably stratified fluids in which the maximum vertical motion takes place below the surface of the fluid. The restoring force is mainly due to gravity; when light fluid from upper layers is depressed into the heavy lower layers, buoyancy forces tend to return the layers to their equilibrium positions. Internal waves have been found in the atmosphere as lee waves (waves in the wind stream downwind from a mountain) and as waves propagated along an inversion layer (a layer of very stable air). They are also associated with wind shears at the lower

boundary of the jet stream. In the oceans, internal oscillations have been observed wherever suitable measurements have been made. The observed oscillations can be analyzed into a spectrum with periods ranging from a few minutes to days. At a number of locations in the oceans, internal tides, or internal waves having the same periodicity as oceanic tides, are prominent.

Internal waves are important to the economy of the sea because they provide one of the few processes that can redistribute kinetic energy from near the surface to abyssal depths. When they break, they can cause turbulent mixing despite the normally stable density gradient in the ocean. Internal waves are known to cause time-varying refraction of acoustic waves because the sound velocity profile in the ocean is distorted by the vertical motions of internal waves. The result is that quasi-horizontal propagation of sound shows phase incoherence and large changes of intensity with time at ranges where the refraction has led to divergence or convergence of rays.

The vertical distribution of motions and phase velocity of internal waves depends on the vertical gradient of density in the fluid and the frequency of the generating forces. There is a simple density distribution that is illustrative: The fluid consists of two homogeneous layers, a lighter one on top of a heavier one, such as kerosene over water. The internal waves in this system are sometimes called boundary waves, because the maximum vertical motion occurs at the discontinuity of density at the boundary between the two fluids. Internal waves move at a slow speed, of the order of a few knots in the deep oceans. The effect of the rotation of the Earth is to increase the phase velocity of waves having periods long enough to approach one pendulum day.

When there is a continuous distribution of density in a fluid, as in the ocean or atmosphere, internal waves are possible only for frequencies that are lower than the maximum value given by a mathematical relationship known as the Väisälä-Brunt frequency, which is related to the downward rate of increase of the density and the velocity of sound in the fluid. In the ocean this maximum frequency occurs in the thermocline, where it commonly amounts to about one-fifth cycle per minute. At any frequency lower than this limit, there is an infinity of possible modes of internal waves. In the first mode, the vertical motion has a single maximum somewhere in the body of the fluid; in the second mode, there are two such maxima ( $180^\circ$  out of phase), with a node between; and so on. The actual motion usually consists of a superposition of modes. In the ocean, the Väisälä-Brunt frequency varies from a maximum, commonly near 0.2 cycle per minute in the steepest part of the thermocline to negligible values at the bottom of the deep seas and in mixed layers such as those at the sea surface. Amplitudes of waves are oscillatory with depth and appreciable only where their frequency is less than the local value of the Väisälä-Brunt frequency. For this reason, internal waves of high frequency are limited in depth to the thermocline. Long-

period (low-frequency) internal waves are affected by Earth rotation, and the limiting periodicity is the local value of the pendulum day (one-half sidereal day divided by the sine of the latitude). These longest-period oscillations are inertial motions, in which the orbital path no longer has any vertical component but forms a circle in the horizontal.

In a continuously stratified fluid such as the ocean below the thermocline, internal waves propagate diagonally upward and downward, thus distributing energy throughout the ocean depths. Their ray paths are easily distorted because the group velocity of internal waves is comparable to the differences in velocity in the vertical of ocean currents and inertial motions. As a consequence, bundles of internal wave rays suffer severe refraction, sometimes to the extent that they form caustics known as critical layers, where the wave energy is absorbed by breaking. Where the sea bed is level, internal waves can be reflected by surface and sea floor to form normal modes within the body of the ocean. On a sloping sea floor, reflection at the bottom causes a change of wave length.

Internal waves in the atmosphere have been detected by a variety of instruments: microbarographs and wind recorders at ground level, and long-term recordings of the scattering of radar or sonar beams by sharp density gradients in the high atmosphere. In the ocean, internal waves have been found by recording fluctuating currents in middepths by moored current meters, by acoustic backscatter Doppler methods, and by studies of the fluctuations of the depths of isotherms as recorded by instruments repeatedly lowered from shipboard or by autonomous instruments floating deep in the water.

Internal waves are thought to be generated in the sea by variations of the wind pressure and stress at the sea surface, by the interaction of surface waves with each other, and by the interaction of tidal motions with the rough sea floor. Their importance is that they can transmit energy and momentum throughout the ocean, not only laterally but also vertically. They can, therefore, transmit energy from the surface to all depths. In this way the otherwise sluggishly moving water at great depths can be agitated.

Charles S. Cox

**Long-period waves.** Long-period waves are those that exist when the period ( $T$ ) is longer than one-half of a pendulum day, that is,  $12 \text{ h}/\sin \theta$ , where  $\theta$  is the latitude of the location of interest. The main types of low-frequency ocean waves are Rossby waves, topographic Rossby waves, coastal Kelvin waves, and equatorial Kelvin waves. *See* CORIOLIS ACCELERATION.

*Rossby waves.* Rossby waves, named after meteorologist C. Rossby, are fundamental to the large-scale dynamics of both atmosphere and ocean. They can exist at periods from a few days to several years and help to describe, for example, seasonal and climatic fluctuations in the oceans. Since they exist at long periods, the Earth rotates several times during a wave period, and the rotation of the Earth therefore plays a central role in Rossby-wave dynamics. In order to

understand Rossby waves, it is necessary to consider rotation, angular velocity, and vorticity.

The angular velocity of the Earth is defined as a vector of magnitude  $\Omega$  and direction northward along the axis of rotation. An angular velocity can be similarly defined for all solid rotating bodies: the magnitude of the angular velocity is  $2\pi$  radians divided by the period of rotation, and the direction is along the axis of rotation in the direction analogous to that for the rotating Earth. The vorticity of a particle of water in solid-body rotation is defined to be twice the angular velocity. In general, the velocity shear for a water particle will not correspond to that for solid-body rotation, and the effective rotation and vorticity will consequently be modified. However, this does not affect the following explanation for the Rossby-wave mechanism.

Because the gravitational force perpendicular to the Earth's surface is so dominant, water particles tend to remain on the same horizontal level; that is, long-period ocean flows tend to be parallel to the Earth's surface. For this reason, rotational effects in the horizontal plane are of prime significance; consequently the vertical component of vorticity is of greatest importance to ocean dynamics. This vertical component of vorticity has two contributions. One, which exists even when the water is at rest relative to the Earth, is the local vertical component of the Earth's rotation at latitude  $\theta$ . The remaining contribution is due to rotational effects in the horizontal currents, and its value is positive when rotation is counterclockwise as viewed locally above the Earth's surface.

*Topographic Rossby waves.* Another type of long-period ocean wave is the topographic Rossby wave, whose mechanism depends on variations of bottom topography and hence on the water depth. An example is the continental shelf wave in which water depth varies strongly perpendicular to the coastline across the continental shelf and slope. Continental shelf waves propagate with the coast on their right in the Northern Hemisphere and with the coast on their left in the Southern Hemisphere.

Like Rossby waves, continental-shelf waves also exist when the water is stratified. Then they are known as coastally trapped waves. These propagate along the coast in the same direction as shelf waves. Continental-shelf and coastally trapped waves play an important role in the dynamics of sea level and current fluctuations on the continental shelf. Much of the wind-forced ocean energy on the continental shelf is associated with these waves.

*Coastal Kelvin waves.* Coastal Kelvin waves, first discussed by Lord Kelvin, are low-frequency waves quite distinct from Rossby waves. For Kelvin waves, gravity is the essential restoring force; coastal Kelvin waves are long gravity waves trapped to a coastal wall by the Coriolis force. The amplitude of the wave decays exponentially from the coast. The water velocity perpendicular to the coast is zero everywhere. Like continental shelf waves, coastal Kelvin waves propagate with the coast on their right in the Northern Hemisphere and on their left in the Southern Hemisphere. The flow underneath a wave crest is in the

direction of wave propagation, so the only way that the Coriolis force can balance the pressure-gradient force tending to flatten the sea surface is if the wave propagates with the coast on its right. Similar arguments show that propagation is with the coast on the left in the Southern Hemisphere.

Since a Kelvin wave is a long-period gravity wave, it can exist as either an external or internal gravity wave. In the external case, the wave behaves as though the water is of constant density, and the phase speed ( $c$ ) is given by the expression  $\sqrt{gb}$ , where  $g$  is the acceleration due to gravity and  $b$  is water depth. In the internal case, the phase speed also depends on the rate of density variation with depth. In general, internal Kelvin waves travel much more slowly than surface Kelvin waves.

Winds and tidal forces effectively generate oceanic coastal Kelvin waves. Strictly speaking, coastal Kelvin waves exist only when water of constant depth is bounded by a vertical wall, and such topography in reality never occurs. However, for cases in which the decay scale is much greater than the distance across the shelf and slope to the constant-depth deep sea, the wave is dynamically coastal Kelvin. Such a condition is most often satisfied in the external Kelvin case.

*Equatorial Kelvin waves.* Equatorial Kelvin waves are long gravity waves with phase speed trapped to the Equator by the Coriolis force. The north-south amplitude variation is symmetric about the Equator and bell-shaped. The wave's north-south velocity is zero, and the wave must propagate eastward. Equatorial Kelvin waves exist at very low frequencies, and the internal ones appear to play a fundamental role in the dynamics of climate fluctuations like those associated with El Niño. See EL NIÑO; OCEAN CIRCULATION.

Allan J. Clarke

*Bibliography.* T. P. Barnett and K. E. Kenyon, Recent advances in the study of wind waves, *Rep. Prog. Phys.*, 38:667-729, 1975; L. J. Duckers (ed.), *Wave Energy Devices, Solar Energy Society Conference Proceedings*, C57, 1990; M. D. Earle and J. M. Bishop, *A Practical Guide to Ocean Wave Measurement and Analysis*, 1984; D. V. Evans and A. F. de O. Falcão (eds.), *Hydrodynamics of Ocean Wave-Energy Utilization*, 1985; J. B. Herbich (ed.), *Handbook of Coastal and Ocean Engineering*, vol. 1: *Wave Phenomena and Coastal Structures*, 1990; B. Kinsman, *Wind Waves*, 1965, reprint 1984; P. H. LeBlond and L. A. Mysak, *Waves in the Ocean*, 1978; A. W. Lewis, Sea-surface variations and energy: Tidal and wave power, in R. J. N. Devoy (ed.), *Sea Surface Studies: A Global View*, pp. 589-625, 1987; W. G. Van Dorn, *Oceanography and Seamanship*, 2d ed., 1993.

## Oceanic islands

Islands rising from the deep sea floor. Oceanic islands range in size from mere specks of rock or sand above the reach of tides to large masses such as Iceland (39,800 mi<sup>2</sup> or 103,000 km<sup>2</sup>). Excluded are islands that have continental crust, such as the Seychelles,

Norfolk, or Sardinia, even though surrounded by ocean; and with this exclusion, all oceanic islands surmount volcanic foundations. A few of these have active volcanoes, such as on Hawaii, the Galápagos islands, Iceland, and the Azores, but most islands are on extinct volcanoes. On some islands, the volcanic foundations have subsided beneath sea level, while coral reefs growing very close to sea level have kept pace with the subsidence, accumulating thicknesses of as much as 5000 ft (1500 m) of limestone deposits between the underlying volcanic rocks and the present-day coral islands. On some of these, for example Enewetak Atoll in the western Pacific, the foundations are as much as  $8 \times 10^9$  years old. *See* REEF; VOLCANO.

Oceanic islands owe their existence to volcanism that began on the deep sea floor and built the volcanic edifices, flow on flow, dike on dike, up to sea level and above. The highest of the oceanic islands is Hawaii, where the peak of Mauna Kea volcano reaches 14,000 ft (4200 m). Most volcanic islands are probably built from scratch in less than  $10^6$  years, but minor recurrent volcanism may continue for millions of years after the main construction stage. Although the main construction of the island of Oahu was completed  $2\text{--}4 \times 10^6$  years ago, a younger phase of volcanism is responsible for volcanoes such as Diamond Head formed during the past  $10^6$  years. *See* VOLCANOLOGY.

**Geology.** Oceanic volcanoes are kept in isostatic balance (floating equilibrium) during growth. The upper, more rigid part of the underlying lithosphere flexes to accommodate the load, and the lower, more ductile part flows. The result is that the edifice subsides during construction, and subsidence continues even after loading, owing to the long-term normal cooling of the supporting lithosphere. Wells drilled on Hawaii show the blocky lavas characteristic of subaerial flows, and soil zones between flows, even at depths far below sea level—supporting the occurrence of subsidence. The long-term fate of oceanic islands is thus to subside completely below sea level on time scales of millions of years.

As the edifice is built up close to sea level, the summit regions are attacked by waves, and during subaerial exposure the normal processes of erosion by waves, streams, and glaciers sculpt and reduce the landscapes, at rates dependent on the local climates, and with forms reflecting the relative erosional resistance of rock masses within the volcanic pile. Steep cliffs around the seaward edges are common; and spectacularly steep, fluted, and pinnacled relief may develop where rainfall is heavy, generally on the windward sides of islands. Erosion may at times be rapid enough to counterbalance thermal subsidence, resulting in isostatic uplift. Generally, the subaerial landscapes are reduced to a surface of low relief by the time the summit region subsides beneath the waves, and the seamount then has a nearly flat top and is termed a guyot. *See* MARINE GEOLOGY; SEAMOUNT AND GUYOT.

**Associated reefs.** Volcanic islands within the latitudinal zone of coral-reef growth (mainly between

about  $25^\circ\text{N}$  and  $25^\circ\text{S}$  latitude) are generally fringed by reefs, which tend to protect the volcanic terranes from direct attack by ocean waves. As predicted originally by Charles Darwin, and amply confirmed by deep drilling through the reefal deposits on several oceanic islands, while the volcanic foundations of the islands subside, the corals, algae, and other shelly organisms keep piling up, layer on layer, keeping pace with subsidence. During subsidence, the reefal region maintains an area close to that enclosed by the original fringing reef, while the volcanic center diminishes in area. The final result should be a flat coral bank mainly at sea level, whereas in fact modern reef-surrounded islands commonly have a lagoon of variable width and depth between the barrier reef and the island, and many reefs have no central island at all, consisting simply of a perimeter reef encircling a deeper lagoon, a form termed an atoll. Although differential growth rates between the biotas on the perimeter reef and those in the lagoon account for some of the relief, the actual forms may result in large part from drowning of landscapes developed during the successive lowerings of global sea level to 300–600 ft (100–200 m) below its present level during the ice ages. The exposed limestone is attacked by rainwater, which dissolves it and may result in the basic rim-and-lagoon morphology by dissolution. Passes of variable width and depth, probably modified relics of stream valleys during lowered sea level, give access from the open sea to the lagoon for boats. Few barrier reef islands or atolls have passes deep enough for large ships, and those islands generally became commercial and administrative centers because of the accessibility of safe anchorage in the lagoon. *See* ATOLL; REEF.

**Resources.** Islands in regions of high oceanic fertility are commonly host to colonies of sea birds, and the deposits of guano have been an important source of phosphate for fertilizer. On some islands, for example, Nauru in the western equatorial Pacific, the original guano has been dissolved and phosphate minerals reprecipitated in porous host limestone rocks. The principal crop on most tropical oceanic islands is coconuts, exploited for their oil content, but some larger volcanic islands, with rich soils and abundant water supplies, are sites of plantations of sugarcane and pineapple. Atoll and barrier-reef islands have very limited water supplies, depending on small lenses of ground water, augmented by collection of rainwater. Tourism, mainly on tropical islands with balmy climates, is concentrated on high islands where water supplies are adequate. *See* ISLAND BIOGEOGRAPHY.

Edward L. Winterer  
Bibliography. H. W. Menard, *Islands*, 1986; P. D. Nunn, *Oceanic Islands*, 1994.

## Oceanographic vessels

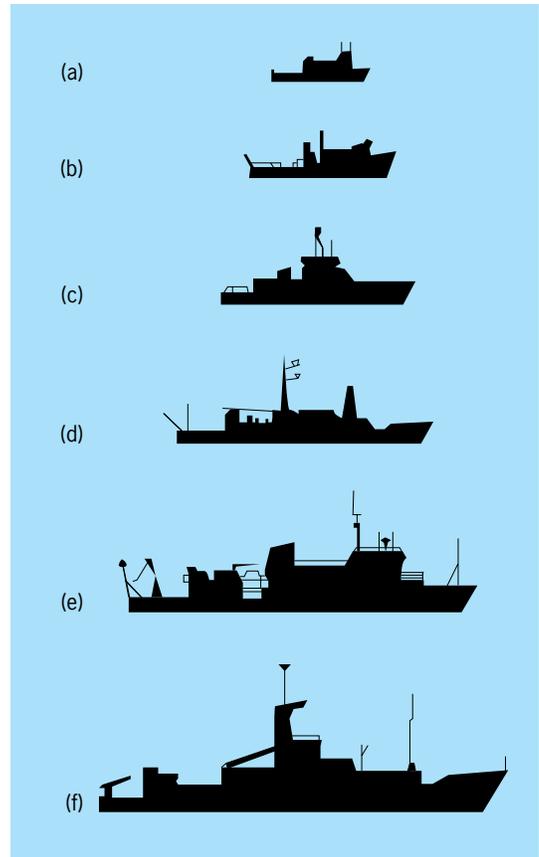
Research vessels designed to collect quantitative data from the sea's surface, its depths, the sea floor, and the overlying atmosphere. The primary tool for obtaining oceanographic data has been the oceanographic vessel, which has a variety of forms and sizes.

The ships can be general purpose or special purpose, and a conventional or radical design. Their primary purpose is to carry scientists and increasingly sophisticated equipment to and from study sites on the ocean's surface, and in some cases below the surface. The ships must have the ability to lower and retrieve instruments by using winches and wires. The ship's equipment and instrumentation must determine precisely the location on the sea surface, and provide suitable communication, data gathering, archiving, and computational facilities for the scientific party.

The requirements list includes seakeeping (seakindliness, a measure of a ship's response to severe seas; and station-keeping, the ability of a ship to maintain its fixed location on the sea surface); work environment; endurance (range, days at sea); scientific complement (number of researchers accommodated); operating economy and scientific effectiveness; subdued acoustical characteristics; payload (scientific storage, weight handling); speed; and ship control. These requirements often conflict, necessitating compromise.

**Types.** Ships typically can be considered in three major groups based on use: general purpose (classical biological, physical, chemical, geological, and ocean engineering research, or a combination); dedicated special purpose (hydrographic survey, mapping, geophysical, or fisheries); or unique (deep-sea drilling, crewed spar buoy, or support of submersible operations). They can be used simply as delivery and support systems for exploratory devices, such as floats and bottom landers, as well as crewless remote operating vehicles (ROVs—tethered, powered, surface-controlled robots), or autonomous underwater vehicles (AUVs—freely operating robots, using computer programmed guidances). They may be of a standard classical ship configuration or an experimental design. They can be built from the keel up as research ships, or they can be adapted from available vessel hulls. Some tourist ships have been built in laboratories for research and tourist education. Even ships of opportunity (a ship not normally utilized for scientific activities) have been used to collect some forms of oceanographic data, including release of untended instrumentation, thus partially qualifying as oceanographic research vessels.

Ocean research ships are typically grouped in terms of overall length (see **table**; **Fig. 1**). In general, the larger ships are used for worldwide expeditionary work and for larger-scale, often multidisciplinary research. Some, however, are special pur-



**Fig. 1.** Representative oceanographic ship silhouettes (profiles). (a) *Barnes*, 66 ft (20 m). (b) *Pelican*, 105 ft (32 m). (c) *Cape Hatteras*, 135 ft (41 m). (d) *Endeavor*, 185 ft (56 m). (e) *Ewing*, 239 ft (73 m). (f) *Revelle*, 274 ft (84 m).

pose. Most make long cruises away from their home port, staying at sea between port calls from 2 to 6 weeks and operating 300 days/yr. The intermediate classes of ships tend to be general purpose, adaptable to any type of research. Many of these also work worldwide on occasion. These classes operate at sea about 220–250 days each year. The smaller ships are generally used for continental shelf, near-shore, or coastal research and tend to make short trips of 1 day to 2 weeks. Main propulsion can be single (or multiple) open or nozzle-surrounded screws, vertical axis propellers, or waterjets. Bow thrusters are generally used for station-keeping. Complexity, maintainability, efficiency, cost, and maneuverability are factors in selection. Power can be derived from direct-drive diesel engines or from electric motors driven by current supplied by individual diesel generators or a central diesel-electric power plant. See MARINE ENGINE.

A standard-configuration expeditionary vessel carries a scientific party of 35 and operates with a crew of 22. The key features include science laboratories, office and storage areas, a staging area, portable vans, cranes, a suite of winches and wires, a workshop, a darkroom, a dive locker, a library, a mess area, and staterooms. This vessel has the normal ship features of a main propulsion unit and bow thruster, pilot house, chart room, aft-control, and galley, as well as

Classification of ocean research ships		
Class	Size (length overall)	Description
I	Over 250 ft (76 m)	Large, high endurance
II	200–250 ft (61–76 m)	Large, medium endurance
III	150–199 ft (45.7–61 m)	Intermediate
IV	100–149 ft (30.5–45.7 m)	Small
V	Less than 100 ft (30.5 m)	Nearshore, coastal

mast-located instruments, warning lights, electronic and satellite navigation, and data and communication antennas.

A marine geology and geophysical ship uses multichannel seismic profiling to probe the deep geological structure underlying the ocean floor. Profiling uses some method of creating a high-level noise, such as that created by a towed airgun sound source. The returning sounds from the subsea layers of rock and sediment are received by long hydrophone-streamer arrays. Properly analyzed, the information from the individual hydrophones forms a detailed picture of the arrangement and materials that underlie the sea floor. These ships are used for oil and other mineral surveys and are generally very large and expensive. They must have large reels for the towed sound arrays, extensive rigging and swing-out booms, and long electromechanical wires or fiber-optic cable deployed by appropriate winches. They constitute a very specialized configuration and are seldom used for any other purpose.

New fisheries vessels are being designed to isolate ship's noise, providing acoustic quieting. Fisheries research and stock assessment vessels look very much like standard fishing vessels, incorporating such capabilities as trawling, gill-netting, purse-seining nets, stern ramps and gallows, trawl winches, and fish-handling, processing, and preservation facilities. They also include scientific winches, laboratories, and various types of oceanographic instrumentation capabilities. See MARINE FISHERIES.

The design concept of the small-waterplane-area twin hull (SWATH) vessel (Fig. 2) is based on having all working spaces elevated above the sea's surface, supported by two in-line submerged hulls. The thin struts connecting the two experience only very small displacement forces with the passage of waves, thus effectively decoupling the ship from surface conditions while at rest and under way and providing a very stable platform from which to work. SWATH designs are becoming more accepted in the oceanographic research community as more are built and placed in service.

The floating instrument platform (FLIP) [Fig. 3] is essentially a 355-ft (108-m) spar buoy with a crew; it is towed in a horizontal mode to a work site and tipped into its vertical attitude, where it is used to make subsurface measurements. After flooding, its ballast tanks place it in vertical position; it extends 300 ft (91 m) underwater, where because of its design stability it is almost unaffected by wave action, moving only about 3 ft (0.9 m) in a 30-ft (9-m) wave. It houses a crew of 5 and a science party of 11, two internal electronic laboratories, and an outside laboratory.

A new Arctic research vessel, USCGC *Healy*, recently entered into service (Fig. 4). This 420-ft (129-m) 17,000-ton icebreaker is outfitted for Arctic research. It is capable of conducting a wide range of research activities, and providing more than 4200 ft<sup>2</sup> of laboratory space. It also has numerous electronic sensor systems, oceanographic winches, and accommodations for up to 50 scientists. *Healy* is designed

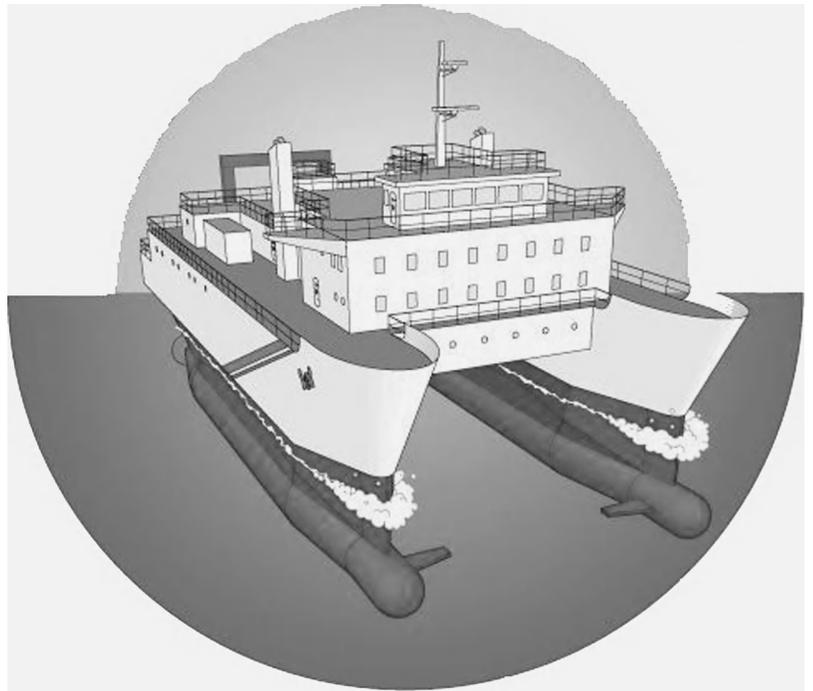


Fig. 2. Small-waterplane-area twin hull (SWATH) vessel, the 180-ft (55-m) *AGOR 26*, University of Hawaii.

to break 4½ ft (1.4 m) of ice continuously at 3 knots and can operate in temperatures as low as -50°F.

**Research requirements.** Two approaches to ocean research are typically used: direct collection of



Fig. 3. Deployed FLIP—floating instrument platform. (Scripps Institution of Oceanography)



Fig. 4. Arctic research vessel *USCGC Healy*, commissioned November 10, 1999. (U.S. Coast Guard)

samples (water, sea-floor rocks and sediment, plants and animals) for topside evaluation and study, and in-place measurement of parameters of interest (directly or from analysis by remote or autonomous instrumentation).

To deploy and retrieve research tools requires a suite of powerful, sensitive winches and strong wires (some with electrical conducting elements). The tools include nets (some simple, others incorporating environmental sensors and cameras), trawls and dredges, coring devices, remote bottom landers, sediment traps, continuously recording conductivity/temperature/depth sensors in rosettes with water collection bottles, and water and organism pumps. Packages such as large floats, drifting buoys, and moored arrays with current meter strings require this capability as well, as does the launching and recovery of small craft, submersibles, ROVs, and AUVs. Large, open working deck areas (with convenient control areas and clear observation vistas) as well as highly trained crews are required.

Winch types in common use include trawl winches (driven by a powered drum), traction winches (driven by a powered traction-head system), hydrographic winches (most containing one or more conducting cable elements), and smart winches (that compensate for ship motion, or even allow instrument packages to change depth to follow ocean parameters of interest). Some carry up to full-ocean-depth cable (33,000 ft or 10,000 m). Most cable is made from twisted wire bundles torque-balanced to reduce kinking when deployed. Fiberglass optical cable (for high rates of transmission of data and images) and synthetic cable made from noncontaminating materials (for sophisticated chemical studies) are increasingly being used. A-frames and J-frames along with specially adapted cranes (such as the knuckleboom crane) guide the wire rope used to handle instrument and other packages over the side or stern of the ship into and out of the sea. Cranes must also be capable of loading and transferring material in port and at sea. *See* HOISTING MACHINES.

Types of laboratories include the general-purpose main laboratory, the wet laboratory (for water and net specimen handling and preservation), the clean Special-Purpose laboratory (for high-quality chem-

ical or radioactive work), and, on large ships, a separate computer laboratory (for main frames and support of unique instrumentation). Sample storage (in some cases cold or refrigerated) is provided. Vans, most of standard configuration, are frequently used. They function as additional work space (special chemistry or dive support, for instance), storage space, and on rare occasions living space.

Hull-mounted multibeam sonar bathymetric systems provide sea-floor surface contour images. The systems installed on most modern large oceanographic vessels have two arrays, one to send sonic energy and the other to receive the reflected returns. Advanced on-board systems can combine data from more than one type of sonar system (multibeam, side-scan, chirp) into three-dimensional imagery. *See* SONAR.

Data from the multitude of scientific devices on board is recorded, archived, and merged with ship parameters to provide a continuous cruise record. Computers in scientific spaces are linked with each other and generally have access to shore-based large computers through satellite transmission. Sophisticated computers on the ship capable of image manipulation are available on board for scientists. Satellite communication is used for communication between and among ships and their home ports. Ships also use satellite reception and linkages for long-range communication, precise navigation (Global Positioning System; GPS), and for transmittal of signals and data streams from buoys and moorings and between vessels. Some data sensed from satellites are used in real time by the ship to locate centers of biological activity, surface currents and eddies, sea surface slopes, storms, and ice floes. Cruise plans can be adapted to changing conditions based on the information. *See* INSTRUMENTED BUOYS; SATELLITE NAVIGATION SYSTEMS.

Altogether there are close to one thousand oceanographic vessels worldwide. The number varies with the classification technique used and the state of activity of the vessels. Russia, the United States, the United Kingdom, Sweden, Japan, France, and Canada operate the major fleets. Increasingly, large-scale ocean research is both international and multidisciplinary. Much more sophisticated tools, especially satellite and computer related, as well as autonomous submersibles, bottom landers, drifters, and moorings, will continue to evolve and be deployed, thereby changing, replacing, or enhancing traditional ship functions. *See* MARINE ENGINEERING; NAVAL ARCHITECTURE.

J. J. Griffin/J. F. Bash

**Bottom samplers.** Historically, the tools used by ocean scientists to collect oceanic rock and sediment samples from oceanographic vessels have been confined to three basic family groups: grabs, dredges, and coring devices. From these three groupings a wide variety of unique tools have been developed for use by the scientific community.

*Grab samplers.* Traditionally, grabs have been reserved for use in shallow water depths where large numbers of individual samples from the upper few centimeters of the bottom sediments are required.

Within this group of samplers, the Van Veen grab is perhaps the most commonly used device. It is produced in a variety of sampling sizes ranging from 0.82-ft (0.25-m) to over 3.3-ft (1-m) surface area sampling capability, making it an ideal tool for the estimation of benthic populations as well as the determination of sediment distribution patterns over wide areas. The size of a Van Veen is determined by the actual square area that the open grab will cover when it first contacts the bottom. One point to remember is that the recovered Van Veen sample will have experienced some slight distortion due to the closure of the grab jaws.

A second, and unique, grab device is the underway sampler, which is deployed on a wire from a moving vessel. The sampler is lowered to the bottom until it contacts the sediment; at this time a pin or light line is sheared, a small volume of sediment recovered, and the sampler returned to the vessel. The design of the device is such that the pin or line will shear before the sampler has an opportunity to dig deeply into the bottom and part the ship's wire.

**Dredges.** The standard chain bag dredge is constructed in a variety of sizes and weights up to a heavy-duty model with a 3.9 ft by 23 in. (1.2 m by 60 cm) mouth opening. Attached to the dredge mouth is either a chain or wire mesh bag used to retain the sample. Occasionally the dredge bag is lined with a canvas or burlap sack in order to retain small fragments of rock, volcanic glass, or sediment. The mouth of the dredge is of a welded steel construction with an articulated bail that is attached to the leading edge. When the dredge is rigged, it is customary

to employ both a weak link and a swivel above the dredge in order to protect the ship's cable.

During actual dredging operations the water depth, dredge weight, and wire strength must be carefully considered and protective measures taken, through the use of a series of weak links, to ensure the integrity of the ship's cable. Two weak links, capable of varying shear loads, are used on each dredge: the first located at the dredge bail and a second at the rear of the dredge mouth. Under this scheme, should the dredge become fouled, the first link would part, transferring the load to the second link, pulling the dredge free, and preserving the sample and instrument. In the event that the dredge cannot be pulled free, the second link will part before damage can be done to the ship's cable. The use of dredges to collect oceanic rocks, manganese nodules, and other sea-floor deposits is virtually non-depth-limited; however, this technique is restricted to recovering only those materials lying on or near the sediment water interface.

**Coring devices.** These tools for studying the structure of oceanic sediments have developed into highly specialized instruments since they were first conceived in the early 1940s. Although individual coring devices may differ as to complexity, weight, sample diameter, and length of sample recovered, their main purpose still remains the collection of an undisturbed section of the sea-floor sediments.

**Gravity cores.** Gravity cores, of lengths up to 11 ft (3.5 m) and weights of 990 lb (450 kg), are frequently used as survey tools in the deep sea. The barrels of these cores may be lined with a plastic tube that is

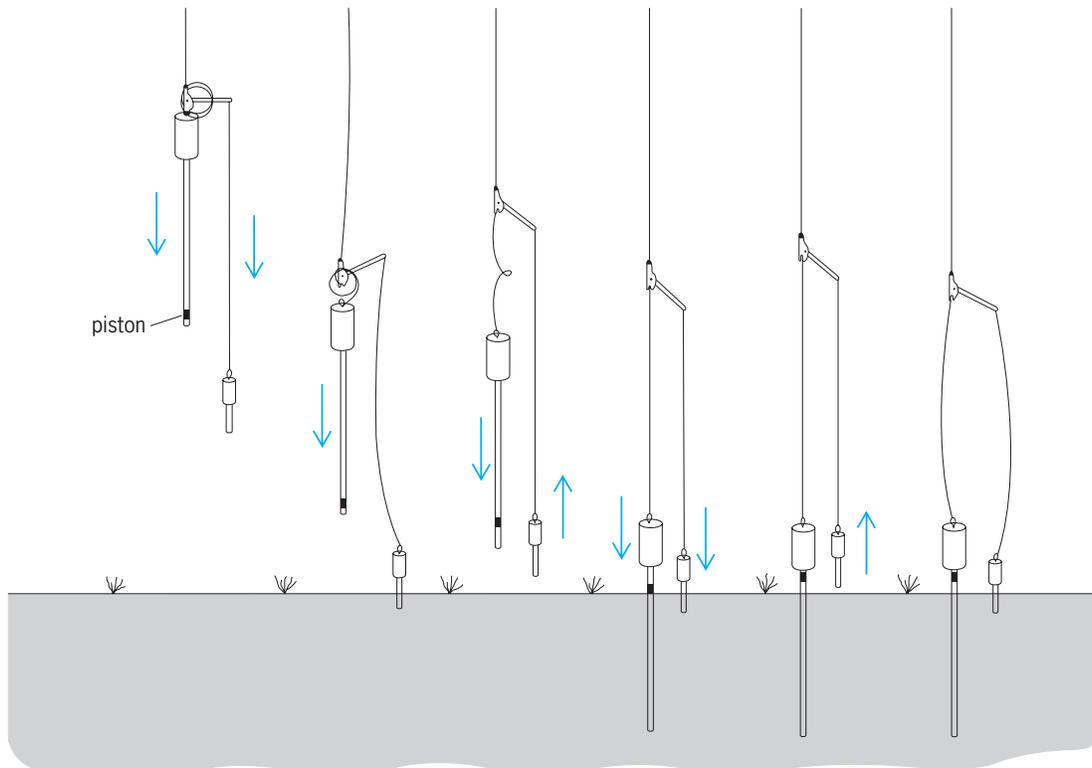


Fig. 5. Piston coring sequence (left to right) showing the effects of cable rebound on the trigger core.

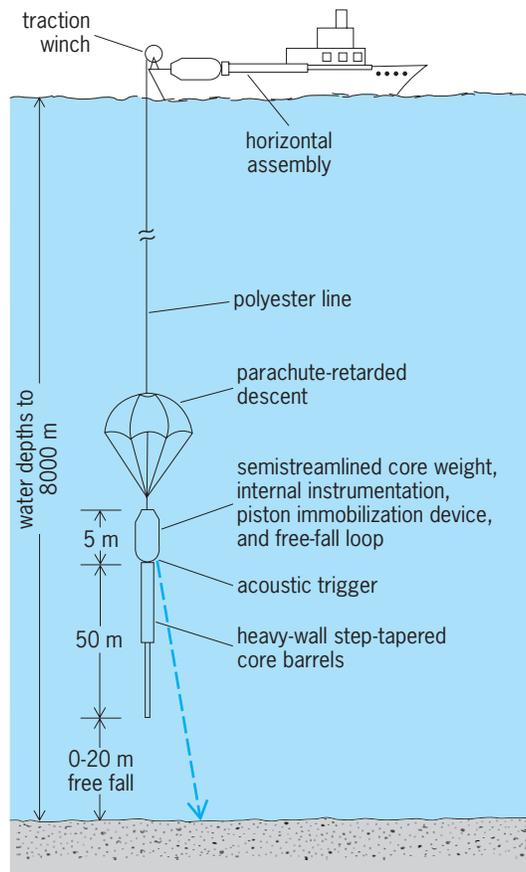


Fig. 6. General configuration of long coring system. (Drawing by A. H. Driscoll)

used as a sample container for the recovered sediment. In theory, the gravity core is dependent upon the speed of the ship's winch and its own weight to achieve penetration and, as a result of this, is limited to shorter sample recovery lengths than are possible with a piston core.

One exception to the gravity core operational theory is the free-fall core. This device is dropped di-

rectly from a moving vessel and allowed to free-fall through the water column to the sea floor. Upon contact with the bottom and recovery of a 3.3-ft (1-m) sample, the sample is pulled free of the ballast portion of the core by two glass flotation spheres. The positive buoyancy of the flotation is sufficient to return the sample to the surface, where it can be retrieved by the surface vessel. By using this type of device it is possible to survey a relatively large area of the sea floor in a minimum of time.

**Box cores.** Another type of coring device capable of recovering undisturbed sections of the sea floor is the box core. This device, although bulkier than the simple gravity core, has been specially designed to recover a square section of the upper meter of sediment at the interface. Penetration of the coring box is achieved via a heavy driving weight mounted directly above the box. As the box core framework settles to the sea floor, slack in the ship's wire releases the weighted box and allows it to sink gently into the sediment, taking the core. Upon completion of the coring sequence, the ship's cable is slowly retrieved, causing a spade, attached to the framework, to cut into the sediment, stopping directly below the implanted box. As the cable becomes taut, this spade rises and seals the bottom of the box, preserving the sample as it is returned to the surface vessel. The general quality of samples taken with this device is usually very high, and it is even possible to see worm tubes and small brittle stars preserved on the top of the sample when the box is opened.

**Piston core.** The piston core has been designed to recover sediment sequences longer than would be possible by using a gravity coring technique. Essentially, the sequence of events surrounding the operation of the piston core is as follows (Fig. 5). The piston core and its accompanying trigger core are lowered to the sea floor by the ship's cable until the trigger core contacts the sediment. At this point the tension on the tripping arm is relaxed and the piston core allowed to fall free. As the piston core free-falls toward the bottom, it pays out a coil of cable behind it; when the cutter contacts the sediment, the cable becomes taut, immobilizing the piston and allowing the core barrels to slide past the piston and obtain the sample. It is this force generated by the action of the barrels bypassing the piston that overcomes the internal wall friction of the sediment and allows a long sample to be taken. Even though the process appears complicated, the entire sequence occurs within 3 s of the trigger core's first striking the bottom.

In general, the piston cores range in weights and sizes from a 3100-lb (1400-kg), 59-ft (18-m), 2.5-in.-diameter (6.5-cm) standard device to a 31,000-lb (14,000-kg), 160-ft (50-m), 5.0-in.-diameter (12.7-cm) long coring system (Fig. 6). The addition of instrumentation capabilities in the core weight has made the piston core more of a deep-ocean experiment platform than the simple collecting device originally developed in the late 1940s.

**Ranging.** Relative to the effective operation of all bottom sampling equipment is a requirement that the user know where and how far away the device

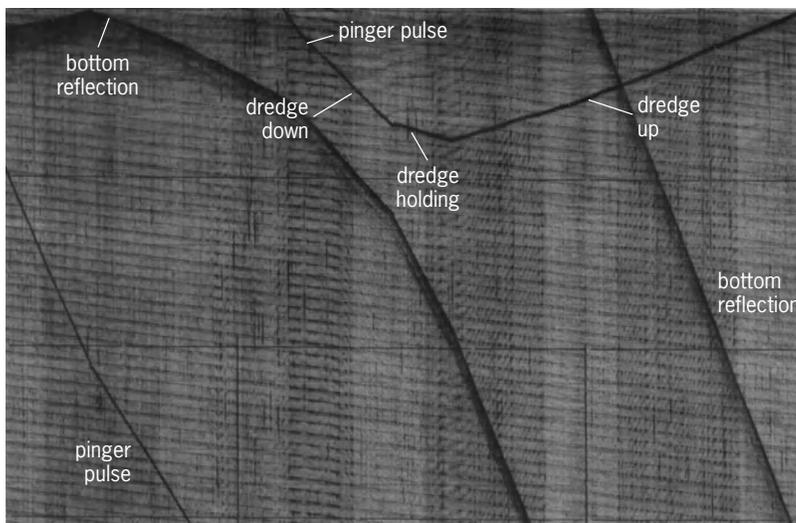


Fig. 7. Typical echo-sounding record of dredge lowering showing descent, bottom contact, and recovery. (Woods Hole Oceanographic Institution)

is from the sea floor. This is usually accomplished by attaching a telemetering pinger to the ship's cable directly above the sampling device. The pinger emits a 12-kHz pulse once per second that, radiated downward, bounces off the sea floor and returns to the surface, where it is received aboard the surface vessel. By measuring the distance between the outgoing pulse at the pinger and the bottom reflection, the observer can accurately calculate the distance between the instrument and the sea floor (**Fig. 7**). *See* ECHO SOUNDER; OCEANOGRAPHY.

Alan H. Driscoll

**Bibliography.** J. F. Bash, UNOLS Research Fleet, *Sea Technology*, June 1998; C. A. Bookman, *Trends and Opportunities in Ocean Technology 1992-1997*, Marine Board, National Academy of Sciences, 1992; J. J. Griffin, *Science at Sea*, 1996; *Marine Technology Society Journal*, quarterly; R. Schmitt et al., *The Academic Research Fleet*, May 1999; UNOLS, *Scientific Mission Requirements for Oceanographic Research Vessels*, 1988; UNOLS and V. Cullen (eds.), *The Research Fleet*, 2000.

## Oceanography

The science of the sea; including physical oceanography (the study of the physical properties of sea water and its motion in waves, tides, and currents), marine chemistry, marine geology, and marine biology. The need to know more about the impact of marine pollution and possible effects of the exploitation of marine resources, together with the role of the ocean in possible global warming and climate change, means that oceanography is an important scientific discipline. Improved understanding of the sea has been essential in such diverse fields as fisheries conservation, the exploitation of underwater oil and gas reserves, and coastal protection policy, as well as in national defense strategies. The scientific benefits include not only improved understanding of the oceans and their inhabitants, but important information about the evolution of the Earth and its tectonic processes, and about the global environment and climate, past and present, as well as possible future changes. *See* CLIMATE HISTORY; COASTAL ENGINEERING; MARINE MINING; MARINE SEDIMENTS; MARITIME METEOROLOGY; OIL AND GAS, OFFSHORE.

**The modern discipline.** The traditional basis of modern oceanography is the hydrographic station. Hydrographic studies are still carried out at regular intervals, with the research vessel in a specific position. Seawater temperature, depth, and salinity can be measured continuously by a probe towed behind the ship. The revolution in electronics has provided not only a new generation of instruments for studying the sea but also new ways of collecting and analyzing the data they produce. Computers are employed in gathering and processing data in all fields, and are also used in the creation of mathematical models to aid in understanding. Much information can also be gained by remote sensing using satellites, which are also a valuable navigational aid. These provide data on sea surface temperature and currents, and on marine

productivity. Satellite altimetry gives information on wave height and winds and even bottom topography (because this affects sea level). Scientists look forward to a day when observations can be made in the deep sea by autonomous vehicles. However, the ship remains a fundamental tool for many observations. *See* ALTIMETER; COMPUTER; DIGITAL COMPUTER; HYDROGRAPHY; MODEL THEORY; OCEANOGRAPHIC VESSELS; REMOTE SENSING; SATELLITE NAVIGATION SYSTEMS; SEAWATER.

**Physical oceanography.** Physical oceanography, and in particular ocean circulation studies, forms the core of oceanographic research. The movements of seawater—ocean currents—are powerful agents in the distribution of heat throughout the world, influencing both weather and climate. The continual renewal of water bearing dissolved nutrients is essential to most marine organisms, which will be abundant only where such supplies are available.

Twentieth-century dynamical oceanographers have shown how the deflecting effect of the Earth's rotation influences water movements. Geostrophic forces are responsible for the intensification of surface currents, such as the Gulf Stream, on the western sides of oceans (western boundary currents). In the 1950s the existence of a southward-flowing countercurrent under the Gulf Stream was predicted. Neutrally buoyant floats, to be tracked by radio signals picked up by hydrophones on board ship, were deployed to confirm this prediction. However, further out in the Atlantic the floats moved unexpectedly fast, with frequent changes in direction. This was the first indication of vigorous eddies in the ocean that have since been shown to be comparable to atmospheric weather systems. Further investigations of these phenomena, studies of equatorial currents and undercurrents, and transport between oceans are the principal topics occupying physical oceanographers in the latter part of the twentieth century. The World Ocean Circulation Experiment, a large-scale international program of cooperation on research projects and data-collecting expeditions, was designed to throw further light on the ocean's influence on world climate. Among the techniques being employed are the use of Swallow floats, arrays of current meters, that can be moored to the sea bed for a period of time to measure deep-water movements and then released for retrieval by acoustic signals, and the use of geochemical tracers, including chlorofluorocarbons, to obtain data on the distribution and age of water masses. *See* CORIOLIS ACCELERATION.

**Marine biology.** Biologists seek to classify the great diversity of life, from microscopic bacteria and phytoplankton to the great whales. To learn how the food web operates, they must examine the constraints on marine productivity and the distribution of species in the surface and midwater layers, and the vertical migrations between them, and in the bottom-living (benthic) fauna. Deep-sea cameras and submersibles now permit visual evidence of creatures in these remote depths to be obtained. *See* DEEP-SEA FAUNA; FISHERIES ECOLOGY; MARINE BIOLOGICAL

SAMPLING; MARINE ECOLOGY; MARINE FISHERIES; MARINE MICROBIOLOGY; SEAWATER FERTILITY.

**Marine geology.** Since the early 1900s, all recorded ocean depths have been incorporated in the General Bathymetric Chart of the Ocean. The amount of data available increased greatly with the introduction of continuous echo sounders; subsequently, side-scan sonar permitted very detailed topographical surveys to be made of the ocean floor. The features thus revealed, in particular the midocean ridges (spreading centers) and deep trenches (subduction zones), are integral to the theory of plate tectonics. An important discovery made toward the end of the twentieth century was the existence of hydrothermal vents, where hot mineral-rich water gushes from the Earth's interior. The deposition of minerals at these sites and the discovery of associated ecosystems make them of potential economic as well as great scientific interest. Possibly life on Earth began in similar situations in the remote past. Investigation of such areas can be made directly by scientists using submersibles and by underwater cameras, as well as by instrumentation. Even the sediments and other rocks of the sea floor are being sampled by the international deep-sea drilling program to provide information on how the present oceans evolved and on past climate change. *See* ECHO SOUNDER; HYDROTHERMAL VENT; MARINE GEOLOGY; MID-OCEANIC RIDGE; PLATE TECTONICS; SONAR; SUBDUCTION ZONES; UNDERWATER VEHICLE; UNDERWATER PHOTOGRAPHY; UNDERWATER TELEVISION. Margaret Deacon

**Bibliography.** A. Colling, *Ocean Circulation*, 2d ed., 2001; T. S. Garrison, *Oceanography: An Invitation to Marine Science*, 5th ed., 2004; J. A. Knauss, *Introduction to Physical Oceanography*, 2d ed., 1996; C. M. Lalli, *Biological Oceanography*, 2d ed., 1997; M. E. Q. Pilson, *An Introduction to the Chemistry of the Sea*, 1998; H. V. Thurman and A. P. Trujillo, *Essentials of Oceanography*, 7th ed., 2004.

## Octane number

A standard laboratory measure of a fuel's ability to resist knock during combustion in a spark-ignition engine. A single-cylinder four-stroke engine of standardized design is used to determine the knock resistance of a given fuel by comparing it with that of primary reference fuels composed of varying proportions of two pure hydrocarbons, one very high in knock resistance and the other very low. A highly knock-resistant isooctane (2,2,4-trimethylpentane,  $C_8H_{18}$ ) is assigned a rating of 100 on the octane scale, and normal heptane ( $C_7H_{16}$ ), with very poor knock resistance, represents zero on the scale. Octane number is defined as the percentage of isooctane required in a blend with normal heptane to match the knocking behavior of the gasoline being tested. *See* SPARK KNOCK.

The CFR (Cooperative Fuel Research) knock-test engine used to determine octane number has a compression ratio that can be varied at will and a knockmeter to register knock intensity. In the clas-

sic method of knock rating, the engine is run on the fuel to be tested, and its compression ratio is adjusted to give a standard level of knock intensity. Without changing the compression ratio, this knock level is then bracketed by running the engine on the primary reference fuel blends, one of which knocks a little more than the test fuel, and the other a little less. The octane number of the fuel being rated is then determined by interpolation from the knockmeter readings of the bracketing reference fuels.

Alternatively, the engine's compression ratio can be adjusted to close to the limit for the fuel being tested, and then, while the engine is run on each of two closely bracketing test fuels, the ratio can be readjusted to a standard intensity reading on the knockmeter. Finally, the engine is again run on the test fuel, and its compression ratio is adjusted to give the same knockmeter reading. The octane number of the test fuel is then interpolated from the compression ratio settings.

Knock tests are performed under one of two sets of engine operating conditions. Results of tests using the so-called Motor (M) method correlated well with the fuels and automobile engines of the 1930s, when the method was developed. The Research (R) method was developed later when improved refining processes and engines gave gasolines better road performance than their M ratings indicated. Today, the arithmetic average of a gasoline's R and M ratings usually is a good indicator of its performance in a typical car on the road. Thus, that average,  $(R + M)/2$ , is posted at service stations to show the antiknock quality of a fuel.

For fuels with a rating higher than 100 octane, the rating is usually obtained by determining the amount of tetraethyllead compound that needs to be added to pure isooctane to match the knock resistance of the test fuel. For example, if the amount is 1.3 ml, the fuel's rating is expressed as  $100 + 1.3$  or extrapolated above 100 (in this case, about 110 octane) by means of a correlation curve.

John C. Lane

**Bibliography.** American Society for Testing and Materials, *Annual Book of ASTM Standards*, pt. 47, annually.

## Octocorallia (Alcyonaria)

A subclass of Anthozoa. These cnidarians are benthic as adults, living attached to firm substrata or burrowed into soft sediments, from the intertidal zone to great depths. As in all anthozoans, the adult form is a polyp—a cylindrical organism that has, at its free end, the single body opening, the mouth, which is surrounded by eight tentacles. They are colonial (**Fig. 1**), with a few possible exceptions.

**Morphology.** Each of the eight tentacles has many small or elongate side branches (pinnules) oriented perpendicular to the axis of the hollow, flexible tentacle. The opposite end of the polyp is attached to the substratum or, more commonly in alcyonarians, emerges from the tissue that unites the members of

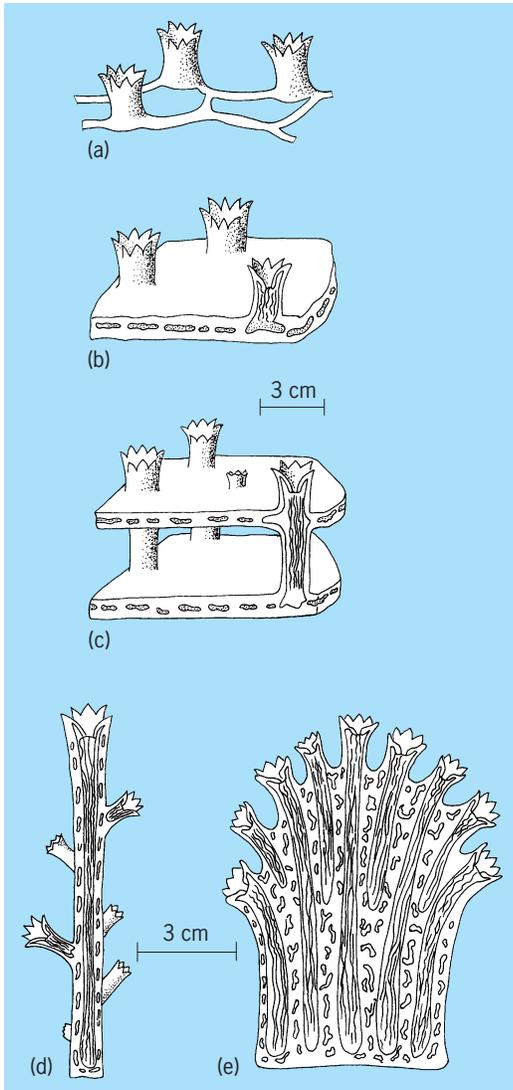


Fig. 1. Diagrammatic figures of the colonies of Alcyonaria. (a) *Cornularia*. (b) *Clavularia*. (c) *Tubipora*. (d) *Telesto*. (e) *Alcyonium*.

the colony. Sexual reproduction typically results in a planula larva, which eventually metamorphoses into a polyp. Asexual (vegetative) reproduction typically involves budding a new polyp from an existing one or from the tissue connecting members of the colony (the coenenchyme); in a colony, the resulting polyps remain attached to one another.

**Autozoid.** All alcyonarians have a type of polyp known as the autozoid (Figs. 2 and 3). The internal body space (the gastrovascular cavity) is divided by eight longitudinally oriented sheets of tissue known as mesenteries (Fig. 3). The best-developed pair of mesenteries (the asulcal pair) bears large, heavily ciliated filaments. The other six mesenteries bear filaments having many gland cells. Gametes form and mature in the mesenteries, which rupture to release them when they are ripe; some kinds of alcyonarians are hermaphroditic and some are gonochoric (with separate sexes).

Typically gametes are spawned freely into the sea, where egg and sperm meet and the planula devel-

ops; in some alcyonarians, however, embryos are brooded internally or on the surface of the colony, so a free-swimming stage of the life cycle is lacking. Longitudinal retractor muscles are strongly developed on the sulcal surface of each mesentery; thus, the retractor muscles of the sulcal pair face each other (Fig. 3).

The actinopharynx (stomodeum) has a single, ciliated siphonoglyph (longitudinal groove) located on the sulcal side (Figs. 2 and 3). Thus, although the polyp exhibits octamerous radial symmetry, it has elements of bilaterality; such biradial symmetry is typical of anthozoans. The flexible oral end of the polyp is termed the anthocodium. The basal portion, or anthostele, which is typically more rigid, is embedded in the coenenchyme, which is permeated by canals (the smaller of which are called solenia) that interconnect the gastrovascular cavities of polyps.

**Siphonozooid and mesozooid.** Some types of alcyonarians are dimorphic, having a second type of polyp

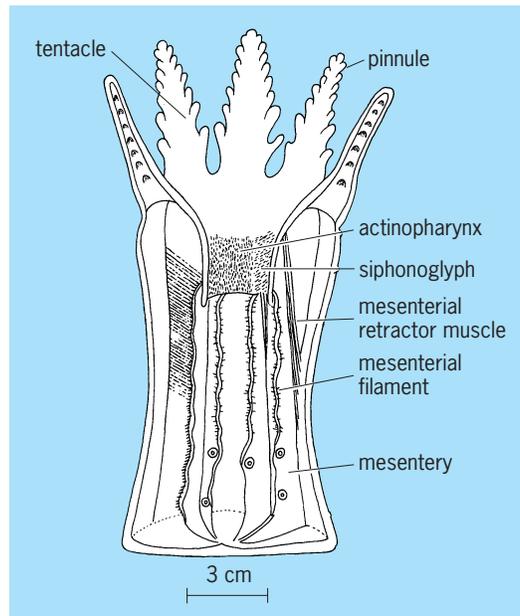


Fig. 2. Autozoid of Alcyonaria.

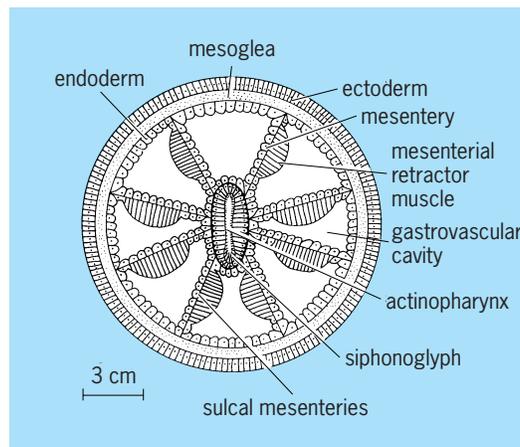


Fig. 3. Mesenterial arrangement of Alcyonaria.

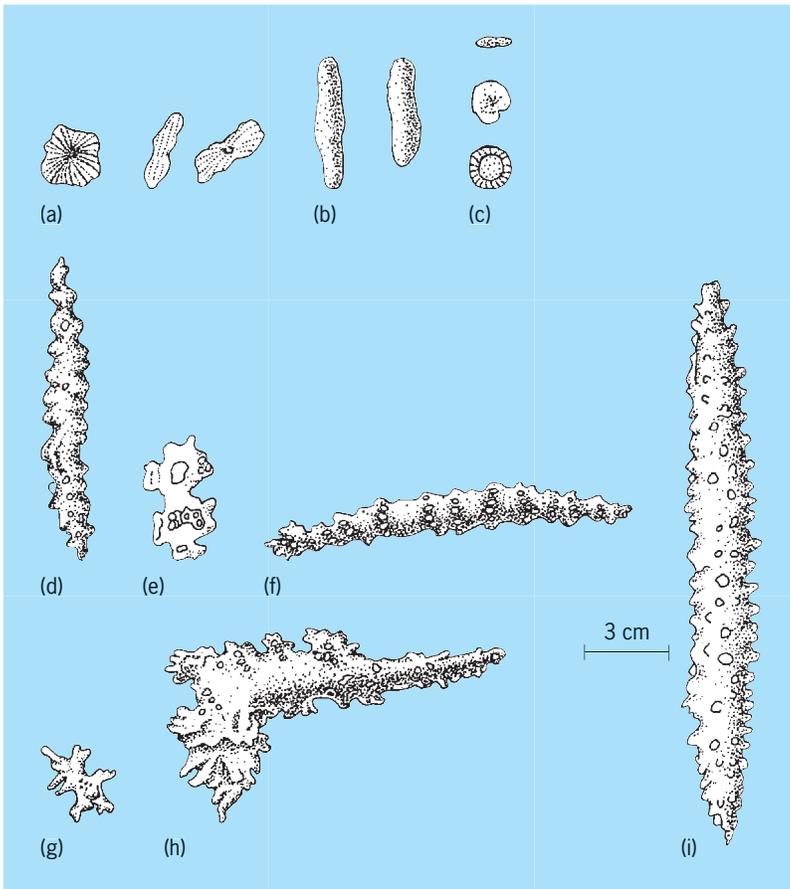


Fig. 4. Sclerites of Alcyonaria. (a) *Xenia uniserta*. (b) *Anthelia fuliginosa*. (c) *Sympodium coeruleum*. (d) *Clavularia chunni*. (e) *Sarcophyton crassocaule*. (f) *Sarcophyton acutangulum*. (g) *Anthomastus antarcticus*. (h) *Capnella rugosa*. (i) *Nephthya pacifica*.

known as a siphonozooid. Typically smaller than an autozooid, it functions to pump water through the canal system. It is by this means that a sea pen can extend from the sand in which the siphonozooid anchors it. Fewer alcyonarians still have a third type of polyp, the mesozooid, which is intermediate in morphology between the other two. Siphonozooids and mesozooids typically do not form eggs or sperm, and have reduced or no tentacles and mesenterial filaments.

**Skeleton.** The skeleton that functions to protect and support the octocoral is difficult or impossible to see while the animal is alive.

In two types of octocorals, the skeleton is solid and composed of the mineral calcium carbonate. One of these is the organ-pipe coral, *Tubipora* (Fig. 1c), whose red skeleton is in the form of elongate pipes (in which the living animal resides) connected intermittently by transverse platforms. The other is blue coral (or Helioporacea), whose skeleton, which resembles that of scleractinian corals (members of order Scleractinia, in the other anthozoan subclass), is blue (the skeleton of scleractinians is white).

In the rest of octocorals, the hard parts are sclerites made of calcium carbonate that range from microscopic to a few millimeters in length, and may look like lumps, discs, or rods (straight or curved) that are thick in the middle, pointed at the ends, and deco-

rated with knobs (Fig. 4). Sclerites may be clustered or fused together to form a cup around each polyp, scattered in the coenenchyme and polyp body wall and tentacles, or embedded in an organic matrix.

Animals with a large amount of coenenchyme typically are fleshy and flexible; they are commonly known as “soft corals.” In sea fans and sea whips (formerly constituting order Gorgonacea), sclerites are embedded in the thin living tissue that coats a skeleton in life and that quickly rots away after death, leaving the skeleton, which consists of an organic matrix that may be mineralized to some extent, and that is typically dendritic—in the form of a tree—or whiplike.

**Classification.** Previously, alcyonarians were divided into six or more orders, but now only three are recognized. The two that have long been recognized are Helioporacea, the blue corals and Pennatulacea, the sea pens and sea pansies. Pennatulaceans are unique among alcyonarians in having a colony with a definite symmetry: the primary (axial) polyp, which anchors the colony in the soft substratum, forms the main axis from which autozooids and siphonozooids branch bilaterally. Since intermediates have been discovered between what had been considered members of separate orders (including Stolonifera and Telestacea), all other alcyonarians are now considered to belong to order Alcyonacea. See ANTHOZOA; CNIDARIA.

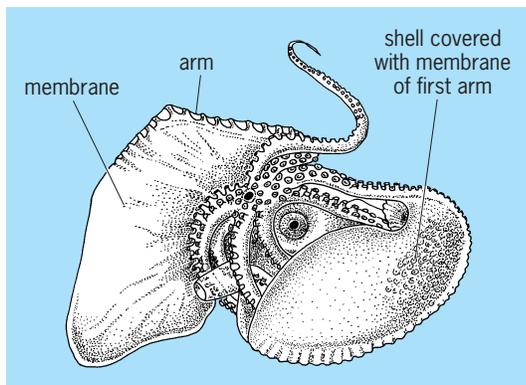
Daphne G. Fautin

**Bibliography.** H. Erhardt and D. Knop, *Corals: Indo-Pacific Field Guide*, Ikan, Frankfurt, 2005; K. Fabricius and P. Alderslade, *Soft Corals and Sea Fans: A Comprehensive Guide to the Tropical Shallow-water Genera of the Central-West Pacific, the Indian Ocean and the Red Sea*, Australian Institute of Marine Science, Townsville, 2001; M. Grasshoff and G. Bargibant, *Coral Reef Gorgonians of New Caledonia*, IRD, Paris, 2001; Several well-illustrated articles on octocorals, *Koralle* [a German-language magazine for aquarium hobbyists], issue 12 (December 2001/January 2002).

## Octopoda

An order of the class Cephalopoda (subclass Coleoidea), characterized by eight appendages that encircle the mouth, a saclike body, and an internal shell that is much modified or reduced from that of its ancestors. One or two rows of suckers without chitinous rings occur along the eight highly flexible contractile arms. The approximately 200 species of octopods include shallow-water forms like the common octopus, *Octopus vulgaris*; open-ocean species like the paper argonaut, *Argonauta argo* (see **illus.**); and deep-sea forms with fins like the flapjack devilfish, *Opisthoteuthis californica*. Two suborders divide the Octopoda into those with paddle-shaped fins on the body and tendrillike cirri on the arms (Cirrata) and those without fins or cirri (Incirrata). See OCTOPUS.

The cirrate or finned octopods generally are deep-sea species that live at depths from about 320 ft



Lateral view of female (*Argonauta*) Argonautidae.

(100 m) to over 16,000 ft (5000 m) either on (benthic) or near (epibenthic) the ocean bottom, such as *Opisthoteuthis* and *Cirrotheuthis*, respectively. They swim with their fins and the gentle jetting of water taken into their mantle and ejected through the funnel. The arms are connected with a very deep web. Most finned octopods are somewhat gelatinous, and the cirri along the arms are thought to be sensory organs that enable them to detect the presence of prey in the nearly total darkness of their habitat. The deep-living *Cirrothauma murrayi* has minute eyes that detect light but do not form images. All other octopods have extremely well-developed eyes, and the incirrate forms like the common *Octopus vulgaris* have acuity of vision that rivals or exceeds that of fishes.

The incirrates are shallow-living species, mostly benthic, but some groups are epipelagic (near-surface, open ocean) or mesopelagic (middepths, open ocean). Such octopuses have highly developed and enlarged brains and a complex nervous system to support their superior visual and tactile senses. Experiments have shown that octopuses learn complicated tasks and have long-term memories. Behavioral, neurological, and anatomical studies of the octopus brain have contributed immeasurably to understanding the human brain.

Benthic incirrate octopuses live worldwide in holes and crevices in rocks and coral reefs, or in burrows on muddy or sandy bottoms. They prey primarily upon crabs, bivalves (clams, mussels), and snails, which they subdue by injecting a poisonous saliva from their jaws, which are like parrot beaks. Species of the Indo-West Pacific genus *Hapalochlaena* have toxic saliva that has caused death in humans. The sexes are separate, and the young hatch out of strands of eggs attached in caves, in empty bivalve shells, or, in *Argonauta*, in the thin-shelled egg case carried by the female. Most species are small, from a few inches (1 in. = 2.5 cm) to 2–3 ft (0.6–1 m) across the outstretched arms, but *Octopus dofleini* from the North Pacific (from northern California to Alaska to Japan) grows to a 32-ft (10-m) arm span. Octopuses are captured and eaten by humans in many parts of the world. See CEPHALOPODA; COLEOIDEA.

Clyde F. E. Roper

Bibliography. P. R. Boyle (ed.), *Cephalopod Life Cycles*, vol. 1, 1983, vol. 2, 1987; F. W. Lane, *Kingdom of the Octopus*, 1960; K. N. Nesis, *Cephalopods of the World*, transla. from Russian, 1987; C. F. E. Roper, M. J. Sweeney, and C. E. Nauen, *Cephalopods of the World*, FAO Fisheries Synopsis 125, vol. 3, 1984; M. J. Wells, *Octopus: Physiology and Behavior of an Advanced Invertebrate*, 1978; K. M. Wilbur et al. (eds.), *The Mollusca*, vol. 11: *Form and Function*, 1988; K. M. Wilbur and M. R. Clarke (eds.), *The Mollusca*, vol. 12: *Paleontology and Neontology of Cephalopods*, 1988.

## Octopus

Any species in the subclass Coleoidea, order Polypoidea, of the cephalopod mollusks or, more particularly, any of the approximately 120 species of the genus *Octopus* (see **illus.**). Widely distributed in coastal areas of all oceans, such octopods have a saclike body without any skeletal support and eight arms of equal length which bear suckers. They are predacious carnivores, crawling on the sea bottom and lurking in submarine caves and crevices. They can swim by jet propulsion but rarely do so, except as an escape mechanism. The body ranges in size from 1.2 to 14 in. (3 to 35 cm) across, and there are no "giant" octopuses.

As in all cephalopods, the mantle cavity in *Octopus* is modified from the usual molluscan pattern by a reversal of the water circulation through the pair of ctenidia (gills). The importance of cilia is greatly reduced, with those of the ctenidia and pallial wall being functionally replaced by powerful muscles which pump water continuously for respiration and intermittently for jet propulsion. The efferent circulation from the gills leads to the two auricles of the systematic heart, but there are also two accessory pumps (sometimes called branchial hearts) receiving



The octopus, a species of the subclass Coleoidea, order Polypoidea, of the cephalopod mollusks.

blood from the closed venous system and driving it into the afferent branchial vessels. The gut is still based on the molluscan pattern, with ciliated leaflets in the posterior part of the stomach sorting indigestible material, but several glands secrete a wide range of powerful enzymes. Prey, such as living crabs, are captured by the arms with their suckers and pressed in toward the mouth, which is armed with both a chitinous parrotlike beak and a powerful radular apparatus. The saliva of *Octopus* contains a neurotoxin.

Rapid and patterned color changes are possible in octopuses, cuttlefish, and their allies because their chromatophores (unlike those of other invertebrates) are miniature bags of pigment expanded by the pull of muscles under nervous control. As well as being used in concealment and for terrorizing displays, rapid color changes along with tactile caresses form part of the courtship behavior of such cephalopods. The male conveys the sperm in a spermatophore packet by a specialized arm (the hectocotylus).

The sense organs of octopuses are complex and efficient in discrimination; they include eyes, olfactory pits, chemotactile and mechanotactile sensillae on the arms, and statocysts capable of detecting direction and angle of acceleration as well as static posture. The eyes have a cornea and iris diaphragm, a movable lens, and a retina, capable of light and dark adaptation, which is not inverted. They represent an outstanding example of convergent evolution when compared in detail to the eyes of higher vertebrates such as mammals, and experiments have shown octopus eyes to be capable of considerable image formation and shape discrimination.

The discrete neural ganglia found in other mollusks are fused into a massive brain in modern cephalopods. The brain and associated neural lobes in an adult *Octopus* may contain over 300 million cells, and the brain-body ratio is much greater than in lower vertebrates such as fish or amphibians. Anatomically and histologically, 30 brain centers can be distinguished, each of which can be experimentally shown to be the functional center for a particular kind of central nervous system activity. Octopuses are remarkably tough experimental animals and rapidly recover from brain lesions. Under totally controlled experimental conditions, octopuses (and cuttlefish of genus *Sepia*) can readily be trained to distinguish visually between pairs of geometric figures, and tactilely between textures and tastes of different surfaces. Thus investigations linking behavioral training to the precise local nerve cells involved in learning and memory have been possible in *Octopus*.

Memory modules consisting of small numbers of nerve cells can have certain circuit biases altered by training, and Young has developed unit models of them with wiring diagrams which correspond to the actual neural microanatomy in the brain of *Octopus*. Experimental lesions in octopuses, both already trained and undergoing training, have established that such neural memory systems are spatially limited to specific lobes. Some progress has

been made in defining in physicochemical terms the unit changes which take place during learning. Significantly, as in other brains, short-term and long-term memory systems are spatially separated in *Octopus*. See CEPHALOPODA; MOLLUSCA; NAUTILUS; OCTOPODA.

W. D. Russell-Hunter

Bibliography. S. P. Parker (ed.), *Synopsis and Classification of Living Organisms*, 2 vols., 1982; M. J. Wells, *Octopus: Physiology and Behaviour of an Advanced Invertebrate*, 1978; J. Z. Young, *The Anatomy of the Nervous System of Octopus vulgaris*, 1971; J. Z. Young, *A Model of the Brain*, 1964.

## Odonata

An order of the class Insecta, commonly known as dragonflies and damselflies (Fig. 1). Odonates generally are conspicuous insects with strong flying ability and a complex range of behaviors. They occur on all continents except Antarctica, although more species occur near the Equator than at higher latitudes. The larvae are aquatic, inhabiting all types of freshwater, including streams, rivers, ponds, and marshes. The adults fly over or near these localities, though some species may disperse and are occasionally found far from water. See INSECTA.

**Morphology.** Adult Odonata range in size from small (with a wingspan of 15 mm, or 0.6 in.) to very large (with a wingspan greater than 150 mm, or 6 in.). They have large compound eyes on either a globular or transversely elongate head. The antennae are short and bristlelike. There are two pairs of elongate, membranous wings, which often are clear or transparent; however, in some species, the wings have patches or bands of color. The wings have a distinctive venation, with five main longitudinal veins that originate from the base and are connected by a network of many crossveins (Fig. 1). Adult males have secondary genital organs beneath the second abdominal segment, or the second and third segments—the organs are unique to this order. Females of some families have a conspicuous ovipositor.



Fig. 1. An aeshnid dragonfly (note the distinctive wing venation) and a lested damselfly.



Fig. 2. Larvae of (a) a dragonfly and (b) a damselfly.

The larvae of Odonata are readily distinguished by a strongly developed, hinged labium, or lower lip, which is used to grasp prey (Fig. 2). This is held underneath the head when not in use; in some families it forms a mask covering the face.

**Classification.** Odonata contains more than 6000 species worldwide, and is divided into three suborders: the Anisoptera, or dragonflies; the Zygoptera, or damselflies; and an intermediate suborder called the Anisozygoptera. The term dragonfly is also used for the order as a whole. All the commonly encountered species belong to either Anisoptera or Zygoptera.

*Anisozygoptera.* Anisozygoptera contains just two living species, one from the Himalayas and the other from Japan, although additional fossil species are known. Anisozygopterans look like dragonflies but with damselfly wings.

*Anisoptera and zygoptera.* Anisoptera has been divided into nine families and Zygoptera into 18. Anisoptera can be distinguished from Zygoptera by differences in the wing venation, structural details of the male secondary genitalia, the grasping organs (claspers) at the end of the abdomen, and the larval gills. In dragonflies the hindwings are broader than the forewings and the venation develops differently; however, in damselflies the two pairs of wings are more or less equal. Dragonfly larvae “breathe” by means of gill folds inside the rectum, whereas larval damselflies have three (rarely two) relatively broad external gills at the end of the abdomen.

Although most dragonflies hold their wings horizontally when at rest, and most damselflies hold their wings closed above the abdomen, exceptions occur within each suborder. This trait is an unreliable guide to subordinal placement except in North America and Europe.

**Developmental stages.** There are three general stages in the life history: the egg, the larva (some-

times incorrectly called a nymph), and the adult.

*Egg.* Eggs are laid into emergent or submerged vegetation or in or on waterside vegetation, or they are dropped on to the water surface, where they sink to the bottom. Most eggs hatch after a few days, but some may overwinter or, in dry climates, may not hatch until it rains.

*Larva.* Larvae typically live in and around submerged vegetation or among the bottom detritus (Fig. 2). All species are predacious, the food being all types of aquatic invertebrates. Occasionally, a large dragonfly larva may attack and eat a tadpole or small fish. Larval development takes from about 5 weeks in some pond-dwelling species to several years in species that inhabit cold, upland streams. Many temperate species have one generation per year. During growth a larva undergoes some 10–15 molts. Wing buds begin to appear half-way through development and are prominent in later instars (stage between two molts). When ready for its final molt, the larva climbs out of the water and attaches itself to a support. The larval cuticle splits and the adult emerges. See INSECT PHYSIOLOGY; MOLTING (ARTHROPODA).

*Adult.* After its wings have dried, the young adult spends several days away from water before returning to breed. During this time, and for several days or weeks afterward, the adult colors gradually develop. Adults may emerge in a synchronized fashion, usually in spring, or else continuously throughout the warmer months. Adult lifespan ranges from about one week to several months, depending on the species and climate. In cold climates, some adults may survive the winter in sheltered spots; in hot climates, some estivate during the dry season.

Adult Odonata chiefly divide their time between feeding and breeding. As with the larvae, all species are predacious at the adult stage. They catch flying insects or pick them from the foliage. The females of many species spend long periods feeding away from water and may appear at breeding sites only to mate and lay eggs. Males typically congregate at or near the water's edge, where they stake out small territories, fight other males, and wait for females to appear. Courtship is not common except in some species of



Fig. 3. Libellulid dragonflies mating.

damselflies. In most species, the male simply grabs the female by the back of the head or the neck (prothorax), and the pair fly off together. In these tandem pairs, the male is in front and the female is behind. If the female is willing to mate, she bends her abdomen forward to collect sperm from the male's accessory genitalia, where he has previously deposited it (Fig. 3). Complicated sperm displacement activities, in which the male first removes any sperm previously inside the female, are common, and matings can last from several minutes to several hours. The subsequent act of egg laying is often accompanied by mate-guarding behavior. See REPRODUCTIVE BEHAVIOR.

**Coloration.** Males, especially of perching species, often are brightly colored. This coloration may be largely a warning to other males that the bearer is male and not worth pursuing. Some of the common colors, especially the bright blues and reds, are produced by movable color granules within the epidermal cells, and the resulting color varies with temperature. Unlike some other insects, the bright colors of Odonata do not indicate distastefulness or danger. Odonates are entirely harmless and without defenses beyond their superb eyesight and flying skills. The males of species that fly back and forth on a fixed beat and rarely perch are often the same color as females of their species. Camouflage colors are common in these species. See PROTECTIVE COLORATION.

**Enemies.** Common enemies of the larvae include other predacious aquatic invertebrates, wading birds, and fish. Small larvae also are at risk of attack from larger larvae of their own or other species. Birds prey on the adults and some are taken by other odonates. Birds, reptiles, frogs, and fish prey heavily on newly emerged and ovipositing adults.

**Interaction with humans.** Odonata are wholly beneficial insects, which (in large numbers) can ameliorate outbreaks of aquatic pest species. However, a large dragonfly can look frightening, and Western folklore offers several negative terms for these insects, including horse stinger and devil's darning needle. In China and Japan, dragonflies (including damselflies) are regarded as benign and auspicious insects. Odonate larvae are a useful natural source of food for many fish, and they are sometimes used by anglers as a fishing bait. Many Odonata are now at risk from pollution and the destruction of their habitats. See ENTOMOLOGY, ECONOMIC.

John Trueman

**Bibliography.** P. S. Corbet, *Dragonflies: Behaviour and Ecology of Odonata*, Harley Books, Colchester, UK, 1999; R. J. Tillyard, *The Biology of Dragonflies (Odonata or Paraneuroptera)*, Cambridge University Press, 1917.

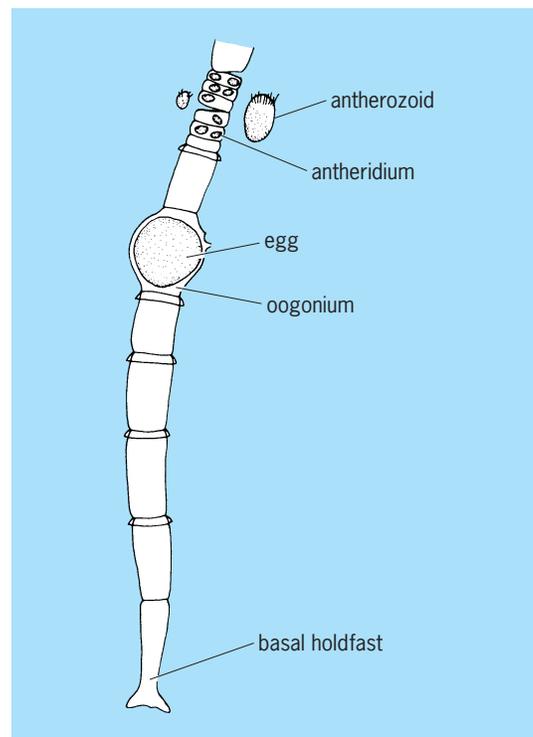
## Oedogoniales

An order of filamentous fresh-water green algae (Chlorophyceae) with unique morphological features including (1) an elaborate method of cell division that results in the accumulation of apical caps,

(2) zoospores and antherozoids with a subapical crown of flagella, and (3) a highly specialized type of oogamy. There is a single family, Oedogoniaceae, comprising three genera. See CHLOROPHYCEAE.

*Oedogonium* has the largest number of species (several hundred) and is the most common of the three genera. Its unbranched filaments are initially attached by a holdfast cell to submerged vegetation, stones, or wood, usually in permanent ponds or pools, but at maturity they may form free-floating masses. The cells are cylindrical, each containing a reticulate chloroplast with numerous pyrenoids. At the beginning of mitosis, a ring of wall material is deposited on the inner face of the lateral wall a short distance below the distal end of the cell. When mitosis is complete, the wall external to the ring ruptures and the ring is stretched to form the new portion of the daughter cell wall. That part of the parental wall lying above the ring is visible as an apical cap. Successive divisions result in an accumulation of caps.

Vegetative multiplication by fragmentation is common. In asexual reproduction, zoospores with a subapical crown of up to 120 flagella are formed singly within a cell. In sexual reproduction, there is a highly specialized interplay between female and male elements. The egg is a metamorphosed protoplast of an enlarged spherical cell (oogonium), which is the upper of two cells resulting from division of an oogonial mother cell, the lower being the supporting (sustentory) cell (see illustration). Antherozoids, with a subapical crown of about 30 flagella, are produced in pairs or tetrads in very small discoid antheridia. The egg is fertilized in place, the antherozoid entering through a pore or fissure in the oogonial wall.



Nannandrous species of *Oedogonium*, showing sexual reproductive structures.

The two types of sex organs may occur on the same filament (homothallic species) or on different filaments (heterothallic species), but in either case certain species have an indirect development of antherozoids. These species, termed nannandrous in distinction to those with direct development (macrandrous), form short cylindrical cells which initially appear like antheridia, but in which the protoplast metamorphoses into a single swarmer bearing a subapical crown of flagella. These swarmers swim around the filament until they are attracted chemically to the oogonial mother cell, the suffultory cell, or less often the oogonium. Upon settling, a swarmer forms a one-celled germling (dwarf male filament) that acts as an antheridial mother cell, cutting off one or more antheridia. The swarmers that give rise to the dwarf male filaments are called androspores, while the short cylindrical cells from which they arise are called androsporangia. The biochemical basis of this indirect development of the male gamete is only partly understood. The zygote secretes a thick, frequently ornamented wall and often accumulates a reddish-orange pigment (hematochrome). After a period of dormancy, it germinates meiotically to form four zoospores. The zoospores reestablish the vegetative filaments, which thus are haploid.

Filaments of *Bulbochaete* are unilaterally branched, and most cells bear a long hyaline bristle with a bulbous base. Additional axial cells are usually cut off by the basal cell, but any vegetative cell can initiate a branch. *Bulbochaete* is found in the same habitats as *Oedogonium*, but is less common and has fewer species (about 125). Filaments of *Oedocladium* are branched but lack bristles. The 13 species, which are terrestrial or aquatic, are seldom encountered. They seem to be most common in warm regions of America, India, and Australia.

Paul C. Silva; Richard L. Moe

Bibliography. H. C. Bold and M. J. Wynne, *Introduction to the Algae: Structure and Reproduction*, 1985; G. M. Smith, *The Fresh-Water Algae of the United States*, 1950; L. H. Tiffany, *The Oedogoniaceae*, 1930.

## Ohmmeter

A portable instrument for measuring relatively low values of electrical resistance. The ohmmeter is ideal for use in the workshop, field, or test laboratory. The range of resistance measured is typically from 0.1



Fig. 1. Electronic digital ohmmeter. (Tinsley)

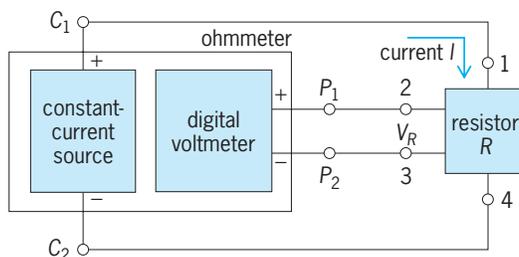


Fig. 2. Diagram of ohmmeter, showing operating principles.  $C_1$ ,  $C_2$  = current terminals;  $P_1$ ,  $P_2$  = potential terminals.

microhm to 1999 ohms ( $\Omega$ ) [Fig. 1].

**Operating principles.** The ohmmeter may be either an analog or a digital instrument. It uses a four-terminal measurement method (Fig. 2) in which a constant-current source is applied to the current terminals 1 and 4 of the resistor  $R$  via the current terminals  $C_1$  and  $C_2$  on the ohmmeter. In the digital type of ohmmeter, a digital voltmeter is then connected across the potential terminals 2 and 3 of the resistor  $R$  via the terminals  $P_1$  and  $P_2$  on the ohmmeter.

The error caused by the shunting effect of the digital voltmeter is almost negligible because the digital voltmeter can have an input resistance in excess of 10 gigohms ( $10^{10} \Omega$ ) on its basic ranges. The contribution due to the shunting error is less than 0.01% when measuring a test resistor of up to 1 megohm ( $10^6 \Omega$ ).

By Ohm's law, the voltage drop  $V_R$  across the resistor (Fig. 2) is related to its resistance  $R$  and the current  $I$  passing through it by Eq. (1). If the value

$$R = \frac{V_R}{I} \quad V_R = IR \quad (1)$$

of the resistor  $R$  is chosen to be 1 k $\Omega$  and the current is set to be a constant value of 1 mA, then the voltage drop across the resistor will be given by Eq. (2). The digital voltmeter connected across the

$$V_R = 0.001 \text{ A} \times 1000 \Omega = 1 \text{ V} \quad (2)$$

resistor will indicate 1.0000 V, assuming that a  $4^{1/2}$ -digit voltmeter is used ( $^{1/2}$  digit refers to the fact that the digit before the decimal point can only be 0 or 1). If the value of the resistor  $R$  is changed to, say, 500  $\Omega$  and the current is constant at 1 mA, then the digital voltmeter will read  $0.001 \text{ A} \times 500 \Omega = 0.5000 \text{ V}$ . See ELECTRICAL RESISTANCE; OHM'S LAW; VOLTMETER.

The digital voltmeter indication is clearly directly proportional to the value of the resistance  $R$ , and so the digital voltmeter display can be annotated to read in kilohms. Other resistance ranges can be added simply by changing the value of the constant current or by changing the full-scale range of the digital voltmeter.

In a practical ohmmeter, the current is not adjusted separately, but the instrument is calibrated by using known values of test resistance; therefore any errors in the digital voltmeter are compensated for. Once the instrument is calibrated, an unknown resistor  $R_x$

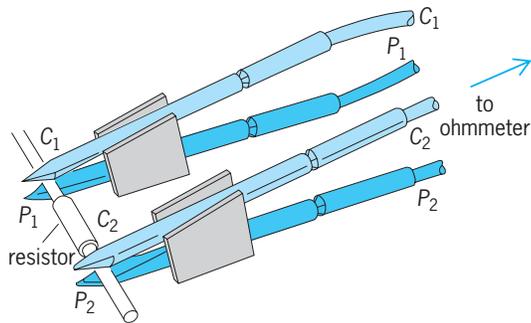


Fig. 3. Special measuring probes (Kelvin clips), consisting of a set of leads with combined current probes ( $C_1$ ,  $C_2$ ) and potential probes ( $P_1$ ,  $P_2$ ).

can be connected to it in place of  $R$ , and its value will be measured and indicated on the digital voltmeter.

The resistance measurement thus carried out provides a true four-terminal value, but resistors with only two terminals or wires can be measured by connecting the current and potential leads together in pairs. For most ohmmeters, special test leads are provided for the purpose of measuring two-terminal resistors, and Fig. 3 shows a typical set of leads with combined current and potential probes, often known as Kelvin clips.

**Applications.** The ohmmeter solves quickly and easily a variety of difficult measurement problems, including measuring the resistance of cladding and tracks on printed circuit boards, electrical connectors, and switch and relay contacts, as well as determining the quality of ground-conductor continuity and bonding, cables, bus-bar joints, and welded connector tags. See RESISTANCE MEASUREMENT.

A. Douglas Skinner

## Ohm's law

The statement that the current  $I$  flowing in an electrical circuit is often, to a very good approximation, proportional to the voltage  $V$  of its source; that is,  $V = IZ$ . If the current is steady direct current (DC),  $Z$  equals the total resistance of the circuit,  $R$ .

Ohm's law, although closely obeyed in most metallic and some other conductors, is an approximation, and does not have the same physical status as, say, Maxwell's equations. See MAXWELL'S EQUATIONS.

If  $Z$  or  $R$  is independent of  $I$  to a sufficient approximation, the circuit and the elements in it are said to be linear and network theory can be applied. If  $V$  and  $I$  are time-dependent, alternating-current circuit theory applies to their frequency components. The current is derived from the voltage by considering not only the resistances but also the vector impedances composed of self- and mutual inductances and capacitances together with their losses. These latter components combine into the total impedance  $Z$ , and are often also linear to a sufficient approximation.

See ALTERNATING-CURRENT CIRCUIT THEORY; CAPACITANCE; ELECTRICAL IMPEDANCE; INDUCTANCE; NETWORK THEORY.

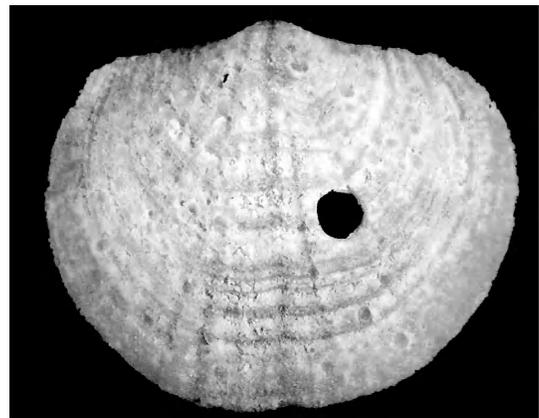
This 'law' was first described by Georg Simon Ohm in 1827 as a result of his experiments with metallic conductors. See ELECTRICAL RESISTANCE; ELECTRICAL RESISTIVITY; RESISTANCE MEASUREMENT; SEMICONDUCTOR; SKIN EFFECT (ELECTRICITY); THERMAL CONDUCTION IN SOLIDS.

Bryan P. Kibble

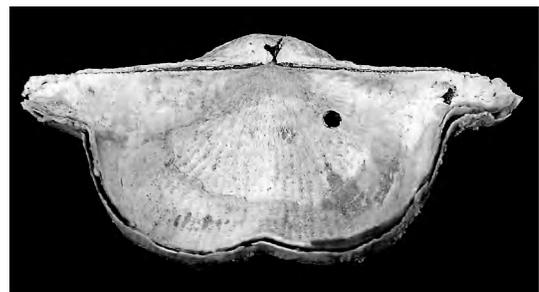
## Oichnus

A morphologically distinct type of trace fossil left by boring organisms in hard biogenic substrates. *Oichnus* Bromley 1981 is a formal taxonomic name with the rank of ichnogenus (a trace fossil genus). *Oichnus* is a Latin-derived term that literally means a trace (ichnus) shaped like the letter "O"; it denotes the small circular to subcircular holes (borings) found primarily in fossilized skeletal remains of invertebrate animals, such as mollusk shells, ostracode valves, or echinoid tests.

**Classification.** *Oichnus* includes several ichnospecies (trace fossil species) that are classified based on differences in the vertical cross section of a boring (cylindrical, conical, etc.), horizontal outline of its outer and inner openings (circular or oval), and



(a)



(b)

Fig. 1. Two examples of *Oichnus simplex* drilled by some unknown predators or parasites in shells of Late Paleozoic brachiopods extracted from the Permian rocks of West Texas. The specimens are from the Copper and Grant Brachiopod Collection housed in the National Museum of Natural History, the Smithsonian Institution. (Courtesy of Finnegan Marsh, National Museum of Natural History)

the degree of penetration (holes or pits). Seven ichnospecies have been named formally so far, including the type ichnospecies *O. simplex* Bromley 1981 (Fig. 1), as well as six other forms: *O. paraboloides* Bromley 1981, *O. ovalis* Bromley 1991, *O. asperus* Nielsen and Nielsen 1991, *O. coronatus* Nielsen and Nielsen 1991, *O. gradates* Nielsen and Nielsen 1991, and *O. excavatus* Donovan and Jagt 2002. The ichnogenus *Tremichnus* Brett 1985 is viewed by some as the junior synonym of *Oichnus*. See TRACE FOSSILS.

**Shell drilling.** In modern ecosystems *Oichnus*-producing drilling organisms are widespread and diverse in many marine habitats. These drillers are often predators that bore to access the soft tissue of their shelled prey. However, externally parasitic (ectoparasitic) borers (which affect hosts negatively without killing them), commensal borers (which benefit without harming their hosts), and even amensal borers (which do not benefit from boring but are detrimental to their hosts), have been documented as well. Shell drilling is a highly convergent behavior that evolved multiple times independently in unrelated groups of organisms; however, the most well-known and studied drillers are boring gastropods: naticids and muricids. In addition, many other groups of snails, as well as some species of octopods, nematodes, and flatworms, are capable of drilling holes in their victims. Drilling ability has also been demonstrated in an extinct group of Paleozoic gastropods (Platyceratidae). In modern oceans, drillings of *Oichnus* type are found in a broad range of marine invertebrate shells, including (most frequently) mollusk prey and hosts and (less commonly) other organisms with biomineralized skeletons, such as echinoderms, arthropods, brachiopods, bryozoans, and foraminiferans.

**Paleoecology of drilling predation.** *Oichnus* is often easy to identify in the fossil record, and its ecological genesis can be evaluated rigorously using a series of qualitative and quantitative criteria. Most widespread and well-studied fossil types of *Oichnus* borings, such as *O. simplex* and *O. paraboloides*, are attributed typically (based on extrapolation from modern ecosystems) to predatory and/or ectoparasitic organisms that drilled holes in external skeletons of live victims. Because of its high informative value and frequent occurrence in the fossil record, *Oichnus* provides arguably some of the best direct paleontological evidence for ecological interactions between ancient marine predators (or parasites) and their prey (or hosts).

*Oichnus* borings occur continuously through the entire fossil record of marine macroorganisms, since the dawn of metazoan animals in the Late Proterozoic Era (approximately 600 million years ago) until present day (Fig. 2). Even the oldest known skeletal remains from the Late Proterozoic, the enigmatic wormlike tubes *Cloudina*, bear traces that are believed to represent the oldest documented case of biotic interactions between drilling predators and their skeletonized prey. Numerical studies of drilling frequencies, based on *Oichnus* trace fossils, have offered some of the most compelling evidence

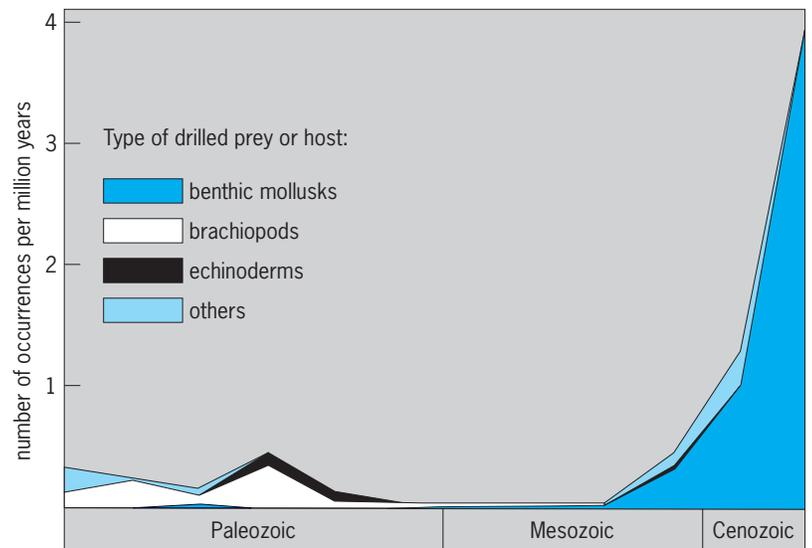


Fig. 2. Phanerozoic history of drilling predation based on fossil occurrences of the trace fossil *Oichnus* in shells of various marine invertebrates. A sharp increase in the Late Mesozoic and Cenozoic coincides with the Mesozoic Marine Revolution—a rapid radiation of many predatory groups coupled with an increase in substrate burrowing by marine benthic animal prey such as bivalve mollusks and echinoids. (Based on the literature compilation of M. Kowalewski et al., 1998)

for a long-term increase in the intensity of predation through evolutionary time. See FOSSIL; PALEOECOLOGY; PALEONTOLOGY; PREDATOR-PREY INTERACTIONS.

Michal Kowalewski

**Bibliography.** R. G. Bromley, *Trace Fossils: Biology and Taphonomy*, Unwin Hyman, London, 1990; P. H. Kelley, M. Kowalewski, and T. A. Hansen, *Predator-Prey Interactions in the Fossil Record*, Topics in Geobiology Series 20, Plenum Press-Kluwer, New York, 2003; M. Kowalewski, A. Dulai, and F. T. Fuersich, A fossil record full of holes: The Phanerozoic history of drilling predation, *Geology*, 26: 1091–1094, 1998; G. J. Vermeij, *Evolution and Escalation: An Ecological History of Life*, Princeton University Press, 1987.

## Oil analysis

Analysis of petroleum, or crude oil, to determine its value in modern refinery operations. In addition, procedures have been developed for analysis of lubricating oil.

For refinery operations, oil analysis, or assay, must provide the refinery planner with the data needed to predict yields, qualities, and operating costs for a wide variety of refinery operating conditions and product demands. In a refinery, crude oil is distilled and separated into products according to the boiling points of the crude oil components (see table). See PETROLEUM; PETROLEUM PRODUCTS.

**Procedure.** A crude assay follows much the same procedure. The oil is distilled and separated into up to 40 narrow-boiling-range cuts. Cuts with normal boiling points higher than about 370°F (190°C) are distilled at greatly reduced pressure. This allows the oil to vaporize at temperatures below 370°F (190°C) and avoids thermal reactions which would alter the

Typical products derived from crude oil		
Product	Carbon atom range	Boiling point range, °F (°C)
Gas	C <sub>1</sub> and C <sub>2</sub>	
Liquefied petroleum gas	C <sub>3</sub> and C <sub>4</sub>	
Gasolines	C <sub>4</sub> to C <sub>10</sub>	59–370 (15–190)
Kerosines	C <sub>9</sub> to C <sub>15</sub>	300–540 (150–280)
Middle distillates	C <sub>12</sub> to C <sub>20</sub>	390–640 (200–340)
Gas oils	C <sub>20</sub> to C <sub>45</sub>	640–1040 (340–560)
Bottoms	Unvaporized residua	

chemical properties of the oil. Each cut is then subjected to a variety of tests sufficient to characterize it for the products the cut could be included in. The refinery planners may then calculate yield and qualities of any product by blending the yields and qualities of the cuts that are included in the product.

Petroleum consists primarily of compounds of carbon and hydrogen containing from 1 to about 60 carbon atoms. Carbon atoms in natural petroleum occur in straight and branched chains (paraffins), in single or multiple saturated rings (cycloparaffins or naphthenes), and in cyclic structures of the aromatic type such as benzene, naphthalene, and phenanthrene. Cyclic structures may have attached to them side chains of paraffinic carbons. In lubricating oil it is usual to have naphthene rings built onto the aromatic rings and side chains attached. In products produced by cracking in the refinery, olefins or compounds with carbon-carbon double bonds not in aromatic rings are also found. The high-boiling fractions of petroleum contain increasing amounts of oxygen, nitrogen, and sulfur compounds, as well as traces of organic compounds of metals such as vanadium, nickel, and iron.

**Tests.** Three types of tests are used to characterize a crude oil and its narrow-boiling-range cuts: (1) tests for physical properties such as specific gravity, refractive index, freeze temperature, vapor pressure, octane number, and viscosity; (2) tests for specific chemical species such as sulfur, nitrogen, metals, and total paraffins, naphthenes, and aromatics; (3) tests for determining actual chemical composition.

*Physical properties.* The physical properties are used to predict how petroleum products will perform. Octane number and vapor pressure are two important qualities of gasolines. Jet fuels must be formulated so that they remain fluid, and therefore pumpable, at the low temperatures encountered by high-flying airplanes. Lubricating oils are characterized by their viscosity level together with the change of viscosity with temperature. The many different tests used for physical properties are relatively simple. Methods have been standardized and published by the American Society for Testing and Materials (ASTM).

*Chemical species.* The amounts of specific chemical species are also important in evaluating petroleum products. In these days of environmental concern and emission controls, the sulfur content often determines whether or not a fuel can be used. Aromatic components are desirable in gasoline and undesirable in diesel fuels. Fractions with vanadium, nickel,

and iron compounds should not be used in catalytic processes since these metals permanently deactivate most catalysts. Tests are available to measure all the chemical species of interest. Many of the tests involve controlled burning of an oil sample, followed by analysis of the combustion products.

*Chemical composition.* The analysis of petroleum in terms of chemical composition can be done with varying degrees of thoroughness. Anything approaching the complete analysis of a crude oil is so time-consuming and expensive that in the whole world only one sample of crude oil has been analyzed thoroughly. This project, known as Project No. 6 of the American Petroleum Institute, was carried on from 1928 to 1968, when the remaining work was transferred to another project. Obviously, such a detailed analysis will never be done on a routine basis.

Gas chromatography, mass spectroscopy, and nuclear magnetic resonance (NMR) methods are capable of economically determining the detailed chemical composition of petroleum fractions through the C<sub>9</sub> or C<sub>10</sub> boiling range plus the distribution of hydrocarbon types up to about C<sub>40</sub>. The hydrocarbon-type analysis includes distinguishing between one-, two-, and four-ring naphthenes and aromatics. *See* CHROMATOGRAPHY; MASS SPECTROMETRY; NUCLEAR MAGNETIC RESONANCE (NMR).

In a gas chromatography analysis a small sample is introduced into a column packed with an adsorbent material such as silica gel. The sample is swept along the column by an inert carrier gas such as nitrogen. The carrier gas forces all components in the sample to move along the column, but each component moves at a unique rate depending on how strongly it is absorbed by the silica gel. Each component leaves the column as a discrete packet, or peak, of noncarrier gas. The time of each peak identifies the component, and the amount in the peak gives the composition. The operation of the gas chromatography column and the calculation of results have been automated so that very little human effort is required. However, setting up a gas chromatography column is a tedious procedure since each column seems to behave differently and so must be calibrated for all components it will be used to measure. Also, the method has trouble distinguishing between similar isomers that give overlapping peaks. *See* GAS CHROMATOGRAPHY.

In a mass spectroscopy analysis the molecules of the sample being tested are bombarded with an

electron beam. This breaks the molecules into ionized fragments whose weights are determined by measuring how much the ionized fragments bend as they move through a magnetic field. The fragment pattern obtained is then correlated against the known fragment patterns of pure compounds to find the composition of the sample. It is difficult to apply the method to compounds for which calibration patterns are not available. It is, however, not feasible to obtain calibration patterns for the many thousands of compounds which appear in petroleum. Nevertheless, even in the higher-boiling fractions where these uncertainties exist, the mass spectroscopy analysis is the best available method for hydrocarbon type analysis.

An NMR analysis in its simplest form determines the type of hydrogen present in a sample, that is, whether the hydrogen is attached to an aliphatic, naphthenic, or aromatic type carbon atom. This is determined from a measurement of the magnetic resonance spectrum of the diluted sample under the influence of a very strong magnetic field. Information of this type, coupled with other analytical information, makes possible the development of a rather detailed picture of molecular types and structures.

Jack Rees

**Lubricating oil.** Analysis of used lubricating oils is part of the maintenance program for many types of engines and industrial equipment. By conducting baseline oil analysis and subsequent used-oil analysis on a regular schedule as part of a preventive maintenance program, mechanical and lubricant-related failures may be greatly reduced. Each use of lubricating-oil analysis was primarily to detect water in the steam-turbine and propulsion systems of large ships. Later, lubricating-oil analysis for mechanical and lubricating-related problems became part of the preventive maintenance program for helicopters and jet-powered aircraft, reducing engine failures and extending the time interval between major overhauls. Owners and operators of bus and truck fleets, and almost all types of engines and other equipment that use liquid lubricant, may use lubricating-oil analysis to prevent or minimize mechanical failure while extending the useful service life of expensive equipment.

To analyze a lubricating oil, a representative sample of the lubricating flowing in the system at normal operating temperature is collected in a clear container and labeled with pertinent data. The sample is then delivered to the testing laboratory as soon as possible. Tests applied to used engine oils include viscosity, fuel dilution, water, insolubles, and spectrochemical or spectrographic analysis. Wear-metal concentrations are determined in parts per million (ppm); these metals may include silver, aluminum, chromium, copper, iron, nickel, lead, and tin, in addition to silicon. An increase in any of these concentrations may indicate abnormal engine wear. High copper and lead levels usually indicate bearing distress. A high aluminium level usually indicates piston, bearing, or cylinder wear. A high chromium level usually indicates cylinder liner or

piston ring wear. The presence of silicon (dirt) is often accompanied by a corresponding increase in wear-metal concentrations. Boron, sodium, and water indicate an internal coolant leak in the engine. See ENGINE; INTERNAL COMBUSTION ENGINE; SPECTROSCOPY.

Donald L. Anglin

**Bibliography.** American Society for Testing and Materials, *ASTM Standards*, annually; ASTM, *ASTM and Other Specifications for Petroleum Products and Lubricants*, 4th ed., 1985; ASTM, *Calculation of Physical Properties of Petroleum Products from Gas Chromatographic Analysis*, 1975; ASTM, *Miscellaneous ASTM Standards for Petroleum Products*, 15th ed., 1981; ASTM, *Petroleum Products, Lubricants, and Fossil Fuels*, 1986; L. F. Hatch and S. Matar, *Hydrocarbons to Petrochemicals*, 1981; Institute of Petroleum, *Standard Methods of Analysis and Testing of Petroleum and Related Products*, 1987; S. H. Kagler, *Spectroscopic and Chromatographic Analysis of Mineral Oil*, 1973.

## Oil and gas, offshore

Oil and gas prospecting and exploitation are conducted on continental shelves and slopes in the seas around the world. Exploration for offshore petroleum and natural gas expanded rapidly after the end of World War II to include exploration, drilling, and production off the coasts of more than 75 nations. See NATURAL GAS; OIL AND GAS FIELD EXPLOITATION; PETROLEUM.

For many years, petroleum companies stopped at the water's edge or sought and developed oil and gas accumulations only in inland waters or shallow seas bordering onshore producing areas. Exploration deeper under the sea, and production from the continental shelves beyond territorial limits, did not begin in earnest until the world's increasing demand for petroleum energy sources, coupled with a lessening return from land drilling, provided the incentives for the huge investments needed for drilling in the open sea. Recent advances in technology permit drilling and production from ever deeper waters on the continental slopes. See CONTINENTAL MARGIN; OIL AND GAS WELL DRILLING; PETROLEUM RESERVES.

**Historical development.** Initial offshore development was a simple extension of land practice. The first offshore well was drilled in 1897 from a wharf made of wooden piling and timbers that extended about 300 ft (90 m) into the Pacific Ocean near Santa Barbara, California. By the early 1930s, oilfields had been discovered in the inshore and coastal areas of Lake Maracaibo, Venezuela; Louisiana; and the Caspian Sea. In 1938, an oilfield 1 mi (1.6 km) off the coast of Louisiana was discovered by directional drilling from land. The first platforms out of sight of land were installed in the Gulf of Mexico in 1947.

Since then, the offshore oil industry has moved progressively into deeper water and farther away from shore. Today offshore oil exploration and production is a worldwide industry. By the early 1990s, offshore sources accounted for 30% of worldwide

crude oil production and 14% of worldwide natural gas. Until now, most offshore production came from reservoirs located under the continental shelf, in water depths up to 600 ft (180 m). Spurred by technology and a need to find additional secure sources of energy, exploration and production are now moving even farther from shore and into the deeper waters of the continental slopes. Exploration wells have been drilled in water deeper than 9000 ft (2750 m), and hydrocarbons are being produced from offshore fields in waters deeper than 5000 ft (1500 m).

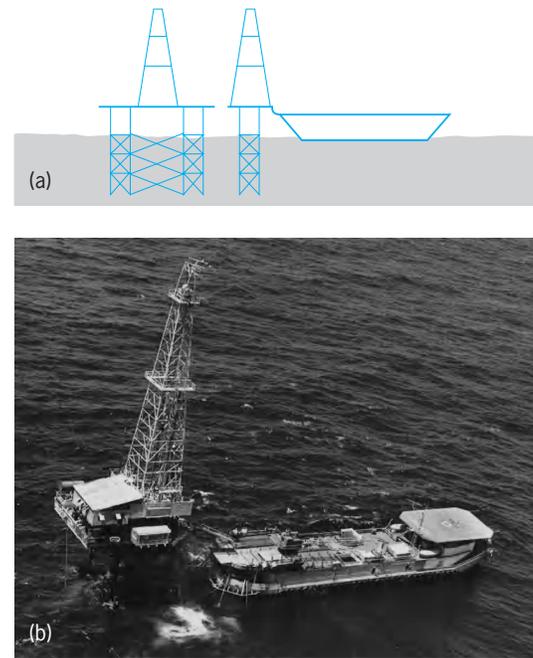
**Geology and the sea.** There is a sound geologic basis for the petroleum industry turning to the continental shelves and slopes in search of needed reserves. Favorable sediments and structures exist beneath the present seas of the world in geologic settings that have proven highly productive onshore. In fact, the subsea geologic similarity, or in some cases superiority, to geologic conditions on land has been a vital factor in the expansion of the world's investment in offshore exploration and production. See MARINE GEOLOGY.

Drilling in deep water has alleviated early fears regarding geochemistry and reservoir quality in deep-water environments. Many young, shallow reservoirs have proved to be oil-productive, with high-quality reservoirs capable of producing at the rates required to justify investment, and virtually all of the world's continental shelves have received some geologic study.

By the early 1990s, major offshore petroleum provinces had been established in the Beaufort Sea; Cook Inlet; offshore California; Gulf of Mexico; West Africa offshore Nigeria, Gabon, and Angola; Campos Basin offshore Brazil; North Sea; Arabian Gulf; Red Sea; and offshore India, Thailand, Indonesia, Malaysia, and Australia. The recent industry move into deeper water has greatly extended the potential of many of these areas.

Significant exploration is taking place in water depths up to 10,000 ft (3050 m) in the Gulf of Mexico, offshore Brazil, offshore West Africa, and in the northern Atlantic off the coasts of Norway, Ireland, and the United Kingdom. Oil and natural gas fields in waters as deep as 5500 ft (1675 m) are already on-stream in the Gulf of Mexico and offshore Brazil. In the North Atlantic west of the Shetland Islands, where weather conditions are particularly severe, production has started from reservoirs in water depths up to 1500 ft (450 m). See PETROLEUM GEOLOGY.

In the United States in the early 1990s, the Atlantic coast, south Florida, the Pacific coast, and most of offshore Alaska were placed under a 10-year moratorium preventing new exploration. This moratorium has been extended and is due to expire in 2012. The move into deep water has therefore been limited to the Gulf of Mexico. Here it is fueled by the geological attractiveness of the area, and has been greatly assisted by advances in geophysical technology such as three-dimensional seismic tomography. It has also been encouraged by the partial royalty relief provided for developments in water depths greater than



**Fig. 1. Offshore fixed drilling platform. (a) Underwater design (World Petroleum). (b) Rig on a drilling site (Marathon Oil Co.).**

650 ft (200 m), which helps offset the considerable development cost. More than 25% of all discoveries in the Gulf of Mexico are now being made in waters deeper than 1000 ft (300 m). In 1998 more than 20% of the total barrel-of-oil-equivalent production in the Gulf of Mexico came from deep-water fields, and there are already many more prospects for possible development in water depths up to 6500 ft (2000 m). The U.S. Minerals Management Service estimated that by year-end 2000, almost two-thirds of the oil and one-third of the gas produced in the Gulf of Mexico would be from reservoirs located in water depths greater than 1000 ft. See SEA-FLOOR IMAGING; SEISMOLOGY.

**Mobile drilling platforms.** The underwater search has been made possible only by vast improvements in offshore technology. Drillers first took to sea with land rigs mounted on barges towed to location and anchored, or with fixed platforms accompanied by a tender ship (**Fig. 1**). As the search for oil and natural gas advanced worldwide and farther away from shore, types of exploration rigs evolved which could move easily between locations and operate in a wide range of water depths.

**Self-elevating platform.** In shallow water, the most widely used mobile platform is the self-elevating, or jack-up, unit (**Fig. 2**). It is towed to location, where the legs are lowered to the sea floor, and the platform is jacked-up above wave height. These self-contained platforms are especially suited to wildcat and delineation drilling, but have also been converted to fixed production use. They are best on firmer sea bottoms, with a depth limit out to 300 ft (90 m) of water for all but a few of the largest units.

**Submersible platforms.** These were developed from earlier submersible barges which were used in

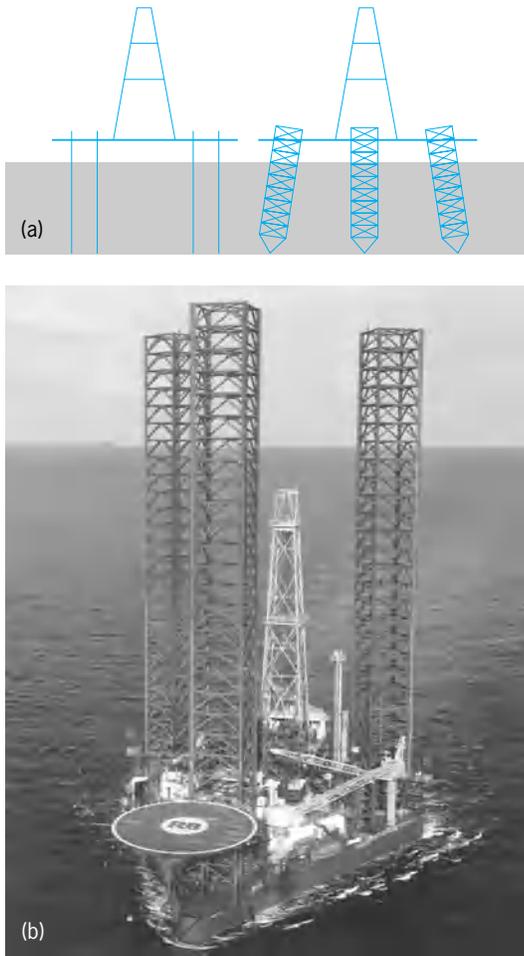


Fig. 2. Offshore self-elevating drilling platform. (a) Underwater design (World Petroleum). (b) Jack-up drilling rig (Reading and Bates).

shallow inlet drilling along the United States Gulf coast. The platforms are towed to the location and then submerged to the sea bottom. They are very stable and can operate in areas with soft sea floors. Difficulty in towing is a disadvantage, but this is partially offset by the rapidity with which they can be raised or lowered, once on location. They are not suitable for deep water and therefore now have restricted use.

**Semisubmersible platforms.** These are a version of submersibles (Fig. 3). They can work as bottom-supported units or in deep water as floaters. They can operate in a wide range of water depths, and when working as floaters, their primary buoyancy lies below the action of the waves, thus providing greater stability.

**Floating drill ships.** These are capable of drilling in the deepest waters (Fig. 4). They are built as self-propelled ships, or with a ship configuration that requires towing. Floating drill ships use anchoring, or in very deep water, dynamic-positioning systems, to stabilize their position. Floaters cannot be used in waters much shallower than 70 ft (20 m). This is because the special drilling equipment is then subject to excessive vertical movement from waves and tidal changes, and horizontal shifts, due to stretch

and play in anchor lines, are too large compared with the water depth.

Exploration in deeper water requires the construction of new semisubmersibles and floating drill ships,

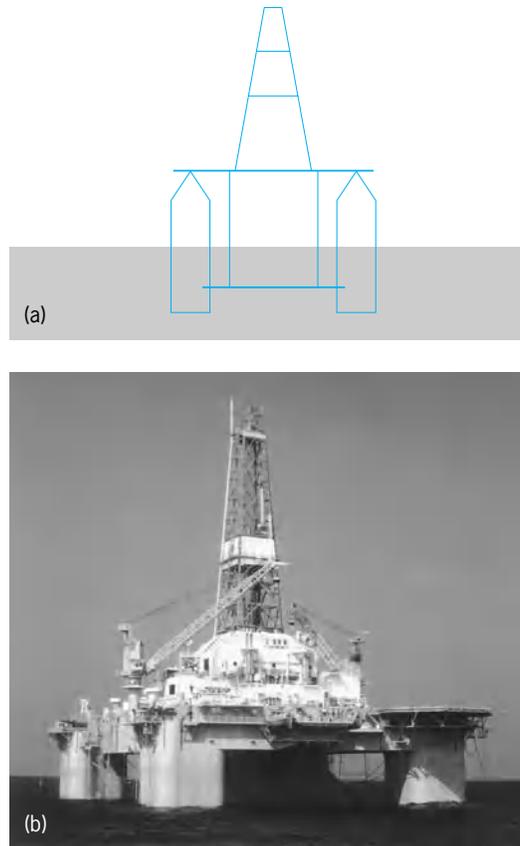


Fig. 3. Offshore semisubmersible drilling platform. (a) Underwater design (World Petroleum). (b) Dynamically positioned fourth-generation semisubmersible (Reading and Bates).

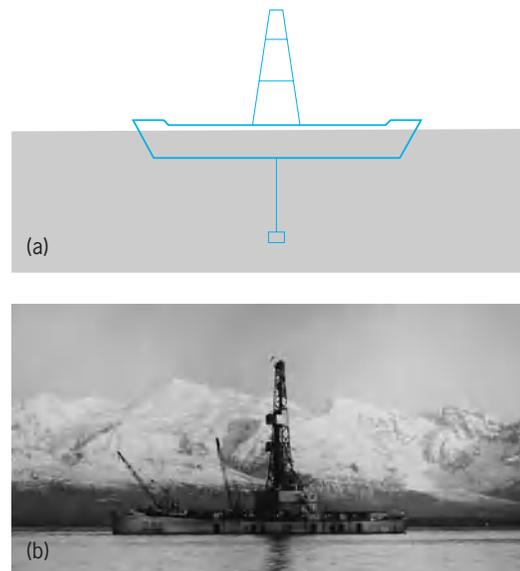


Fig. 4. Floating drill ship. Such ships can drill in depths from 60 to 1000 ft (18 to 300 m) or more. (a) Underwater design (World Petroleum). (b) Floating drill ship on a drilling site (Marathon Oil Co.).



Fig. 5. Offshore drilling and production platform.

which can operate in rough seas far from shore and take the large deck loads required for drilling wells in water depths up to 10,000 ft (3000 m).

The *Glomar Challenger* has drilled stratigraphic holes in water depths of 20,000 ft (6 km) to a depth of 3334 ft (1016 m) below the sea floor, but the technology for stratigraphic investigation is simpler and less demanding than that required for oil exploration and development. See OCEANOGRAPHIC VESSELS.

**Production and well-completion technology.** The move of exploration into the open and often hostile sea has required not only the development of drilling vessels but also a host of auxiliary equipment and techniques. An entire industrial complex has developed to serve the offshore industry, including construction of fixed platform structures from which the majority of the world's offshore oil and gas production is presently drilled and produced (Fig. 5). More than 7000 fixed platforms have

been fabricated and installed around the world. There are approximately 3900 fixed platforms in the waters of the United States outer continental shelf in the Gulf of Mexico, of which 1400 are larger structures with more than five well slots. Most of these are located in less than 350 ft (100 m) of water, and as recently as 1994 fewer than 20 were in waters deeper than 600 ft (180 m).

The industry has traditionally modified established onshore practices for offshore operations, and wherever possible until now has preferred surface wellheads on fixed platforms. In deep water the structural cost of the surface facility is increasingly significant, leading to large platforms with heavy topsides and many well slots. Directional drilling is necessary in order to reach all the well targets within the reservoir from a single platform location. More recently, technology has been developed which enables the last part of the wellbore to be drilled horizontally, further extending platform reach and the reserves which can be recovered from each location. See DRILLING, GEOTECHNICAL.

Since traditional fixed platforms directly resist the force of the wind and waves, they become massive and increasingly expensive in deeper water. Therefore, the deepest traditional fixed platform installed to date worldwide is in 1353 ft (412 m) of water in the Gulf of Mexico. Compliant towers rely on the damping effect of the ocean to partly resist the horizontal forces, and extend the economic limit of bottom-founded steel structures to around 3000 ft (900 m). The first two compliant production towers were installed in the Gulf of Mexico in 1998, the taller in a water depth of 1754 ft (535 m). As the top of the drilling rig is 2094 ft (638 m) above the sea bed, this is also the tallest free-standing structure in the world.

Driven by the high cost of fixed platforms, alternative technologies have been developed for marginal fields and deeper water. These include floating production systems, which rely on subsea completions where oil and gas is produced from "Christmas trees" located on the sea floor, and the tension-leg platform (Fig. 6), which is a buoyant structure fixed to the seabed by means of tensioned cables which permits surface trees. Such systems are now used to develop deep-water and marginal fields in all major offshore areas around the world.

Recent advances in seabed production and processing systems permit deep-water and marginal fields to be developed 50 mi (80 km) or more away from a fixed- or floating-platform host facility. If production must flow long distances, there is a potential for hydrates and waxes to plug the line; therefore, methods to clean the line and pressure boosting pumps may be required. Subsea control umbilical units are used to transmit hydraulic and electric power signals across the sea floor from the main process facility to the subsea well-control equipment. See HYDRATE.

In water depths out of the range of practical diving, remotely operated vehicles equipped with robotic arms and specialty tools for moving, torquing, winching, and applying hydraulic power must now



Fig. 6. Tension-leg platform. (U.S. Minerals Management Service)

do all the work that divers once managed. The depth range of remotely operated vehicles is virtually unlimited, whereas diver work ceases to be practical and cost-effective beyond 600–800 ft (180–250 m). Drilling and production units that are totally enclosed and sit on the sea bottom have been considered, but none has yet proved practical or cost-effective when compared to surface-based operations using remotely operated vehicles to perform the subsea intervention tasks. See DIVING; OIL AND GAS WELL COMPLETION; UNDERWATER VEHICLE.

A number of issues need to be resolved as drilling and production operations move into waters deeper than 5000 ft (1500 m) into an environment that is still comparatively unknown. For instance, the hydrostatic head of the water column is such that it makes good sense to drill from the sea floor rather than from the surface. If satellite fields are to be brought on-stream miles from their host platforms, then well-heads, pumps, and other process equipment are required to operate reliably on the sea bed for long periods of time and yet be retrievable for maintenance and repair. Hydraulic power has limitations over long distances and in deep water, so high-voltage electrical distribution and switching equipment will need to be developed for deep-water use.

**Hazards at sea.** The United States Gulf coast, where a large percentage of the world's offshore drilling has taken place, is regularly hit by hurricanes that damage structures in their path, and a number of platforms were lost early on due to high wind and waves. However, improved understanding of the environment, as well as more advanced materials and structural analysis techniques, has significantly reduced failures due to environmental forces. The major causes of accident are now due to human error, process hazards, and transportation to and from offshore facilities by helicopter. See HURRICANE.

Improved reservoir management and further discoveries have increased production from onshore fields, where costs are less than for offshore. However, despite the high cost involved in extracting oil and gas from beneath the sea, the world's increasing demand for petroleum energy continues to force the search for new reserves into even deeper waters and more remote corners of the world.

G. R. Schoonmaker; J. A. Turley; Cyril Arney

Bibliography. M. T. Halbouty (ed.), *Future Petroleum Provinces of the World*, 1986; M. T. Halbouty (ed.), *Giant Oil and Gas Fields of the Decade 1978-1988*, 1992; *Offshore Atlas of World Oil and Gas Theatres*, Pennwell Publishing, 1996; J. A. Pratt, T. Priest, and C. Costaneda, *Offshore Pioneers*, 1997.

## Oil and gas field exploitation

In the petroleum industry, a field is an area underlain without substantial interruption by one or more reservoirs of commercially valuable oil or gas, or both. A single reservoir (or group of reservoirs which cannot be separately produced) is a pool. Several

pools separated from one another by barren, impermeable rock may be superimposed one above another within the same field. Pools have variable areal extent. Any sufficiently deep well located within the field should produce from one or more pools. However, each well cannot produce from every pool, because different pools have different areal limits.

### Development

Development of a field includes the location, drilling, completion, and equipment of wells necessary to produce the commercially recoverable oil and gas in the field.

**Related oil field conditions.** Petroleum is a generic term which, in its broadest meaning, includes all naturally occurring hydrocarbons, whether gaseous, liquid, or solid. By variation of the temperature or pressure, or both, of any hydrocarbon, it becomes gaseous, liquid, or solid. Temperatures in producing horizons vary from approximately 60°F (16°C) to more than 300°F (149°C), depending chiefly upon the depth of the horizon. A rough approximation is that temperature in the reservoir sand, or pay, equals 60°F (16°C), plus 0.017°F/ft (0.031°C/m) of depth below surface. Pressure on the hydrocarbons varies from atmospheric to more than 11,000 lb/in.<sup>2</sup> (76 megapascals). Normal pressure is considered as 0.465 (lb/in.<sup>2</sup>)/ft (10.5 kilopascals/m) of depth. Temperatures and pressure vary widely from these average figures. Hydrocarbons, because of wide variations in pressure and temperature and because of mutual solubility in one another, do not necessarily exist underground in the same phases in which they appear at the surface.

Petroleum occurs underground in porous rocks of wide variety. The pore spaces range from microscopic size to rare holes 1 in. (2.5 cm) or more in diameter. The containing rock is commonly called the sand or the pay, regardless of whether the pay is actually sandstone, limestone, dolomite, unconsolidated sand, or fracture openings in relatively impermeable rock. See PETROLEUM.

**Development program.** After discovery of a field containing oil or gas, or both, in commercial quantities, the field must be explored to determine its vertical and horizontal limits and the mechanisms under which the field will produce. Development and exploitation of the field proceed simultaneously. Usually the original development program is repeatedly modified by geologic knowledge acquired during the early stages of development and exploitation of the field.

Ideally, tests should be drilled to the lowest possible producing horizon in order to determine the number of pools existing in the field. Testing and geologic analysis of the first wells sometimes indicates the producing mechanisms, and thus the best development program. Very early in the history of the field, step-out wells will be drilled to determine the areal extent of the pool or pools. Stepout wells give further information regarding the volumes of oil and gas available, the producing mechanisms, and the desirable spacing of wells.

The operator of an oil and gas field endeavors to select a development program which will produce the largest volume of oil and gas at a profit. The program adopted is always a compromise between conflicting objectives. The operator desires (1) to drill the fewest wells which will efficiently produce the recoverable oil and gas; (2) to drill, complete, and equip the wells at the lowest possible cost; (3) to complete production in the shortest practical time to reduce both capital and operating charges; (4) to operate the wells at the lowest possible cost; and (5) to recover the largest possible volume of oil and gas.

*Selecting the number of wells.* Oil pools are produced by four mechanisms: dissolved gas expansion, gas-cap drive, water drive, and gravity drainage. Commonly, two or more mechanisms operate in a single pool. The type of producing mechanism in each pool influences the decision as to the number of wells to be drilled. Theoretically, a single, perfectly located well in a water-drive pool is capable of producing all of the commercially recoverable oil and gas from that pool. Practically, more than one well is necessary if a pool of more than 80 acres (32 hectares) is to be depleted in a reasonable time. If a pool produces under either gas expansion or gas-cap drive, oil production from the pool will be independent of the number of wells up to a spacing of at least 80 acres per well (1860 ft or 570 m between wells). Gas wells often are spaced a mile or more apart. The operator accordingly selects the widest spacing permitted by field conditions and legal requirements.

*Major components of cost.* Costs of drilling, completing, and equipping the wells influence development plans. Having determined the number and depths of producing horizons and the producing mechanisms in each horizon, the operator must decide whether a well will be drilled at each location to each horizon or whether a single well can produce from two or more horizons at the same location. Clearly, the cost of drilling the field can be sharply reduced if a well can drain two, three, or more horizons. The cost of drilling a well will be higher if several horizons are simultaneously produced, because the dual or triple completion of a well usually requires larger casing. Further, completion and operating costs are higher. However, the increased cost of drilling a well of larger diameter and completing the well in two or more horizons is 20–40% less than the cost of drilling and completing two wells to produce separately from two horizons.

In some cases, the operator may reduce the number of wells by drilling a well to the lowest producible horizon and taking production from that level until the horizon there is commercially exhausted. The well is then plugged back to produce from a higher horizon. Selection of the plan for producing the various horizons obviously affects the cost of drilling and completing individual wells, as well as the number of wells which the operator will drill. If two wells are drilled at approximately the same location, they are referred to as twins, three wells at the same location are triplets, and so on.

*Costs and duration of production.* The operator wishes to produce as rapidly as possible because the net income from sale of hydrocarbons is obviously reduced as the life of the well is extended. The successful operator must recover from the productive wells the costs of drilling and operating those wells, and in addition must recover all costs involved in geological and geophysical exploration, leasing, scouting, and drilling of dry holes, and occasionally other operations. If profits from production are not sufficient to recover all exploration and production costs and yield a profit in excess of the rate of interest which the operator could secure from a different type of investment, he is discouraged from further exploration.

Most wells cannot operate at full capacity because unlimited production results in physical waste and sharp reduction in ultimate recovery. In many areas, conservation restrictions are enforced to make certain that the operator does not produce in excess of the maximum efficient rate. For example, if an oil well produces at its highest possible rate, a zone promptly develops around the well where production is occurring under gas-expansion drive, the most inefficient producing mechanism. Slower production may permit the petroleum to be produced under gas-cap drive or water drive, in which case ultimate production of oil will be two to four times as great as it would be under gas-expansion drive. Accordingly, the most rapid rate of production generally is not the most efficient rate.

Similarly, the initial exploration of the field may indicate that one or more gas-condensate pools exist, and recycling of gas may be necessary to secure maximum recovery of both condensate and of gas. The decision to recycle will affect the number of wells, the locations of the wells, and the completion methods adopted in the development program.

Further, as soon as the operator determines that secondary oil-recovery methods are desired and expects to inject water, gas, steam, or, rarely, air to provide additional energy to flush or displace oil from the pay, the number and location of wells may be modified to permit the most effective secondary recovery procedures.

**Legal and practical restrictions.** The preceding has assumed control of an entire field under single ownership by a single operator. In the United States, a single operator rarely controls a large field and this field is almost never under a single lease. Usually, the field is covered by separate leases owned and operated by different producers. The development program must be modified in consideration of the lease boundaries and the practices of the other operators.

Oil and gas know no lease boundaries. They move freely underground from areas of high pressure toward lower-pressure situations. The operator of a lease is obligated to locate the wells in such a way as to prevent drainage of the lease by wells on adjoining leases, even though the adjoining leases may be owned by that operator. In the absence of conservation restrictions, an operator must produce petroleum from wells as rapidly as it is produced from

wells on adjoining leases. Slow production on one lease results in migration of oil and gas to nearby leases which are more rapidly produced.

The operator's development program must provide for offset wells located as close to the boundary of the lease as are wells on adjoining leases. Further, the operator must equip the wells to produce as rapidly as the offset produces, and must produce from the same horizons which are being produced in offset wells. The lessor who sold the lease to the operator is entitled to a share of the recoverable petroleum underlying the land. Negligence by the operator in permitting drainage of a lease makes the operator liable to suit for damages or cancellation of the lease.

A development program acceptable to all operators in the field permits simultaneous development of leases, prevents drainage, and results in maximum ultimate production from the field. Difficulties may arise in agreement upon the best development program for a field. Most states have enacted statutes and have appointed regulatory bodies under which judicial determination can be made of the permissible spacing of the wells, the rates of production, and the application of secondary recovery methods.

*Drilling unit.* Commonly, small leases or portions of two or more leases are combined to form a drilling unit in whose center a well will be drilled. Unitization may be voluntary, by agreement between the operator or operators and the interested royalty owners, with provision for sharing production from the well between the parties in proportion to their acreage interests. In many states the regulatory body has authority to require unitization of drilling units, which eliminates unnecessary offset wells and protects the interests of a landowner whose acreage holding may be too small to justify the drilling of a single well on that property alone.

*Pool unitization.* When recycling or some types of secondary recovery are planned, further unitization is adopted. Since oil and gas move freely across lease boundaries, it would be wasteful for an operator to repressure, recycle, or water-drive a lease if the adjoining leases were not similarly operated. Usually an entire pool must be unitized for efficient recycling, or secondary recovery operations. Pool unitization may be accomplished by agreement between operators and royalty owners. In many cases, difference of opinion or ignorance on the part of some parties prevents voluntary pool unitization. Many states authorize the regulatory body to unitize a pool compulsorily on application by a specified percentage of interests of operators and royalty owners. Such compulsory unitization is planned to provide each operator and each royalty owner his fair share of the petroleum products produced from the field regardless of the location of the well or wells through which these products actually reach the surface.

### Exploitation Practices

Oil and gas production necessarily are intimately related, since approximately one-third of the gross gas production in the United States is produced

from wells that are classified as oil wells. However, the naturally occurring hydrocarbons of petroleum are not only liquid and gaseous but may even be found in a solid state, such as asphaltite and some asphalts.

Where gas is produced without oil, the production problems are simplified because the product flows naturally throughout the life of the well and does not have to be lifted to the surface. However, there are sometimes problems of water accumulations in gas wells, and it is necessary to pump the water from the wells to maintain maximum, or economical, gas production. The line of demarcation between oil wells and gas wells is not definitely established since oil wells may have gas-oil ratios ranging from a few cubic feet (1 cubic foot =  $2.8 \times 10^{-2} \text{ m}^3$ ) per barrel to many thousand cubic feet of gas per barrel of oil. Most gas wells produce quantities of condensable vapors, such as propane and butane, that may be liquefied and marketed for fuel, and the more stable liquids produced with gas can be utilized as natural gasoline.

**Factors of method selection.** The method selected for recovering oil from a producing formation depends on many factors, including well depth, well-casing size, oil viscosity, density, water production, gas-oil ratio, porosity and permeability of the producing formation, formation pressure, water content of producing formation, and whether the force driving the oil into the well from the formation is primarily gas pressure, water pressure, or a combination of the two. Other factors, such as paraffin content and difficulty expected from paraffin deposits, sand production, and corrosivity of the well fluids, also have a decided influence on the most economical method of production.

Special techniques utilized to increase productivity of oil and gas wells include acidizing, hydraulic fracturing of the formation, the setting of screens, and gravel packing or sand packing to increase permeability around the well bore.

**Aspects of production rate.** Productive rates per well may vary from a fraction of a barrel [1 barrel (bbl) =  $0.1590 \text{ m}^3$ ] per day to several thousand barrels per day, and it may be necessary to produce a large percentage of water along with the oil.

*Field and reservoir conditions.* In some cases reservoir conditions are such that some of the wells flow naturally throughout the entire economical life of the oil field. However, in the great majority of cases it is necessary to resort to artificial lifting methods at some time during the life of the field, and often it is necessary to apply artificial lifting means immediately after the well is drilled.

*Market and regulatory factors.* In some oil-producing states of the United States there are state bodies authorized to regulate oil production from the various oil fields. The allowable production per well is based on various factors, including the market for the particular type of oil available, but very often the allowable production is based on an engineering study of the reservoir to determine the optimum rate of production.

**Useful terminology.** A few definitions of terms used in petroleum production technology are listed below to assist in an understanding of some of the problems involved.

**Porosity.** The percentage porosity is defined as the percentage volume of voids per unit total volume. This, of course, represents the total possible volume available for accumulation of fluids in a formation, but only a fraction of this volume may be effective for practical purposes because of possible discontinuities between the individual pores. The smallest pores generally contain water held by capillary forces.

**Permeability.** Permeability is a measure of the resistance to flow through a porous medium under the influence of a pressure gradient. The unit of permeability commonly employed in petroleum production technology is the darcy. A porous structure has a permeability of 1 darcy if, for a fluid of 1 centipoise ( $10^{-3}$  Pa · s) viscosity, the volume flow is  $1 \text{ cm}^3/(\text{s} \cdot \text{cm}^2)$  [ $10^{-2} \text{ m}^3/(\text{s} \cdot \text{m}^2)$ ] under a pressure gradient of 1 atm/cm (10 MPa/m).

**Productivity index.** The productivity index is a measure of the capacity of the reservoir to deliver oil to the well bore through the productive formation and any other obstacles that may exist around the well bore. In petroleum production technology, the productivity index is defined as production in barrels per day per pound per square inch drop in bottom-hole pressure. For example, if a well is closed in at the casinghead, the bottom-hole pressure will equal the formation pressure when equilibrium conditions are established. However, if fluid is removed from the well, either by flowing or pumping, the bottom-hole pressure will drop as a result of the resistance to flow of fluid into the well from the formation to replace the fluid removed from the well. If the closed-in bottom-hole pressure should be  $1000 \text{ lb/in.}^2$  (7 MPa), for example, and if this pressure should drop to  $900 \text{ lb/in.}^2$  (6 MPa) when producing at a rate of 100 bbl/day (a drop of  $100 \text{ lb/in.}^2$  or 0.7 MPa), the well would have a productivity index of one.

**Barrel.** The standard barrel used in the petroleum industry is 42 U.S. gallons (approx.  $0.1590 \text{ m}^3$ ).

**API gravity.** The American Petroleum Institute (API) scale that is in common use for indicating specific gravity, or a rough indication of quality of crude petroleum oils, differs slightly from the Baumé scale commonly used for other liquids lighter than water. The **table** shows the relationship between de-

grees API and specific gravity referred to water at  $60^\circ\text{F}$  ( $16^\circ\text{C}$ ) for specific gravities ranging from 0.60 to 1.0.

**Viscosity range.** Viscosity of crude oils currently produced varies from approximately 1 centipoise ( $10^{-3}$  Pa · s) to values above 1000 cP (1 Pa · s) at temperatures existing at the bottom of the well. In some areas it is necessary to supply heat artificially down the wells or circulate lighter oils to mix with the produced fluid for maintenance of a relatively low viscosity throughout the temperature range to which the product is subjected.

In addition to wells that are classified as gas wells or oil wells, the term gas-condensate well has come into general use to designate a well that produces large volumes of gas with appreciable quantities of light, volatile hydrocarbon fluids. Some of these fluids are liquid at atmospheric pressure and temperature; others, such as propane and butane, are readily condensed under relatively low pressures in gas separators for use as liquid petroleum gas fuels or for other uses. The liquid components of the production from gas-condensate wells generally arrive at the surface in the form of small droplets entrained in the high-velocity gas stream and are separated from the gas in a high-pressure gas separator.

### Production Methods in Producing Wells

The common methods of producing oil wells are natural flow; pumping with sucker rods; gas lift; hydraulic subsurface pumps; electrically driven centrifugal well pumps; and swabbing.

Numerous other methods, including jet pumps and sonic pumps, have been tried and are used to slight extent. The sonic pump is a development in which the tubing is vibrated longitudinally by a mechanism at the surface and acts as a high-speed pump with an extremely short stroke.

A discussion of production methods, in approximate order of relative importance, follows.

**Natural flow.** Natural flow is the most economical method of production and generally is utilized as long as the desired production rate can be maintained by this method (**Fig. 1**). It utilizes the formation energy, which may consist of gas in solution in the oil in the formation; free gas under pressure acting against the liquid and gas-liquid phase to force it toward the well bore; water pressure acting against the oil; or a combination of these three energy sources. In some areas the casinghead pressure

Degrees API corresponding to specific gravities of crude oil at  $60^\circ\text{F}$  ( $16^\circ\text{C}$ )

Specific gravity (tenths)	Specific gravity (hundredths)*									
	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.60	104.33	100.47	96.73	93.10	89.59	86.19	82.89	79.69	79.59	73.57
0.70	70.64	67.80	65.03	62.34	59.72	57.17	54.68	52.27	49.91	47.61
0.80	45.38	43.19	44.06	38.98	36.95	34.97	33.03	31.14	29.30	27.49
0.90	25.72	23.99	22.30	20.65	19.03	17.45	15.90	14.38	12.89	11.43
1.00	10.00									

\*Add to specific gravity (tenths) to read comparable value in degrees API.

may be of the order of  $10,000 \text{ lb/in.}^2$  (70 MPa), so it is necessary to provide fittings adequate to withstand such pressures. Adjustable throttle valves, or chokes, are utilized to regulate the flow rate to a desired and safe value. With such a high-pressure drop across a throttle valve the life of the valve is likely to be very short. Several such valves are arranged in parallel in the tubing head “Christmas tree” with positive shutoff valves between the chokes and the tubing head so that the wearing parts of the throttle valve, or the entire valve, can be replaced while flow continues through another similar valve.

An additional safeguard that is often used in connection with high-pressure flowing wells is a bottom-hole choke or a bottom-hole flow control valve that limits the rate of flow to a reasonable value, or stops it completely, in case of failure of surface controls. The packer, while not essential, is often used to reduce the free gas volume in the casing (Fig. 1).

Flow rates for United States wells seldom exceed a few hundred barrels per day because of enforced or voluntary restrictions to regulate production rates and to obtain most efficient and economical ultimate recovery. However, in some countries, especially in the Middle East, it is not uncommon for natural flow rates to exceed  $10,000 \text{ bbl}/(\text{day})(\text{well})$  [ $1590 \text{ m}^3/(\text{day})(\text{well})$ ].

**Lifting.** Most wells are not self-flowing. The common types of lifting are outlined here.

*Pumping with sucker rods.* Approximately 90% of the wells made to produce by some artificial lift method in the United States are equipped with sucker-rod-type pumps. In these the pump is installed at the lower end of the tubing string and is actuated by a string of sucker rods extending from the surface to the subsurface pump. The sucker rods are attached to a polished rod at the surface. The polished rod extends through a stuffing box and is attached to the pumping unit, which produces the necessary reciprocating motion to actuate the sucker rods and the subsurface pump (Fig. 2). The two common variations are mechanical and hydraulic long-stroke pumping.

*Mechanical pumping.* The great majority of pumping units are of the mechanical type, consisting of a suitable reduction gear, and crank and pitman arrangement to drive a walking beam to produce the necessary reciprocating motion. A counterbalance is provided to equalize the load on the upstroke and downstroke (Fig. 3). Mechanical pumping units of this type vary in load-carrying capacity from about 2000 to about 43,000 lb (900 to 19,500 kg), and the torque rating of the low-speed gear which drives the crank ranges from 6400 in.-lb (720 N · m) in the smallest API standard unit to about 1,500,000 in.-lb (170,000 N · m) for the largest units now in use. Stroke length varies from about 18 to 192 in. (46 to 488 cm). Usual operating speeds are from about 6 to 20 strokes/min. However, both lower and higher rates of speed are sometimes used.

Production rates with sucker-rod-type pumps vary from a fraction of 1 bpd in some areas, with part-time pumping, to approximately 3000 bpd

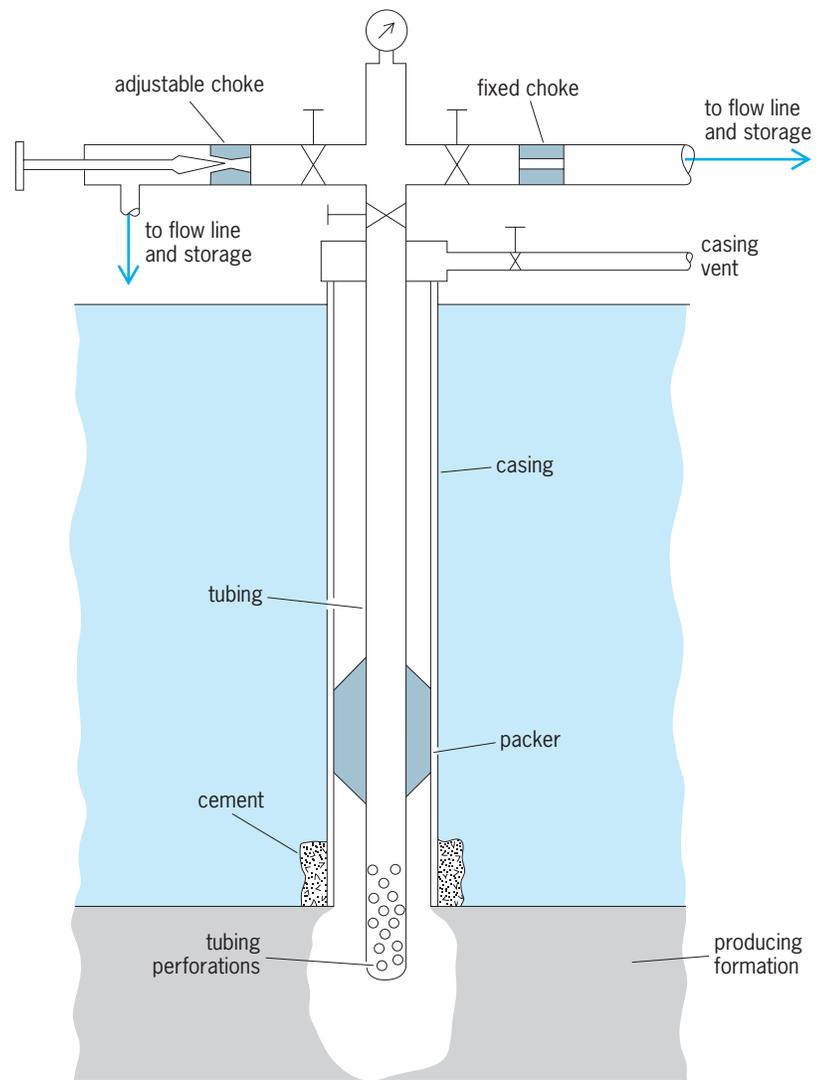


Fig. 1. Schematic view of well equipped for producing by natural flow.

( $480 \text{ m}^3/\text{day}$ ) for the largest installations in relatively shallow wells.

*Hydraulic long-stroke pumping.* For this the units consist of a hydraulic lifting cylinder mounted directly over the well head and are designed to produce stroke lengths of as much as 30 ft (9 m). Such long-stroke hydraulic units are usually equipped with a pneumatic counterbalance arrangement which equalizes the power requirement on the upstroke and downstroke.

Hydraulic pumping units also are made without any provision for counterbalance. However, these units are generally limited to relatively small wells, and they are relatively inefficient.

*Gas lift.* Gas lift in its simplest form consists of initiating or stimulating well flow by injecting gas at some point below the fluid level in the well. With large-volume gas-lift operations the well may be produced through either the casing or the tubing. In the former case, gas is conducted through the tubing to the point of injection; in the latter, gas may be conducted to the point of injection through the casing or

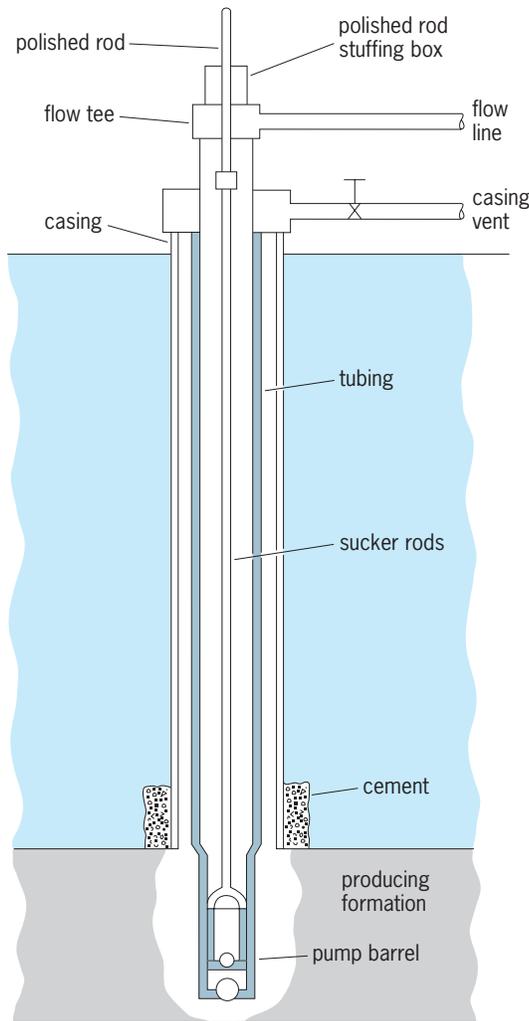


Fig. 2. Schematic view of a well which is equipped for pumping with sucker rods.

through an auxiliary string of tubing. When gas is injected into the oil column, the weight of the column above the point of injection is reduced as a result of the space occupied by the relatively low-density gas. This lightening of the fluid column is sufficient to permit the formation pressure to initiate flow up the tubing to the surface. Gas injection is often utilized to increase the flow from wells that will flow naturally but will not produce the desired amount by natural flow.

There are many factors determining the advisability of adopting gas lift as a means of production. One of the more important factors is the availability of an adequate supply of gas at suitable pressure and reasonable cost. In a majority of cases gas lift cannot be used economically to produce a reservoir to depletion because the well may be relatively productive with a low back pressure maintained on the formation but will produce very little, if anything, with the back pressure required for gas-lift operation. Therefore, it generally is necessary to resort to some mechanical means of pumping before the well is abandoned, and it may be more economical to

adopt the mechanical means initially than to install the gas-lift system while conditions are favorable and later replace it.

This discussion of gas lift has dealt primarily with the simple injection of gas, which may be continuous or intermittent. There are numerous modifications of gas-lift installations, including various designs for flow valves which may be installed in the tubing string to open and admit gas to the tubing from the casing at a predetermined pressure differential between the tubing and casing. When the valve opens, gas is injected into the tubing to initiate and maintain flow until the tubing pressure drops to a predetermined value; and the valve closes before the input gas-oil ratio becomes excessive. This represents an intermittent-flow-type valve. Other types are designed to maintain continuous flow, proper pressure differential, and proper gas injection rate for efficient operation. In some cases several such flow valves are spaced up the tubing string to permit flow to be initiated from various levels as required.

Other modifications of gas lift involve the utilization of displacement chambers. These are installed on the lower end of the well tubing where oil may accumulate, and the oil is displaced up the tubing with gas injection controlled by automatic or mechanical valves.

*Hydraulic subsurface pumps.* The hydraulic subsurface pump has come into fairly prominent use. The subsurface pump is operated by means of a hydraulic reciprocating motor attached to the pump and installed in the well as a single unit. The hydraulic motor is driven by a supply of hydraulic fluid under pressure that is circulated down a string of tubing and through the motor. Generally the hydraulic fluid consists of crude oil which is discharged into the return line and returns to the surface along with the produced crude oil.

Hydraulically operated subsurface pumps are also arranged for separating the hydraulic power fluid from the produced well fluid. This arrangement is



Fig. 3. Pumping unit with adjustable rotary counterbalance. (Oil Well Supply Division, U.S. Steel Corp.)

especially desirable where the fluid being produced is corrosive or is contaminated with considerable quantities of sand or other solids that are difficult to separate to condition the fluid for use as satisfactory power oil. This method permits use of water or other nonflammable liquids as hydraulic power fluid to minimize fire hazard in case of a failure of the hydraulic power line at the surface.

**Centrifugal well pumps.** Electrically driven centrifugal pumps have been used to some extent, especially in large-volume wells of shallow or moderate depths. Both the pump and the motor are restricted in diameter to run down the well casing, leaving sufficient clearance for the flow of fluid around the pump housing. With the restricted diameter of the impellers the discharge head necessary for pumping a relatively deep well can be obtained only by using a large number of stages and operating at a relatively high speed. The usual rotating speed for such units is 3600 rpm, and it is not uncommon for such units to have 50 or more pump stages. The direct-connected electric motor must be provided with a suitable seal to prevent well fluid from entering the motor housing, and electrical leads must be run down the well casing to supply power to the motor.

**Swabs.** Swabs have been used for lifting oil almost since the beginning of the petroleum industry. They usually consist of a steel tubular body equipped with a check valve which permits oil to flow through the tube as it is lowered down the well with a wire line. The exterior of the steel body is generally fitted with flexible cup-type soft packing that will fall freely but will expand and form a seal with the tubing when pulled upward with a head of fluid above the swab. Swabs are run into the well on a wire line to a point considerably below the fluid level and then lifted back to the surface to deliver the volume of oil above the swab. They are often used for determining the productivity of a well that will not flow naturally and for assisting in cleaning paraffin from well tubing. In some cases swabs are used to stimulate wells to flow by lifting, from the upper portion of the tubing, the relatively dead oil from which most of the gas has separated.

**Bailers.** Bailers are used to remove fluids from wells and for cleaning out solid material. They are run into the wells on wire lines as in swabbing, but differ from swabs in that they generally are run only in the casing when there is no tubing in the well. The capacity of the bailer itself represents the volume of fluid lifting each time since the bailer does not form a seal with the casing. The bailer is simply a tubular vessel with a check valve in the bottom. This check valve generally is arranged so that it is forced open when the bailer touches bottom in order to assist in picking up solid material for cleaning out a well.

**Jet pumps.** A jet pump for use in oil wells operates on exactly the same principle as a water-well jet pump. Advantage is taken of the Bernoulli effect to reduce pressure by means of a high-velocity fluid jet. Thus oil is entrained from the well with this high-velocity jet in a venturi tube to accelerate the fluid and assist in lifting it to the surface, along with any



Fig. 4. Lease tank battery with four tanks and two gas separators. (Gulf Oil Corp.)

assistance from the formation pressure. The application of jet pumps to oil wells has been insignificant.

**Sonic pumps.** Sonic pumps essentially consist of a string of tubing equipped with a check valve at each joint and mechanical means on the surface to vibrate the tubing string longitudinally. This creates a harmonic condition that will result in several hundred strokes per minute, with the strokes being a small fraction of 1 in. (2.5 cm) in length. Some of these pumps are being used in relatively shallow wells.

**Lease tanks and gas separators.** A typical lease tank battery consists of four 1000-bbl (159-m<sup>3</sup>) tanks and two gas separators (Fig. 4). Such equipment is used for handling production from wells produced by natural flow, gas lift, or pumping. In some pumping wells the gas content may be too low to justify the cost of separators for saving the gas.

**Natural gasoline production.** An important phase of oil and gas production in many areas is the production of natural gasoline from gas taken from the casinghead of oil wells or separated from the oil and conducted to the natural gasoline plant. The plant consists of facilities for compressing and extracting the liquid components from the gas (Fig. 5). The natural gasoline generally is collected by cooling and



Fig. 5. Modern natural gasoline plant in western Texas. (Gulf Oil Corp.)

condensing the vapors after compression or by absorbing in organic liquids having high boiling points from which the volatile liquids are distilled. Many natural gasoline plants utilize a combination of condensing and absorbing techniques.

#### Production Problems and Instruments

To maintain production, various problems must be overcome. Numerous instruments have been developed to monitor production and to control production problems.

**Corrosion.** In many areas the corrosion of production equipment is a major factor in the cost of petroleum production.

For practical consideration, corrosion in oil and gas-well production can be classified into four main types.

1. Sweet corrosion occurs as a result of the presence of carbon dioxide and fatty acids. Oxygen and hydrogen sulfide are not present. This type of corrosion occurs in both gas-condensate and oil wells. It is most frequently encountered in the United States in southern Louisiana and Texas, and other scattered areas. At least 20% of all sweet oil production and 45% of condensate production are considered corrosive.

2. Sour corrosion is designated as corrosion in oil and gas wells producing even trace quantities of hydrogen sulfide. These wells may also contain oxygen, carbon dioxide, or organic acids. Sour corrosion occurs in the United States primarily throughout Arbuckle production in Kansas and in the Permian basin of western Texas and New Mexico. About 12% of all sour production is considered corrosive.

3. Oxygen corrosion occurs wherever equipment is exposed to atmospheric oxygen. It occurs most frequently in offshore installations, brine-handling and injection systems, and in shallow producing wells where air is allowed to enter the casing.

4. Electrochemical corrosion is designated as that which occurs when corrosion currents can be readily measured or when corrosion can be mitigated by the application of current, as in soil corrosion.

Corrosion inhibitors are used extensively in both oil and gas wells to reduce corrosion damage to subsurface equipment. Most of the inhibitors used in the oil field are of the so-called polar organic type. All of the major inhibitor suppliers can furnish effective inhibitors for the prevention of sweet corrosion as encountered in most fields. These can be purchased in oil-soluble, water-dispersible, or water-soluble form. See CORROSION.

**Paraffin deposits.** In many crude-oil-producing areas paraffin deposits in tubing and flow lines and on sucker rods are a source of considerable trouble and expense. Such deposits build up until the tubing or flow line is partially or completely plugged. It is necessary to remove these deposits to maintain production rates. A variety of methods are used to remove paraffin from the tubing, including the application of heated oil through tubular sucker rods to mix with and transfer heat to the oil being produced and raise the temperature to a point at which the deposited paraffin will be dissolved or melted. Paraffin solvents may also be applied in this manner without the necessity of applying heat.

Mechanical means often are used in which a scraping tool is run on a wire line and paraffin is scraped from the tubing wall as the tool is pulled back to the surface. Mechanical scrapers that attach to sucker rods also are in use. Various types of automatic scrapers have been used in connection with flowing wells. These consist of a form of piston that will drop freely to the bottom when flow is stopped but will rise back to the surface when flow is resumed. Electrical heating methods have been used rather extensively in some areas. The tubing is insulated from the casing and from the flow line, and electric current is transmitted through the tubing for the time necessary to heat the tubing sufficiently to cause the paraffin deposits to melt or go into solution in the oil in the tubing. Plastic coatings have been utilized inside tubing and flow lines to minimize or prevent paraffin deposits. Paraffin does not deposit readily on certain plastic coatings.

A common method for removing paraffin from flow lines is to disconnect the line at the wellhead and at the tank battery and force live steam through the line to melt the paraffin deposits and flow them out. Various designs of flow-line scrapers have also been used rather extensively and fairly successfully. Paraffin deposits in flow lines are minimized by insulating the lines or by burying the lines to maintain a higher average temperature.

**Emulsions.** A large percentage of oil wells produce various quantities of salt water along with the oil, and numerous wells are being pumped in which the salt-water production is 90% or more of the total fluid lifted. Turbulence resulting from production methods results in the formation of emulsions of water in oil or oil in water; the commoner type is oil in water. Emulsions are treated with a variety of demulsifying chemicals, with the application of heat, and with a combination of these two treatments. Another method for breaking emulsions is the electrostatic or electrical precipitator type of emulsion treatment.



Fig. 6. Two pumping wells with tank battery. (Oil Well Supply Division, U.S. Steel Corp.)



Fig. 7. Numerous offshore wells located in Lake Maracaibo, Venezuela. (Creole Petroleum Corp.)

In this method the emulsion to be broken is circulated between electrodes subjected to a high potential difference. The resulting concentrated electric field tends to rupture the oil-water interface and thus breaks the emulsion and permits the water to settle out. A typical tank battery is equipped with a wash tank, or gun barrel, and a gas-fired heater for emulsion treating and water separation before the oil is admitted to the lease tanks (Fig. 6). See OIL AND GAS STORAGE.

**Gas conservation.** If the quantity of gas produced with crude oil is appreciably greater than that which can be efficiently utilized or marketed, it is necessary to provide facilities for returning the excess gas to the producing formation. Formerly, large quantities of excess gas were disposed of by burning or simply by venting to the atmosphere. This practice is now unlawful. Returning excess gas to the formation not only conserves the gas for future use but also results in greater ultimate recovery of oil from the formation.

**Salt-water disposal.** The large volumes of salt water produced with the oil in some areas present serious disposal problems. The salt water is generally pumped back to the formation through wells drilled for this purpose. Such salt-water disposal wells are located in areas where the formation already contains water. Thus this practice helps to maintain the formation pressure as well as the productivity of the producing wells.

**Offshore production.** Offshore wells present additional production problems since the wells must be serviced from barges or boats. Wells of reasonable depth on land locations are seldom equipped with derricks for servicing because it is more economical

to set up a portable mast for pulling and installing rods, tubing, and other equipment. However, the use of portable masts is not practical on offshore locations, and a derrick is generally left standing over such wells throughout their productive life to facilitate servicing. There are a considerable number of offshore wells along the Gulf Coast and the Pacific Coast of the United States, but by far the greatest number of offshore wells in a particular region is in Lake Maracaibo in Venezuela. Figure 7 shows a considerable number of derricks in Lake Maracaibo with pumping wells in the foreground. These wells are pumped by electric power through cables laid on the lake bottom to conduct electricity from power-generating stations onshore. An overwater tank battery is visible at the extreme right. All offshore installations, such as tank batteries, pump stations, and the derricks and pumping equipment, are supported on pilings in water up to 100 ft (30 m) or more in depth. There are approximately 2300 oil derricks in Lake Maracaibo. Semipermanent platform rigs and even bottom storage facilities are used in Gulf of Mexico waters at depths of more than 100 ft (30 m). See OIL AND GAS, OFFSHORE.

**Instruments.** The commoner and more important instruments required in petroleum production operations are included in the following discussion.

Gas meters, which are generally of the orifice type, are designed to record the differential pressure across the orifice, and the static pressure.

Recording subsurface pressure gages small enough to run down 2-in. (5-cm) ID (inside diameter) tubing are used extensively for measuring pressure gradients down the tubing of flowing wells, recording pressure buildup when the well is

closed in, and measuring equilibrium bottom-hole pressures.

Subsurface samplers designed to sample well fluids at various levels in the tubing are used to determine physical properties, such as viscosity, gas content, free gas, and dissolved gas at various levels. These instruments may also include a recording thermometer or a maximum reading thermometer, depending upon the information required.

Oil meters of various types are utilized to meter crude oil flowing to or from storage.

Dynamometers are used to measure polished-rod loads. These instruments are sometimes known as well weighers since they are used to record the polished-rod load throughout a pumping cycle of a sucker-rod-type pump. They are used to determine maximum load on polished rods as well as load variations, to permit accurate counterbalancing of pumping wells, and to assure that pumping units or sucker-rod strings are not seriously overloaded.

Liquid-level gages and controllers are used. They are similar to those used in other industries, but with special designs for closed lease tanks.

A wide variety of scientific instruments find application in petroleum production problems. The above outline gives an indication of a few specialized instruments used in this branch of the industry, and there are many more. Special instruments developed by service companies are valued for a wide variety of purposes and include calipers to detect and measure corrosion pits inside tubing and casing and magnetic instruments to detect microscopic cracks in sucker rods. See PETROLEUM RESERVOIR ENGINEERING.

Roy L. Chenault

**Bibliography.** R. Baker, *A Primer of Offshore Operations*, 3d ed., 1998; W. D. Berger and K. E. Anderson, *Modern Petroleum: A Basic Primer of the Industry*, 3d ed., 1992; R. A. Dawe and A. G. Lucas (eds.), *Modern Petroleum Technology, vols. 1 and 2*, 6th ed., 2000; M. J. Economides, A. D. Hill, and C. Ehlig-Economides, *Petroleum Production Systems*, 1993; N. J. Hyne, *Nontechnical Guide to Petroleum Geology; Exploration, Drilling and Production*, 2d ed., 2001; M. A. Mian, *Petroleum Engineering Handbook for the Practicing Engineer*, vol. 1, 1992; T. E. W. Nind, *Hydrocarbon Reservoir and Well Performance*, 1989.

## Oil and gas storage

Storage, usually in great quantities, of crude oil and natural gas after production from natural reservoirs. Large amounts of refined products are stored as well. Storage is necessary to meet seasonal and other fluctuations in demand; for efficient operation of producing equipment, pipelines, tankers, and refineries; and for emergency use.

**Crude oil and refined products.** Oil from producing wells is first collected in welded-steel, bolted-steel, or wooden tanks or 100 bbl (16 m<sup>3</sup>) or greater capacity located on individual leases. These tanks,

upright cylinders with low-pitched conical roofs, provide temporary storage while the oil is awaiting shipment. Several tanks grouped together are a tank battery. Assemblages of large steel tanks, known as tank farms, are used for more permanent storage at pipeline pump stations, points where tankers load and unload, and refineries.

The trend toward giant tankers, accelerated by the closing of the Suez Canal in 1967, created a need for large storage facilities at both the loading and unloading ends of the tanker runs. Large-capacity excavated reservoirs with concrete linings have been used for many years in California to store both crude and fuel oil. One such reservoir with a fixed roof and elliptical in form is 780 ft (238 m) long, 467 ft (142 m) wide, and 23 ft (7.9 m) deep. It covers 9<sup>1</sup>/<sub>4</sub> acres (3.74 ha) and provides storage for more than 1 × 10<sup>6</sup> bbl (1.6 × 10<sup>5</sup> m<sup>3</sup>). Another reservoir has a capacity of 4 × 10<sup>6</sup> bbl (6 × 10<sup>5</sup> m<sup>3</sup>) and covers 16 acres (6.5 ha).

**Offshore storage.** For offshore producing fields a number of unique storage systems have been designed. In several instances old tankers have been adapted for storage, and barges have been constructed especially for offshore storage use. One underwater installation consists essentially of three giant inverted steel funnels. Each unit is 270 ft (82 m) in diameter and 205 ft (62 m) high, weighs 28 × 10<sup>6</sup> lb (13 × 10<sup>6</sup> kg), and has a capacity of 0.5 × 10<sup>6</sup> bbl (8 × 10<sup>4</sup> m<sup>3</sup>). The bottom is open, and the unit is anchored to the sea floor by 95-ft (29-m) pilings. A reinforced concrete installation features a nine-module storage unit with 1 × 10<sup>6</sup> bbl capacity surrounded by a perforated wall 302 ft (92 m) in diameter and serves as a breakwater. The outer wall is about 270 ft (82 m) high and extends about 40 ft (12 m) above the water surface. A submerged floating storage tank 96 ft (29 m) in diameter and 305 ft (93 m) high is held in place by six anchor lines and has a capacity of 300,000 bbl (50,000 m<sup>3</sup>). One relatively small unit consists of a platform with four vertical legs, each holding 4700 bbl (750 m<sup>3</sup>), and four horizontal tanks at the bottom holding 1850 bbl (294 m<sup>3</sup>) each, for a total capacity of 26,200 bbl (4170 m<sup>3</sup>). A second small unit, utilizing bottom tanks as an anchor, holds 2400 bbl (380 m<sup>3</sup>) underwater and 600 bbl (90 m<sup>3</sup>) in a spherical tank above the surface. A third small unit consists of a sea-floor base, connected by a universal joint to a large-diameter vertical cylinder, about 350 ft (110 m) high, which extends above the water surface. One design includes an excavated cavern beneath the sea floor; nuclear cavities have also been suggested.

**Volatility problems.** To minimize vaporization losses, lease tanks are sometimes equipped to hold several ounces pressure. At large-capacity storage sites, special tanks are generally used. Tanks with lifter or floating roofs are used to store crude oil, motor gasoline, and less volatile natural gasoline. Motor and natural gasolines are also stored in spheroid containers. Spherical containers are used for more volatile liquids, such as butane. Horizontal cylindrical containers are used for propane and butane storage.

Refrigerated insulated tank systems enabling propane to be stored at a lower pressure are also in use.

**Underground storage.** Large quantities of volatile liquid-petroleum products, including propane and butane, are stored in underground caverns dissolved in salt formations and in mined caverns, gas reservoirs, and water sands. Liquid-petroleum products are also stored underground in Belgium, France, Germany, and France. In Pennsylvania an abandoned quarry with a capacity of  $2 \times 10^6$  bbl ( $0.3 \times 10^6$  m<sup>3</sup>) was equipped with a floating roof for storing fuel oil. Refrigerated propane is also being stored in excavations in frozen earth and in underground concrete tanks.

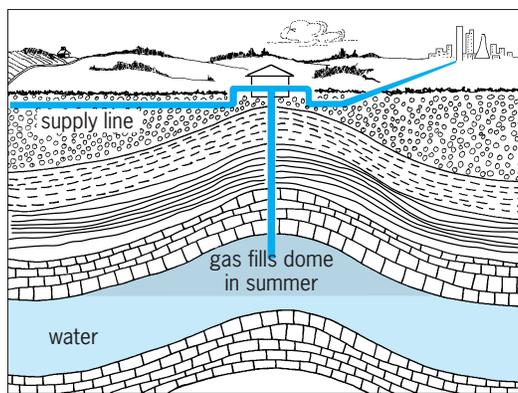
**Natural gas.** Natural gas is stored in low-pressure surface holders, buried high-pressure pipe batteries and bottles, depleted or partially depleted oil and gas reservoirs, water sands, and several types of containers at extremely low temperature ( $-258^\circ\text{F}$  or  $-161^\circ\text{C}$ ) after liquefaction.

Low-pressure holders, which store relatively small volumes of gas, basically use either a water or a dry seal, and variations of each type exist. With the displacement of manufactured gas by natural gas in the United States, the need for surface holders greatly diminished and they have disappeared almost entirely.

**Underground storage.** In the United States gas pipeline and utility companies store large quantities of natural gas in underground reservoirs. In most cases these reservoirs are located near market areas and are used to supplement pipeline supplies during the winter months when the gas demand for residential heating is very high. Since gas can be stored in the summer when the gas demand is low, underground storage permits greatly increased pipeline utilization, resulting in lower transportation costs and reduced gas cost to the consumer. Underground storage is the only economical method of storing large enough quantities of gas to meet the seasonal fluctuations in pipeline loads, and has enabled gas companies to meet market requirements which otherwise could not be satisfied.

Gas was first stored underground in 1915 in a partially depleted gas field in Ontario, Canada. The following year gas was injected into a depleted gas field near Buffalo, New York. Gas reservoirs, which utilize depleted gas and oil fields, water sands, salt caverns, and an abandoned coal mine, are located in 26 states. Gas is also being stored in underground reservoirs in Canada, France, Germany, Austria, Italy, Poland, Romania, Czechoslovakia, and Russia. In the United Kingdom a salt cavern is being used for gas storage.

Gas storage in water sands was first undertaken in 1952, and this method of storage steadily increased, especially in areas lacking gas or oil fields. A cross section of a typical aquifer storage field is shown in the **illustration**. A geologic trap having adequate structural closure and a suitable caprock is needed. The storage sand must be porous and thick enough and under sufficient hydrostatic pressure to hold large quantities of gas. The sand must also be sufficiently permeable and continuous over a wide enough area



Schematic cross section of typical aquifer gas storage field showing injected gas displacing water. (Natural Gas Pipeline Company of America)

so that water can be pushed back readily to make room for the stored gas. In some cases, water removal wells are also utilized.

**Reservoir pressure.** In operating storage reservoirs only a portion of the stored gas, called working gas, is normally withdrawn. The remaining gas, called cushion gas, stays in the reservoir to provide the necessary pressure to produce the storage wells at desired rates. In aquifer storages some water returns to help maintain the reservoir pressure. The percentage of cushion gas varies considerably among reservoirs. In some instances, the original reservoir pressure of the oil and gas field is exceeded in storage operations. This results in storage volumes greater than the original content and increases well deliverabilities. In aquifer storages the original hydrostatic pressure must be exceeded in order to push the water back.

In Russia, aquifer gas storage has been undertaken in one area where no appreciable structural closure exists. A company in the United States also has experimented with storing without structure. In an inconclusive field test, air was injected into a center well with control of the lateral spread of the air bubble attempted by injecting water into surrounding wells. Storage of gas in cavities created by nuclear explosions has been proposed and seriously considered.

**Liquefied gas.** Storage of liquefied natural gas throughout the world is in connection with shipment of liquefied natural gas by tanker, and is located at the loading and unloading ends of the tanker runs as well as at peak sharing facilities operated by gas pipeline and local utility companies. In addition to the United States and Canada, England, France, the Netherlands, Germany, Italy, Spain, Algeria, Libya, and Japan also have installations. Storage is in insulated metal tanks, buried concrete tanks, or frozen earth excavations. In two projects using frozen earth excavations, excessive boil-off of the liquefied gas led to replacement with insulated metal tanks. See LIQUEFIED NATURAL GAS (LNG); OIL AND GAS FIELD EXPLOITATION; PETROLEUM; PIPELINE. Peter G. Burnett

Bibliography. American Gas Association, *The Underground Storage of Gas in the United States and*

Canada, Annual Report on Statistics; D. C. Bond, *Underground Storage of Natural Gas*, Illinois State Geological Survey, 1975; G. Hobson (ed.), *Modern Petroleum Technology*, pts. 1 and 2, 5th ed., 1984; D. L. Katz et al., *Handbook of Natural Gas Engineering*, 1959; D. L. Katz and P. A. Witherspoon, *Underground and Other Storage of Oil Products and Gas*, Proceedings of the 8th World Petroleum Congress, 1971; E. V. Tiratsoo, *Natural Gas*, 3d ed., 1980; Stone and Webster Engineering, *Gas Storage at the Point of Use*, Amer. Gas Ass. Proj. PL-56, 1965.

## Oil and gas well completion

The operations that prepare a well bore for producing oil or gas from the reservoir. The goal of these operations is to optimize the flow of the reservoir fluids into the well bore, up through the producing string, and into the surface collection system. See OIL AND GAS FIELD EXPLOITATION; OIL AND GAS WELL DRILLING.

**Casing and cement.** The well bore is lined (cased) with steel pipe, and the annulus between well bore and casing is filled with cement.

Properly designed and cemented casing prevents collapse of the well bore and protects fresh-water aquifers above the oil and gas reservoirs from becoming contaminated with oil and gas and the oil reservoir brine. Similarly, the oil and gas reservoir is prevented from becoming invaded by extraneous water from aquifers that were penetrated above or below the productive reservoir. See AQUIFER.

The casing string is made up of joints of steel pipe which are screwed together to form a continuous string as the casing is extended into the well bore. The common length of an individual joint is 30 ft (9 m). Such factors as the depth of the well, the pressure, temperature, and corrosivity of the fluids to be produced and those in the reservoirs that are to be cased off (behind pipe) are taken into account in specifying the diameter, wall thickness, strength, and chemical composition of the steel pipe for a particular casing job.

In deep wells, one or more intermediate strings of casing are set (**Fig. 1**) in order to cement off either high-pressure intervals which cannot be controlled by the weight of the drilling fluid, or low-pressure intervals into which large volumes of drilling mud may flow and result in lost circulation, preventing further controlled drilling. When drilling into a high-pressure formation, casing is frequently set on top of it in order to facilitate well control operations if a blowout appears to be imminent.

In order to achieve its objectives, the casing must be securely sealed to the well bore itself with cement, although special formulations may be required for specific wells. For example, high-temperature formations or producing formations which will be extensively fractured will require cement that will not set too rapidly at high temperature or will not

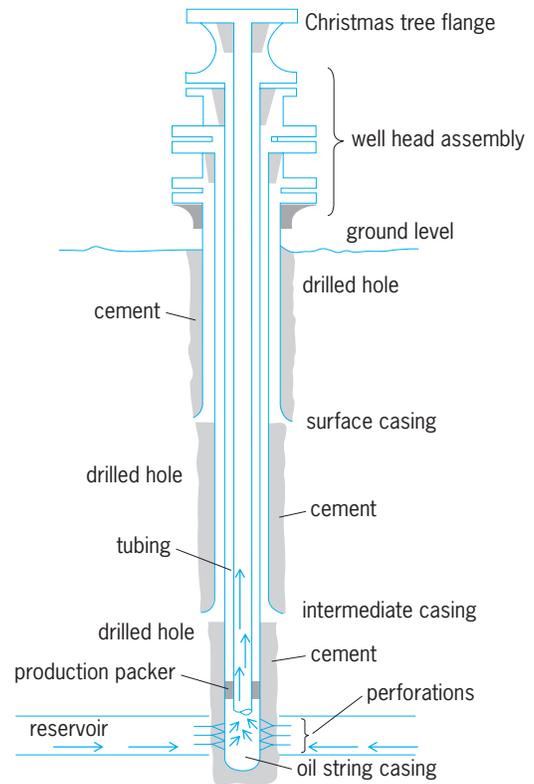


Fig. 1. Casing detail; casing strings in an oil well.

crack too badly as a result of the pressure shock of hydraulic fracturing, respectively. The cement is pumped down the casing and then on up into the annulus to a predetermined height. Cement returns (to the surface) are not universally required. The cement is mixed, pumped, and metered with highly specialized mobile equipment which is supplied by an appropriate service company. In order not to end up with the casing filled with cement, a specially designed plug is inserted after the required amount of cement has been pumped in and displaced with water until the plug hits the bottom of the casing string. The plug and some minor amount of cement will have to be drilled out after the cement has set.

**Well bore–reservoir connection.** The nature of the reservoir, evaluated from a core analysis, cuttings, or logs, or from experience with like productive formations, determines the type of completion to be used: barefoot, casing set through and then perforated, or a shop perforated or slotted liner.

In a barefoot completion, the casing is set just above the producing formation, and the latter is drilled out and produced with no pipe set across it (**Fig. 2**). Such a completion can be used for hard rock formations which are not friable and will not slough, and when there are no opportunities for producing from another, lower reservoir.

Set-through and perforated completions are also employed for relatively well-consolidated formations from which the potential for sand production is small. However, the perforated completion is used

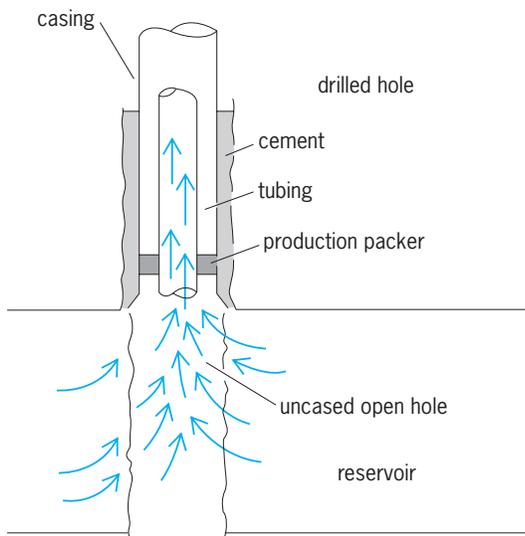


Fig. 2. Diagram of barefoot completion.

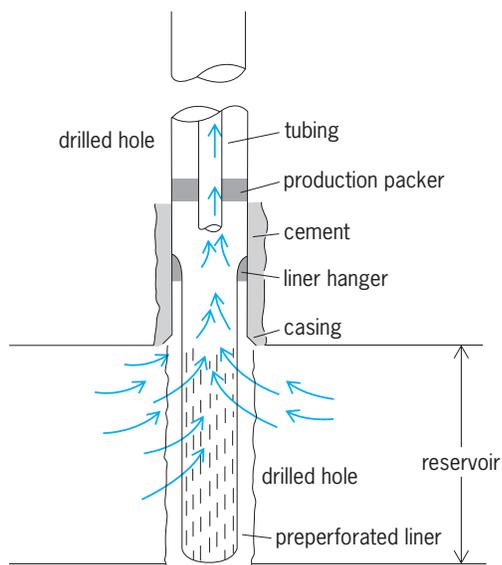


Fig. 3. Liner-type completion; preperforated liner.

when a long producing interval must be prevented from collapse, when multiple intervals are to be completed in the one borehole, or when intervening water sands within the oil-producing interval are to be shut off and the oil-saturated intervals selectively perforated. Perforations are made with bullets or shaped charges (jet perforation). The bullets are fired from a gun with multiple barrels, spaced at desired intervals, which is lowered into the hole on a wire line. An electric impulse detonates the bullets. The holes created by bullets are frequently lined with fused metal and mineral debris and as a result may offer some resistance to fluid influx. See WELL LOGGING.

The charges used in jet perforating are similar to the shaped charges used in bazookas. The shaped charges are run into the hole on a glass gun which disintegrates.

A shop-fabricated liner is used for friable formations from which some of the formation sand may flow into the well bore. The passage of such sand into the well bore may cause scoring of the seats and valves in the pump and its consequent failure to be able to lift produced fluid; or it may result in accumulation of a sand plug in the lower joints of casing through which the flow of fluids would be impeded, or in erosion of surface valves and piping. The holes in the liner are designed to screen out any produced sand, and such liners are gravel-packed. A slurry of gravel is circulated (washed-in) down behind the liner, prior to setting the liner hanger (Fig. 3). The distribution of particle diameters in the gravel pack is chosen so that the pack is an effective screen for the reservoir sand. A prepacked gravel liner may also be used, but since a gap is left between the well bore and the liner, which may fill up with the fine silt that is carried with the produced fluids, it is generally preferable to use a washed-in gravel pack.

**Production.** A string of steel tubing is lowered into the casing string and serves as the conduit for the produced fluids. The tubing may be hung from the well-head or supported by a packer set above the producing zone. The packer is used when it is desirable to isolate the casing string from the produced fluids because of the latter's pressure, temperature, or corrosivity, or when such isolation may improve production characteristics.

**Artificial lift.** The tops of wells from which fluids flow as a result of the indigenous reservoir energy are equipped with a manifold known as the Christmas tree (Fig. 4). However, only some reservoirs have sufficient pressure and sufficient gas in solution (which is released at the lower pressure existing in the well bore and therefore lowers the effective density of the

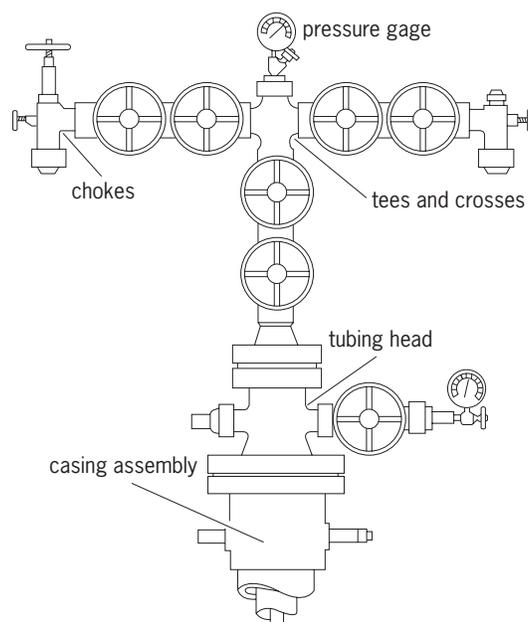


Fig. 4. Typical layout of a Christmas tree manifold.

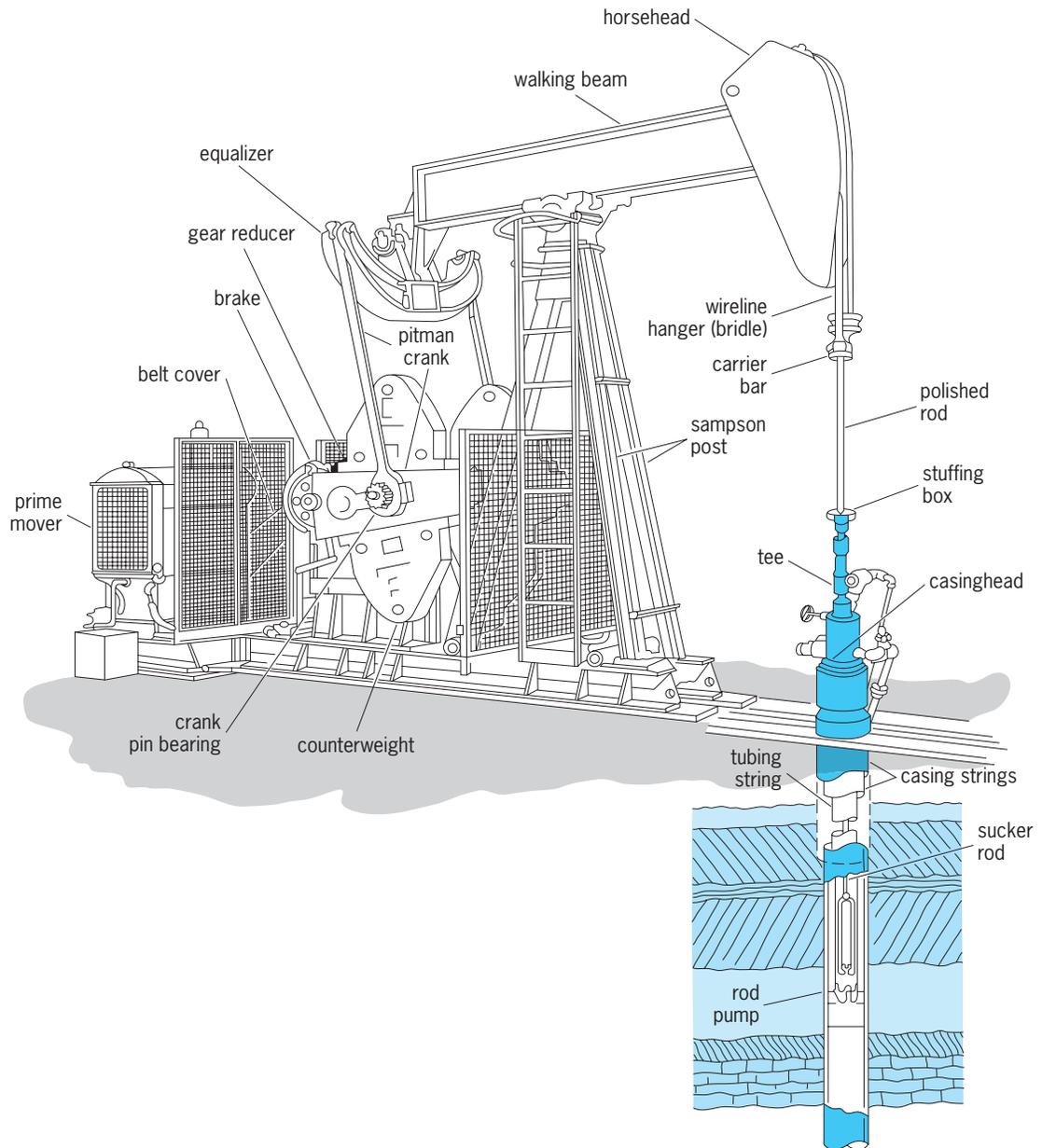


Fig. 5. Schematic diagram of most commonly used downhole pumps.

fluid in the tubing) to permit natural flow to the surface. The reservoir fluids from other reservoirs and, after pressure depletion, even from those which initially flowed must be brought to the surface by one of several methods of artificial lift.

The most common method is the use of a rod pump which is set near the bottom of the hole and operated by reciprocating sucker rods which are in turn attached to the walking beam on the surface (Fig. 5). The walking beam is driven by a motor, and by the use of suitable cams and cranks the beam's seesaw movement raises and lowers the sucker rod string. The cycle and stroke length of the sucker rods are adjustable. Tubing pumps attached to the bottom of the tubing string have a relatively high capacity, but the entire string must be pulled to re-

pair a damaged pump. Insert pumps are set within the tubing. Because of their restricted diameter, they have a limited capacity for lifting reservoir fluids, but they have the advantage that they can be pulled and replaced with a wire line without pulling the entire tubing string. For deeper wells, hydraulic motors can be used for which the actuating fluid (crude oil) is pumped down the tubing and returns with the produced fluid up through the annulus. Wells which produce a large amount of fluid (both water and oil) can economically use a submerged centrifugal pump driven by an electric motor for which an electric cable is run down the annulus. See CENTRIFUGAL PUMP.

For deep wells which produce a significant amount of gas, gas lift can be employed in which

some of the produced gas is compressed and returned to the casing-tubing annulus. A series of pressure-actuated valves inserted in the tubing string permits the gas to enter the string at various levels to lower the effective density of the fluids in the tubing and propel the fluids to the surface (Fig. 6). A plunger lift system to assist with unloading liquids can be easily installed inside tubing without the need to pull the tubing. Such systems are used to produce high-gas-oil-ratio (GOR) wells, water-producing gas wells, or very-low-bottom-hole-pressure oil wells (used with gas lift).

**Multiple completion.** In some geological provinces, several successive but separated intervals are productive of oil and gas. In some instances, the production from all the intervals may be commingled in a single well bore. However, if the properties of the reservoirs or their fluids are different, then commingling may be unacceptable because of the potential for cross flow between the individual reservoirs. Multiple completions in which the producing zones are separated by the use of packers and individual tubing string are then used (Fig. 7).

**Water problems.** Excessive water production increases the cost of oil production since energy must be expended in lifting the water to the surface. Water production may also jeopardize the production of oil and gas by saturating the oil-productive interval with water. Such damage is more likely to occur in low-pressure formations or formations which contain water-sensitive clays that swell in an excess of water.

**Water-exclusion methods.** Water exclusion may be effected by the application of cements of various types. If it is determined that water is entering from the lower portion of a producing sand in a relatively shallow, low-pressure well, a cement plug may be placed in the bottom of the hole so that it will cover the oil-water interface of the reservoir. This technique is called laying in a plug and may be accomplished by placing the cement with a dump-bottom bailer on a wire line or by pumping cement down the drill pipe or tubing. For deeper, higher-pressure, or more troublesome wells, a squeeze method is used. Squeeze cementing is the process of applying hydraulic pressure to force a cement into an exposed formation or through openings in the casing or liner. It is also used for repairing casing leaks; isolating producing zones prior to perforating for production; remedial or secondary cementing to correct a defective condition, such as channeling or insufficient cement on a primary cement job; sealing off a low-pressure formation that causes lost circulation of drilling fluids; and abandoning depleted producing zones to prevent migration of formation effluent and to reduce possibilities of contaminating other zones or wells.

The squeeze tool is a packer-type device designed to isolate the point of entry between or below packing elements. The tool is run into the hole on drill pipe or tubing, and the cement is squeezed out between or below these confining elements into the

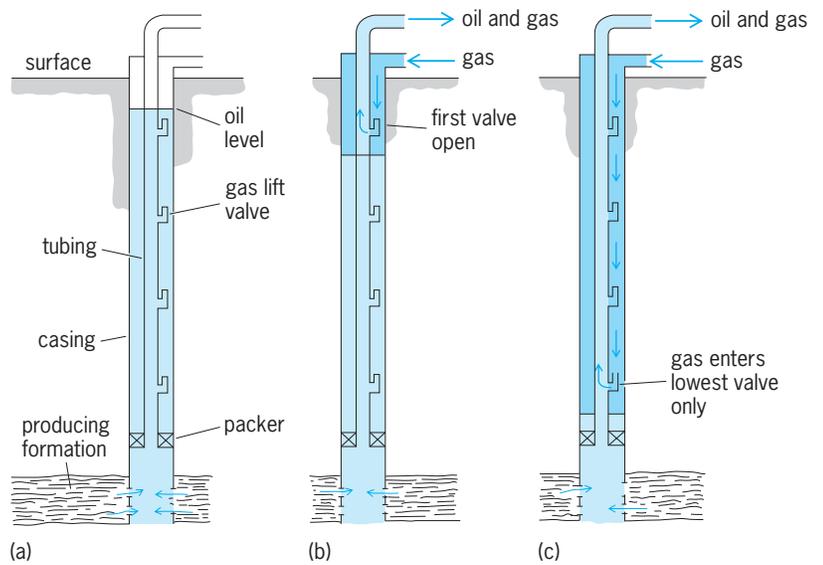


Fig. 6. Schematic representation of operation of a gas lift string. (a) Oil level above first valve. (b) First valve open and gas entering tubing; oil level in casing/tubing annulus moving downward. (c) Oil level has moved downward, and each valve has closed as gas has entered the next lowest valve.

problem area. The well is then recompleted. It may be necessary to drill the cement out of the hole and reperform, depending upon the outcome of the job performed in the squeeze process.

**Water-exclusion plug back.** Simple water shutoff jobs in shallow, deep, or high-pressure wells may also be performed in multizone wells in which the lower

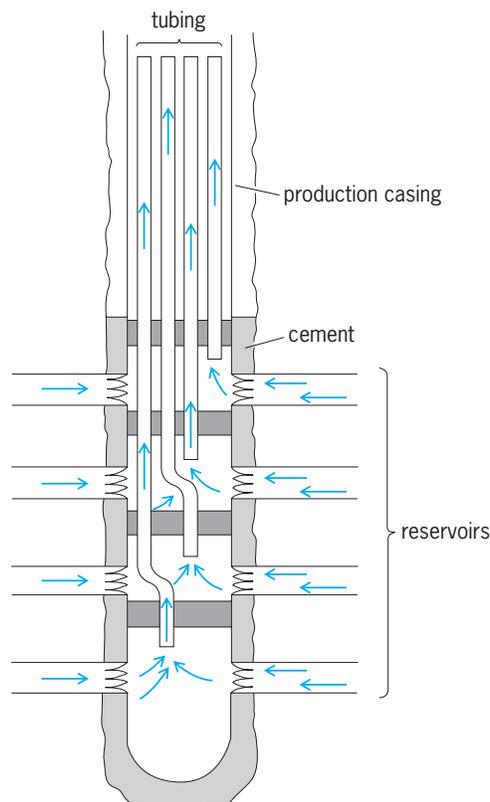


Fig. 7. Schematic diagram of a multiple completion.

producing interval is depleted or the remaining recoverable reserves do not justify recompletion.

Here, water may be excluded by placing a packer-type plug above the interval, then producing formations that are already open or perforating additional intervals that may be present higher up the hole.

**Production stimulation.** Production may be impaired from a well bore as a result of drilling-mud invasion or of accumulation of clays and fine silts carried by the producing fluids to the borehole, or the lithology of the formation itself may have a naturally low permeability to reservoir fluids. Since the permeability to fluids of the formation within the first few feet of the well bore has an exponential effect on limiting the influx of fluid, the productivity of a well can frequently be increased manifold by increasing the permeability of this element of the reservoir or removing the skin just at the face of the producing interval. This is accomplished by acidization and fracturing, and in some instances by the use of surfactants, solvents, and explosives. Specialized service companies conduct the work by using their own specially designed equipment.

*Acidizing.* Inhibited hydrochloric acid contains a chemical additive (an inhibitor) which prevents the acid from attacking steel. In this way the acid can be used for dissolving carbonates, oxides, and other compounds without fear of it attacking the well's steel tubulars. This formulation is used to dissolve limestone and dolomitic matrices and thus enlarge the flow channels in production-impaired reservoirs. Hydrochloric acid is also used to shrink and disperse sheaths of drilling mud on the well bore and to dissolve calcareous cements, which results in larger channels through which fluids can flow to the well bore.

Hydrofluoric acid (released by injecting a mixture of hydrochloric acid and a soluble fluoride salt) is sometimes used in sandstone reservoirs to dissolve and disperse drilling mud that has invaded the reservoir.

*Fracturing.* Formation fracturing is a hydraulic process aimed at the parting of the formation. Vertical fractures most frequently occur. Horizontal fracturing occurs only in relatively shallow formations, in formations where the major tectonic stress is horizontal, or in relatively plastic formations. The fracturing fluid is injected into the well, and the pressure is raised to maintain a given flow rate until formation breakdown occurs. Injection is continued with a slurry of a selected grade of sand or gravel or particles of other material (such as sintered bauxite or ceramic beads). These particles prop the fracture open after the hydraulic pressure is released. Crude oil, acid, and a variety of gelled liquids are used as fracturing fluids. The propping material guarantees that there will be a high-permeability path into the well bore, and the nature of fluid flow in the vicinity of the well bore is changed to being predominantly linear rather than radial with an associated decrease in pressure drop (or higher flow rate at the same pressure drop).

*Other stimulation techniques.* Explosives were the first means used to stimulate oil and gas production. How-

ever, this technique has largely been supplanted by more effective and safe fracturing and acidizing technology.

Solvents are used when the substances believed to be inhibiting production are asphaltenes, waxes, and emulsions stabilized by such organic materials. Surfactants are frequently used with the solvents to aid in the dispersion of the sediments. Surfactants or alcohols (for example, methanol) are also used alone when the cause of impairment is believed to be a high saturation of water that has accumulated in the reservoir near the well bore.

**Sand consolidation.** Sand exclusion techniques using liners and gravel packs are not perfect, and therefore technology has been developed that attempts to consolidate friable formations. The consolidating medium must be capable of cementing the grains together without significantly reducing the permeability of the reservoir to fluid flow. Epoxy and phenolic resins have been developed for such purposes; some techniques use thermally deposited nickel metal and precipitated aluminum oxides. However, liners with properly designed gravel packs continue to be the most economical and useful technique for sand control.

**Coiled tubing.** Many of the well completion or workover techniques can be implemented with a coiled tubing unit that can greatly reduce costs. Instead of moving in a completion rig to lower or pull tubing, a coiled tubing unit may be moved next to the wellhead. With such a unit, instead of having to connect and disconnect stands of tubing, a continuous length of tubing may be uncoiled or coiled into the borehole by using a large spool. In this way, numerous operations may be performed on a well such as acidizing, setting and retrieving bridge plugs or packers, cementing, cleaning out the hole, and even light-duty or slim-hole drilling. A wide variety of remedial operations may be performed. See PETROLEUM RESERVOIR ENGINEERING.

Todd M. Doscher; R. E. Wyman

**Bibliography.** J. Algeroy, Equipment and operation of advanced completions in the M-15 Wytch Farm multilateral well, presented at the *2000 Annual Technical Conference and Exhibition* (Dallas), Pap. SPE 62951, October 1-4, 2000; G. Botto et al., Innovative remote controlled completion for Aquila Deepwater Challenge, *1996 SPE European Petroleum Conference* (Milan), Pap. SPE 36948, October 22-24, 1996; R. A. Dawe and A. G. Lucas (eds.), *Modern Petroleum Technology, vols. 1 and 2*, 6th ed., 2000; M. J. Economides, A. D. Hill, and C. Ehlig-Economides, *Petroleum Production Systems*, 1993; N. J. Hyne, *Nontechnical Guide to Petroleum Geology, Exploration, Drilling and Production*, 2d ed., 2001; V. B. Jackson, Intelligent completion technology improves economics in the Gulf of Mexico, *Amer. Oil Gas Rep.*, June 2000; D. E. Johnson, Reliable and completely interventionless intelligent completion technology: Application and field study, *2002 Offshore Technology Conference* (Houston), Pap. OTC 14252, May 6-9, 2002.

## Oil and gas well drilling

The drilling of holes for exploration and extraction of crude oil and natural gas. Deep holes and high pressures are characteristics of petroleum drilling not commonly associated with other types of drilling. In general, it becomes more difficult to control the direction of the drilled hole as the depth increases, and additionally, the cost per foot of hole drilled increases rapidly with the depth of the hole. Drilling-fluid pressure must be sufficiently high to prevent blowouts but not high enough to cause fracturing of the borehole. Formation-fluid pressures are commonly controlled by the use of a high-density clay-water slurry, called drilling mud. The chemicals used in drilling mud can be expensive, but the primary disadvantage in the use of drilling muds is the relatively low drilling rate which normally accompanies high bottom-hole pressure. Drilling rates can often be increased by using water to circulate the cuttings from the hole; when feasible, the use of gas as a drilling fluid can lead to drilling rates as much as 10 times those attained with mud. Drilling research has the objectives of improving the utilization of current drilling technology and the development of improved drilling techniques and tools.

**Hole direction.** The hole direction must be controlled within permissible limits in order to reach a desired target at depths as great as 25,000 ft (7600 m). Inclined layers of rocks with different hardnesses tend to cause the direction of drilling to deviate; consequently, deep holes are rarely truly straight and vertical. The drilling rate generally increases as additional drill-collar weight is applied to the bit by adjusting the pipe tension at the surface. However, crooked-hole tendency also increases with higher weight-on-bit. A so-called packed-hole technique has been used to reduce the tendency to hole deviation. One version of this technique makes use of square drill collars that nearly fill the hole on the diagonals but permit fluid and cuttings to circulate around the sides. This procedure reduces the rate at which the hole direction can change.

**Directional drilling.** In mountainous terrain, it is difficult to construct well locations over each subsurface drilling target, and from offshore drilling platforms it is necessary to drill many wells from a single surface location (Fig. 1). For these situations, technology has been developed that permits wells to be drilled directionally from the single surface location to the desired subsurface point.

The earliest method used for directional drilling is whipstocking, which involves placing a wedge-shaped piece of steel at the bottom of the hole to force the bit and drill pipe off into the desired direction (Fig. 2). The whipstock is set with the aid of instruments that permit the desired angle and its direction to be initiated. The angle of the drill can then be built up by setting successive whipstocks in place.

Another development is the use of turbines and positive displacement motors that are driven by the circulating mud. The motor is located below a bent

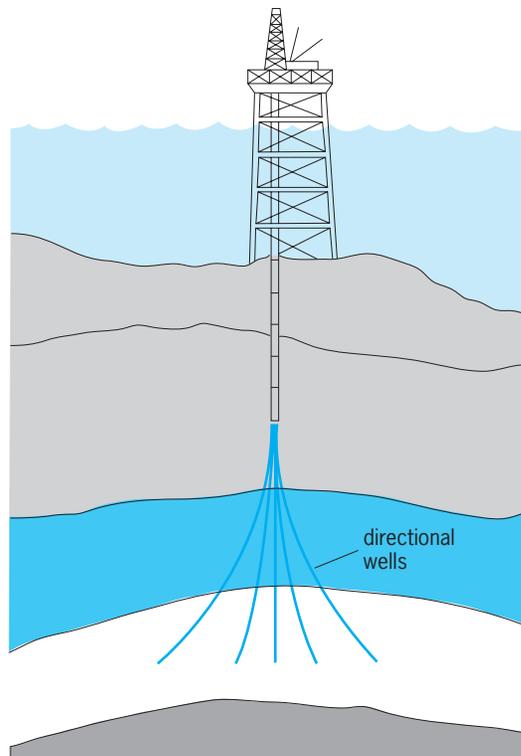


Fig. 1. Directional wells from offshore platform.

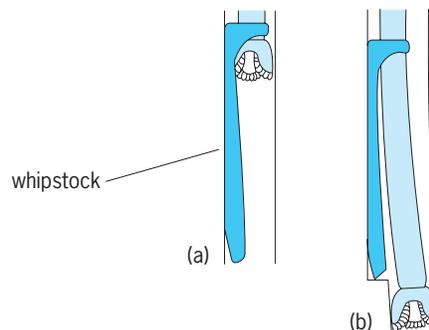


Fig. 2. Whipstocking for deviating a well. (a) Whipstock in position. (b) Bit and drill pipe deviated by whipstock.

sub (Fig. 3), a component of the drill string, locate just above the mud motor, which permits the bit to be set in a predetermined direction. This component initiates the new drilling angle. The motor turns only the bit and not the entire drill string as in conventional rotary operations. Again the angle can be built by successive use of bent subs. Twelve or more directional wells can be drilled from one location and through one leg of a drilling platform by using modern directional drilling techniques.

**Horizontal drilling.** Advances in technology and the need to accomplish special objectives have led to drilling horizontal wells in deep oil or gas reservoirs. The angle of the well is successively built up in order to reach the ultimate horizontal course of the well (Fig. 4). The program for drilling the well is developed before hand, based on the technology for actually increasing the angle and then continuing the

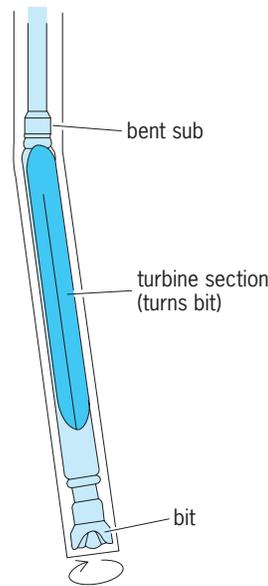


Fig. 3. Turbine for deviating a well.

drilling at the desired angle. In order to be able to drill such a hole without excessive torque and differential sticking, a drilling mud of high lubricity is required. The mud must also have a high yield point and a low fluid loss to prevent hole instability.

One of the shortcomings of horizontal wells is the inability to complete them selectively. Wells have been completed either barefoot (no casing or liner through the producing section) or with noncemented, preslotted lines set through the producing section. The wells are therefore somewhat more vul-

nerable to being damaged or to collapse by formation transport.

The productivity of a horizontal well can be greater than that of a conventional vertical well. Whereas the effective length of a vertical well is limited to that of the producing formation, the length of a horizontal well, theoretically, can be as big as the reservoir itself. The ratio of the productivity of a horizontal well to that of a vertical well increases with the length of the horizontal well, and the effect is more pronounced the greater the value of the parameter  $L/b$ , where  $L$  is the length of the horizontal well and  $b$  is the thickness of the reservoir. The ratio of the horizontal well productivity to that of a conventional well also increases with the value of  $L/r_e$ , where  $r_e$  is the drainage radius of the reservoir.

The true value of horizontal wells is realized in the following special reservoir situations.

1. The oil saturation and fluid conductivity of the formation may be primarily associated with vertical fractures in echelon. A conventional vertical well may or may not intersect such fractures, depending upon the density and orientation of the fractures. An accurately placed horizontal well can intersect a number of such fractures and effectively drain them.

2. This same advantage will pertain in low-permeability formations, which are often stimulated with hydraulic fractures. These artificially induced fractures will tend to propagate perpendicular to the least principal stress, which is the same direction in which vertical natural fractures are oriented. Hence, horizontal drilling may allow greater production from naturally fractured low-permeability formations by being directed to cross the natural fractures.

3. The oil column is bounded by a gas cap or bottom water, or both. A conventional vertical well under such conditions will be perforated somewhere in the middle of the formation in order to restrict gas coning (the downward flow of gas overlying an oil-saturated interval) and water coning (the entry of bottom water, or mobile water existing below the oil-saturated interval, which rises up and enters the well bore that is open opposite the oil-saturated interval). Such coning actually curtails oil production by reducing the relative permeability of the oil around the wellbore.

4. Another application related to the foregoing is production of the reservoir under a steam drive. Because of the low density of the steam, it overrides the oil column so that oil at the top of the column is heated, mobilized by viscosity reduction, and then displaced by the high-velocity steam to the producing well. As the column of oil is depleted from the top, the cross section available for steam flow increases, and as a result the vapor velocity decreases with decreasing steam vapor velocity, and the ratio of (produced) oil to (injected) steam decreases with a correlative decrease in economic efficiency. A steam-stimulated horizontal well drilled at or near the base of the oil column could promote more economically efficient drainage of oil under the influence of a steam drive.

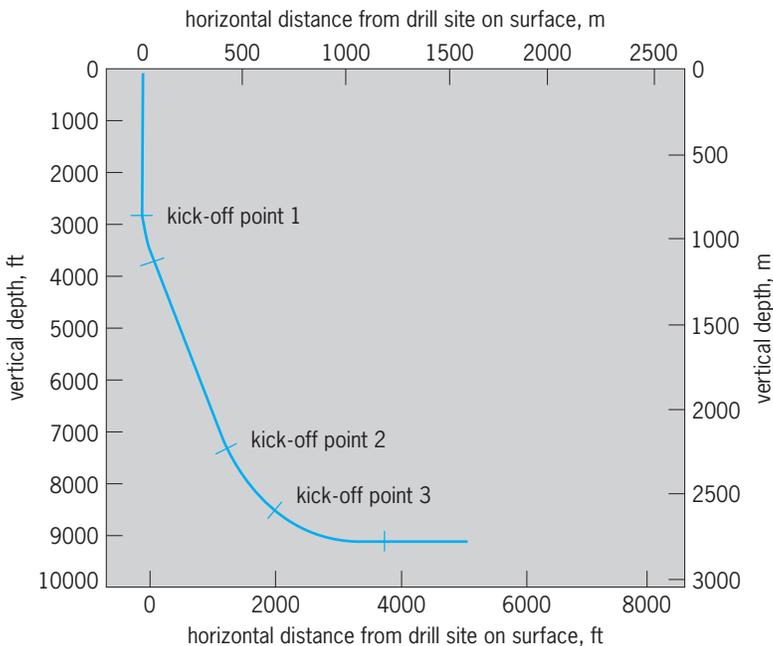


Fig. 4. Trajectory of a horizontal well. At kick-off point 1, the drilling angle is built up  $2.5^\circ$  per 100 ft to a maximum angle of  $22.5^\circ$ . Drilling continues at this angle until kick-off point 2 is reached, where the angle is built at  $1.5^\circ$  per 100 ft to a maximum angle of  $44^\circ$ . At kick-off point 3, the angle is built at  $2^\circ$  per 100 ft to  $90^\circ$ .

5. Horizontal drilling may have advantages in certain geological situations by penetrating enhanced porosity or nearby isolated reservoirs that would not be drained by vertical holes. For instance, a number of isolated sand or conglomerate sequences laid down at a low angle as an ocean retreated from a beach could be reached with one horizontal hole. Wells have been drilled that penetrate several miles of formation in a horizontal direction.

**Drilling fluids.** The increased formation or pore-fluid pressures existing at great depths in the Earth's crust adversely affect drilling. Gushers, blowouts, or other uncontrolled pressure conditions cannot be tolerated; therefore, high-density drilling fluids are used to control well pressures. A normal fluid gradient for salt water is about  $0.5 \text{ lb/in.}^2$  per foot of depth (11 kilopascals/m), and the total stress due to the weight of the overburden increases approximately  $1 \text{ lb/in.}^2$  per foot (23 kPa/m). Under most drilling conditions in permeable formations, the well-bore pressure must be kept between these two limiting values. If the mud pressure is too low, the formation fluid can force the mud from the hole, resulting in a blowout; whereas if the mud pressure becomes too high, the rock adjacent to the well may be fractured, resulting in lost circulation. In this latter case the mud and cuttings are lost into the fractured formation and may also result in a blowout.

High drilling-fluid pressure at the bottom of a borehole impedes the drilling action of the bit. Rock failure strength increases, and the failure becomes more ductile as the pressure acting on the rock is increased. Ideally, cuttings are cleaned from beneath the bit by the drilling-fluid stream; however, relatively low mud pressure tends to hold cuttings in place. In this case, mechanical action of the bit is often necessary to dislodge the chips. Regrinding of fractured rock greatly decreases drilling efficiency by lowering the drilling rate and increasing bit wear.

Drilling efficiency can be increased under circumstances where mud can be replaced by water as the drilling fluid. This might be permissible, for example, in a well in which no high-pressure gas zones are present. Hole cleaning is improved with water drilling fluid because the downhole pressure is lower and no clay filter cake is formed on permeable rock surfaces. So-called fast-drilling fluids provide a time delay for filter cake buildup. This delay permits rapid drilling with no filter cake at the bottom of the hole and, at the same time, prevents excessive loss of fluid into permeable zones above.

In portions of wells where no water zones occur, it is frequently possible to drill using air or natural gas to remove the cuttings. Drilling rates with a gas drilling fluid are often 10 times those obtained with mud under similar conditions. Sometimes a detergent foam is used to remove water in order to permit gas drilling in the presence of limited water inflow. In other instances, porous formations can be plugged with plastic to permit continued gas drilling. However, in many cases it is necessary to revert to either water or mud drilling when a water zone is encountered. Another possibility is the cementing of steel

casing through the zone containing water and then proceeding with gas drilling.

**Blowout prevention.** A drilling well is "kicked" when a high-pressure zone is encountered, and the high-pressure fluids begin to enter the well bore. If the kick is uncontrolled, then the high-pressure fluid continues to enter the well bore and the well may blow out. This is all the more likely to happen if the kick is caused by high-pressure gas, or liquids that contain a significant quantity of dissolved gas, since the gas will continue to expand, displacing more and more drilling mud, as it flows up through the well-bore. The effective density of the drilling mud is thus decreased, and the influx of high-pressure fluid accelerates. The drilling crew will notice the kick by the mud pits' beginning to fill rapidly with the displaced mud.

Drilling operations are immediately stopped, and the blowout preventer (BOP) stacks (Fig. 5) are actuated to close in the well. The blowout parameter stack is located just over the well head beneath the drilling floor. The stack is composed of three different seals. The annular preventer is located at the top of the stack and is the first to be actuated. It consists of an expandable, steel-reinforced rubber seal that shuts the annular space between the well bore and drill pipe. If the annular preventer does not succeed in sealing the well, the pipe rams are then closed. These form a steel gate that fits tightly around the drill pipe and seals off the annulus. The blind rams are similar to the pipe rams except that they are used for sealing off the hole if there is no pipe in it. Many blowout preventer stacks are equipped with shear rams, which can cut off the drill pipe and then form the required seal.

As soon as the preventers are closed, dense drilling mud is pumped into the hole down through a choke or kill line. Weighting material is added to the mud until its bottom-hole pressure is sufficient to counterbalance the pressure of the formation that is kicked.

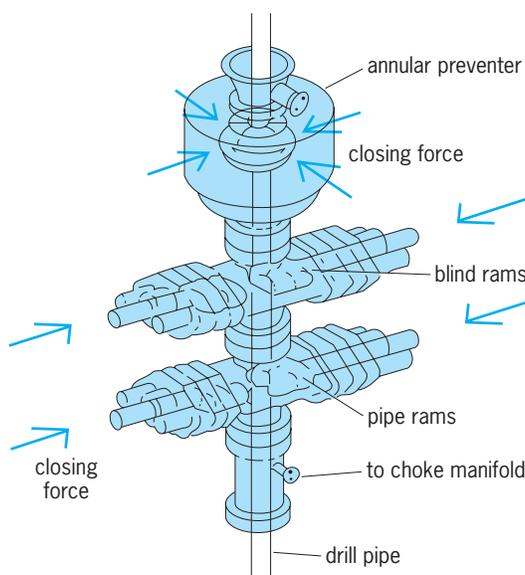


Fig. 5. Components of a BOP stack.

**Blowout control.** If, despite efforts to prevent a blowout, such an event does occur, it becomes imperative to control the blowout as soon as possible. Usually specialists in blowout control are called to the site to assess the situation and implement control procedures. Various methods may be used, but a typical sequence of events includes cooling the well site with water sprays; removing the surface debris and establishing access to the wellhead; extinguishing the fire if the fluids have ignited; removing damaged wellheads or blowout preventers; and installing a new valve assembly to kill the well.

To extinguish the fire of a blowout, two methods that have been used successfully are to discharge an explosive charge directly above the wellhead, or to direct at the flame a high-pressure spray of a dry chemical such as potassium bicarbonate ( $\text{KHCO}_3$ ) in siliconized powder form. After the fire is out, another option to kill a well is to force a hollow tapered tube known as a stinger into the gushing oil. Heavy mud can then be pumped into the well to offset the natural pressures. This method was used extensively to bring oil well fires under control in Kuwait following the Gulf War in 1991.

**Cost factors.** Drilling costs depend on the costs of such items as the drilling rig, the bits, and the drilling fluid, as well as on the drilling rate, the time required to replace a worn bit, and bit life. The cost per foot increases with depth when encountering geopressures, heavy shale, lost circulation, and well-consolidated, hard formations. Operating costs of a large land rig required for deep holes are much less than comparable costs for an off shore drilling platform or floating drilling vessel. Although weak rock at shallow depth can be drilled at rates exceeding 100 ft/h (30 m/h), drilling rates often average about 5 ft/h (1.5 m/h) in deep holes. Conventional rotary bits have an operating life of 10–20 h. Diamond bits may drill for as long as 50–200 h, but the drilling rate is relatively low and the bits are expensive. For more economical drilling, it is desirable to increase both bit life and drilling rate simultaneously.

**Research.** Drilling research includes the study of drilling fluids, the evaluation of rock properties, laboratory simulation of field drilling conditions, and the development of new drilling techniques and tools. Fast-drilling fluids have been developed by selecting drilling-fluid additives that plug the pore spaces very slowly, thereby providing the desired time delay for filter cake buildup. Water-shutoff chemicals have been formulated that can be injected into a porous water-bearing formation in liquid form and then, within a few hours, set to become solid plastics. Well-logging techniques can warn of high-pressure permeable zones so that the change from gas or water drilling fluid to mud can be made before the drill enters the high-pressure zone. Downhole instrumentation can lead to improved drilling operations by providing information for improved bit design and also by feeding information to a computer for optimum control of bit weight and rotary speed. Computer programs can also utilize information from nearby

wells to determine the best program for optimum safety and economy in new wells.

Since rock is a very hard, strong, abrasive material, there is a challenge to provide drill bits that can penetrate rock more efficiently. A better understanding of rock failure can lead to improved use of present equipment and to the development of better tools. Measurements of physical properties of rocks are correlated with methods for the theoretical analysis of rock failure by a drill bit. A small  $1\frac{1}{4}$ -in.-diameter (32-mm) bit, called a microbit, has been used in scale-model drilling experiments in which independent control is provided for borehole, formation-fluid, and over-burden pressures. These tests permit separation of the effects of the various pressures on drilling rates.

Novel drilling methods that have been explored include studies of rock failure by mechanical, thermal, hydraulic, fusion and vaporization, and chemical means. Jet piercing is widely used for drilling very hard, spallable rocks, such as taconite. Other methods include the use of electric arc, laser, plasma, spark, and ultrasonic drills. See DRILLING, GEOTECHNICAL; OIL AND GAS WELL COMPLETION; PETROLEUM GEOLOGY; ROCK MECHANICS; TURBODRILL; WELL LOGGING.

Todd M. Doscher

Bibliography. R. Baker, *A Primer of Oilwell Drilling*, 6th ed., 2000; R. A. Dawe and A. G. Lucas (eds.), *Modern Petroleum Technology*, vols. 1 and 2, 6th ed., 2000; M. J. Economides, A. D. Hill, and C. Ehlig-Economides, *Petroleum Production Systems*, 1993; C. T. Franklin, *Handbook of Oil and Gas Operations: Drilling*, vol. 2, 1994; F. A. Giuliano, *Introduction to Oil and Gas Technology*, 1989; N. J. Hyne, *Nontechnical Guide to Petroleum Geology, Exploration, Drilling and Production*, 2d ed., 2001; J. D. Jansen, *Nonlinear Dynamics of Oilwell Drillstrings*, 1993; J. A. Short, *Introduction to Directional and Horizontal Drilling*, 1993; T. Yonezawa et al., *Robotic Controlled Drilling: A New Rotary Steerable Drilling System for the Oil and Gas Industry*, *Society of Petroleum Engineers*, SPE 74458, 2002.

## Oil burner

A device for converting fuel oil from a liquid state into a combustible mixture. A number of different types of oil burners are in use for domestic heating. These include sleeve burners, natural-draft pot burners, forced-draft pot burners, rotary wall flame burners, and air-atomizing and pressure-atomizing gun burners. The most common and modern type that handles 80% of the burners used to heat United States homes is the pressure-atomizing-gun-type burner (Fig. 1).

**Characteristics.** The sleeve burner, commonly known as a range burner because of its use in kitchen ranges, is the simplest form of vaporizing burner. The natural-draft pot burner relies on the draft developed by the chimney to support combustion. The forced-draft pot burner is a modification of the natural-draft pot burner, since the only significant difference between the two types is the means of

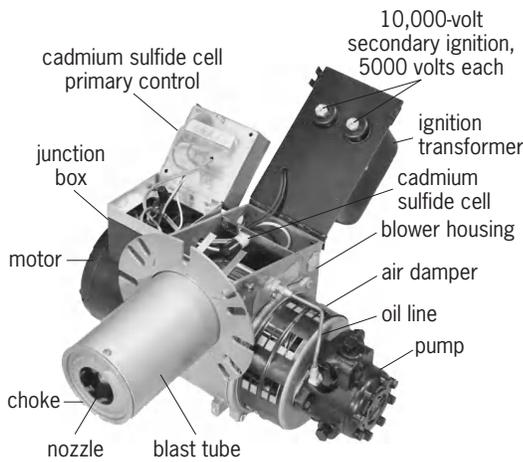


Fig. 1. Oil burner of the pressure-atomizing type. (Automatic Burner Corp.)

supplying combustion air. The forced-draft pot burner supplies its own air for combustion and does not rely totally on the chimney. The rotary wall flame burners have mechanically assisted vaporization. The gun-type burner uses a nozzle to atomize the fuel so that it becomes a vapor, and burns easily when mixed with air.

The most important feature of a high-pressure atomizing gun burner is the method of delivering the air. The most efficient burner is the one which com-

pletely burns the oil with the smallest quantity of air. The function of the oil burner is to properly proportion and mix the atomized oil and air required for combustion (Fig. 2).

**Efficiency.** If a large quantity of excess air is used to attempt to burn the oil, there is a direct loss of usable heat. This air absorbs heat in the heating unit which is then carried away through the stack with the combustion gases. This preheated air causes high stack temperatures which lower the efficiency of the combustion. The higher the CO<sub>2</sub> (carbon dioxide), the less excess air. Overall efficiencies or stack loss can be estimated by the use of the stack loss chart (Fig. 3).

Increased efficiency can be obtained through the use of devices located near the flame end of the burner. CO<sub>2</sub> is not the ultimate factor in efficiency. Burners producing high CO<sub>2</sub> can also produce high smoke readings. Accumulations of 1/8-in. (3-mm) soot layers on the heating unit surface can increase fuel consumption as much as 8%. An oil burner is always adjusting to start smoothly with the highest CO<sub>2</sub> and not more than a number two smoke on a Bacharach smoke scale. For designs in high-efficiency burners commonly known as flame-retention-type burners see Fig. 2.

**Nozzle.** The nozzle is made up of two essential parts: the inner body, called the distributor, and the outer body, which contains the orifice that the oil sprays through. Under this high pump pressure of

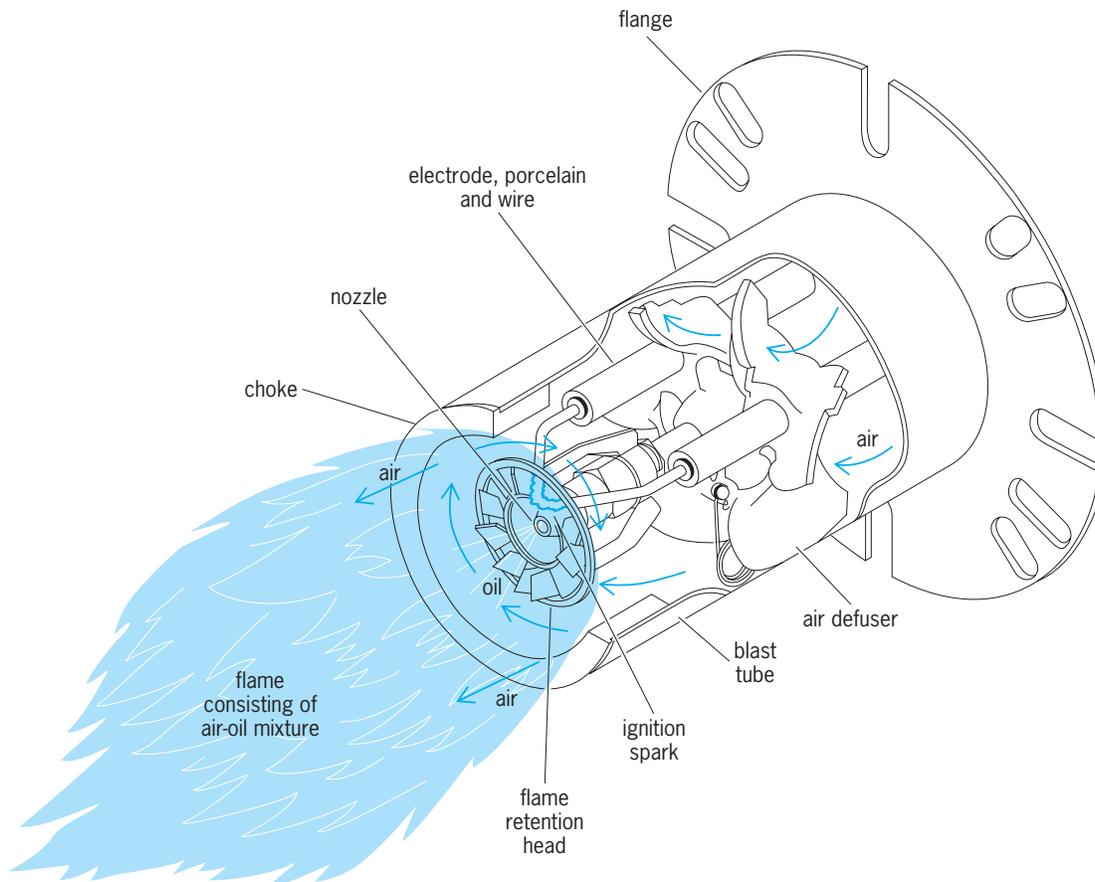
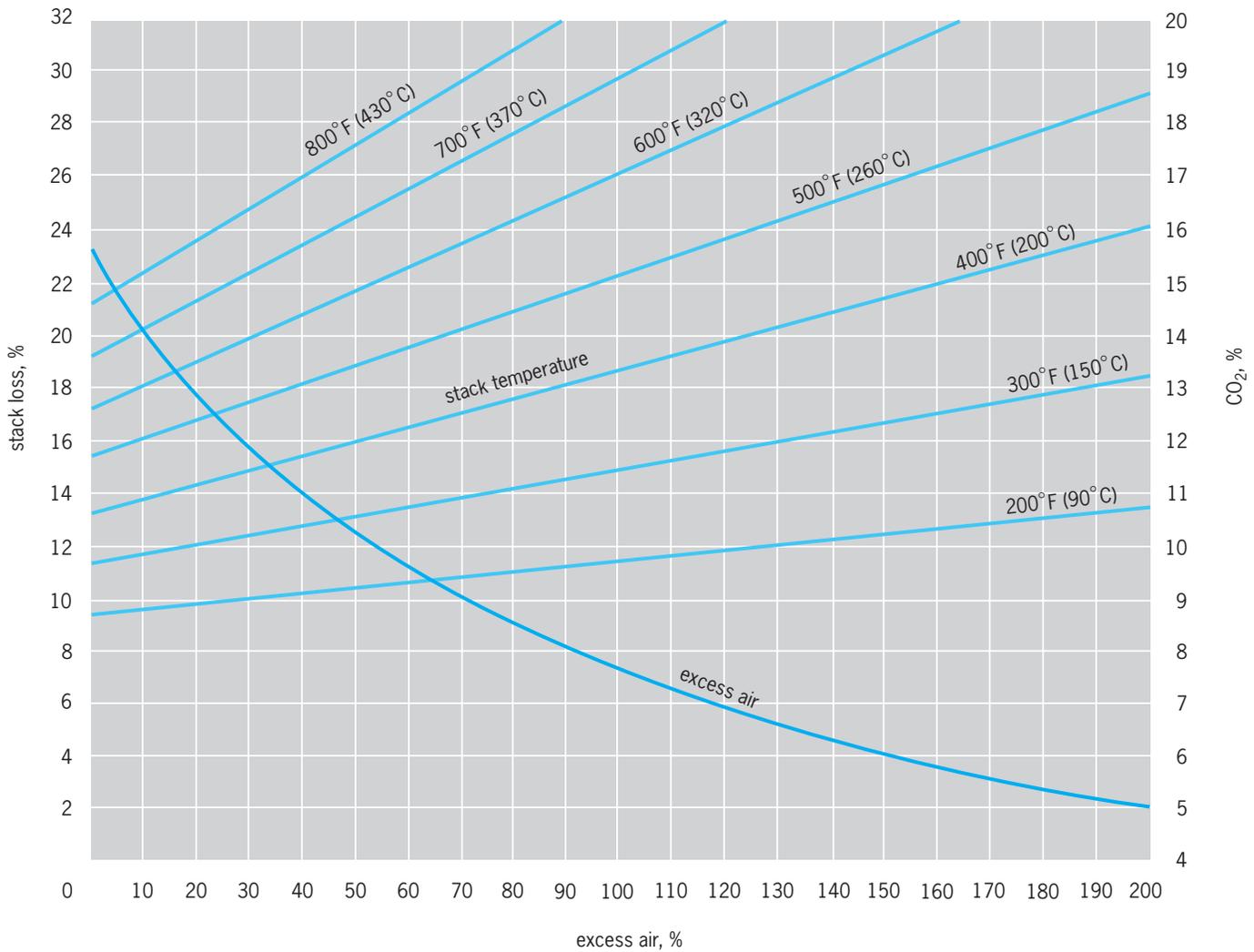


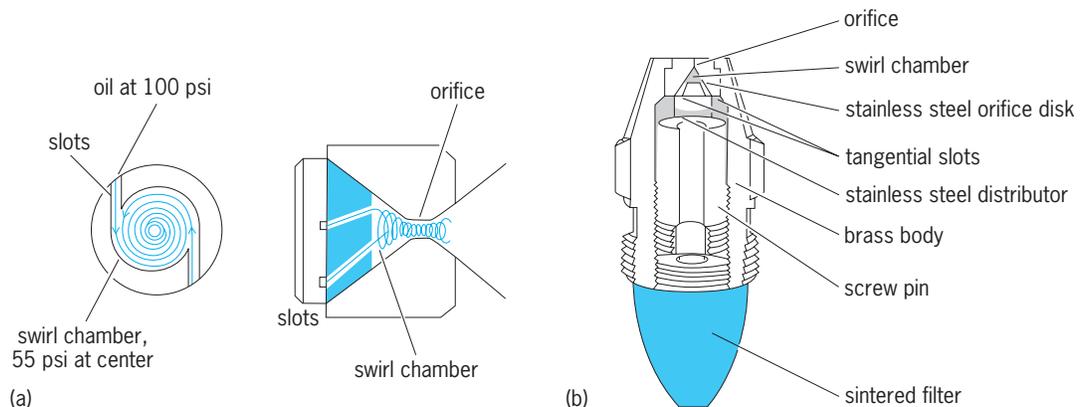
Fig. 2. Flame-retention-type burner, an example of a high-efficiency burner.



**Fig. 3. Typical stack loss chart. To determine stack loss and efficiency, start with correct CO<sub>2</sub> and follow horizontal line to excess air curve, then vertical line to stack temperature, and finally horizontal line to stack loss. Overall efficiency percent = 100 – stack loss.**

100–300 psi (700–2100 kilopascals) the oil is swirled through the distributor and discharged from the orifice as a spray (Fig. 4). The spray is ignited by the spark and combustion is self-sustaining, provided the proper amount of air is supplied by the squirrel cage

blower. Air is delivered through the blast tube and moved in such a manner that it mixes well with the oil spray (Fig. 2). This air is controlled by a damper located on either the intake side or the discharge side of the fan (Fig. 1).



**Fig. 4. Oil burner nozzle. (a) Operation of a nozzle. 1 psi = 7 kPa. (b) Cutaway view. (Delavan Manufacturing Co.)**

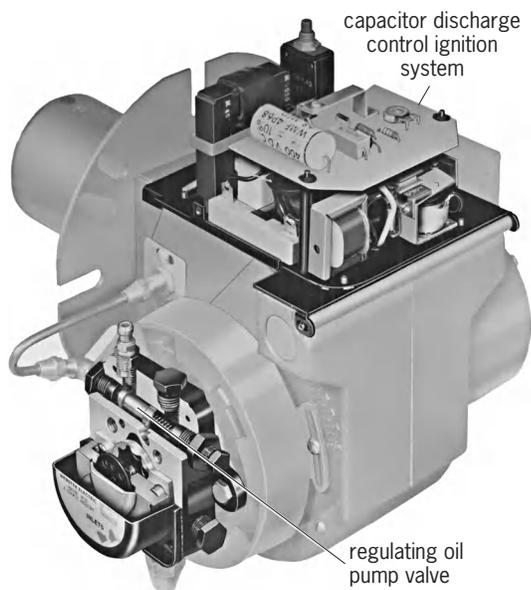


Fig. 5. Oil burner, showing capacitor discharge control ignition system and regulating valve for fuel pump. (Automatic Burner Corp.)

The essential parts of the pressure gun burner are the electric motor, squirrel-cage-type blower, housing, fuel pump, electric ignition, atomizing nozzle, and a primary control (Fig. 1 and Fig. 5).

The fuel oil is pumped from the tank through the pump gears, and oil pressure is regulated by an internal valve that develops 100–300 psi (700–2100 kPa) at the burner nozzle. The oil is then atomized, ignited, and burned.

The fine droplets of oil that are discharged from the nozzle are electrically ignited by a transformer which raises the voltage from 115 to 10,000 V. The recently developed electric ignition system called the capacitor discharge system incorporates a capacitor that discharges voltage into a booster transformer developing as much as 14,000 V. The electric spark is developed at the electrode gap located near the nozzle spray (Figs. 2 and 5).

The high velocity of air produced by the squirrel cage fan helps develop the ignition spark to a point where it will reach out and ignite the oil without the electrode tips actually being in the oil spray. Figure 2 shows a typical gun-type burner inner assembly.

**Installations.** Fuel oils for domestic burners are distilled from the crude oil after the lighter products have been taken off. Consequently, the oil is nonexplosive at ordinary storage temperatures. Domestic fuel oils are divided into two grades, number one and number two, according to the ASTM specification. There are several different methods of installing an oil tank system with an oil burner (Fig. 6).

One of the most important safety items of an automatic oil burner is the primary control. This device will stop the operation when any part of the burner or the heating equipment does not function properly. This control protects against such occurrences as incorrect primary air adjustments, dirt in the atomizing nozzle, inadequate oil supply, and improper

combustion. The most modern type of primary control is called the cadmium cell control. The cadmium sulfide flame detection cell is located in a position where it directly views the flame. If any of the above functional problems should occur, the electrical resistance across the face of the cell would increase, causing the primary control to shut the burner off in 70 s or less.

Draft caused by a chimney or technical means is a very important factor in the operation of domestic oil burners. The majority of burners are designed to fire in a heating unit that has a minus 0.02 in. (0.5 mm) water column draft over the fire. Because the burner develops such low static pressures in the draft tube, over fire drafts are an important factor in satisfactory operation.

If the heating unit is designed to create pressure over the fire, a burner developing high static pressures in the blast tube must be used. These burners develop high static by means of higher rpm on the motor and fan. Burners of this nature are available and are being produced. Flame control is very important. Firing heads similar to Fig. 2 are used on applications of this nature.

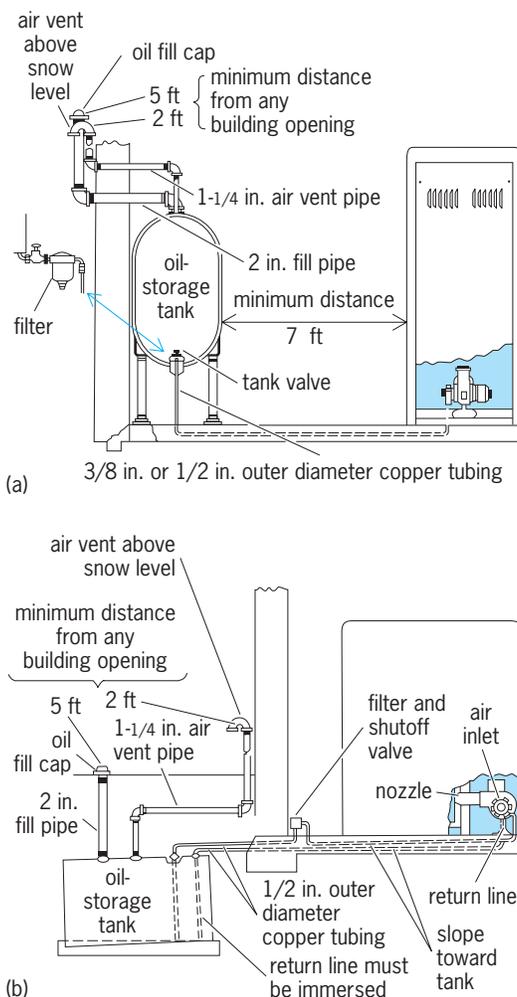


Fig. 6. Two typical installation methods for oil storage tank and oil burner. (a) Inside basement. (b) Outside underground storage tank. 1 ft = 0.3 m. 1 in. = 2.5 cm.

The oil burner is also used for a wide assortment of heating, air conditioning, and processing applications. Oil burners heat commercial buildings such as hospitals, schools, and factories. Air conditioners using the absorption refrigeration system have been developed and fired with oil burners. Oil burners produce CO<sub>2</sub> in greenhouses to accelerate plant growth. They produce hot water for many commercial and industrial applications. See AIR COOLING; COMFORT HEATING; HOT-WATER HEATING SYSTEM; OIL FURNACE.

Robert A. Kaplan

Bibliography. American Society of Heating, Refrigerating, and Air-conditioning Engineers, *ASHRAE Fundamentals Handbook*, 1994, revised 1997; *ASHRAE, Equipment Handbook*, 1992; E. M. Field, *Oil Burners*, 5th ed., 1990.

### Oil furnace

A combustion chamber in which oil is the heat-producing fuel. Fuel oils, having from 18,000 to 20,000 Btu/lb (42 to 47 megajoules/kg), which is equivalent to 140,000 to 155,000 Btu/gal (39 to 43 MJ/liter), are supplied commercially. The lower flash-point grades are used primarily in domestic and other furnaces without preheating. Grades having higher flash points are fired in burners equipped with preheaters. The ease with which oil is transported, stored, handled, and fired is advantageous in small installations. The fuel burns almost completely so that, especially in a large furnace, combustible losses are negligible.

Domestic oil furnaces with automatic thermostat control usually operate intermittently, being either off or operating at maximum capacity. The heat-absorbing surfaces, especially the convection surface, should therefore be based more on maximum capacity than on average capacity if furnace efficiency is to be high. The combustion chamber should provide at least 1 ft<sup>3</sup> for each 1.5–2 lb (1 m<sup>3</sup> for each 24–32 kg) of fuel burned per hour. Gas velocity should be below 40 ft/s (12 m/s). The shape of the chamber should follow the outline of the flame. See FUEL OIL; OIL BURNER; STEAM-GENERATING FURNACE.

Frank H. Rockett

### Oil sand

A loose-to-consolidated sandstone or a porous carbonate rock, impregnated with bitumen, a heavy asphaltic crude oil with a viscosity under reservoir conditions greater than 10,000 millipascal seconds (10,000 centipoise); also known as tar sand or bituminous sand. See ASPHALT AND ASPHALTITE.

Heavy-oil reservoirs are distributed throughout the world and were formerly categorized in many different and sometimes conflicting ways. Under the leadership of the United Nations Institute for Training and Research (UNITAR), modern definitions for these heavy substances have emerged: (1) Bitumens (oil sand hydrocarbons) possess viscosities greater

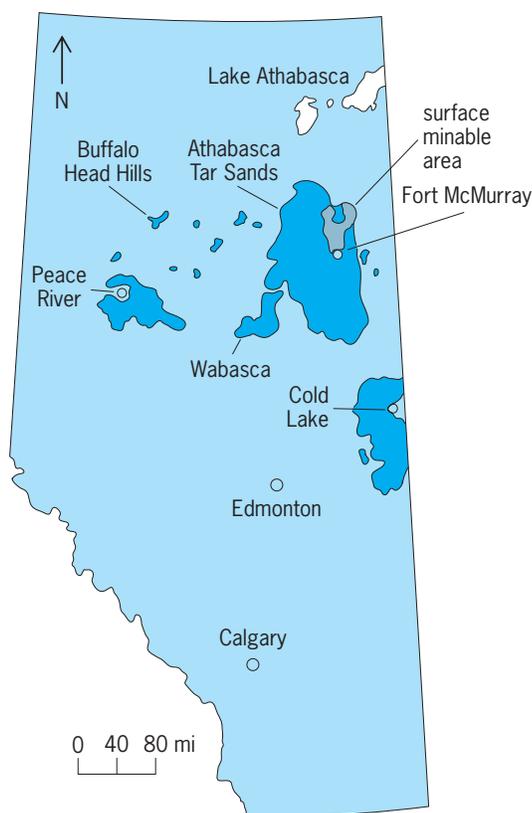
than 10,000 mPa · s (10,000 cP) and cannot be produced by conventional methods. (2) Extra-heavy crude oils possess viscosities equal to or less than 10,000 mPa · s (10,000 cP), and an American Petroleum Institute (API) gravity less than 10°. They are heavier than water but have some mobility under reservoir conditions. (3) Heavy crude oils possess API gravities ranging from 10 to 20°.

About 175 experimental projects for the enhanced recovery of heavy oil and bitumen are in operation around the world and, because similar technologies are being used, reserve and recovery estimates for the three categories—heavy crude, extra-heavy crude, and bitumen—frequently are lumped together.

Estimates of the combined world reserves of heavy crude and bitumen exceed  $5 \times 10^{12}$  bbl ( $8 \times 10^{11}$  m<sup>3</sup>), with bitumen accounting for about half the total. The seven largest of these heavy crude and bitumen deposits contain almost as much oil in place as the world's 300 largest conventional oil fields. The largest proven accumulation of bitumen-containing oil sands occurs in Alberta, Canada, but deposits in Russia rank close behind. Much of the huge accumulation in Venezuela's Orinoco Basin has been reclassified as heavy and extra-heavy crude oil, but this area also, seemingly, contains substantial reserves of bitumen. Smaller reserves occur, in descending order, in the United States, Madagascar, Italy, Albania, Trinidad, and Romania. Estimates of the bitumen that will prove ultimately to be recoverable—generally of the order of 5–25%—are speculative and depend on the development of successful technologies at competitive costs.

**Geological factors.** The major oil sand and heavy oil deposits are remarkable not only for their size but also for their unconventional geologic settings. The deposits generally exhibit certain characteristics. (1) They tend to occur in a deltaic or nonmarine environment, featuring extensive, fingerlike, fluvial sands with large porosities and permeabilities. Porosities, often of the order of 25–35%, tend to be considerably greater than in most petroleum reservoir sandstones (5–20%). A lack of mineral cement, which occupies the void space in most sandstones, permits these high porosities and gives rise to the term oil sands. (2) The deposits are covered by a widespread regional shale cap which restrained upward escape of the oil migrating into the ancient river channels, the efficient systems that gathered the subsurface fluids. (3) The deposits are emplaced in a gently dipping homocline with updip stratigraphic convergence. (4) Degradation of the original, lighter crude to heavy oil has been effected by water washing and bacterial action. Considerable debate continues about the origin of the heavy oils and whether they are thermally mature or immature. The evidence suggests, however, that the oils were degraded in place.

While most of the world's heavy-oil reservoirs are found in deltaic settings, there are significant exceptions. These exceptions include the Cambrian and Lower Permian carbonate rocks, which host the bulk of Russia deposits, and the Paleozoic Carbonate



Tar sand deposits in Alberta. 1 mi = 1.6 km.

Triangle in northern Alberta. See PETROLEUM GEOLOGY.

**Alberta deposits.** Canada's Alberta Oil Sands, comprising four enormous and a number of lesser deposits, contain the best known of the world's oil sand reservoirs. Established resources are of the order of  $1.25 \times 10^{12}$  bbl ( $2 \times 10^{11}$  m<sup>3</sup>). The Athabasca Tar Sands deposit is the largest known petroleum accumulation in the world, with a total area of 13,000 mi<sup>2</sup> (34,000 km<sup>2</sup>) and in-place reserves of  $9.19 \times 10^{11}$  bbl ( $1.46 \times 10^{11}$  m<sup>3</sup>). Cold Lake, rapidly gaining prominence for its in-place projects, contains  $2.05 \times 10^{11}$  bbl ( $3.26 \times 10^{10}$  m<sup>3</sup>); Peace River,  $7.5 \times 10^{10}$  bbl ( $1.2 \times 10^{10}$  m<sup>3</sup>); and Wabasca,  $4.26 \times 10^{10}$  bbl ( $6.8 \times 10^9$  m<sup>3</sup>). The deposits range across the northern part of the province (see *illus.*).

The Athabasca Oil Sands are located in the McMurray Formation, of Lower Cretaceous age. The formation outcrops along the Athabasca River, north of the city of Fort McMurray. Elsewhere, the deposit is buried under a variable layer of overburden which reaches up to 1720 ft (525 m) in thickness. Fortunately, some of the richest parts of the deposit are covered by the thinnest overburden.

**Russian deposits.** Hundreds of bitumen occurrences have been charted in Russia; about 70% are found in carbonate rock. Estimates of reserves approximate  $1.15 \times 10^{12}$  bbl ( $1.8 \times 10^{11}$  m<sup>3</sup>). Of considerable interest is the Olenek anticline in northeastern Siberia where, it is speculated, reserves total  $6 \times 10^{11}$  bbl ( $9.5 \times 10^{10}$  m<sup>3</sup>). Huge deposits have also been discovered near Baku in Kazakhstan

and at Melekes in Volga-Ural. Yarega, in Siberia's Timan-Pechora province, is the only place in the world where underground mining for commercial production of heavy oil is carried out.

**Venezuela deposits.** Enormous reserves of heavy oil occur in a 36-mi-wide (60-km) band stretching along the northern bank of the Orinoco River for a distance of 400 mi (700 km). The deposits, for many years known collectively as the Orinoco Tar Belt, contain between  $7 \times 10^{11}$  bbl ( $1.1 \times 10^{11}$  m<sup>3</sup>) and  $1 \times 10^{12}$  bbl ( $1.6 \times 10^{11}$  m<sup>3</sup>) of heavy hydrocarbons. Some of the crude possesses an API gravity of between 10 and 20°, but most is between 4 and 10°. Most of the heaviest oil, however, shows some mobility because of the below-normal asphaltene content, the temperature of the reservoir, and the presence of dissolved gases. For this reason, the Venezuelan government encourages the use of the term Orinoco Heavy Oil Belt, instead of Tar Belt, to describe these deposits. It is believed, though, that substantial reserves, perhaps as much as one-third of the total hydrocarbons, are too viscous to flow to the well bore and may therefore be categorized as bitumen, or tar oil.

**United States deposits.** In the United States, 550 tar sand occurrences are known to exist in 22 states, and it is expected the number will increase significantly in future surveys. Seven states (Alabama, California, Kentucky, New Mexico, Texas, Utah, and Wyoming) possess deposits aggregating  $1 \times 10^6$  bbl ( $1.6 \times 10^5$  m<sup>3</sup>) or more. Total United States reserve estimates have remained relatively constant since the mid-1970s at approximately  $3 \times 10^{10}$  bbl ( $4.8 \times 10^9$  m<sup>3</sup>), but considerable changes have occurred in the amounts ascribed to the various states.

Utah is the most important tar sand state with an estimated  $1.47 \times 10^{10}$  bbl ( $2.3 \times 10^9$  m<sup>3</sup>) of bitumen in place, or about half of the United States total. At one time, though, it was believed that more than 90% of United States tar sand reserves were located there. Over 50 deposits have been identified, but the vast bulk of the oil occurs in six giant deposits, four of which are in the Uinta Basin of northeastern Utah. Three of these (P. R. Spring, Hill Creek, and Sunnyside) occur in the Green River Formation (Eocene), but Asphalt Ridge, probably the best-known tar sand deposit in the country, occurs in older strata.

The Tar Sand Triangle, with reserves of approximately  $6.1 \times 10^9$  bbl ( $9.6 \times 10^8$  m<sup>3</sup>), the largest deposit in the United States, lies in a remote and very rugged area of northeastern Utah, in the White Rim Sandstone Formation. Most of the deposit underlies Federal Recreation and Wilderness areas. Federal agencies also control major sections of the P. R. Spring, Hill Creek, and Sunnyside deposits. The other major ore body, Circle Cliffs, is also located in southeastern Utah.

The thickness of the Utah reservoirs ranges between 0 and 330 ft (0 and 100 m), and the depth ranges from 0 to 2200 ft (0 to 670 m). Some interesting characteristics distinguish the reservoirs: the bitumen occurs in consolidated sandstone; the connate water, at 0.6%, and the sulfur content of the

hydrocarbon, at 0.5%, are about one-tenth of the corresponding values in the Canadian and Venezuelan oil sands. However, the porosity of the formations and the continuity and magnitude of the bitumen saturation are far less. Economic exploitation is therefore in question.

In terms of United States reserves, Alabama has moved into second place, with  $4.3 \times 10^9$  bbl ( $6.8 \times 10^8$  m<sup>3</sup>) attributed to the North Area. Texas (Uvalde area) and California are next, each with  $3 \times 10^9$  bbl ( $4.8 \times 10^8$  m<sup>3</sup>).

In California, the reserves are contained in a number of well-known deposits, Santa Maria, with  $2 \times 10^9$  bbl ( $3.2 \times 10^8$  m<sup>3</sup>), being the largest. The Edna deposit, located midway between Los Angeles and San Francisco, is a rarity among oil sand deposits in that it is considered a marine facies. The Sisquoc deposit is amenable to the hot water extraction process used by the large Canadian projects. The fossiliferous McKittrick deposit is more properly considered a diatomaceous oil shale. See OIL SHALE.

**Other major world deposits.** The Bemolanga deposit in western Madagascar is noteworthy. Covering approximately 154 mi<sup>2</sup> (400 km<sup>2</sup>), it contains reserves estimated at  $2.1 \times 10^{10}$  bbl ( $3.3 \times 10^9$  m<sup>3</sup>). Another country with multibillion-barrel oil sand deposits is Italy; the Ragusa area in Sicily contains  $1.4 \times 10^{10}$  bbl ( $2.2 \times 10^9$  m<sup>3</sup>) of bitumen. G. Ronald Gray

**Bibliography.** C. J. Borregales, Production characteristics and oil recovery in the Orinoco Oil Belt, *UNITAR 1st International Conference on the Future of Heavy Crude and Tar Sands*, Edmonton, Alberta, 1979; M. Danyluk, B. Galbraith, and R. Omana, Department of Defense, Toward definitions for heavy crude oil and tar sands, *United Nations Institute of Training and Research (UNITAR) 2d International Conference on the Future of Heavy Crude and Tar Sands*, Caracas, Venezuela, 1982; G. J. Demaison, Tar Sands and Supergiant Oil Fields, Can. Inst. Min. Metallurgy Spec. Vol. 17, 1977; V. A. Kuuskraa, S. Chalton, and T. M. Doscher, *The Economic Potential of Domestic (U.S.) Tar Sands*, U.S. Department of Defense, January 1978; R. F. Meyer, P. A. Fulton, and W. D. Dietzman, A preliminary estimate of world heavy crude oil and bitumen resources, *UNITAR 2d International Conference on the Future of Heavy Crude and Tar Sands*, Caracas, Venezuela, 1982; G. D. Mossop, *Geology of the Athabasca Oil Sands*, SCIENCE Reprint Series, January 1980.

## Oil shale

A sedimentary rock containing solid, combustible organic matter in a mineral matrix. The organic matter, often called kerogen, is largely insoluble in petroleum solvents, but decomposes to yield oil when heated. Although "oil shale" is used as a lithologic term, it is actually an economic term referring to the rock's ability to yield oil; oil shale appears to be the cheapest source after natural petroleum for large amounts of liquid fuels. No real minimum oil yield or content of organic matter can be established to dis-

tinguish oil shale from sedimentary rocks. Additional names given to oil shales include black shale, bituminous shale, carbonaceous shale, coaly shale, cannel shale, cannel coal, lignitic shale, torbanite, tasmanite, gas shale, organic shale, kerosine shale, coorongite, maharahu, kukersite, kerogen shale, algal shale, and "the rock that burns." See KEROGEN.

**Origin.** Oil shale is lithified from lacustrine or marine sediments relatively rich in organic matter. Most sedimentary rocks contain small amounts of organic matter, but oil shales usually contain substantially more. Specific geochemical conditions are required to accumulate and preserve organic matter, and these were present in the lakes and oceans whose sediments became oil shale. R. M. Garrels and C. L. Christ define these conditions in terms of oxidation-reduction potential and acid-base condition (pH) of the water in and around the sediment. Organic matter accumulates under the strongly reducing conditions and neutral or basic pH present in euxinic marine environments and organic-rich saline waters. The organic-rich sediments which became oil shale accumulated slowly in water isolated from the atmosphere. This isolation was achieved through the stratification of the water body. The stratification limited clastic mineral input to the sediment and also protected the sediment from oxidation and from physical disturbance. See MARINE SEDIMENTS; PH.

**Mineral composition.** Quartz, illite, and pyrite (sometimes with marcasite and pyrrhotite) occur in virtually every oil shale. Other clays (particularly montmorillonite) are found in many oil shales, as are feldspar minerals. Most oil shale deposits contain small amounts of carbonate minerals, but some, most notably the Green River Formation in Colorado, Utah, and Wyoming, contain large amounts of dolomite, calcite, and other carbonates. The oil shale minerals were formed probably in the sediment by chemical processes related, at least in part, to the presence of organic matter. See CALCITE; CARBONATE MINERALS; DOLOMITE.

Some oil shales, particularly those called black shales because of the coal-like color of their organic matter, have tended to become enriched in trace metals. The reducing conditions necessary to preserve organic matter were conducive to precipitating available trace metals, frequently as sulfides. The Kupferschiefer of Mansfield, Germany, contains an unusually large content of copper, and the Swedish Alum Shale has been exploited for its uranium content. The Devonian Chattanooga Shale of Tennessee and neighboring states contains an average of about 0.006 wt % uranium and has been extensively studied as a potential low-grade source of this element. Vanadium in potentially commercial amounts occurs in the Permian Phosphoria Formation of Wyoming and Idaho. Enrichment of arsenic, antimony, molybdenum, nickel, cadmium, silver, gold, selenium, and zinc has also been noted in black oil shales.

**Physical properties.** Oil shales are fine-grained rocks generally with low porosity and permeability. Many are thinly laminated and fissile. The colors of

oil shales, which range from black to light tan, are produced or altered by organic matter.

The physical properties of oil shale are influenced strongly by the proportion of organic matter in the rock. The rock's decrease in density with increasing organic content illustrates this most graphically. The mineral components possess densities of about 2.6–2.8 g/cm<sup>3</sup> for silicates and carbonates and 5 g/cm<sup>3</sup> for pyrite, but the density of organic matter is near 1 g/cm<sup>3</sup>. Larger fractions of organic matter produce rocks with appreciably lower density. Equation (1)

$$D_T = \frac{D_A D_B}{A(D_B - D_A) + D_A} \quad (1)$$

quantifies this relationship. Here  $D_T$  = density of the rock,  $A$  = weight fraction of organic matter,  $D_A$  = average density of organic matter, and  $D_B$  = average density of mineral matter.

The volume of organic matter in oil shale rock strongly influences its physical properties. The marked increase in volume fraction of organic matter in oil shale with increasing organic weight fraction can be demonstrated by using the above equation. For example, in Green River Formation oil shale the average density of the organic matter ( $D_A$ ) is about 1.07 g/cm<sup>3</sup>, and the average density of the mineral matter ( $D_B$ ) is about 2.72 g/cm<sup>3</sup>. Thus the density of an oil shale containing 4 wt % organic matter (a shale yielding about 6 gal of oil per ton of shale) is calculated to be 2.56 g/cm<sup>3</sup>. The organic matter occupies about 10 vol % of this rock. An oil shale containing 15 wt % organic matter (a shale yielding about 26 gal per ton or 110 liters per metric ton) possesses a density of 2.21 g/cm<sup>3</sup>, and the organic matter occupies 31 vol % of this rock. An oil shale containing 39 wt % organic matter (a shale yielding about 52 gal per ton or 220 liters per metric ton) possesses a density of 1.86 g/cm<sup>3</sup>, and the organic matter makes up about 52 vol % of the rock's volume. In the Green River Formation, in oil shales containing 15 wt % or more organic matter, the organic material is the largest component by volume, and the physical properties of the organic matter predominate in determining the physical properties of the rock. The organic matter makes this rock tough, resilient, and difficult to crush. The richer Green River Formation oil shales tend to deform plastically under load. The volume of organic matter in oil shales can be used to estimate mechanical properties of rocks in oil shale deposits.

Equation (1) can yield a relationship between oil yield and oil shale density by incorporating a factor for conversion of the organic matter to oil. The resulting relationship is useful in calculating resources and reserves in an oil shale deposit. Such relationships have been worked out for Green River Formation oil shales. Similar relationships can be developed for most oil shale deposits.

**Organic composition.** The organic matter in oil shales and other sedimentary rocks has been studied extensively by organic geochemists, but a specific description of it has not been produced. Although

**Relationship between the organic carbon-to-organic hydrogen ratio and the conversion of oil shale organic matter to oil by heating**

Deposit sampled	Carbon-hydrogen ratio	Organic carbon converted, wt %
Pictou County, Nova Scotia, Canada	12.8	13
Top Seam, Glen Davis, Australia	11.5	26
New Albany Shale, Kentucky, United States	11.1	33
Ermelo, Transvaal, South Africa	9.8	53
Cannel Seam, Glen Davis, Australia	8.4	60
Garfield County, Colorado, United States	7.8	69

some oil shales contain recognizable organic fragments like spores or algae, most do not, because the reducing conditions associated with oil shale deposition digested and homogenized the organic debris. The resulting organic matter (kerogen) is best described as a high-molecular-weight organic mineraloid of indefinite composition. This composition varies from deposit to deposit and is influenced by the depositional conditions and the nature of the organic debris. Variations in the hydrogen content of this organic matter are significant, because the fraction of organic carbon converted to oil on heating increases as the amount of hydrogen available in the organic matter increases. To illustrate this relationship, the **table** compares the proportion of organic carbon converted and recovered as oil during Fischer assay with the weight ratio of organic carbon to organic hydrogen in several oil shales. For petroleum, the carbon-hydrogen values range from 6.2 to about 7.5; for coal, they range upward from 13. The carbon-hydrogen values for organic matter in the world's oil shales range from near petroleum to near coal.

Analytical determination of the elemental composition of the organic matter has been difficult because of the heterogenous nature of oil shales. Carbon, hydrogen, sulfur, oxygen, and nitrogen are the major elements of the organic matter; but they also occur in the mineral matter of oil shales. The organic matter and the mineral matter in oil shales are difficult to separate from each other either physically or chemically. Analytical techniques designed to distinguish between organic and mineral forms of elements, specialized organic matter enrichment techniques, and other specialized evaluation techniques have been developed to aid in the study of oil shales.

**Oil production.** The Fischer assay is the best known of the specialized analytical procedures used for oil shale. It was developed by the U.S. Bureau of Mines for oil shale resource evaluation and has become a standard ASTM (American Society for Testing and Materials) method. This method, employing a modified Fischer retort, determines quantities of oil products recoverable from an oil shale sample heated under

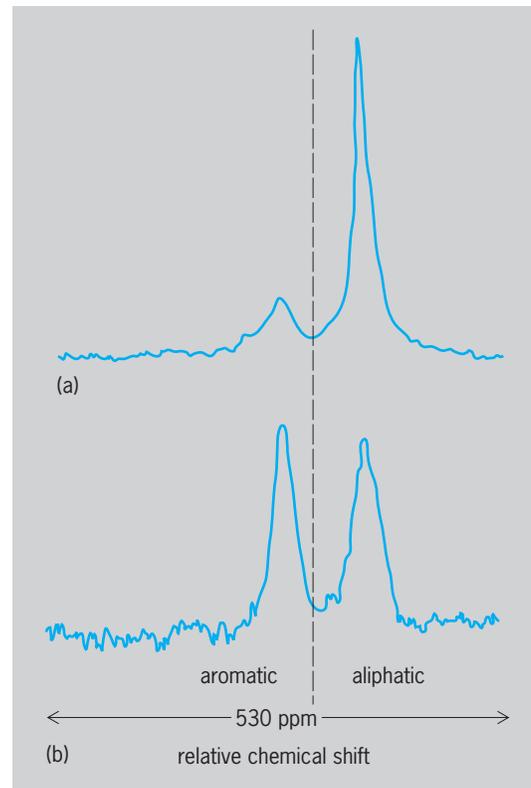
prescribed conditions. Although the procedure does not measure the total amount of organic matter in the sample, it approximates the oil available by commercial operations. This simple procedure has proved suitable for oil shale evaluation purposes. Resource information for United States oil shales is based on the Fischer assay oil-yield data accumulated by the U.S. Bureau of Mines.

Nuclear magnetic resonance (NMR) on solid oil shale samples has been applied to evaluating the fraction of organic carbon in an oil shale which can be converted to oil by heat. This conversion fraction, correlated above with the laboriously determined organic hydrogen, works with the quantity of organic matter in an oil shale resource to define how much rock must be heated to generate production quantities of oil. The carbon-13 NMR test divides the organic carbon into hydrogen-rich (aliphatic) and hydrogen-poor (aromatic) fractions. The proportion of aliphatic carbon correlates very well with the proportion of organic carbon recovered in the oil produced under the standard heating conditions like Fischer assay. Equation (2) is used for this correla-

$$\text{CONV} = 1.37 \cdot \text{AL} - 0.41 \quad (2)$$

tion in United States oil shales, where CONV is the fraction of organic carbon converted to oil and AL is the fraction of organic carbon defined by NMR as occurring in aliphatic compounds. **Figure 1** illustrates this easily interpreted NMR characterization of organic carbon for a Green River Formation oil shale from Colorado and a Devonian Black Shale from Kentucky. The  $x$  axis measures the relative chemical shift generated by the organic matter in the oil shale; the shift is a function of proportion of available hydrogen in the organic matter of the oil shale. The Colorado sample shows much greater aliphatic content and a correspondingly greater conversion of organic carbon to oil. See NUCLEAR MAGNETIC RESONANCE (NMR).

**Oil shale units.** The United States expresses oil yield determined by Fischer assay in the volume unit of gallons of oil per ton of shale. This unit is converted from weight percent oil yield by multiplying by a conversion factor, 2.4, and dividing the result by absolute oil specific gravity, determined at 15.6°C (60.1°F). The metric volume unit, liters per metric ton, corresponds to gallons per ton. One gallon per ton is equivalent to 4.172 liters/metric ton. Oil shale resource values are expressed in barrels of oil in the United States. A barrel of oil, the 42-gal volume unit used in Western petroleum commerce, has no direct equivalent in metric countries. However, the SI unit equivalent to a barrel of oil represents 0.159 m<sup>3</sup>. Specification of oil density, a variable, is necessary to convert the volume unit, barrels, into the metric weight trade unit, metric ton. The 1975 World Energy Conference agreed to define a barrel of oil as 0.145 metric ton, and 1 gal/ton  $\times$  0.29 as 1 kg/metric ton, approximations ignoring density variations in oil. Gallons of oil per ton of shale will be the unit of oil yield used in the remainder of this article.



**Fig. 1.** C-13 nuclear magnetic resonance characterization of organic matter in two United States oil shale samples. (a) Colorado oil shale, 80% aliphatic carbon content with 69% conversion of organic carbon to oil. (b) Kentucky oil shale, 52% aliphatic carbon content with 30% conversion of organic carbon to oil. The relative area under the two peaks provides the proportion of the aliphatic carbon in the oil shale's organic carbon. The  $x$  axis is a measurement of the chemical shift of the organic carbon in the oil shale; the shift can be related to the proportion of available hydrogen in the organic matter of the oil shale. (After F. P. Miknis and J. W. Smith, *An NMR survey of United States oil shales*, *Geochem.*, 5:93-201, 1984)

**World resources.** The world's oil shale deposits represent a tremendous store of fossil energy. It has been estimated that the organic matter in sedimentary rocks contains  $1.2 \times 10^{16}$  tons ( $1.1 \times 10^{16}$  metric tons) of organic carbon, nearly 1000 times that found in coals. Although part of that organic carbon has matured to produce oil and gas, most of it is still oil shale. Unfortunately, most of this tremendous resource is not well known. Oil shales occur on every continent in sediments ranging in age from Cambrian to Tertiary. **Figure 2** illustrates this by showing distribution and geologic age of oil shales in the United States. More than 20% of the United States land area is underlain by oil shales, but only the Tertiary Green River Formation and the Devonian black shales, particularly those in Kentucky, have been evaluated in any detail. The balance of the deposits shown are known from associated geologic studies but not as oil shale resources. With the limited amount of information available, one estimate of known oil shale resources of the world was made, and to it was added possible extensions of known resources and geologically based estimates of unappraised and undiscovered

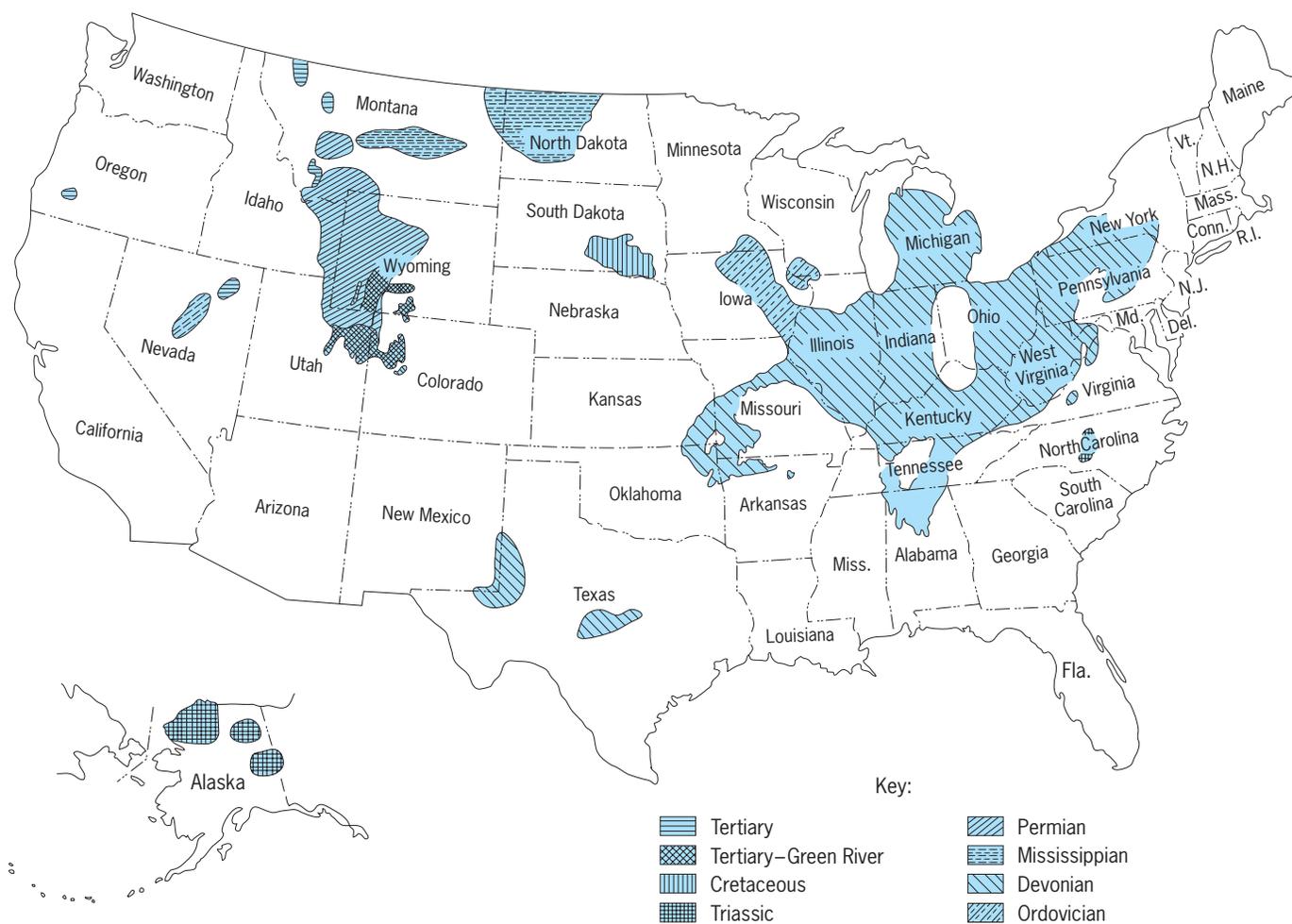


Fig. 2. Oil shale deposits of the United States. (Courtesy of John Ward Smith)

resources to obtain order-of-magnitude numbers for the total in-place oil resource in the world's oil shale deposits. These estimates for the total oil resource in shales of all grades reached  $1.75 \times 10^{15}$  barrels. How much of this tremendous resource could actually be produced economically is a variable which changes continuously with the supply of natural oil and developments in technology. However, just 1% of that total shale oil represents more oil than the world is expected to produce as natural petroleum ( $2 \times 10^{12}$  bbl). Oil shale represents a tremendous supply of liquid fuels.

**World developments.** Although the oil potential of the world's oil shales is great, commercial production of this oil has been considered uneconomic generally. Oil shales are lean ores, producing only limited amounts of oil which historically has been low in price. Mining and heating 1 ton of relatively rich oil shale yielding 25 gal/ton produces only 0.6 bbl of oil.

In special situations when other fuels were in short or uncertain supply, or when energy transportation was difficult, energy development from oil shales has been carried out commercially. World War II caused sharp increases in petroleum demand and disrupted both petroleum production and petroleum distribution. Oil shale production operations dur-

ing and since World War II have been conducted in Germany, France, Spain, Manchuria (China), Estonia and other areas of the former Soviet Union, Sweden, Scotland, South Africa, Australia, and Brazil. *See ENERGY SOURCES; MINING.*

**Shale oil.** Shale oil is produced from the organic matter in oil shale when the rock is heated in the absence of oxygen (destructive distillation). This heating process is called retorting, and the equipment that is used to do the heating is known as a retort. The rate at which the oil is produced depends upon the temperature at which the shale is retorted. Most references report retorting temperatures as being about  $500^{\circ}\text{C}$  ( $930^{\circ}\text{F}$ ).

Retorting temperature affects the nature of the shale oil produced. Low retorting temperatures produce oils in which the paraffin content is greater than the olefin contents; intermediate temperatures produce oils that are more olefinic; and high temperatures produce oils that are nearly completely aromatic, with little olefin or saturate content.

In general, shale oils can be refined to marketable products in modern petroleum refineries. There is no really typical shale oil produced from Green River oil shale, but the oils do share many properties in common. They usually show high pour points,  $20\text{--}32^{\circ}\text{C}$

(68–90°F); high nitrogen contents, 1.6–2.2 wt %; and moderate sulfur contents, about 0.5 wt %. High pour points make necessary some processing before the oils are amenable to pipeline transportation. The high nitrogen contents make hydrogenation necessary to reduce the nitrogen contents so that the oils can be processed into fuels. Hydrogenation also reduces the sulfur content. See DESTRUCTIVE DISTILLATION; HYDROGENATION.

**United States technology.** The two general approaches to recovering shale oil from Green River Formation oil shales are (1) mining, crushing, and aboveground retorting, called conventional processing; and (2) in-place processing. The basic problems facing conventional processing are handling and heating huge amounts of low-grade ore and disposing of huge volumes of spent shale. The in-place approach largely avoids the problems of handling and disposal but faces a different basic problem—the impermeability of the oil shale beds.

With the conventional approach, oil shale mining by the room-and-pillar technique developed by the U.S. Bureau of Mines appears capable of producing the huge amounts of ore necessary to operate a large production plant. The procedure has also been tested by industry in Mahogany Zone shales. Outputs on the order of 2500 tons (2250 metric tons) of oil shale per worker shift have been reported from the highly mechanized operation. Crushing technology is well demonstrated.

Retorting must be done continuously in order to reach the throughput necessary for economic production of shale oil. Two general systems for heating a continuous stream of oil shale are outlined in Fig. 3. In the internally heated system the oil shale furnishes its own heat because part of its organic matter is burned inside the retort. Examples of the internally fired retort system include: the Bureau of Mines gas combustion retort; one form of the Paraho retort of Development Engineering, Inc.; one form of the Union Oil Company's rock-pump retort; and Superior Oil Company's moving-grate retort. The TOSCO Corporation's TOSCO II retort, in which preheated

ceramic balls heat the oil shale; the SGR rock-pump retort of Union Oil; one form of the Paraho retort and the Petrosix pilot retort, in which preheated gas heats the oil shale; and the Lurgi retort, in which preheated sand or spent shale heats the oil shale, are all examples of the externally fired retort system. Several retorting systems, including both internally and externally heated designs, have been tested on pilot or semiworks scales.

Spent shale disposal has been studied intensively. More than 80% of the mined oil shale remains as residue after oil production. An oil shale plant which produces 50,000 bbl of oil per day from Mahogany Zone shale might mine 75,000 tons (68,000 metric tons) of rock and dispose of 60,000 tons (54,000 metric tons) of spent shale daily. In the vast and largely unpopulated areas of the Green River Formation, dumping these volumes of spent shale is not as large an environmental problem as it appears to be. The spent shale resulting from most surface retorts is largely insoluble, consisting largely of the same native minerals being exposed in huge quantities to natural weathering by the existing extensive erosion of the Green River Formation. Contoured dumping to control water flow will minimize leaching, already low in an arid region. Native vegetation will establish itself on spent shale dumps, and revegetation procedures can accelerate this process. Returning spent shale to the mine to furnish roof support may permit recovery of additional ore, but this approach has not been tested.

Research and development efforts toward in-place processing have concentrated on creating permeability in the impermeable oil shale. In-place processing may be accomplished by two means: (1) a borehole technique in which oil shale is first fractured underground and heat is applied, and (2) a process in which some rock is first removed by mining, then the remaining oil shale is fragmented into the voids created by mining, and finally heat is applied. These two methods are referred to as in-place (no mining) and modified in-place (some mining) processing. Several investigators have field-tested in-place methods. Both of the heating methods outlined for retorts (Fig. 3) have also been applied to in-place processing. Proposals for in-place processing include solvent extraction and radio-frequency electrical heating. See SOLVENT EXTRACTION.

Only one-third of the total Green River Formation oil shale resource is considered minable for conventional processing. In-place processing may make much of the remainder of the resource available to production. Conventional processing offers process control advantages, including ready adaptation of many existing industrial procedures, but it is capital-intensive, requiring huge investments before production begins. In-place processing is less easily controlled and evaluated but requires less capital outlay before production begins. See ENERGY SOURCES; MINING.

John Ward Smith; Howard B. Jensen

Bibliography. V D. Allred (ed.), *Oil Shale Processing*, 1982; American Society for Testing and Materials, *Standard Test Method for Oil from Oil*

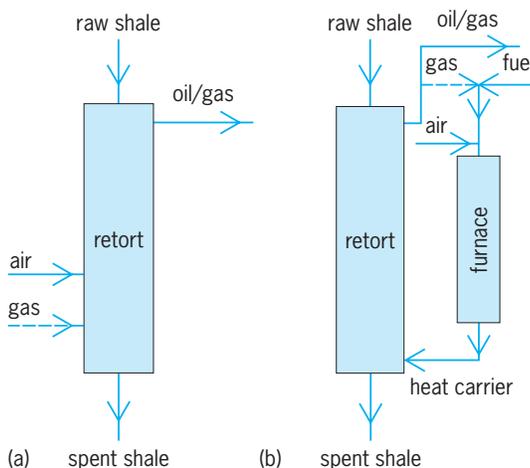


Fig. 3. Oil shale retorting systems: (a) internally heated, (b) externally heated.

*Shale*, Procedure D3904-80, 1980; K. P. Chong and J. W. Smith, *Mechanics of Oil Shale*, 1984; Colorado School of Mines, *1st through 17th Oil Shale Symposium Proceedings*, 1963-1984; B. Durand (ed.), *Kerogen: Insoluble Organic Matter from Sedimentary Rocks*, Technip, Paris, 1980; R. M. Garrels and C. L. Christ, *Solutions, Minerals, and Equilibria*, 1965, reprint 1985; R. A. Meyers (ed.), *Handbook of Syn-fuels Technologies*, 1984; F. P. Miknis and J. F. McKay (eds.), *Geochemistry and Chemistry of Oil Shales*, ACS Symp. Ser. 230, 1983; *Oil Shale: Prospects and Constraints*, Federal Energy Administration Project Independence Blueprint, GPO 4118-00016, 1974; L. C. Ruedisili and M. W. Firebaugh (eds.), *Perspectives on Energy*, 3d ed., 1982; J. W. Smith, *Oil Shale Resources of the United States*, Colorado School of Mines Publications, 1981; J. W. Smith, *Theoretical Relationship Between Density and Oil Yield for Oil Shales*, U. S. Bur. Mines Rep. Inv. 7179, 1969.

## Oil field waters

Waters of varying mineral content which are found associated with petroleum and natural gas or have been encountered in the search for oil and gas. They are also called oilfield brines, or brines. They include a variety of underground waters, usually deeply buried, and have a relatively high content of dissolved mineral matter. These waters may be (1) present in the pore space of the reservoir rock with the oil or gas, (2) separated by gravity from the oil or gas and thus lying below it, (3) at the edge of the oil or gas accumulation, or (4) in rock formations which are barren of oil and gas. Brines are commonly defined as water containing high concentrations of dissolved salts. Potable or fresh waters usually are not considered oilfield waters but may be encountered, generally at shallow depths, in areas where oil and gas are produced.

Oilfield waters or brines differ widely in composition and concentration. They may differ from one geologic province to another, from one formation to another within a given geologic province, or from one part of a specific geologic horizon to another. They range from slightly salty water with 1000-3000 parts of dissolved substances in 1 million parts of solution to very nearly saturated brines with dissolved mineral content of more than 270,000 parts per million (ppm).

The most common and abundant mineral found in oilfield waters is sodium chloride, or common table salt. Calcium chloride is next in order of abundance. Carbonates, bicarbonates, sulfates, and the chlorides of magnesium and potassium are present in lesser quantities. In addition to the above mentioned salts, salts of bromine and iodine are also found. Traces of strontium, boron, copper, manganese, silver, tin, vanadium, and iron have been reported. Barium has been reported in many of the Paleozoic brines of the Appalachian region. The commercial value of a brine depends upon the concentration of salts, purity of the products to be recovered, and value and prac-

ticability of by-product recovery. Concentrations less than 200,000 ppm are seldom of commercial interest.

Slightly salty waters, while not suitable for human consumption, may be used in some industrial processes or may be amenable to beneficiation for municipal supplies in areas lacking fresh waters.

Classified genetically, oilfield waters are generally considered connate; that is, they are seawaters which (presumably) originally filled the pore spaces of the rock in which they are now confined. However, few analyses of these waters correspond to present-day seawater, thus indicating some mixing and modification since confinement. Dilute solutions suggest that rainwater has percolated into the rocks along bedding planes, fractures, faults, and other permeable zones. Presence of carbonates, bicarbonates, and sulfates in an oilfield water further suggests that at least some of the water had its origin at the surface. Concentrations of dissolved solids greater than that of modern seawater suggest partial evaporation of the water or addition of soluble salts from the adjacent or enclosing rocks.

Waters in most sedimentary rocks increase in mineral concentration with depth. This increase may be due to the fact that, since salt water is heavier than fresh water, the more dense solution will eventually find a position as low as possible in the aquifer. An additional factor would be the longer exposure of the deeper waters to the mineral-bearing rocks. Exceptions have been noted and probably are due to the presence of larger quantities of soluble salts in some geological formations than in others.

Probably the most important geological use of oilfield water analyses is their application to the quantitative interpretation of electrical and neutron well logs, particularly micrologs. In order to compute the connate water saturation of a formation in a quantitative manner from electrical data, it is necessary to know with accuracy the connate water resistivity.

Naturally mineralized waters are frequently the only waters available for water-flooding operations. Water analyses are useful in predicting the effect of the water on minerals in the reservoir rock and on the mechanical equipment employed on the project. Waters which exert a corrosive action on the lines and pumps or which tend to plug up the pay zone are not suitable for water-flooding operations.

Oil field water composition may be an important factor in the determination of the source of water in oil wells which have leaky casings or improper completions with resulting communication between wells, and in identifying and correlating reservoirs in multipay oil pools, particularly in those containing lenticular sand bodies.

Industrial wastes, including mineralized water produced with oil, may be disposed of in underground reservoirs. Between the zone of potable water and the horizon of commercial brines, there commonly are rock formations, the waters of which contain chemicals in amounts sufficient to make the waters unsuitable for domestic, municipal, industrial, and livestock consumption, but not in sufficient

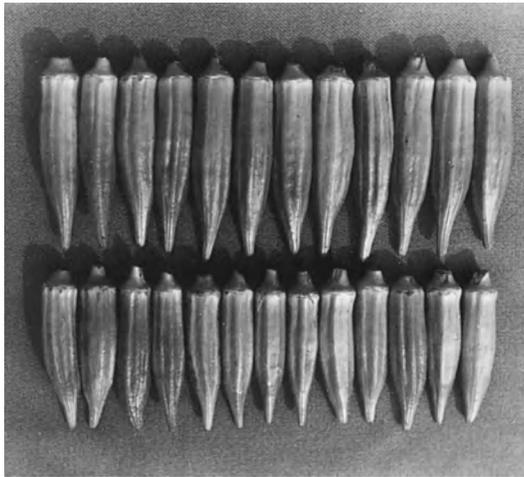
quantity to be considered as a source for recovery of chemicals. Provided there is sufficient porosity and permeability, these rock formations could receive industrial wastes which would otherwise contaminate surface streams and shallow, fresh groundwater horizons into which they might be discharged. See GEOPHYSICAL EXPLORATION; PETROLEUM GEOLOGY.

Preston McGrain

Bibliography. A. G. Collins, *Geochemistry of Oil-field Waters (Developments in Petroleum Science)*, 1975; J. Hunt, *Petroleum Geochemistry and Geology*, 2d ed., 1995.

## Okra

A warm-season annual, *Hibiscus esculentus*, of Ethiopian origin. Okra, also called gumbo, is grown for its immature pods (see *illus.*), which are



Okra pods. (Asgrow Seed Co., subsidiary of The Upjohn Co.)

generally used for preparing soups but are also eaten as a freshly cooked vegetable. It is a member of the order Malvales and is related to cotton. Propagation is by seed. Popular varieties are Clemson Spineless and Green Velvet. Okra is sensitive to low temperatures; commercial production in the United States is primarily in the South. Harvesting begins when the pods are 3–4 in. (7.5–10 cm) long, usually 50–60 days after planting. Georgia, Florida, and Louisiana are important producing states. See COTTON; MALVALES.

H. John Carew

## Olbers' paradox

The riddle of why is the sky is dark at night. This celebrated riddle originated in the sixteenth century. In 1823, Wilhelm Olbers presented it in the simplest terms: In an infinite universe, populated everywhere with stars, a line of sight in any direction when extended out into space must ultimately intercept the surface of a star. Hence stars should cover the entire sky. And if all stars are sunlike, the sky at

every point should blaze as brightly as the disk of the Sun. Olbers has been credited incorrectly with the discovery of the riddle; he did, however, express it in this lucid form, and showed that the riddle still holds even when stars are irregularly distributed in clusters.

**Nature of the riddle.** A boundless universe of either flat or curved space, in which luminous stars stretch away endlessly, seems a not unreasonable cosmological picture. Yet it leads to a conclusion in total contradiction with experience. The sky at night is dark and not a continuous blaze of bright stars. A calculation, first made in 1744 by Jean-Philippe Loys de Chéseaux, showed that starlight should be 90,000 times brighter than sunlight at the Earth's surface. Chéseaux showed that in an infinite universe, uniformly populated with stars, only a finite number of stars are seen. Visible stars cover the sky out to a distance (the "background distance") where they fuse together and form a continuous background; the stars at greater distances are covered over by the foreground stars and cannot be seen.

A forest analogy helps to understand the riddle. Every (horizontal) line of sight in an endless forest terminates at a tree trunk. Wherever we stand we find ourselves surrounded by trees stretching away into a continuous background. The distance of the background is the average area occupied by a tree and its surroundings halfway to the nearest neighboring trees divided by the typical thickness of a tree trunk at eye level. (For example, the background distance is 50 m if trees are separated from one another by distances of 5 m and have trunk diameters of 0.5 m.)

**Solutions.** More than a dozen solutions have been proposed since the riddle was first discovered by Thomas Digges in 1576. The riddle has two interpretations, and the proposed solutions are therefore of two kinds.

*Missing star light.* The first interpretation assumes that the argument is correct and the sky is indeed covered with stars, most of which cannot be seen. The riddle then asks: What happened to the missing starlight? Both Chéseaux and Olbers adopted this interpretation and proposed that starlight is slowly absorbed as it travels large distances in space. The analogy in this case is the foggy forest; only foreground trees are seen and fog obscures background trees. This solution fails because the interstellar absorbing medium quickly heats up and then emits the absorbed radiation. Hermann Bondi in the 1950s also adopted this first interpretation and argued that the light from distant stars is redshifted into invisibility by the expansion of the universe. The redshift solution contributed greatly to the popularity of cosmology, and for several years it was believed that darkness at night proved that the universe is expanding. But Bondi's redshift solution is incorrect in a big bang universe and applies only in the steady-state expanding universe that was disproved in 1965 by the discovery of the radiation from the big bang. See BIG BANG THEORY; COSMIC BACKGROUND RADIATION; HEAT RADIATION.

*Missing stars.* The second interpretation assumes that the sky is not covered with stars, in agreement with observation, and therefore the assertion that every line of sight intercepts the surface of a star is misleading. The riddle, in effect, now asks: What do lines of sight intercept when we look at the dark gaps between stars? Johannes Kepler in the early seventeenth century adopted this second interpretation and argued that we look out between a finite number of stars and see a dark surface enclosing the universe. The analogy is a finite forest enclosed by a high wall. Another solution, consistent with this interpretation and popular with astronomers in the late nineteenth and early twentieth centuries, was the argument that our Galaxy—the Milky Way—is the only galaxy in the universe, and we look out between the stars to an empty, dark, and infinite space beyond. The corresponding analogy is a finite forest, and we look out between the trees to a treeless plain beyond. Edgar Allen Poe suggested in 1848 that the universe is not old enough for the light from very distant stars to have reached us; hence lines of sight that fail to intercept stars reach back to the darkness that prevailed before the birth of stars. This was investigated by Lord Kelvin in 1901. He assumed that stars are no older than 100 million years, and showed that out to a distance traveled by light in 100 million years, visible stars are insufficient in number to cover the whole sky. The missing stars are actually out there, but we look out in space and back in time to a prenatal era before the stars were born.

Modern calculations confirm and extend Kelvin's conclusions, and the second interpretation of Olbers' paradox is hence correct: Most lines of sight do not intercept stars but extend back to the beginning of the universe. Light travels at approximately 300,000 km/s (186,000 mi/s), and a static universe  $10\text{--}20 \times 10^9$  years old is not old enough for starlight to reach the Earth from regions sufficiently distant for visible stars to cover the sky. If the sky at night is dark in a static universe, then obviously in an expanding universe of the same age the night sky is even darker because of the redshift of starlight. Calculation shows that the redshift effect is in fact relatively small.

**Big bang universe.** Olbers' paradox, night has been solved by modern cosmology. In the big bang universe,  $10\text{--}20 \times 10^9$  years old, we cannot see sufficient stars to cover the sky. Instead, on looking out in space, we look far back in time to the beginning of the universe and see in all directions the big bang covering the sky. The expansion of the universe has reduced the incandescent glare of the big bang and redshifted its radiation into the invisible infrared. See COSMOLOGY; UNIVERSE. Edward Harrison

Bibliography. H. Bondi, *Cosmology*, 2d. ed., Cambridge University Press, 1960; E. R. Harrison, *Cosmology: The Science of the Universe*, 2d. ed., New Cambridge University Press, 2001; E. R. Harrison, *Darkness at Night: A Riddle of the Universe*, Harvard University Press, Cambridge, MA, 1987.

## Olfaction

One of the chemical senses, specifically the sense of smell. Olfaction registers chemical information in organisms ranging from insects to humans, including marine organisms. For terrestrial animals, its stimuli comprise airborne molecules. The typical stimulus is an organic chemical with molecular weight below 300 daltons; about a half million such substances exist. A few inorganic chemicals can also stimulate olfaction, notably hydrogen sulfide, ozone, ammonia, and the halogens. For marine organisms, amino acids and proteins, which derive largely from the decomposition of organic matter, form particularly good stimuli.

**Anatomy and physiology.** The anatomy of olfactory structures and the neurophysiology of olfaction differ significantly among different animal groups.

*Insects.* Insect olfactory receptors exist within sensory hairs on the antennae. Each hair contains thousands of pores in its cuticle. Stimulating molecules must pass through the pores in order to stimulate the outer segment of a receptor cell. This process of transduction has been studied through the measurement of electric potentials from the receptor neurons. The best known potential, the electroantennogram, reflects the action of many receptor neurons.

Certain insect olfactory receptors are specialists, and respond to a very narrow range of relevant chemicals, such as sex attractants secreted by females but registered only by males. Generalist receptors, on the other hand, respond to a wide range of chemicals. One generalist cell differs somewhat from another in the variety of materials to which it responds, permitting discrimination of different materials in an across-receptor pattern of induced neural activity. See CHEMICAL ECOLOGY; INSECT PHYSIOLOGY.

*Fishes.* The olfactory organ of fishes resides typically in tubular chambers on either side of the mouth (Fig. 1). As the fish swims, water passes through the chamber and comes into contact with the olfactory epithelium that lines the folds of a structure called

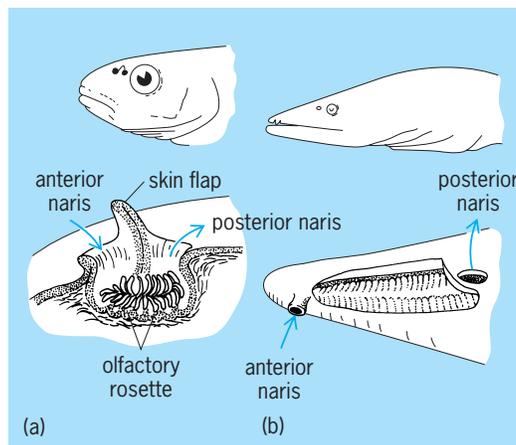


Fig. 1. Position and internal structure of the nose in fishes. (a) Minnow (*Phoxinus phoxinus*). (b) Eel (*Anguilla anguilla*). (After H. Teichmann, *Was leistet der Gervchssinn bei Fischen?*, *Umschau Wiss. Tech.*, 62:588–591, 1962)

the olfactory rosette. This epithelium has great similarity to that of terrestrial vertebrates. In addition to proteins and amino acids, fishes respond to many of the odorants that form the stimuli for terrestrial olfaction. Hence, fishes respond to stimuli that humans characterize as floral, fruity, putrid, and so on.

*Terrestrial vertebrates.* In terrestrial vertebrates, the olfactory receptors reside within a sac or cavity more or less similar to the human nasal cavity (Fig. 2). The olfactory mucosa patch in the cavity characteristically contains millions of receptor cells, though

in some olfactory-dominated (or macrosmatic) mammals, such as the dog and rabbit, it contains tens of millions. The location of the olfactory mucosa relative to air currents in the cavity plays some role in the ongoing olfactory vigilance of the organism. In the human, a microsmatic organism, the mucosa sits out of the main airstream. During quiet breathing eddy currents may carry just enough stimulus to evoke a sensation, whereupon sniffing will occur. Sniffing amplifies the amount of stimulus reaching the receptors by as much as tenfold.

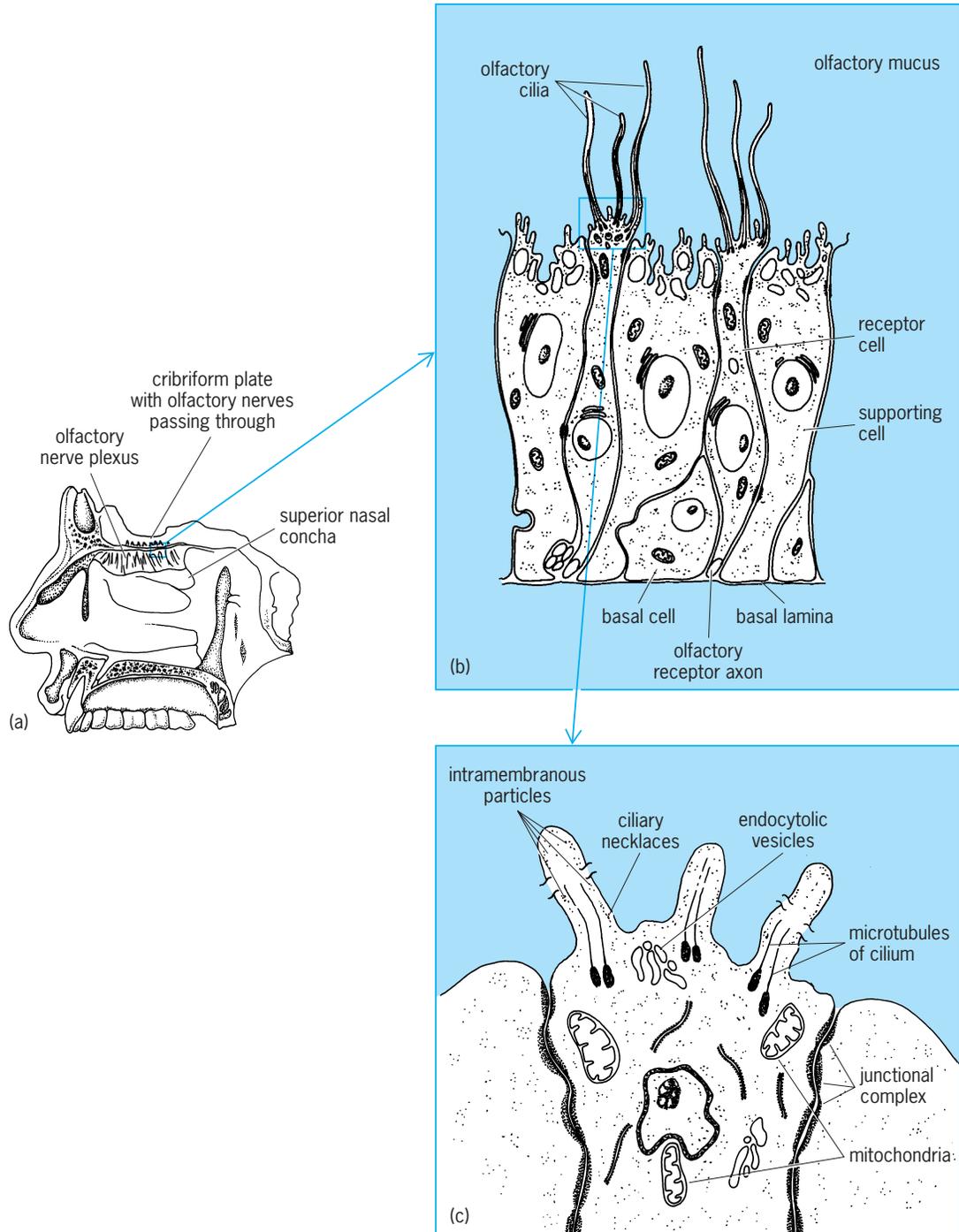


Fig. 2. Human peripheral olfactory system. (a) Position of olfactory epithelium. (b) Olfactory epithelium. (c) Olfactory receptor. (After H. B. Barlow and J. D. Mollon, eds., *The Senses*, Cambridge University Press, 1982)

Reception of the chemical stimulus and transduction into a neural signal apparently occur on the olfactory receptor cilia (Fig. 2). The ciliary membrane contains receptor protein molecules that interact with stimulating molecules through reversible binding. Nonprotein portions of the membrane may also have some chemoreceptive role, through modulation of the conformation or orientation of the receptor proteins.

Vertebrate receptor cells, like the generalist cells of insects, show broad tuning, that is, they respond to many odors. The breadth of tuning depends in part on cell age—receptor cells have a life-span of just a few weeks, and show broader tuning in their first two to three weeks than thereafter. Since new cells emerge from basal cells beneath the receptor cells, a receptor cell of any age may appear anywhere in the olfactory mucosa. However, there do exist some regional differences in the tuning. In frogs and the tiger salamander, organisms commonly used to study vertebrate olfaction, cells in the anterior portion of the mucosa show a bias toward water-soluble odorants, whereas those in the posterior portion show a bias toward lipid-soluble odorants. An anterior-posterior gradation, seen easily in regional recordings of the receptor potential known as the electroolfactogram, also reveals itself in recordings from the olfactory bulb of the brain.

Adjacent points in the mucosa project generally, though not exclusively, to adjacent points in the olfactory bulb (Fig. 3). The synapses between the incoming olfactory nerve fibers and the second-order cells, mitral cells, occur in basketlike structures called glomeruli. On average, a glomerulus receives about 1000 receptor cell fibers for each mitral cell. The tangled arrangement of fibers within a glomerulus invites the conclusion that it must operate as a functional unit. Second-order cells, which are also influenced by collateral neurons, such as the interior granule cell, nevertheless show very similar breadth

of tuning to that seen in individual receptor cells. The location of cells within the bulb seems to play a role in encoding odor quality: each odorant stimulates a more or less unique spatial array.

Second-order cells, unlike first-order cells, may change their responsiveness with the state of need of the organism. For instance, mitral cells will respond more vigorously to food odors in a hungry organism. Such modulation presumably results from interplay of incoming neural signals with outgoing activity which originates in more central structures.

The central neural pathways of the olfactory system have a complexity unmatched among the sensory systems. One pathway carries information to the pyriform cortex (paleocortex of the temporal lobe), to a sensory relay in the thalamus (dorsomedial nucleus), and to the frontal cortex (orbitofrontal region) [Fig. 4]. This pathway seems rather strictly sensory. Another pathway carries information to the pyriform cortex, the hypothalamus, and other structures of the limbic system. The latter have much to do with the control of emotions, feeding, and sex. The strong affective and motivational consequences of olfactory stimulation seem compatible with projections to the limbic system and with the role of olfaction in certain types of physiological regulation.

In many organisms, odors from other members of the same species have powerful motivational effects. Such stimuli are called pheromones. Sex attractants, trail pheromones, and maternal pheromones are examples. In many vertebrate species, reception of pheromones occurs via an important accessory olfactory organ, known as the vomeronasal organ, which characteristically resides in the hard palate of the mouth or floor of the nasal cavity (Fig. 4). See PHEROMONE.

**Sensitivity and functional properties.** At its best, human absolute sensitivity exceeds that of the most sensitive physical instruments, but sensitivity varies from odorant to odorant over several orders of

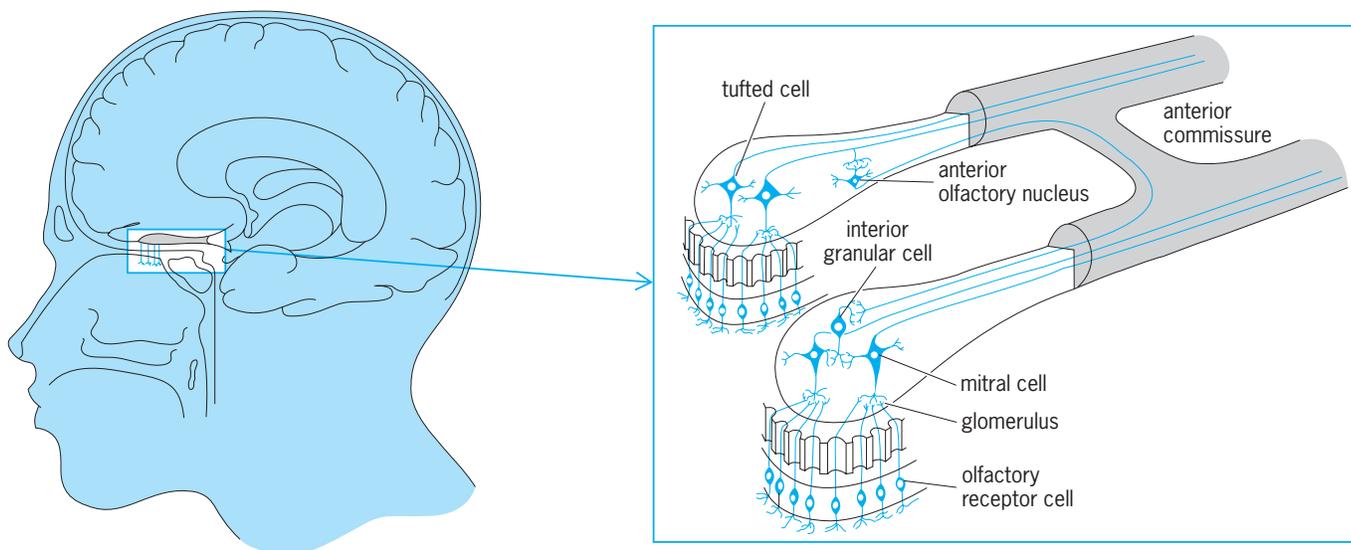
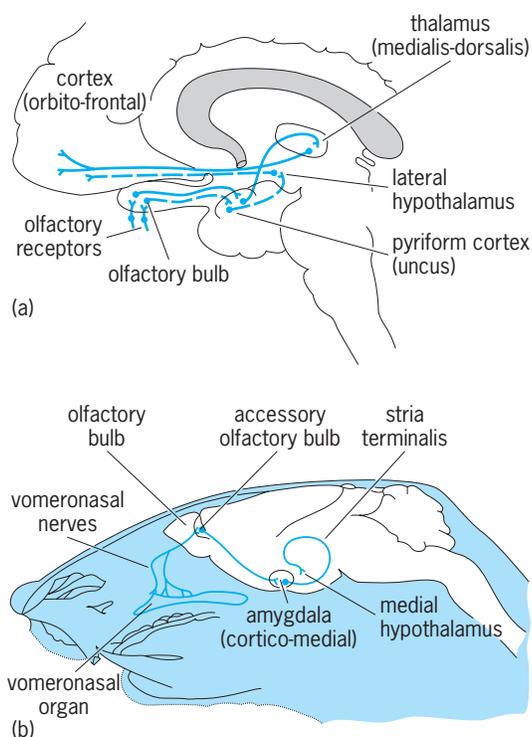


Fig. 3. Location of the olfactory bulbs at the interior surface of the human brain and their connections via the anterior commissure. (After D. Ottoson, *Physiology of the Nervous System*, Oxford University Press, 1983)



**Fig. 4.** Olfactory neural pathways in terrestrial vertebrates. (a) Central olfactory pathways characteristic of higher primates. Broken and solid lines represent distinct pathways. (b) Accessory olfactory pathways characteristic of rodents. (After H. B. Barlow and J. D. Mollon, eds., *The Senses*, Cambridge University Press, 1982)

magnitude. For instance, the threshold for the potent odorant 3-methoxy-3-isobutyl pyrazine (green bell pepper odor) equals about 1 part per  $10^{12}$  parts of air, whereas that for the far less effective odorant methanol (wood alcohol) equals about 100 parts per  $10^6$ . A common range of thresholds for materials used in fragrances and flavors is 1 to 100 parts per  $10^9$ . Physicochemical determinants of molecular deposition and penetration through mucus correlate rather well with this threshold range. In aliphatic series in which physicochemical properties vary progressively with molecular chain length, threshold characteristically changes progressively. Such a correlation holds true for terrestrial vertebrate species other than humans, though absolute values differ. The available data suggest that rats and dogs have best sensitivity, about a thousandfold better than that of humans.

Thresholds gathered from various groups of human subjects permit certain generalities about how the state of the organism affects olfaction. For instance, persons aged 70 and above are about tenfold less sensitive than young adults. Males and females have about equal sensitivity, except perhaps in old age, where females are more sensitive. Females show a small fluctuation in sensitivity during the menstrual cycle. Sensitivity increases prior to ovulation. Smokers have about the same sensitivity as nonsmokers. Persons with certain medical disorders, such as mul-

iple sclerosis, Parkinson's disease, paranasal sinus disease, Kallmann's syndrome, and olfactory tumors, exhibit decreased sensitivity (hyposmia) or complete absence of sensitivity (anosmia).

Above its threshold, the perceived magnitude of an odor changes by relatively small amounts as concentration increases. A tenfold increment in concentration will cause, on average, about a twofold change in perceived magnitude. The perceived magnitude of an odor is often greatly influenced by olfactory adaptation, a process whereby during continuous short-term exposure to a stimulus its perceived magnitude falls to about one-third of its initial value. Concentrations below this conditioning level may then prove imperceptible, but concentrations above the conditioning level will be affected relatively little. Recovery from such adaptation begins as soon as the stimulus is removed, and is generally complete in a few minutes.

Adaptation is largely specific to the conditioning odorant. That is, exposure-induced changes in sensitivity to one odorant will leave sensitivity to most other odorants roughly unchanged. Instances where exposure to one odorant does alter the sensitivity to another exemplify cross-adaptation. Experiments on cross-adaptation serve as a means to explore chemical or structural commonalities among stimuli.

The stimuli for olfaction are commonly complex, that is, they are mixtures. Such products as coffee, wine, cigarettes, and perfumes contain at least hundreds of odor-relevant constituents. Only rarely does the distinctive quality of a natural product, such as a vegetable, arise from only a single constituent. A chemical analysis of most products will not usually allow a simple prediction of odor intensity or quality. One general rule, however, is that the perceived intensity of the mixture falls well below the sum of the intensities of the unmixed components. This suppression of intensity works against the weaker components, which may lose their separate identity in the mixture but may still contribute to overall perceptual quality. A pleasant-smelling mixture of unknown composition may actually be found upon analysis to contain some very unpleasant constituents in small amounts. Their presence is often partially masked by stronger-smelling constituents, yet may make an essential contribution. Perfumery, which plays a wide role in the "deodorization" (often through addition of odor) and reodorization of many manufactured products, entails a careful blending of pleasant- and sometimes unpleasant-smelling materials into a final acceptable odor.

**Structure-activity relations.** General notions about the properties that endow a molecule with its quality have spawned more than two dozen theories of olfaction, including various chemical and vibrational theories. Most modern theories hold that the key to quality lies in the size and shape of molecules, with some influence of chemical functionality. For molecules below about 100 daltons, functional group has obvious importance: for example, thiols smell skunky, esters fruity, amines fishy-uriny, and carboxylic acids

rancid. For larger molecules, the size and shape of the molecule seem more important. Shape detection is subtle enough to enable easy discrimination of some optical isomers (for example, D-carvone versus L-carvone, which smell like spearmint and caraway, respectively). Progressive changes in molecular architecture along one or another dimension often lead to large changes in odor quality. No current theory makes testable predictions about such changes. See CHEMICAL SENSES; CHEMORECEPTION.

William Cain

**Bibliography.** W. Breipohl (ed.), *Ontogeny of Olfaction*, 1986; T. Engen, *Odor Sensation and Memory*, 1992; A. I. Farbman, *Cell Biology of Olfaction*, 1992; T. L. Payne, M. C. Birch, and C. E. Kennedy (eds.), *Mechanisms in Insect Olfaction*, 1986; M. J. Serby and K. L. Chober (eds.), *The Science of Olfaction*, 1992; E. T. Theimer (ed.), *Fragrance Chemistry: The Science of the Sense of Smell*, 1982.

## Oligocene

The third oldest of the seven geological epochs of the Cenozoic Era. It corresponds to an interval of geological time (and rocks deposited during that time) from the close of the Eocene Epoch to the beginning of the Miocene Epoch. The most recent geological time scales assign an age of 34 to 24 million years before present (m.y. B.P.) to the Oligocene Epoch. See CENOZOIC; EOCENE; MIOCENE.

In his early subdivision of the Tertiary published in 1833, Charles Lyell established four stratigraphic units, from the oldest to the youngest: Eocene, Miocene, and Older and Younger Pliocene. Stratigraphers in the Netherlands and Germany, however, kept describing strata that they considered to be intermediate in position and characteristics between Eocene and Miocene. They concluded that these sediments represented a major marine transgression in northern Europe at the close of the Eocene Epoch. This eventually led the German stratigrapher E. Beyrich in 1854 to propose the Oligocene as an independent subdivision of the Tertiary based on a sequence of marine, brackish-water, and nonmarine sediments of the Mainz Basin in Germany. He proposed a new Oligocene Epoch and constructed it out of the younger part of Eocene and the older part of Miocene epochs of the Lyellian subdivisions of Tertiary.

An important event that characterizes the Oligocene Epoch was the development of extensive glaciation on the continent of Antarctica. Prior to that time, the world was largely ice-free through much of the Mesozoic and early Tertiary. A significant amount of ice is now known to have existed on the Antarctic continent since at least the beginning of the Oligocene, when the Earth was ushered into its most recent phase of ice-house conditions. This in turn created revolutions in the global climatic and

CENOZOIC	QUATERNARY	Holocene
		Pleistocene
	TERTIARY	Pliocene
		Miocene
		Oligocene
		Eocene
	Paleocene	
MESOZOIC	CRETACEOUS	

hydrographic systems, with important repercussions for the marine and terrestrial biota. The changes include steepened latitudinal and vertical thermal gradients affecting major fluctuations in global climates, and the shift in the route of global dispersal of marine biota from an ancestral equatorial Tethys seaway, which had become severely restricted by Oligocene time, to the newly initiated circum-Antarctic circulation. See GLACIAL EPOCH.

**Subdivisions.** In modern time scales, this epoch is subdivided into two series, a Lower and an Upper Oligocene. Northern European Rupelian and Chattian stratigraphic stages are designated to be time-equivalent to the Early and Late Oligocene, respectively (Fig. 1). Worldwide, the epoch represents an overall regressive sequence when there was a draw-down of global sea level, with relatively deeper, marine facies in the Early Oligocene and shallower-water to nonmarine facies in the Late Oligocene. See FACIES (GEOLOGY).

The Rupelian Stage was proposed by A. Dumont in 1849, based on sedimentary strata in Belgium. The Tongrian Stage, also described from Belgium by Dumont in 1839, was originally regarded to be, in part, older than the Rupelian by Beyrich, who considered that the Tongrian and Rupelian together

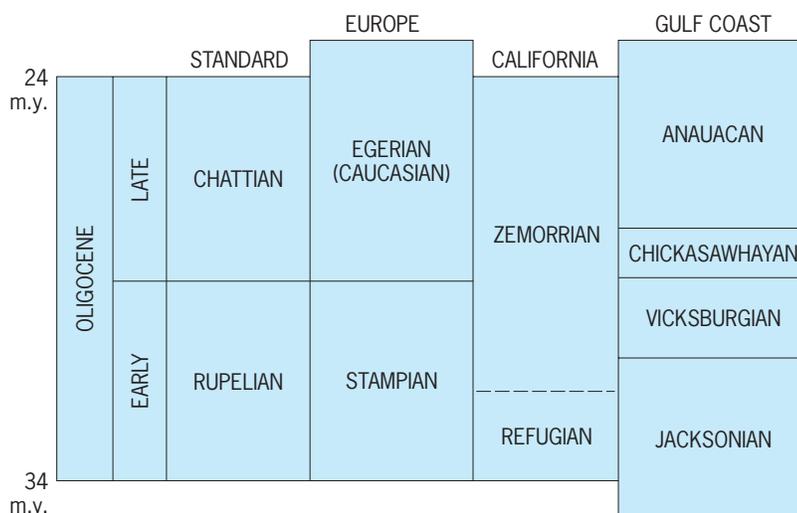


Fig. 1. Oligocene stages and their temporal equivalents.

constituted his new epochal unit, the Oligocene. In 1983 K. Mayer-Eymar erected the Lattorfian Stage with a type locality in Saxony, Germany. Mayer-Eymar considered the Lattorfian to be temporally equivalent to the Early Oligocene. Both the Tongrian and Lattorfian strata were later shown to range downward into the uppermost Eocene, and their use has largely been abandoned in favor of an expanded notion of the Rupelian. Because the Rupelian at its type section spans somewhat less than the time interval included in the Early Oligocene, some French stratigraphers prefer the Stampian Stage to Rupelian. The Stampian was described by A. D'Orbigny in 1852 from marine and lacustrine beds at Etampes in France.

The Chattian, elected by consensus as the standard stage for the Late Oligocene, was based on marine sand beds near Kassel, Germany, and proposed by T. Fuchs in 1894. Fuchs considered the Chattian to be younger than Rupelian and older than the Aquitanian Stage (which was later placed in the Lower Miocene). Later studies have shown that there is a temporal gap between the top of Chattian and the base of Aquitanian as defined at their stratotype sections. However, in stratigraphy it is normal practice to extend the formal concept of standard stages to include such gaps. Thus, to accommodate the gap between the uppermost Eocene stage of Priabonian and the Oligocene Rupelian, and the gap between Chattian and Aquitanian, the formal concepts of Rupelian and Aquitanian were extended downward by later stratigraphers to bridge the intervening lacunae.

Other regional stages of the Oligocene include the Egerian, sometimes used in central and eastern European countries, which is in part equivalent to the Chattian, but may extend into the lower part of Miocene. In Russia the Caucasian Stage is considered to be an equivalent to the Egerian. Marine Oligocene in California is subdivided into the Refugian and Zemorrian stages, the former stage encompassing only the earliest Oligocene, and the latter spanning the bulk of the epoch. In the Gulf coast of the United States the Oligocene extends from the upper part of the Jacksonian Stage to Vicksburgian, Chickasawhayan, and lower part of Anahuacan stages. Temporally, the latter stage extends into the Miocene. Australian stratigraphers include the Willungian, Janjukian, and lower part of Longfordian stages in their Oligocene, and the New Zealanders now consider the epoch to span three local stages: Whaingaroan, Duntroonian, and Waitakian.

Paleontologists who study mammal and other vertebrate fossils often use their own subdivisions to express the ages of terrestrial assemblages. Workers in Europe consider the Oligocene to span three subdivisions, Suevoian, Arvernian, and lower part of Agenian. North American vertebrate paleontologists include Orellan, Whiteneyan, and the lower two-thirds of the Arikarean ages in their concept of the Oligocene. See PALEONTOLOGY; STRATIGRAPHY.

**Tectonics, oceans, and climate.** Radical changes occurred in the Oligocene Epoch that revolutionized global oceanographic and climatic conditions. The

most prominent change was the shift of circum-global circulation (and a major means of biotic dispersal) from the equatorial to the southern high-latitude regions. Both the restriction of the equatorial flow between the Indian and the Atlantic oceans through the ancestral Tethys seaway, and the opening of the oceanic gateway at Drake Passage occurred during the Oligocene. These changes had important repercussions that led to the entry of the Earth into predominantly ice-house conditions that continue to the present. The cooling of the polar regions in the Oligocene led to the accentuation of latitudinal and vertical thermal gradients and an increase in seasonality. This resulted in a fundamental shift in the bottom-water regime, from density-driven warm, saline bottom waters in the Cretaceous and Paleocene-Eocene, to cold bottom waters, largely driven by thermal contrast. The accentuated latitudinal thermal contrast also resulted in the expansion of the erosive activity of bottom waters. See CRETA-CEOUS.

The separation of Svalbard and Greenland in the latest Eocene and the breaching of the Rio Grande Rise in the South Atlantic in the earliest Oligocene set the stage for the crossing of the important climatic threshold near the Eocene-Oligocene boundary. It has been suggested that after the development of the connection between the Arctic and the Norwegian Sea that allowed the supply of cold deep water to the Atlantic and the southern high latitudes, more moisture became available around Antarctica. This, combined with the partial isolation of Antarctica, may have led to large-scale freezing at sea level and the initiation of the formation of Antarctic Bottom Water. There is evidence that the temperature of the bottom water dropped by 4–5°C (7–9°F) near this boundary, and the change from warm to cold bottom water may date back to this time. The development of the psychrosphere (deeper, cold layer of the ocean) manifests itself in the presence of widespread erosional hiatuses in the eastern Indian and southwestern Pacific oceans, where scouring by cold bottom waters has stripped away much of the older sediments. The vigorous activity of the deep water is also indicated by widespread drift sediments in the North Atlantic.

Another important tectonic event that had major significance for the oceanic-climatic conditions was the breaching of the straits between South America and Antarctica at the Drake Passage in the mid-Oligocene and complete geographic isolation of Antarctica. This eliminated the last barrier in the path of the circum-Antarctic circulation. The development of the Circum-Antarctic Current led to further thermal isolation of Antarctica that was conducive to the development of an extensive ice cap on the continent. Overall, there was a dramatic increase in the areas of the desert through the Oligocene and younger epochs that can be ascribed to the cooling higher latitudes and the development of polar ice caps. See PALEOCLIMATOLOGY; PALEOGEOGRAPHY; PLATE TECTONICS.

In Oligocene time the global surface circulation patterns had essentially evolved the major features

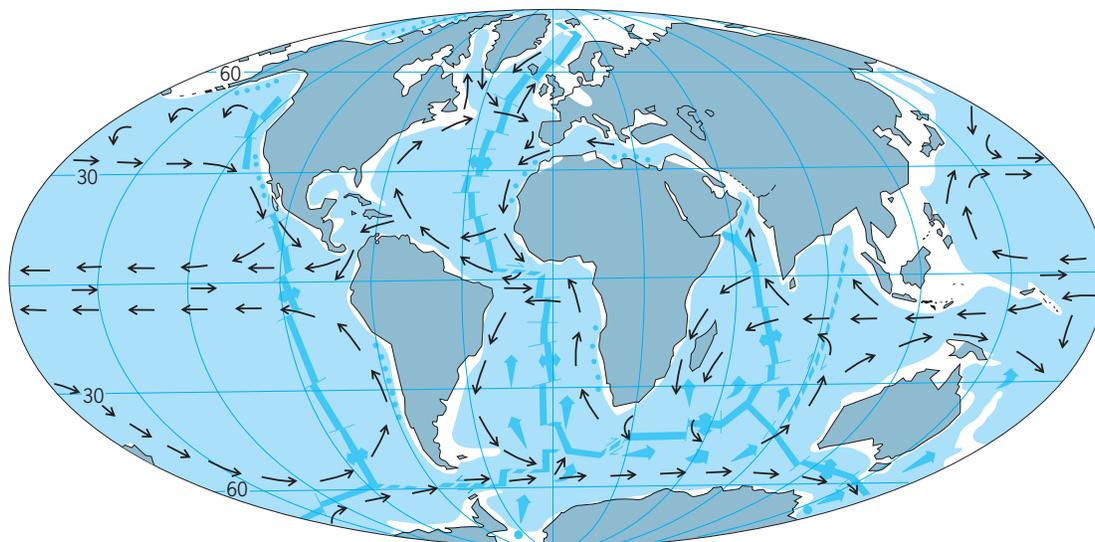


Fig. 2. Paleogeography and oceans of the Oligocene. (After B. U. Haq and F. W. B. van Eysinga, *Geological Time Table*, Elsevier, 1998)

of the modern oceans (Fig. 2). One major difference was that in the equatorial region an open Panama isthmus permitted the exchange of waters between the Pacific and the Atlantic oceans. However, by the Early Oligocene the flow of the Tethys Current from east to west had already become sharply reduced, intermittent, and restricted to a narrow passage southwest of the Indian plate, following the collision of the subcontinent with the Asian mainland, initial uplift of the Himalayas in the Eocene, and subsequent raising of the Tibetan Plateau. The convergence between India and Asia and between Africa and Europe closed the Tethys seaway, leaving behind smaller remnants that include the Mediterranean, Black, and Caspian seas.

The worldwide cooling trend that began in the latest Eocene but became accentuated in the Oligocene may have been in part caused by the long-term effect of the uplift of the Tibetan Plateau, which was significant by Oligocene time. The raised plateau is thought to be able to deflect the atmospheric jet stream more vigorously, leading to the strengthening of the summer monsoon, and increased rainfall and weathering in the Himalayas. Since increased weathering and dissolution of carbonates results in greater carbon dioxide drawdown from the atmosphere, the reduced levels of carbon dioxide may have provided a condition that helped the Earth enter into a renewed ice-house phase.

By mid-Oligocene the Circum-Antarctic Current was well established. The ice shelves around Antarctica were the dominant sources of cold bottom waters. Northern sources of cold, deep water had also become operative, and the surface exchange between North Atlantic and the Arctic through Norwegian-Greenland and Labrador passages had become active. There is indirect evidence that suggests the initiation of an ice cover in the formerly ice-free Arctic by Oligocene time. By this time the central Arctic Ocean had become largely landlocked, and the

connections to the open sea were restricted enough that the input of fresh water from Eurasian rivers, combined with salinity stratification characteristic of isolated basins, may have allowed winter ice to form. This, however, does not seem to have affected the deeper Arctic waters, as indicated by the presence of benthic faunas well up to the late Miocene.

As a consequence of the cooling higher latitudes and expanding ice sheets, the sea-level history of the Oligocene is one of repeated regressions that were followed by smaller transgressions that successively covered less epicontinental areas than before. The first major sea-level fall occurred near the Eocene-Oligocene boundary, which most likely dates back to the first extensive ice accumulation on Antarctica. A second more pronounced global sea-level fall is indicated in the mid-Oligocene, around 30 m.y. B.P., when the sea level dropped about 150 m (500 ft) and continental margins around the world were laid bare to extensive erosion. This eustatic fall event was also in response to a significant amount of accumulation on the Antarctic ice cap. Later Oligocene eustatic fluctuations were of lesser magnitude, indicating less extensive ice-sheet fluctuations. In the early Miocene a major transgression occurred, and the sea covered much of the coastal areas that were exposed during the Oligocene.

The mid-Oligocene sea-level lowering was accompanied by extensive shelf erosion and stream incision on many continental margins. Prominent canyons were initiated on many of the exposed shelves that remained operative for several million years. The mid-Oligocene eustatic retreat of the seas was characteristically long-lasting. Although this event was followed by minor eustatic recoveries, the sea level never rose back to pre-Oligocene levels, and recovered fully only after the close of Oligocene Epoch. This implies that during the long lowstand of the sea the submarine canyons were kept active continuously, to be filled back only after the full eustatic

recovery in the early Miocene. See CONTINENTAL MARGIN; PALEOCEANOGRAPHY; SUBMARINE CANYON.

**Life.** Due to accentuated thermal gradients and seasonality in the Oligocene (compared to previous Tertiary epochs), marine biotic provinces became more fragmented. Extreme climates, with greater diurnal and seasonal temperature contrasts, are held responsible for reduced diversities in marine plankton. The marine micro and macro fauna and flora of the Oligocene have strong affinities with those of the late Eocene. The typically Paleogene assemblages gradually become extinct during the Oligocene. The boundary between Oligocene and Miocene represents a complete changeover to Neogene faunal elements. Thus, the Oligocene was characterized by transitional faunal features between the Paleogene and the Neogene.

Planktonic foraminifera, which had diversified rapidly in the Eocene, were severely reduced in diversity in the Oligocene. Calcareous nannofossils that had also proliferated earlier, suffered reduction in diversities, though somewhat less severe than the planktonic foraminifera. Benthic fauna fared better and some lineages actually diversified, perhaps due to greater niche fragmentation. However, the last of the *Nummulites* disappeared in the Late Oligocene. The discocyclinids, dominant in the Eocene, were followed by lepidocyclinids in the Oligocene, and later, in the Miocene, gave way to the miogypsinids.

An evolutionary leap was made by the mammals across the Eocene-Oligocene boundary. The Eocene perissodactyls and prosimians saw their last days, giving way to rhinocerids, tapirs, and wild boarlike hog species with strong incisors and large canines. *Hyaenodon* was an Oligocene carnivore with strong canines and sharp molars, much like those of the modern cat species. Grasslands still supported large browsers, and a new group of perissodactyls, typified by the *Titanotherium* (Fig. 3), appeared and climaxed, to later die out near the close of Oligocene. The horses that had first appeared in the Eocene continued to increase in size and became three-toed, typified by *Mesobippus*. Other hoofed mammals also diversified during the Oligocene. Elephants made their first appearance near the Eocene-Oligocene boundary and developed a short trunk and two pairs of

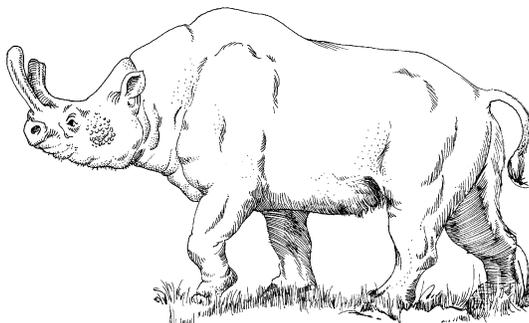


Fig. 3. Oligocene *Titanotherium* (= *Brontotherium*) of North America, a large browsing mammal. The titanotheres disappeared abruptly at the close of the Oligocene Epoch. (After R. A. Stirton, *Time, Life and Man*, John Wiley, 1959)

tusks. An early simian, *Propliopithecus*, made its first appearance in Oligocene, and is considered ancestral to the modern family of gibbons. The general uniformity of mammalian fauna in the Oligocene suggests that the widespread regressions of the sea most likely resulted in land bridges that reconnected some of the Northern Hemispheric land masses, which may have led to transmigrations of some families of mammals between North America, Asia, and Africa. Birds had achieved some of their modern characteristics, and at least 10 modern genera had already made their appearance by the close of Oligocene time. See AVES; GEOLOGIC TIME SCALE; MAMMALIA; PALEOECOLOGY; PERISSODACTYLA.

Bilal U. Haq

**Bibliography.** B. U. Haq, J. Hardenbol, and P. R. Vail, Chronology of fluctuation sea levels since the Triassic, *Science*, 235:1156-1167, 1987; B. U. Haq and F. W. B. van Eysinga, *Geological Time Table*, Elsevier, 1998; C. Pomerol, *The Cenozoic Era: Tertiary and Quaternary*, Ellis Horwood, 1982.

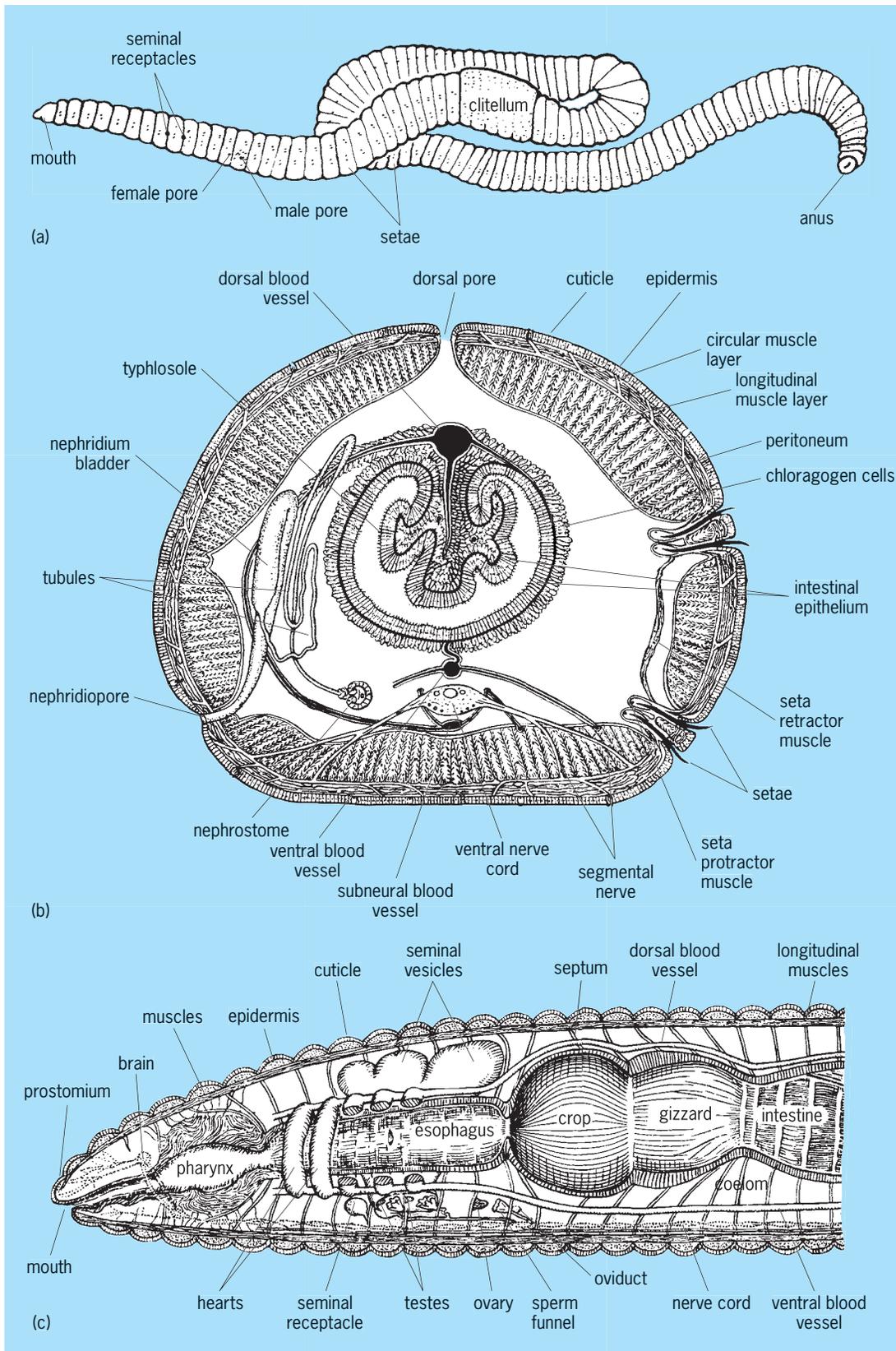
## Oligochaeta

A class or subclass (depending on the classification system used) of worms, including the earthworms, within the phylum Annelida. These animals exhibit both external and internal segmentation—that is, furrows in the body wall and transverse partitions (septa) dividing the coelom into chambers. They usually possess chaetae (setae), or bristles, made of chitin, which are not borne on parapodia and are few in number (compared to the Polychaeta). Oligochaetes are hermaphroditic. The gonads are few in number and situated in the anterior part of the body, the male gonads being anterior to the female gonads. The gametes are discharged through the oviducts and sperm ducts. At maturity, a portion of the body wall is swollen into a secretory area called the clitellum (illus. a). There is no larval stage during the development of oligochaetes. See POLYCHAETA.

The oligochaetes are primarily freshwater and burrowing terrestrial animals. About 200 species are marine, mostly occurring in the intertidal zone. They range in size from aquatic species that are 0.02 in. (0.5 mm) long to the giant Australian earthworm (*Megascolides*), which is up to 3 m (10 ft) long. About 25 families and over 3000 species are presently recognized.

**Body form and anatomy.** Oligochaetes are cylindrical, elongated animals with the anterior mouth usually overhung by a fleshy lobe, the prostomium, and the anus located terminally. The body plan is that of a tube within a tube. The chaetae, usually a pair of ventrolateral bundles, are borne on most segments. Other external features are the pores of the reproductive systems opening on certain segments, the openings of the excretory organs (metanephridia), and in many earthworms dorsal pores which open externally from the coelom (see illus. b).

**Body wall.** The body wall consists of (from interior to exterior) a layer of thin epithelial cells (peritoneum) lining the coelom, muscle tissue arranged



**Lumbricidae. (a) Earthworm (*Lumbricus terrestris*), external features; (b) cross section (left half shows an entire metanephridium and a dorsal pore, and right half includes chaetae but no nephridium); (c) internal structures of anterior portion from the left side. (After T. I. Storer, *General Zoology*, 3d ed., McGraw-Hill, 1957)**

in inner longitudinal and outer circular layers, and a layer of columnar epithelial tissue (epidermis) covered by a thin secreted cuticle. The epidermis contains interspersed glandular and sensory cells. Movement in terrestrial forms is accomplished by alternate contractions of the muscle layers, which produce a peristaltic motion in which the chaetae aid forward progression by gripping the substrate. Other, smaller muscles protract and retract the chaetae and are associated with the intestine and reproductive organs.

**Coelom and gut.** The fluid-filled coelom is segmentally partitioned in most cases by the septa. The alimentary canal passes through this series of spaces and is composed of a buccal cavity, a muscular-walled pharynx, a narrow esophagus, and the intestine proper. In the earthworms the posterior portion of the esophagus is often enlarged into a thin-walled chamber (crop) that stores ingested food and a thicker-walled portion (gizzard) that grinds the food (see *illus. c*). The pharynx may contain glands that secrete mucus and digestive enzymes, while the esophagus often demonstrates swellings (calciferous glands) that function in regulation of calcium and carbonate levels. In many of the larger oligochaetes the epithelium lining the gut wall, composed of glandular and absorptive cells, projects downward from the dorsum as a longitudinal fold (typhlosole) into the lumen of the intestine. There are thin muscle layers in the wall of the intestine, and it is invested with a layer of modified peritoneal chloragogen cells; these pigmented cells seem to be important in the synthesis and storage of glycogen and deamination of proteins. The food of most oligochaetes is plant material ingested by a sucking action of the pharynx. Some aquatic forms are carnivorous; a few others may be parasitic.

**Nephridia and excretion.** Excretion and osmoregulation in the oligochaetes are accomplished by segmentally arranged paired metanephridia. Typically each organ is composed of a funnel (nephrostome), which opens into the coelom, a variously coiled tubule sometimes with an expanded portion (bladder), and a nephridiopore which discharges ventrolaterally to the outside. Nitrogenous wastes and salts enter the metanephridium through the nephrostome and from capillaries surrounding it. Along the tubule, some salts and water may undergo reabsorption into the capillaries, depending on habitat, but products of protein breakdown (such as ammonia and urea) are mostly released at the nephridiopore.

**Respiration.** Exchange of respiratory gases is in general a function of the body surface. Some aquatic forms (for example, *Dero* and *Branchiura*) have posterior extensions of the body wall—filamentous gills—to increase surface area for diffusion of these gases.

**Circulation.** The circulatory system consists of a main dorsal vessel, often arising from a perienteric sinus, in which the blood flows anteriorly, and various lateral, ventral, and subneural vessels. The blood contains the respiratory pigment hemoglobin dissolved in it, and phagocytic corpuscles. Contraction of the dorsal vessel or lateral vessels (“hearts”) propels the blood. See RESPIRATORY PIGMENTS (INVERTEBRATE).

**Nervous system.** The nervous system is composed of dorsal cerebral ganglia, circumesophageal connectives, and paired ventral nerve cords with segmentally arranged ganglia which give off nerves to the body wall and intestine. Sensory cells of various types are located in the epidermis. Oligochaetes react to a wide range of stimuli.

**Reproduction.** All oligochaetes are hermaphroditic. The testes are located anterior to the ovaries, and both are derived from the septal peritoneum. The gametes are liberated into the coelom or its pouches and reach the outside through variously differentiated ducts. In copulation, spermatozoa are received in spermatheca (seminal receptacles), and at oviposition, which occurs later, the clitellum secretes a cocoon into which eggs and spermatozoa are discharged. Embryonic development occurs within the cocoon. In the families Aeolosomatidae and Naididae, asexual reproduction by a type of binary fission (paratomy) occurs. See INVERTEBRATE EMBRYOLOGY; REPRODUCTION (ANIMAL).

**Economic importance.** The oligochaetes, especially the common large terrestrial “earthworm” *Lumbricus* (family Lumbricidae) and, more recently, the transparent aquatic “California blackworm” *Lumbriculus* (of the more primitive family Lumbriculidae), have been used in studies of anatomy, physiology, regeneration, and metabolic gradients. Some aquatic forms are important in studies of stream pollution as indicators of organic contamination. Earthworms are important in turning over the soil and reducing vegetable material into humus. It is likely that fertile soil furnishes a suitable habitat for earthworms, rather than being a result of their activity.

**Classification.** The nature of the reproductive system and the absence of parapodia sharply set the oligochaetes apart from the polychaetes. The Hirudinea (leeches) are much closer to the oligochaetes but constitute a homogeneous group long considered a separate class. The Oligochaeta historically have been treated as a class of the phylum Annelida, coordinate in rank with the Polychaeta and Hirudinea (or Hirudinomorpha). More recently, however, this taxon, regarded by some specialists as polyphyletic, has been merged with the Hirudinea (Hirudinomorpha) into a group (generally given class rank), known as the Clitellata, based on the mutual presence of the clitellum and other common features.

That the oligochaetes are descended from marine polychaete-like ancestors seems certain, but there is no agreement as to the number and relationships of subordinate taxa. Classification into orders and families has largely been based on placement of male gonopores relative to the segment the testes are located in, or similar traits of the reproductive system that are difficult to discern. Two of the three currently recognized orders—Lumbriculida and Moniligastrida—each contain a single family (the aquatic Lumbriculidae and the terrestrial Moniligastridae); both are characterized by four pairs of chaetae per segment. Remaining families are generally included in the order Haplotaxida. Of the aquatic

families, the most important ones ecologically are the larger Tubificidae, often reddish "sludge worms" of over 40 segments; Naididae, smaller worms usually with less than 40 segments (chaetae lacking on the anterior ones), often possessing eyespots but lacking prostomial cilia; and the even smaller (few millimeter) Aeolosomatidae, characterized by a ventrally ciliated prostomium. Tubificid worms generally live in vertical tubes embedded in bottom mud. Some, such as *Tubifex*, can tolerate very low oxygen concentrations, assisting the acquisition of oxygen from the water by waving their tails. Naidids (such as *Dero* and *Stylaria*, the latter bearing a long prostomium) are usually associated with submergent (entirely underwater) vegetation. The Aeolosomatidae (regarded by some as Polychaeta, since representatives lack the clitellum) are usually found gliding around the water-sediment interface. Families with high numbers of large terrestrial species include the widely distributed Lumbricidae (such as *Lumbricus*), the Glososcolecidae of South America and the Caribbean, and the Megascolecidae of Asia and Australia. See ANNELIDA.

Perry C. Holt; Robert Knowlton

**Bibliography.** R. O. Brinkhurst and B. G. M. Jamieson, *Aquatic Oligochaeta of the World*, 1971; C. A. Edwards and P. Bohlen, *Biology and Ecology of Earthworms*, 3d ed., 1996; M. S. Laverack, *The Physiology of Earthworms*, 1963; E. E. Ruppert et al., *Invertebrate Zoology*, 7th ed., 2004; J. A. Wallwork, *Earthworm Biology*, 1983.

## Oligoclase

A plagioclase feldspar with composition in the range  $Ab_{90}An_{10}$  to  $Ab_{70}An_{30}$ , where Ab represents the composition of albite,  $NaAlSi_3O_8$ , and An represents the composition of anorthite,  $CaAl_2Si_2O_8$ . The diagnostic properties are hardness on Mohs scale, 6–6.5; density, 2.65 g/cm<sup>3</sup>; two good cleavages that intersect at approximately 90°; and color usually white or colorless, transparent to translucent. The presence of minute, mutually parallel inclusions of hematite ( $Fe_2O_3$ ) causes a golden play of color in the variety of oligoclase called aventurine or sunstone. Repeated twinning is common and results in closely spaced striations visible with a hand-held magnifier. Mean refractive index is 1.545; the mineral may be optically positive or negative, depending on composition. Oligoclase is triclinic. The mineral is common in both plutonic and volcanic silicic igneous rocks, as well as in quartzofeldspathic and pelitic metamorphic rocks. See ALBITE; ANORTHITE; CRYSTAL STRUCTURE.

Like all feldspars, oligoclase is a tectosilicate. Each aluminum (Al) and each silicon (Si) atom is surrounded by four oxygen atoms that form a tetrahedron around it; each tetrahedron shares the four oxygen atoms at its corners with four other tetrahedra, forming a three-dimensional atomic framework containing open voids that accommodate the sodium-calcium (Na,Ca) atoms. At the highest temperatures typical of igneous processes, the Al and

Si atoms are statistically disordered among the symmetrically distinct tetrahedra. As cooling proceeds, Al and Si order preferentially into symmetrically different tetrahedra. If the composition is more Ab-rich than about  $Ab_{83}An_{17}$ , then exsolution or spinodal decomposition to an intergrowth of  $Ab_{76\pm6}An_{24\pm6}$  and completely ordered  $Ab_{99\pm1}An_{1\pm1}$  also occurs. The more calcic lamellae of this so-called peristerite intergrowth are e-plagioclase, consisting of alternating nanometer-scale lamellar domains of albitelike and anorthitelike structure. Oligoclases with bulk compositions more calcic than  $Ab_{83}An_{17}$  do not form peristerite intergrowths, but they nonetheless develop into e-plagioclases upon cooling. See FELDSPAR; IGNEOUS ROCKS; METAMORPHIC ROCKS.

Dana T. Griffen

**Bibliography.** W. A. Deer, R. A. Howie, and J. Zussman, *An Introduction to the Rock-Forming Minerals*, 2d ed., 1992; D. T. Griffen, *Silicate Crystal Chemistry*, 1992; J. V. Smith and W. L. Brown, *Feldspar Minerals*, 2d ed., 1988.

## Oligonucleotide

A single-stranded short polymer of deoxyribonucleic acid (DNA) or ribonucleic acid (RNA). Although the number of nucleic acid monomers (known as nucleotides) in the polymer varies widely, practical uses for oligonucleotides ranging in size from 4 to over 100 monomers have been reported.

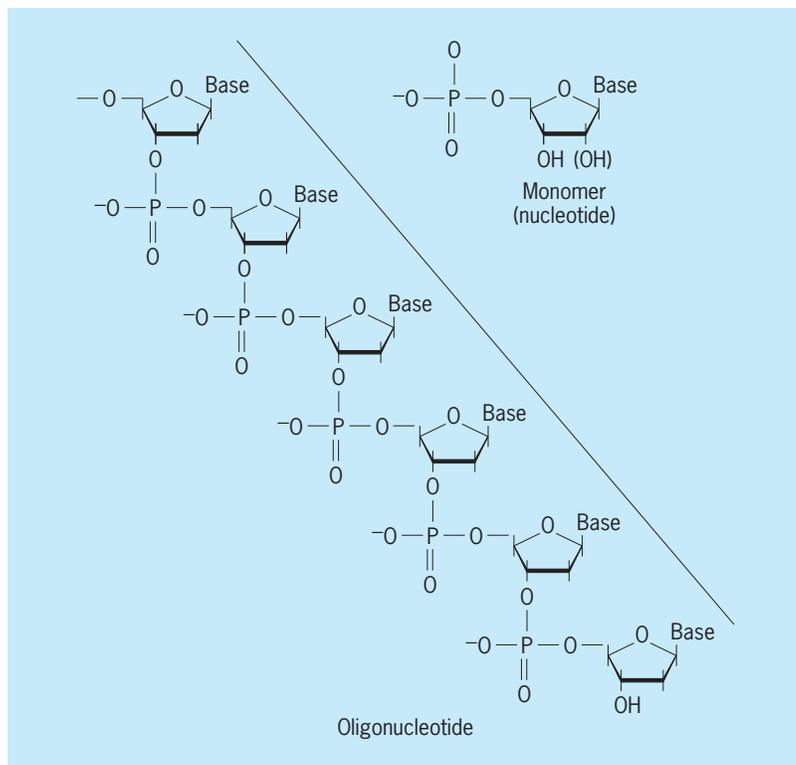
**Composition.** Each nucleotide in the nucleic acid polymer is composed of three parts: a five-carbon sugar, a nitrogenous base, and a phosphate group (**Fig. 1**).

The five-carbon sugar (ribose or 2'-deoxyribose) is covalently attached to the nitrogenous base through a glycosidic linkage to create a nucleoside. If the sugar is ribose, the resulting oligonucleotide is known as an RNA oligonucleotide, or oligoribonucleotide; if the sugar is 2'-deoxyribose, the resulting polymer is called either a DNA oligonucleotide or an oligodeoxyribonucleotide.

There are two classes of bases: purines (adenine and guanine) and pyrimidines (thymine, cytosine, and uracil). Guanine, cytosine, adenine, and uracil are the four bases found in RNA oligonucleotides; DNA oligonucleotides contain guanine, cytosine, adenine, and thymine. The sequence, or ordering, of these bases determines the specificity of the interaction of one oligonucleotide with a second polymeric strand (either DNA, RNA, or a protein) to form a duplex (**Fig. 2**).

The phosphate molecule covalently links the 5' hydroxyl group of one nucleoside to the 3' hydroxyl group of the next nucleoside. This creates a phosphodiester bond between successive sugar molecules, forming the backbone of the nucleic acid strand (**Fig. 1**). See DEOXYRIBONUCLEIC ACID (DNA); NUCLEIC ACID; RIBONUCLEIC ACID (RNA).

**Synthesis.** Oligonucleotides are synthesized from high-molecular-weight DNA or RNA in a manner that defines the order, or sequence, of purines and



**Fig. 1.** Oligonucleotides are composed of monomers called nucleotides. The most common chemical form of monomers has either deoxyribose or ribose (an additional OH) attached to a purine (adenine and guanine) or pyrimidine base [thymine (in DNA oligonucleotides), cytosine, uracil (in RNA oligonucleotides)]. Monomers are joined in a consistent and directional manner through the formation of phosphodiester bonds.

pyrimidines. This synthesis is most often carried out chemically using automated instruments, but may also be carried out enzymatically using RNA or DNA polymerase.

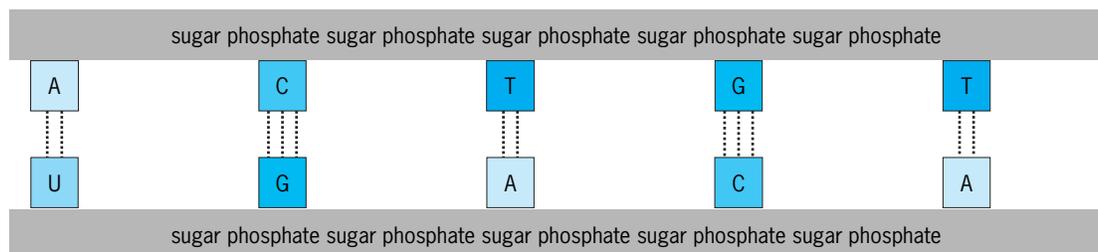
**Modifications.** The basic chemical structure of the oligonucleotide can be modified in a number of ways to enhance desired properties. For instance, purines or pyrimidines can be modified by alkylation (addition or substitution of an alkyl group) or deamination (removal of an amino group) to alter their interaction with other nucleic acids. A methyl group can be added to the sugar at the 2' carbon to reduce susceptibility to nucleases (enzymes that cleave nucleic acids by attacking the phosphodiester bond), or replaced entirely with other molecules, such as a morpholino. The bridge between monomers can be changed from a phosphodiester linkage to a phosphorothioate (P-S

bond) or phosphoramidate (P-N bond). These modifications are made to provide resistance against the numerous cellular nucleases. However, the vast majority of the oligonucleotides currently synthesized have unmodified purine and pyrimidines bases and 2'-deoxyribose as the sugar, and are linked through a normal phosphodiester bond.

**Applications.** Oligonucleotides are commonly used to bind to other nucleic acids or proteins through base-specific hydrogen bonds. Generally, they are used to form antiparallel duplexes, in which the sugar-phosphate backbones run in opposite directions (one strand runs in the 5' to 3'-direction and the other runs 3' to 5'). Oligonucleotides may also bind, in a sequence-specific manner, to a nucleic acid duplex to form three-stranded structures known as a triplexes. Oligonucleotides used to form triplexes with genomic DNA have been used experimentally to inhibit gene expression and to create site-directed mutations.

*In vitro: DNA oligonucleotides.* Oligonucleotides have been used to assemble genes, or even a small genome, as was demonstrated by the assembly of the complete poliovirus genome in a research laboratory. Perhaps the most common current use of chemically synthesized oligonucleotides is as part of a nucleic acid amplification technique called the polymerase chain reaction (PCR). In a PCR reaction, two different DNA oligonucleotides interact, in a sequence-specific manner, with a long segment of denatured DNA to form two antiparallel duplexes a defined distance apart (a few hundred to a few thousand base-pairs). A unique DNA polymerase known as Taq, which has been isolated from the bacteria *Thermophilus aquaticus*, uses the oligonucleotides as primers to synthesize DNA *in vitro* in a series of reactions that doubles the amount of newly synthesized DNA with each cycle of denaturation, annealing, and elongation. By repeating this cycle from 35 to 50 times, one copy of a starting DNA template can be amplified to produce useful amounts of a DNA fragment. See GENE AMPLIFICATION.

The cost of chemically synthesizing DNA oligonucleotides has decreased dramatically. A result of this important technological breakthrough has been the development of oligonucleotide gene chips, in which tens of thousands of DNA oligonucleotides (representing different potentially expressed genes) are attached to a solid matrix. Messenger RNA (mRNA) molecules from specific cells are first



**Fig. 2.** Oligonucleotides make sequence-specific, noncovalent dimers (or duplexes) with DNA or RNA through Watson-Crick base pairing.

amplified and labeled and then added to the oligonucleotide gene chip. Some of the labeled nucleic acid hybridizes to the complementary DNA oligonucleotides on the gene chip, and the remainder is washed away. A computer determines the location and intensity of the signal generated on the chip, thus determining a pattern of genes expressed in the original cells. This type of transcriptional profiling has revealed groups of genes that are coordinately regulated in some cancers and other disease states.

*In vivo: RNA oligonucleotides.* In 1998, it was found that long double-stranded RNA oligonucleotide duplexes injected into worms could alter gene expression. This process, known as RNA interference (or RNA silencing), has been described in cells from a number of diverse species. In cells, long RNA oligonucleotide duplexes are first cut into shorter, biologically active 21–25 base-pair fragments. (Alternatively, duplexes can first be made from RNA oligonucleotides 21–25 bases in length, and then introduced into the cell.) After associating with a number of specific proteins, an RNA oligonucleotide: protein complex is formed, in which one strand of the RNA duplex is degraded and the other serves as a guide, leading to the degradation of a complementary mRNA.

*In vivo: DNA oligonucleotides.* The therapeutic use of DNA oligonucleotides to treat diseases in humans has taken two main forms. The first is used in clinical trials as antiviral or anticancer therapeutic agents. Oligonucleotides act by serving as a specific inhibitor of messenger mRNA translation inside cells. The effect of the oligonucleotide is to reduce the level of proteins (encoded by mRNAs) that allow a virus to propagate or a cancer cell to divide in an unregulated way. The DNA oligonucleotide first binds to a complementary mRNA. This approach is often called “antisense,” which refers to the anti-parallel binding rules of guanine binding to cytosine (G-C), and adenine binding to uracil (A-U). The fate of the formed DNA-RNA heteroduplex depends upon the type of DNA oligonucleotide used. If the oligonucleotide has regions with native phosphodiester bonds, the RNA in the heteroduplex will be cleaved by the ubiquitous enzyme ribonuclease H, thus preventing translation of the mRNA. If the oligonucleotide is modified so that the heteroduplex is not recognized by ribonuclease H, then any inhibition of translation must result from preventing the proper assembly of the translational machinery (that is, by blocking ribosome binding).

A second therapeutic use for DNA oligonucleotides is to activate the immune system. The sequence cytosine-guanine (CpG) is usually methylated in vertebrate, but not bacterial, DNA. DNA oligonucleotides with unfamiliar cytosine methylation patterns provide a signal in humans to initiate cellular processes that activate both innate and acquired immunological responses. CpG immune modulators are currently being studied as agents to treat a number of human diseases. See GENE; GENETIC CODE; GENETIC ENGINEERING; IMMUNITY.

Daniel Weeks; John Dagle

*Bibliography.* J. Cello, A. V. Paul, and E. Wimmer, Chemical synthesis of poliovirus cDNA: Generation of infectious virus in the absence of natural template, *Science*, 297:1016–1018, 2002; J. M. Dagle and D. L. Weeks, Oligonucleotide-based strategies to reduce gene expression, *Differentiation*, 69:75–82, 2001; A. Fire et al., Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*, *Nature*, 391:806–811, 1998; A. M. Krieg, CpG motifs in bacterial DNA and their immune effects, *Annu. Rev. Immunol.*, 20:709–760, 2002; R. J. Lipshutz et al., High density synthetic oligonucleotide arrays, *Nat. Genet.*, 21:20–24, 1999; R. J. Lipshutz et al., Using oligonucleotide probe arrays to access genetic diversity, *Biotechniques*, 19:442–447, 1995; A. C. Pease et al., Light-generated oligonucleotide arrays for rapid DNA sequence analysis, *Proc. Natl. Acad. Sci. USA*, 91:5022–5026, 1994; K. H. Rubins et al., The host response to smallpox: analysis of the gene expression program in peripheral blood cells in a nonhuman primate model, *Proc. Natl. Acad. Sci. USA*, 101:15190–15195, 2004; R. K. Saiki et al., Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia, *Science*, 230:1350–1354, 1985; U. Scherf et al., A gene expression database for the molecular pharmacology of cancer, *Nat. Genet.*, 24:236–244, 2000; J. Soutschek et al., Therapeutic silencing of an endogenous gene by systemic administration of modified siRNAs, *Nature*, 432:173–178, 2004; C. A. Stein and A. M. Krieg, *Applied Antisense Oligonucleotide Technology*, Wiley, 1998; J. Summerton and D. Weller, Morpholino antisense oligomers: design, preparation, and properties, *Antisense Nucleic Acid Drug Dev.*, 7:187–195, 1997.

## Oligopygoida

An order of irregular echinoids in the superorder Neognathostomata resembling clypeasteroids but lacking the accessory ambulacral pores characteristic of that group. Oligopygoids have well-developed petals, and there are characteristic small demiplates present below the petals. These demiplates are wedge-shaped and usually do not reach the inner surface of the test. Each has a simple ambulacral pore. The apical disk is monobasal and the mouth oval and usually deeply sunken. Oligopygoids have a lantern, which closely resembles that of clypeasteroids, and their lantern muscle-attachment structures are a mixture of ambulacral and interambulacral processes.

There are two genera, *Oligopygus* and *Haimea*, containing about 25 species, all from the middle and upper Eocene of the Caribbean and Gulf of Mexico regions. They were probably infaunal deposit feeders like present-day laganiids. See ECHINODERMATA; NEOGNATHOSTOMATA.

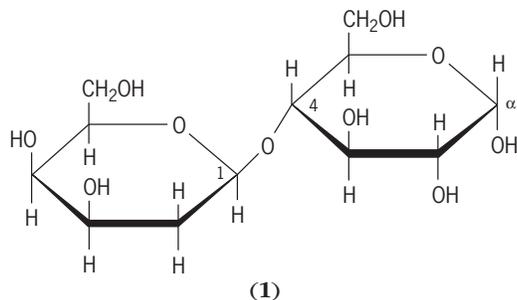
Andrew B. Smith

*Bibliography.* A. B. Smith, *Echinoid Palaeobiology*, 1984.

## Oligosaccharide

Oligo(few)saccharide(sugar): A carbohydrate molecule composed of 3–20 monosaccharides. Generally, free oligosaccharides do not quantitatively constitute a significant proportion of naturally occurring carbohydrates. Most carbohydrates that occur in nature are in the form of monosaccharides (such as blood sugar, or glucose), disaccharides (such as table sugar, or sucrose, and milk sugar, or lactose), and polysaccharides (such as starch and glycogen, polyglucose molecules, or chitin). See GLUCOSE; LACTOSE; MONOSACCHARIDE; POLYSACCHARIDE.

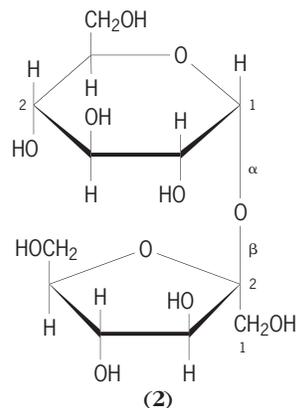
**Composition.** The monosaccharides of multiple sugar units such as disaccharides, oligosaccharides, and polysaccharides are connected with each other through bonds called glycosidic linkages. Monosaccharides are generally classified as aldose (an aldehyde sugar) or ketose (a ketone sugar). Chemically, these sugar molecules are found mainly as cyclic molecules, which are referred to as hemiacetals. Cyclic hemiacetals exist as five-membered rings (called furanoses) or six-membered rings (called pyranoses). Both classes of sugars are reducing sugars, meaning that the aldehyde or ketone group can become oxidized and in turn will reduce certain compounds. Simple sugars (monosaccharides) are linked primarily to other sugars and to other molecules through these reducing groups. For example, lactose (1) has the structure of galactosyl-



1,4-glucose, meaning that carbon-1 of galactose is linked to carbon-4 of glucose. The proper name for this structure of lactose is  $\beta$ -D-galactopyranosyl-(1-4)- $\alpha$ -D-glucopyranose. In ring structures, the aldehyde or ketone carbon is referred to as the anomeric carbon, and the cyclized aldose or ketose exists in one of two anomeric configurations designated as  $\alpha$  or  $\beta$ . The complexity of oligosaccharide chemistry is appreciated when the possible combinations are considered in forming a disaccharide between galactose and glucose. See GALACTOSE.

Carbon-1 of galactose can be linked to any of the hydroxyl groups of glucose, giving four possible disaccharides. Since carbon-1 of galactose exists as the anomeric carbon ( $\alpha$  or  $\beta$ ), the number is multiplied by 2, so that there are eight possible structures based only on these assumptions. The number of possible isomeric oligosaccharides that could exist for a trisaccharide composed of three different monosaccharides with the typical structure given in (1) is 1056. If the reducing group of one monosaccharide is linked

to the reducing group of another monosaccharide, such as in sucrose (2), the resulting disaccharide is nonreducing.



**Glycoconjugates.** Most naturally occurring oligosaccharides are linked either to proteins (glycoproteins) or to lipids (glycolipids). A number of different sugars can occur in oligosaccharides of glycoproteins and glycolipids. For example, in eukaryotic glycoproteins, seven sugars predominate: these are D-galactose, D-glucose, D-mannose, and L-fucose; the hexosamines *N*-acetyl-D-glucosamine and *N*-acetyl-D-galactosamine; and the nine-carbon sugars called sialic acids (in humans, the sialic acid is primarily *N*-acetylneuraminic acid). The amino sugars have the hydroxyl group on carbon-2 replaced by an amino group ( $\text{NH}_2$ ), which is in turn acetylated. Two of the common linkages of sugars to proteins are O-linked glycoproteins and N-linked glycoproteins. Examples are *N*-acetylgalactosamine, linked to the hydroxyl group of the amino acid serine, and *N*-acetylglucosamine, linked to the amide nitrogen of the amino acid asparagine. See GLYCOLIPID; GLYCOPROTEIN.

Glycoconjugates are present in essentially all life forms and particularly in cell membranes and cell secretions. Many hormones are glycoproteins, and an increasing number of enzymes have been shown to have sugars attached. Antigenic properties of the human red blood cell ABO blood group system are determined by glycolipid oligosaccharides. In fact, all the major protein components of blood serum, with the exception of serum albumin, are glycoproteins. See BLOOD GROUPS; CELL MEMBRANES.

Many changes in the structures of oligosaccharides of glycoconjugates have been detected in cancer cells. Often these alterations resemble structures present in fetal tissues. Changes or differences in oligosaccharide structures are generally the result of differences in biosynthetic pathways or of degradative pathways.

**Biosynthesis.** The biosynthesis of essentially all macromolecules such as proteins, nucleic acids, and polysaccharides requires an activated form of the monomeric units. The nature of the first activated sugar was established as a glucose molecule linked to a nucleotide, commonly called uridine diphosphate glucose (UDPglucose). Over 100 different nucleotide

sugars are known wherein both the nucleotide and the sugar can vary.

Specific enzymes called glycosyltransferases transfer one sugar at a time to acceptor molecules. The acceptors can be highly divergent substances such as proteins, lipids, and other sugars. For example, the blood type A glycoprotein of pig salivary secretion (an O-linked glycoprotein) is synthesized by adding the following sugars in sequence: *N*-acetylgalactosamine, galactose, fucose, and *N*-acetylgalactosamine from their respective nucleotide derivatives. Oligosaccharides of N-linked glycoproteins are mechanistically synthesized the same as O-linked glycoproteins; that is, one sugar is added at a time from nucleotide sugars. However, oligosaccharides of N-linked glycoproteins are first formed on a lipid intermediate as a dolichol pyrophosphate-oligosaccharide, and then the oligosaccharide is transferred to an asparagine group of the protein moiety by oligosaccharide transferase. Specific glycosidases (removal of certain sugars) and glycosyltransferases (addition of specific sugars) (together called oligosaccharide processing) are involved in the synthesis of an extremely large number of different oligosaccharide side chains.

**Biological functions.** There is a great interest in determining the functions of glycoconjugates, and especially the roles of the oligosaccharide side chains. Sugar-mediated glycoprotein clearance from blood serum was first established in studies on Wilson's disease, an inherited disorder of copper metabolism. The normal half-life of most glycoproteins in serum is generally about 56 h; but when the terminal sialic acid was removed from ceruloplasmin (the copper-carrying glycoprotein) by the glycosidase sialidase, liver plasma membranes recognized the exposed galactose moiety and essentially immediately removed the asialoceruloplasmin from circulation. Receptor molecules for hormones, growth factors, and other biological regulators as membrane-bound glycoconjugates have been identified and characterized. Proteins produced by recombinant deoxyribonucleic acid (DNA) technology, or biotechnology, which are potential glycoproteins, require specific glycosylation to elicit a proper biological response. For example, if erythropoietin, a glycoprotein which regulates the production of red blood cells, is not glycosylated properly, it is rapidly removed from the serum. An understanding of glycoconjugates in normal biological systems and in certain disease states is currently of great importance. See BIOTECHNOLOGY.

Don M. Carlson

Bibliography. B. Alberts et al., *Molecular Biology of the Cell*, 3d ed., 1994; A. L. Lehninger, D. L. Nelson and M. M. Cox, *Principles of Biochemistry*, 3d ed., 2000.

## Olive

Olive fruits, which are produced by a small to medium-sized evergreen tree (*Olea europaea*), can be eaten, after processing, as table olives, or can be



Fig. 1. Olive tree with fruits.

extracted for oil that is used on salads, for cooking, for body lotions, or for medicinal purposes. The olive tree (Fig. 1) is a historically ancient cultivated plant, having been domesticated by early civilizations in the eastern Mediterranean regions. Olive culture later spread to all the Mediterranean countries and subsequently, during the age of exploration, to South America, California, South Africa, and Australia. The major olive-producing countries are Spain, Italy, and Greece, which provide about 65% of the world's olives. Olives are grown commercially in the United States only in California, and are used primarily as table olives. The olive tree has very precise climatic requirements for good fruit production. First, the trees are killed outright if winter temperatures drop much below 14°F (−10°C). Second, trees of most varieties require a certain amount of winter cold in order to flower in the spring, ruling out the tropics, and third, they need a long, hot, dry summer growing period to properly mature the fruit. See EVERGREEN PLANTS; FRUIT, TREE.

Olive trees are very drought-resistant, but they produce better if they are given summer irrigation. The trees also respond markedly to nitrogen fertilizers with increased growth and yields.

Olive trees are unusually free of insect and disease pests. However, one insect, the olive fly (*Dacax olea*), is a serious problem in the Mediterranean countries, causing the fruits to be wormy. Fungus diseases affecting olives are verticillium wilt and peacock spot. A bacterial disease, olive knot, is also a problem in many countries.

Olive trees bloom from early April to late May in the Northern Hemisphere, depending on the season and the region. The wind-pollinated flowers are small and of two types: having both male and female flower parts, or having only male flower parts (Fig. 2). The trees tend toward alternate bearing—a heavy crop



Fig. 2. Olive branches showing (left) blooming in the spring and (right) mature fruits ready for harvest for oil in the winter.

one year followed by no crop or only a light crop the next.

Olive fruits are harvested for table olives in the autumn, when the fruits change from green to straw or to a slightly red color. Raw fruits contain a bitter glucoside which makes them inedible, but treatment with an alkali such as sodium hydroxide (lye) neutralizes the bitterness. The lye must subsequently be leached out of the fruits with water. In another method most of the bitterness can be removed by leaching with salt water. A lactic acid fermentation process is widely used in the Mediterranean countries to preserve the olives. In California the lye treatment is mostly used, after which the olives are canned in brine and autoclaved at 240°F (115.5°C) for 1 h. If the fruits are aerated during the lye treatments, they become black, and are called black-ripe olives. If they are kept in the solution and not exposed to air, they remain green, and are called green-ripe olives.

For oil production, the fruits are harvested in midwinter when they have become black and have reached their maximum oil content—15 to 25% of the fresh weight, depending on the variety. In producing olive oil, the freshly harvested fruits including pits are ground, after which this material is placed in burlap or cloth bags and placed under high pressure. Oil and water are extracted and transferred to tanks, where the oil rises to the top and is removed. The oil is washed with warm water several times to remove the bitterness, and again separated by gravity. The oil is passed through centrifuges to remove all water, and through filters to remove suspended matter and give the oil clarity. See FAT AND OIL (FOOD).

Some leading table varieties are Sevillano, Manzanillo, Amphissa, Mission, Calamata, and Ascolano. Oil varieties are Picual, Frantojo, Chemlali, and Mission. Many varieties are grown either for table olives or for olive oil.

Hudson T. Hartmann

## Olivine

A name given to a group of magnesium-iron silicate minerals crystallizing in the orthorhombic system. Crystals are usually of simple habit, a combination of the dipyrmaid with prisms and pinacoids. The luster is vitreous and the color olive green, giving rise to the name olivine. Hardness is  $6\frac{1}{2}$ –7 on Mohs scale; specific gravity is 3.27–3.39, increasing with increase in iron content.

Olivine is a nesosilicate with the composition  $(\text{Mg,Fe})_2\text{SiO}_4$ . It comprises a complete solid solution series from the pure iron member fayalite,  $\text{Fe}_2\text{SiO}_4$ , to the pure magnesium member forsterite,  $\text{Mg}_2\text{SiO}_4$ . Minerals of intermediate composition have been given their own names but are usually designated simply as olivine. The magnesium-rich varieties are more common than those rich in iron. The minerals tephroite,  $\text{Mn}_2\text{SiO}_4$ , monticellite,  $\text{CaMgSiO}_4$ , and larsenite,  $\text{PbZnSiO}_4$ , although not in this chemical series, are of the olivine structure type.

**Occurrence.** Olivine is found in some crystalline limestones but occurs chiefly as a rock-forming mineral in igneous rocks. It varies greatly in amount from an accessory to the main rock-forming constituent. Although it may be present in granites and other light-colored rocks, it is found chiefly in the dark rocks such as gabbro, basalt, and peridotite. The rock dunite is composed almost completely of olivine.

Olivine is one of the first minerals to form upon crystallization of a magma. It is believed that this early-formed olivine accumulated through the process of magmatic differentiation to form the large dunite masses. The type locality is at Dun Mountain, New Zealand; the rock is also found with corundum deposits in North Carolina. See MAGMA; PERIDOTITE.

Olivine alters readily to serpentine, a hydrous magnesium silicate. The alteration may take place on a large scale to form great masses of the rock serpentine, or on a small scale to form pseudomorphs of serpentine after single crystals of olivine. See SERPENTINE; SERPENTINITE.

At a few localities, notably on St. John's Island in the Red Sea and in Burma, olivine is found in transparent crystals. These are cut into gemstones which go under the name of peridot. Olivine is a major constituent of many stony meteorites. See METEORITE.

Cornelius S. Hurlbut, Jr.

**Spinel transition.** When subjected to very high pressures, minerals with olivine structure are transformed to denser polymorphs with spinel structure. For the composition  $\text{Fe}_2\text{SiO}_4$ , the univariant curve for the olivine-spinel transformation has been determined experimentally. The divariant transition interval for compositions in most of the series  $\text{Fe}_2\text{SiO}_4$ – $\text{Mg}_2\text{SiO}_4$  has also been determined at pressures up to about 100 kilobars. Olivine is believed to constitute a high proportion of the upper mantle, estimates ranging between 60 and 90%. The composition of the olivine is generally assumed to be about 90% forsterite ( $\text{Mg}_2\text{SiO}_4$ ) and 10% fayalite ( $\text{Fe}_2\text{SiO}_4$ ). The experimental results, combined with estimates of the temperature distribution in the

mantle, indicate that olivine of this composition would begin to transform to a more iron-rich spinel phase at a depth of about 220 mi (370 km), with transformation being completed at a depth of about 270 mi (435 km). Gradients of seismic-wave velocities in the transition zone of the upper mantle are abnormally high in two layers 30–60 mi (50–100 km) thick, one beginning at depth of 220 mi (350 km) and the other at about 390 mi (630 km). The olivine-spinel transformation appears to correlate well with the 220-mi (350-km) layer. The spinel present below this layer may be replaced in turn by even more dense phases at the 390-mi (630-km) depth. See LITHOSPHERE; SPINEL.

Peter J. Wyllie

Bibliography. W. A. Deer, R. A. Howie, and J. Zussman, *Rock-Forming Minerals, Orthosilicates*, Vol. 2a, 2d ed., 1997.

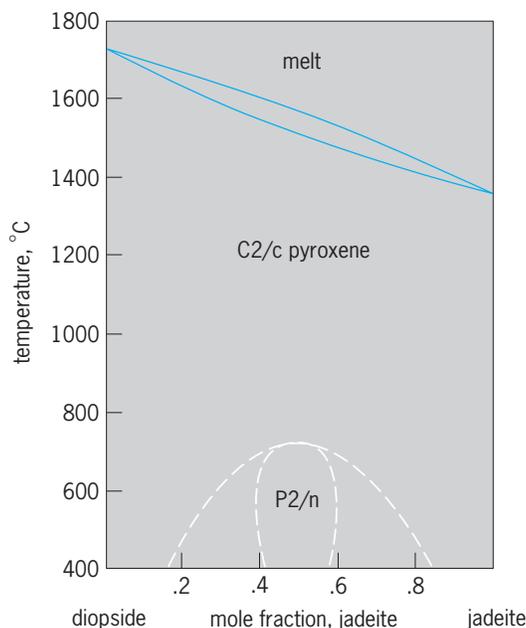
## Omphacite

The pale to bright green monoclinic pyroxene found in eclogites and related rocks. Omphacites are essentially members of the solid solution series between jadeite ( $\text{NaAlSi}_2\text{O}_6$ ) and diopside ( $\text{CaMgSi}_2\text{O}_6$ ), but contain smaller amounts of the components acmite ( $\text{NaFe}^{3+}\text{Si}_2\text{O}_6$ ), hedenbergite ( $\text{CaFe}^{2+}\text{Si}_2\text{O}_6$ ), and Tschermak's pyroxene ( $\text{CaAlAlSiO}_6$ ). The density ranges from 3.16 to 3.43 g/cm<sup>3</sup> (1.8–2.0 oz/in.<sup>3</sup>), hardness is 5–6, cleavage is well developed on {110}, and simple or lamellar twinning on {100} is common. With four formula units to the unit cell, omphacite crystallizes in space groups Cs/c, as well as in the lower symmetry group P2/n. The possibility of space groups P2 and P2/c has also been suggested for some omphacites. C2/c structure is characteristic of omphacites crystallized above a P2/n-C2/c inversion temperature (just over 1300°F or 700°C), although there is some evidence for metastable growth at lower temperatures. In this structure (Ca,Na) and (Al,Mg) are more or less randomly distributed over M2 and M1 sites. See DIOPSIDE; ECLOGITE; JADEITE; SOLID SOLUTION.

Compositions close to 50% jadeite and 50% diopside ( $\text{Jd}_{50}\text{Di}_{50}$ ) have a tendency to order their M1 chains into a regular alternation of Al and Mg octahedra. To preserve local charge balance, the Na and Ca ions are partially ordered on the M2 chains. Substitution of the acmite component tends to inhibit this ordering, so that natural P2/n omphacites cluster about the  $\text{Jd}_{50}\text{Di}_{50}$  composition. The schematic phase diagram (see *illus.*) is based on the uncertain assumption that the ordering transition is first-order in character.

Omphacite is stable only at the relatively high pressures of the blueschist and eclogite facies of metamorphism, where it is associated with minerals such as glaucophane and lawsonite or pyropic garnet, respectively. In such environments it also occurs in veins, either on its own or with quartz. See GLAUCOPHANE.

High pressures favor minerals in which aluminum is octahedrally coordinated, so that for the reaction



Temperature-composition section at 30 kilobars (3 gigapascals) pressure for the jadeite-diopside system. The low-temperature region is shown schematically. °F = (°C × 1.8) + 32.

$\text{CaMgSi}_2\text{O}_6 + x\text{NaAlSi}_2\text{O}_6 = (\text{CaMgSi}_2\text{O}_6 : x\text{NaAlSi}_2\text{O}_6) + x\text{SiO}_2$  jadeite solubility in diopside increases with increasing pressure. Experimental work on this reaction allows an estimate of the pressure of equilibration of omphacite + albite + quartz assemblages if a temperature estimate is available. For feldspar-absent assemblages, the pressure estimated for a particular omphacite composition is a minimum value. See HIGH-PRESSURE MINERAL SYNTHESIS; PYROXENE.

Timothy J. B. Holland

Bibliography. W. A. Deer, R. A. Howie, and J. Zussman, *Rock-Forming Minerals*, Vol. 2B, Double-Chain Silicates, 2d ed., 1997; K. Smulikowski, Chemical variation of clinopyroxenes in eclogites and related rocks, *Bull. Acad. Pol. Sci., Serie Sciences de la Terre*, 24(1):1–7, 1976.

## Oncholaimida

An order of nematodes comprising the single superfamily Oncholaimoidea. These nematodes are principally marine and brackish-water forms with alleged predaceous and carnivorous feeding habits. The external amphidial aperture is an oval or a widened ellipse. Generally, the stoma is armed with one dorsal tooth and two subventral teeth, and its wall may be further fortified with transverse rows of small denticles. The stoma is divisible into the cheilostome (secondary blastocoel invagination) and the esophastome (primary blastocoel invagination). In some species the stoma of the adult male is collapsed or indistinct. The cephalic sensilla are in two whorls: one is circumoral and composed of six papilliform sensilla; the second combines the ancestral two whorls

of six and four into a single whorl of ten setiform sensory organs. In some forms these sensilla are papilliform. The cylindrical-to-conoid esophagus may exhibit a series of muscular bulbs posteriorly. The cuticle is generally smooth, and over the length of the body there are scattered sensory setae or papillae. *See* NEMATA. Armand R. Maggenti

## Oncofetal antigens

Small proteins or carbohydrates (capable of stimulating an immune response) that are present in normal fetal tissues during early development but also are abnormally expressed by adult tumors. Oncofetal antigens are different from tumor-associated antigens, which are found only in tumors and have no role in development.

One of the goals of modern cancer research is to determine biochemical or immunological differences between normal and tumor cells. Although the malignant state is not necessarily characterized by gross aberrations in metabolic pathways or in the biochemical constitution of cell components, a subtle change does occur in a cell during neoplastic transformation. One area of investigation is the study of oncofetal antigens. These antigens, primarily glycoprotein in nature and produced by cancerous cells, are the products of one or more genes that normally are expressed only during fetal development and then are repressed in adult life. Production of these proteins in association with cancer as a result of activation of control genes, by unknown mechanisms, has been termed retrogenetic expression. Advances in tumor immunology have enabled the use of oncofetal antigens in both the diagnosis and treatment of cancer. In particular, fetal antigens in adult body fluids can serve as tumor markers for the detection of early oncogenic processes and can aid in monitoring the efficiency of, cancer treatment as well as in the development of new treatment modalities.

The two best-known oncofetal antigens are  $\alpha$ -fetoprotein and carcinoembryonic antigen (CEA); however, several lesser-known oncofetal antigens, such as pancreatic oncofetal antigen, have been described.

**Alpha fetoprotein.** Alpha fetoprotein is a substance produced by the liver of a healthy fetus. The exact function of this protein is unknown. However, the physiochemical properties of  $\alpha$ -fetoprotein are similar to those of albumin, the major normal serum component in the adult. In fetal serum, an inverse relationship exists between  $\alpha$ -fetoprotein level decreases and the albumin level increases with increasing fetal age. After birth, the infant's liver stops producing  $\alpha$ -fetoprotein, and an adult liver contains only trace amounts. During pregnancy, the fetus excretes  $\alpha$ -fetoprotein in urine, and some of the protein is able to cross the placenta and enter the mother's blood. *See* ALPHA FETOPROTEIN.

*Birth defects.* Analysis of the amount of  $\alpha$ -fetoprotein found in the mother's blood, can aid in the determination of the probability that the fetus is at risk for certain birth defects. Alone,  $\alpha$ -fetoprotein

screening cannot diagnose a birth defect. It is often part of a "triple check" blood test that analyzes three substances as risk indicators of possible birth defects:  $\alpha$ -fetoprotein, estriol, and human chorionic gonadotropin (HCG). When all three substances are measured in the mother's blood, the accuracy of the test results increases. *See* HORMONE; PRENATAL DIAGNOSIS.

Abnormally high  $\alpha$ -fetoprotein may indicate that the fetus has an increased risk of a neural tube defect, the most common and severe type of disorder associated with increased  $\alpha$ -fetoprotein. These include spinal column defects and anencephaly (a severe and usually fatal brain abnormality). If the screening test indicates abnormally high  $\alpha$ -fetoprotein, ultrasound is used for the diagnosis. Levels may also be high if there is too little fluid in the amniotic sac around the fetus, more than one developing fetus, or a pregnancy that is further along than estimated.

For unknown reasons, abnormally low  $\alpha$ -fetoprotein may indicate that the fetus has an increased risk of Down syndrome. Down syndrome is a condition that is linked to an abnormality of chromosome 21 (called trisomy 21). If the screening test indicates an abnormally low  $\alpha$ -fetoprotein, amniocentesis is used in the diagnosis. Abnormally low levels of  $\alpha$ -fetoprotein can also occur when the fetus has died or when the mother is overweight. *See* CONGENITAL ANOMALIES; DOWN SYNDROME.

*Cancer.* Although  $\alpha$ -fetoprotein in human blood gradually disappears after birth, it never disappears entirely. It may reappear in liver disease or in tumors of the liver, ovaries, or testicles. The  $\alpha$ -fetoprotein test is used to screen people at high risk for these conditions. After a cancerous tumor is removed, an  $\alpha$ -fetoprotein test can monitor the progress of treatment. Continued high  $\alpha$ -fetoprotein levels suggest the cancer is growing. *See* CANCER (MEDICINE).

**Carcinoembryonic antigen.** The carcinoembryonic antigen of human digestive system cancer has been the most studied oncofetal antigen. Originally thought to be a specific antigen of the fetal digestive tract as well as cancer of the colon, carcinoembryonic antigen is now known to occur normally in feces and secretions from the pancreas and bile ducts. It also appears in the plasma due to cigarette smoking and a diverse group of neoplastic and non-neoplastic conditions, including cancers of the colon, pancreas, stomach, lung, and breast; alcoholic cirrhosis; pancreatitis; inflammatory bowel disease; and rectal polyps.

*Cancer prognosis.* The carcinoembryonic antigen levels in the blood are one of the factors that doctors consider when determining the prognosis, or most likely outcome, of cancer. The carcinoembryonic antigen test is ordered for patients with known cancers, commonly cancer of the gastrointestinal system. These include cancers of the colon, rectum, stomach (gastric cancer), esophagus, liver, and pancreas. It is also used to predict the outcome of cancers of the breast, lung, and prostate gland. In general, a higher carcinoembryonic antigen level predicts a more severe disease, one that is less likely to be curable. However, the carcinoembryonic antigen

test does not provide clear-cut information; therefore, the results are usually considered along with other laboratory or imaging studies to follow the course of the disease.

**Cancer treatment response.** Once cancer treatment has begun, carcinoembryonic antigen tests have a valuable role in monitoring the patient's progress. The primary use of carcinoembryonic antigen is in monitoring response to treatment of colorectal cancer. A decreasing carcinoembryonic antigen level means therapy is effective in fighting the cancer. A stable or increasing carcinoembryonic antigen level may mean that the treatment is not working, or that the tumor is growing. Serial carcinoembryonic antigen measurements, that is, tests done over a period of time, are the most useful. A single test result is difficult to evaluate, but a number of tests, done weeks apart, show trends in disease progression or regression.

**Cancer recurrence.** Carcinoembryonic antigen tests are also used to help detect recurrence of a cancer after surgery or other treatment has been completed. A rising carcinoembryonic antigen level may be the first sign of cancer return, and may show up months before other test results or patient symptoms would raise concern. Unfortunately, this does not always mean the recurrent cancer can be cured. For example, only a small percentage of patients with colorectal cancers and rising carcinoembryonic antigen levels will benefit from another surgical exploration. Those with recurrence in the same area as the original cancer, or with a single metastatic tumor in the liver or lung, have a chance that surgery will eliminate the disease. Patients with more widespread return of the cancer are generally not treatable with surgery. However, the carcinoembryonic antigen test does not distinguish between the two groups.

**Nonstandard treatments.** Patients who are most likely to benefit from nonstandard treatments, such as bone marrow transplants, may be identified on the basis of carcinoembryonic antigen values combined with other test results. Carcinoembryonic antigen levels also may be one of the criteria for determining whether the patient will benefit from more expensive studies, such as computerized tomography (CT) or magnetic resonance imaging (MRI). See COMPUTERIZED TOMOGRAPHY; MAGNETIC RESONANCE; MEDICAL IMAGING; TRANSPLANTATION BIOLOGY.

**Pancreatic oncofetal antigen.** Several other less defined oncofetal antigens, such as the pancreatic oncofetal antigen, have been described. It is associated with pancreatic cancer and is found in extracts of the fetal pancreas, adult human pancreatic tumors, sera from individuals with carcinoma of the pancreas, and—to a lesser degree—in persons with other malignancies. See ANTIGEN; ONCOLOGY. Zoë Cohen; T. Ming Chu

**Bibliography.** H. Magdelenat, Tumour markers in oncology: Past, present and future, *J. Immunol. Meth.*, 150(1-2):133-43, Jun 24, 1992; R. G. McKinnell et al., *The Biological Basis of Cancer*, 3d ed., Cambridge University Press, 1998, reprint 2000; R. W. Ruddon, *Cancer Biology*, 3d ed., Oxford University Press, 1995.

## Oncogenes

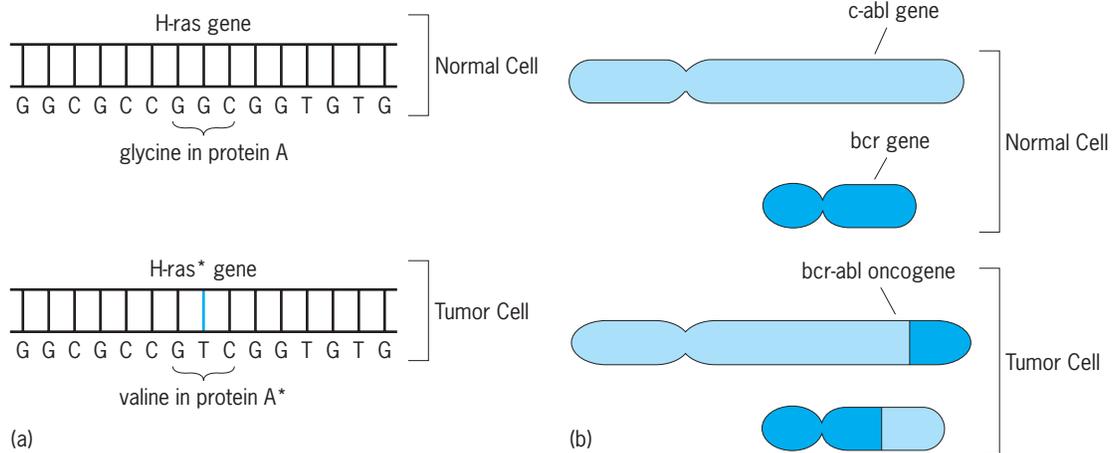
Genes that contribute to the conversion of a normal cell into a cancerous cell. Oncogenes can derive from cellular genes that undergo mutations that alter their expression or activity, or from viruses that carry oncogenes within their genome and are transferred into the cell by infection. See ANIMAL VIRUS; GENE.

**Genetic basis of neoplasia.** Most, if not all, neoplasms (cancers) arise when individual cells within the body suffer irreversible genetic damage that leads to unrestrained cell growth, a block in the normal process of differentiation, or programmed cell death (apoptosis). Since genes consist of segments of deoxyribonucleic acid (DNA) that are linked in a specific order along each chromosome, any agent or process that breaks the DNA or alters the individual chemical subunits of the DNA can cause genetic damage.

Genetic lesions in neoplastic cells can affect two classes of genes. The first genes identified were called oncogenes, and the mutations that alter these genes occur on only one of the two similar chromosomes in a cell. The unaffected chromosome retains the normal cellular gene, whereas the mutant form of the gene on the affected chromosome overrides the normal gene function and promotes neoplastic growth. In contrast to oncogenes, mutations in cancer cells can also inactivate or even delete genes whose function is required for normal cell growth. This second class of genes, called tumor suppressors, appears to restrain cell growth or prevent the accumulation of mutations, and a cell remains normal as long as at least one copy of the tumor suppressor gene is functional. However, if both normal copies of the gene are inactivated by mutation or lost from the cell, neoplastic growth can ensue.

Oncogenic mutations in cellular genes are never inherited from parent to child, presumably because the mutations are so detrimental to normal development that an embryo with the mutation would not survive. Instead, these mutations arise during the lifetime of an individual through spontaneous mutation or exposure to environmental carcinogens, both of which damage DNA. In contrast, a small fraction of human tumors are caused by the inheritance of a mutation in a tumor suppressor, giving some families a predisposition to cancer. See DEOXYRIBONUCLEIC ACID (DNA); MUTAGENS AND CARCINOGENS; MUTATION; PREDICTIVE GENETICS OF CANCER; TUMOR.

**Mechanisms of oncogene formation.** Cellular genes have highly specific functions and patterns of expression in normal cells, and the mutations that convert a normal gene into an oncogene within a tumor cell can alter either of these properties. Four major types of oncogene-forming lesions are commonly found in tumor cells: (1) Point mutations can change individual bases within the segment of the gene that encodes a protein, and the resulting mutant protein has enhanced or altered activity that contributes to cancer cell growth (see **illustration**). (2) Chromosomes can break near specific genes and rejoin in such a



**Two common mechanisms of oncogene activation in cancer cells. (a) DNA mutation.** Mutation of the DNA sequence that encodes a protein involved in growth signals. In this example, the GGC codon directs the amino acid glycine as the 12th amino acid in the H-ras protein. Mutation of a single base changes the codon to GTC and results in a valine as the 12th amino acid of the oncogenic H-ras protein. This mutant H-ras protein was one of the first and most potent oncogenes discovered. **(b) Chromosomal translocation.** Chromosomes can break and rejoin near specific genes. In this example, the *c-abl* tyrosine kinase gene on chromosome 9 breaks and rejoins to the *bcr* gene on chromosome 22. The resulting fusion protein has high tyrosine kinase enzyme activity and contributes to the development of chronic myelogenous leukemia. This chromosomal translocation was the first to be found in human cancer.

way that an oncogene is created through an abnormally high level of gene expression or by the loss of the gene's normal mode of regulation. These lesions are referred to as chromosomal translocations, and they are particularly common in leukemias and lymphomas. Chromosomal translocations can also fuse two genes together so that a new oncogene with altered or enhanced activity is created (see illus.). (3) Oncogenes can be created by the localized amplification of small chromosomal domains, which leads to extra copies of the gene within the amplified segment and elevated levels of oncogene-encoded ribonucleic acid (RNA) and protein. (4) Viruses or mobile genetic elements can be inserted into the chromosome near a specific gene and alter the expression of the gene by imposing new regulatory signals, a process called insertional mutagenesis. See CHROMOSOME ABERRATION.

**Viral oncogenes.** Some viruses contain oncogenes that are part of the viral genetic material and not derived from the cell. Although most human cancer is not associated with infectious agents such as viruses, some cancers have been shown to be caused at least in part by viruses. The most common is cervical carcinoma, where the causative agent is human papilloma virus (HPV). The HPV genome contains two viral oncogenes that contribute to the outgrowth of cancer cells. Rather than providing enzymatic or gene regulatory activities of their own, the HPV oncogenes function by binding to cellular tumor suppressor proteins and inactivating their function. Since the tumor suppressor protein is rendered nonfunctional, the viral infection promotes unrestrained cell growth or genetic instability by the same mechanisms that might otherwise occur through genetic lesions that mutate or delete the tumor suppressor genes themselves. Immunization against HPV has recently become a reality, and elimi-

nating this virus from the population should dramatically reduce the incidence of cervical cancer. See ANIMAL VIRUS; TUMOR VIRUSES.

**Oncogene functions.** Normal cells divide or remain quiescent as the result of diverse extracellular signals (such as growth factors and cell-cell interactions). These signals are interpreted by receptors and cytoplasmic factors, and eventually the signals lead to reprogramming of gene expression in the nucleus. Thus, it is logical that the majority of oncogenes encode proteins that are components of this signal transduction pathway and they inappropriately activate the normal growth signals. Many oncogene-encoded proteins localize to the plasma membrane and are enzymes that phosphorylate other proteins (protein kinases), altering the activity of these substrate proteins. Other oncogenic proteins function as regulatory subunits of enzymes, and the mutations signal the enzymes to be active continuously rather than in a regulated manner. The oncogene-encoded proteins that localize to the nucleus are largely DNA-binding proteins that regulate specific sets of cellular genes involved in growth control. Finally, oncogenes can function not by promoting cell growth but by blocking apoptosis. Since many cells within the body (especially blood cells) are programmed to die after a limited lifespan, overexpression of a protein that blocks this programmed cell death can also contribute to cancerous growth. See CELL SENESCENCE AND DEATH; GENE ACTION; GENETIC ENGINEERING.

**Oncogene-targeted cancer therapy.** The recognition that there are specific genetic mutations that drive the growth of tumor cells provides a new opportunity to develop drugs that target cancer cells. Since these mutations exist only in the tumor, a drug designed to inhibit the mutated or overexpressed oncogenic protein can slow the growth of tumor cells (or even kill them) with reduced side effects on normal

tissues. Several new drugs have been introduced in recent years that selectively target oncogenic mutations, and they are proving to be highly efficacious against certain forms of cancer. For example, antibodies directed against the epidermal growth factor (EGF) receptor selectively target breast cancer cells with amplification and overexpression of the EGF receptor gene. Another successful therapy targets the abnormal enzyme that is created by a chromosomal translocation in chronic myelogenous leukemia. These drugs and many others in the pharmaceutical pipeline will be available to treat the dozens of different cancers, each of which may have a unique collection of causative mutations. Therapies will be chosen based on the underlying molecular pathology of each tumor. Oncogene-targeted therapy promises a new era in the treatment of the cancers that afflict a large fraction of the population. See CANCER (MEDICINE); GENE ACTION; ONCOLOGY. Michael Cole Bibliography. D. W. Kufe et al. (eds.), *Cancer Medicine*, 6th ed., Lewiston, New York [distributor], BC Decker, Hamilton, Ontario, 2003; B. Vogelstein and K. W. Kinzler (eds.), *The Genetic Basis of Human Cancer*, 2d ed., McGraw-Hill, Medical Pub. Div., New York, 2002.

## Oncology

The study of cancer. There are five major areas of oncology: etiology, prevention, biology, diagnosis, and treatment. As a clinical discipline, it draws upon a wide variety of medical specialties; as a research discipline, oncology also involves specialists in many areas of biology and in a variety of other scientific areas. The approach to the study of cancer is multidisciplinary because cancer is a fundamental problem in biology. Oncology has led to major progress in the understanding not only of cancer but also of normal biology.

Cancer defies simple definition. It is a disease that develops when the orderly relationship of cell division and cell differentiation becomes disordered. In general, tissues of the body are composed of cells with distinct capabilities. Some cells are specialized to conduct the functions of the particular tissue or organ and such cells do not divide. Division is normally restricted to less well-differentiated stem cells within tissues that proliferate to replace senescent mature, well-differentiated cells. As a proliferating stem cell differentiates, it loses the capacity to divide. In cancer, dividing cells seem to lose the capacity to differentiate, and they acquire the ability to invade through basement membranes and spread (metastasize) to many areas of the body through the bloodstream or lymphatics. Cancer is usually clonal, that is, it develops initially in a single cell. That abnormal cell then produces progeny that may behave rather heterogeneously. Some progeny continue to divide, some develop the capacity to metastasize, and some develop resistance to therapeutic agents. This single cell and its progeny, if unchecked, typically lead to the death of the host.

## Causes of Cancer

What causes cancer? Cancer is generally thought to result from one or more permanent genetic changes in a cell. In some cells a single mutational event can lead to neoplastic transformation, but for most tumors it appears that carcinogenesis is a multistep process. The notion that cancer reflects a permanent genetic change is not absolute, as some cancer cells can develop into normal, functioning, and differentiated progeny if placed under the inductive influence of a fertilized egg. Thus, some cancer cells retain responsiveness to differentiation signals.

Although some rare congenital conditions lead to cancer in infancy, the vast majority of human cancers arise as a result of the complex interplay between genetic and environmental factors. Therefore, estimates of the amount of cancer attributable to each cause seem fruitless. Without question, there are forms of cancer clearly related to particular environmental exposures (Table 1); it is equally clear, however, that these factors act on a genetic substrate that may be either susceptible or resistant to the development of cancer. It is well known that lung cancer is much more common in smokers than in non-smokers. However, the incidence of lung cancer is about 30 times higher among smokers with high lung levels of aryl hydrocarbon hydroxylase, an enzyme that metabolizes benzo(a)pyrene in tobacco smoke to highly carcinogenic epoxides, than among other smokers. In this instance, the environmental agent, tobacco smoke, interacts with the genetically determined high level of metabolizing enzyme to facilitate the carcinogenesis.

**Genetic factors.** The emergence of cancer appears to involve the accumulation of genetic damage in a target tissue. In one study of genetic changes in colon cancer, it appeared that progression from normal colonic epithelium to small adenomatous polyp to large adenomatous polyp to overt carcinoma was accompanied by mutation in the *ras* oncogene followed by deletion of sequences on chromosomes 5 and then 18 and 17. Such complex, sequential genetic changes specific to tissues appear to underlie the progression to cancer. In some tissues, like colon, each genetic lesion produces a clinical abnormality. In other tissues (such as the ovary), stepwise progression has not been documented.

Multistep progression is quite complicated to study in experimental systems. Much work has focused on the identification, isolation, and characterization of oncogenes, which have the ability to transform non-neoplastic cells into cancer cells. More than 50 bona fide or putative oncogenes have been characterized and mapped throughout the human genome. Many were identified because they accounted for the transforming ability of animal retroviruses. However, over time it has become clear that most oncogene sequences are a normal part of the genome, and such endogenous oncogenes, termed proto-oncogenes, serve important roles in cell proliferation and differentiation. Proto-oncogenes encode for growth factors, growth factor receptors, nuclear

TABLE 1. Selected environmental causes of human cancer

Agent	Site or type of cancer
Life-style risk factors	
Tobacco, smoking	Lung, larynx, mouth, pharynx, esophagus, bladder, pancreas, kidney, cervix, breast
Tobacco, smokeless	Mouth, oral cavity
Alcohol	Mouth, pharynx, esophagus, larynx, liver, breast, colon (?)
Sunlight	Skin, lip
Anabolic steroids	Liver
Phenacetin	Renal pelvis
Occupational risk factors	
Asbestos	Lung, pleura, peritoneum
Benzene	Leukemia
Benzo(a)pyrene	Lung, skin
Arsenic	Lung, skin, liver angiosarcoma
Aromatic amines	Bladder
Chromium	Lung
Nickel dust	Lung, nasal sinuses
Vinyl chloride	Liver angiosarcoma
Wood dust	Nasal sinuses, Hodgkin's disease
Herbicides	Lymphoma, soft tissue sarcoma
Leather	Nasal sinuses
Infectious risk factors	
<i>Helicobacter pylori</i>	Gastric MALT lymphoma, gastric cancer
<i>Schistosoma haematobium</i>	Squamous carcinoma of the bladder
<i>Clonorchis sinensis</i>	Cholangiocarcinoma of the liver
Epstein-Barr virus	Lymphomas, nasopharyngeal carcinoma
Hepatitis-B virus	Hepatocellular carcinoma
Human T-cell lymphotropic virus	Adult T-cell leukemia
Papilloma virus	Cervix, skin, vulva
Iatrogenic risk factors	
Chronic alkylating agents	Leukemia, bladder
Radiation	Nearly any site
Estrogens, conjugated	Endometrium
Estrogens, synthetic	Transplacental cervix, vagina
Immunosuppressive agents	Lymphoma, skin, Kaposi's sarcoma
Thorium oxide	Liver angiosarcoma
Chlornaphazine	Bladder

proteins, regulatory enzymes such as tyrosine kinases, and components of intracellular pathways of signal transduction mediated by growth and differentiation factors acting at the cell membrane. When proto-oncogenes are normally regulated, their functions are essential to the cell. However, there are at least six mechanisms by which proto-oncogenes may be altered so that their expression results in neoplastic transformation. These mechanisms include mutation, amplification, translocation, promoter insertion, recombination, and insertion into the cell membrane. They generally result in the escape of a proto-oncogene from its normal controls and lead to aberrant expression or to expression of an altered product with unregulated functions. Some oncogenes require another, complementary oncogene to transform a normal cell. Their aberrant expression commonly results in autonomous cell proliferation and escape of the cell from the normal mechanisms that control proliferation. See ONCOGENES.

One or more genetic mechanisms may prohibit oncogenes from causing transformation. Many tumor types are associated with nonrandom chromosomal deletions. The best-studied example is the retinoblastoma gene on chromosome 13, whose gene product is capable of interacting with transforming viral proteins from adenovirus and papilloma virus families. It appears that the transforming genes are more efficient at transforming their target if the retinoblastoma gene product is absent. When either the prod-

uct or the gene itself is inserted into malignant cells that lack it, the malignant phenotype of the cell is reversed. Thus, these recessive genes appear to function as tumor suppressor genes or antioncogenes, and their absence creates a cell more vulnerable to malignant transformation. See HUMAN GENETICS.

There is a close link between oncogenes and growth factors. Unlike normal cells, whose proliferation is often regulated by the availability of a trophic growth factor produced by a distinct cell type, many tumor cells are capable of autocrine growth, that is, they secrete their own growth factors and proliferate autonomously. Another mechanism of uncontrolled proliferation is escape from negative growth control. Many cell types can be inhibited by negative growth factors such as transforming growth factor-beta. Some tumor cells fail to activate this factor (which requires cleavage of an inactive precursor to become active), and others fail to express receptors for the growth-inhibitory polypeptides. The failure of some tumor cells to differentiate normally may also be related to alterations in the expression or function of receptors for differentiation factors. See GROWTH FACTOR.

Some cancers are increased in families. About 10% of breast cancers develop in women with a strong family history of breast cancer in first-degree relatives. In about half of these families, germline mutations in *BRCA1*, a gene on chromosome 17 encoding a protein involved in DNA repair, result in a

nonfunctional gene product. Women with *BRCA1* mutations have an 80% lifetime risk of developing breast cancer and a 33% chance of developing ovarian cancer. Women with a strong family history of breast cancer should be encouraged to undergo genetic screening. Those of Ashkenazi Jewish descent may have a specific *BRCA1* mutation, deletion of adenine and guanine at position 185. Currently, women with *BRCA1* mutations are advised to undergo bilateral mastectomy and bilateral salpingo-oophorectomy to prevent the cancers to which they are predisposed. Another breast cancer susceptibility gene has been mapped to *BRCA2* on chromosome 13. *BRCA2* mutations are about half as common as *BRCA1* mutations and predispose both male and female carriers to breast cancer. Familial predispositions to colorectal cancer include familial polyposis coli (mutations in the *APC* gene on chromosome 5) and Lynch syndrome (hereditary non-polyposis colorectal cancer; mutations in mismatch repair genes on chromosomes 2 and 3). These syndromes account for less than 10% of colorectal cancers. Prophylactic total colectomy is often recommended.

**Environmental factors.** Environmental factors involved in the development of cancers can be chemical, physical, or biological carcinogenic agents (Table 1). At least three stages occur in the natural history of cancer development from environmental factors. The first stage is initiation, which is a specific alteration in the deoxyribonucleic acid (DNA) of a target cell; environmental agents may act by inducing expression of oncogenes. Many chemical carcinogens are capable of inducing mutation, which can result in activation of the transforming potential of an oncogene. The second phase, promotion, involves the reversible stimulation of expansion of the initiated cell or the reversible alteration of gene expression in that cell or its progeny. Agents that serve as promoters are often incapable of inducing cancers on their own, but significantly enhance the development of cancers by initiating agents. Because promotion is thought to be reversible, it is a target for prevention. The final phase of carcinogenesis is progression. It is characterized by the development of aneuploidy and clonal variation in the tumor; these in turn result in invasiveness, metastasis and growth advantage.

*Chemical carcinogens.* The first clear demonstration of an environmental risk factor for the development of a cancer was made in 1775, when it was noted that chimney sweeps, a group with a considerable occupational exposure to coal tar, had a high incidence of scrotal cancer. Over 140 years later, the precise cause of the cancer in coal tar was found to be benzo(*a*)pyrene.

Epidemiologic studies attempt to identify factors associated with risk, and experimental studies define the precise nature of the factor. Chemical compounds capable of inducing cancer include polycyclic hydrocarbons, aromatic amines, and alkylating agents. Many chemical carcinogens are actually prodrugs activated to become carcinogenic by the body's metabolic machinery, such as the microsomal

enzymes that, ironically, evolved to detoxify toxic compounds. The active metabolite of many chemical carcinogens is a free-radical compound, and nearly all chemical carcinogens have been found to interact directly with DNA, forming adducts that can result in errors in base sequence during replication.

Tobacco, the most important environmental carcinogen, accounts for an increasing number of deaths each year from cancer as well as heart and nonmalignant respiratory diseases. Perhaps the most disturbing information about smoking is the evidence that nonsmokers who inhale exhaled smoke are also at risk for smoking-associated cancers.

The role of dietary factors in carcinogenesis is considerably more controversial. Ingestion of alcohol has been associated with increased risk of cancers of the esophagus, pharynx, larynx, and mouth—the same sites associated with cigarette smoking. Some data implicate drinking alcohol with breast and colon cancer, but the carcinogenic potential of other foodstuffs is unclear and the data linking dietary carcinogens to cancer are generally weak. Nitrosamines in smoked meat and salted fish are probably carcinogenic. However, there are foods that appear to have the capacity to detoxify the carcinogens that occur in other foods. Epidemiologic evidence suggests that high-fat and high-calorie diets are associated with an increased risk of colon and breast cancer. However, causal association has not been conclusively shown.

*Physical carcinogens.* The three major physical carcinogens are ionizing radiation, ultraviolet radiation, and foreign bodies. Ionizing radiation can originate from many sources and can occur in many forms, but the two major categories are electromagnetic radiation, from x-rays and gamma rays, and particle radiation, originating in electrons, protons, neutrons, and alpha particles. The rate at which energy is deposited in the tissue is characteristic of the type of radiation and is called the linear energy transfer. Electromagnetic radiation is weakly ionizing and has a low linear-energy-transfer value, but particle radiation is densely ionizing and has a high linear-energy-transfer value. The biological damage done by radiation is higher with high-linear-energy-transfer radiations than with low-linear-energy-transfer radiations. The major physical features that determine the risk of radiation-induced cancer are linear-energy-transfer value, dose, dose rate, and fractionation. Radiation may induce cancers at any site. More than 80% of radiation exposure is from natural sources such as cosmic rays, terrestrial gamma rays, and radon. Radon may also emanate from the ground and from building materials, disperse in the air, and decay into short-lived aerosolized alpha particles with high-linear-energy-transfer characteristics that are capable of inducing lung cancer when inhaled. Radon daughters, such as polonium-214 and -218, could cause serious radiation injury. Radon exposure varies markedly across the United States, but on average probably accounts for slightly over half the exposure to radiation. Medical uses of radiation account for about 18% of the total radiation exposure. See LINEAR ENERGY TRANSFER (BIOLOGY); RADON.

It is known that radiation can induce mutations in DNA and can activate oncogenes, but the precise mechanism of radiation-induced carcinogenesis is unclear. Also, it is not clear why different forms of radiation-induced cancer have such a long latency. For example, cancers of the breast and thyroid can develop more than 20 years after radiation therapy of the chest and neck for Hodgkin's disease. For leukemia, the latent period appears to be much shorter (4–6 years), and the risk for leukemia decreases after 10 years following radiation. Certain genetic conditions, such as familial dysplastic nevus syndrome and ataxia-telangiectasia, may be associated with an increased susceptibility to radiation-induced cancer, but mechanisms are not known. Radiation-induced solid tumors may emerge in radiation-exposed patients at about the same age that tumors of the same tissue develop in nonirradiated persons, but with increased frequency, suggesting that a genetically determined series of processes leads to neoplastic transformation in a particular tissue. Radiation acts to increase the probability that a tissue will undergo transformation, rather than to accelerate the transformation process.

Ultraviolet radiation, mainly from the Sun, is carcinogenic to skin. The incidence of skin cancers other than melanoma is much higher in southern latitudes and on sites exposed to the Sun, and the incidence of melanoma may also be augmented. The UVB portion (280–320 nanometers) of the ultraviolet spectrum is most damaging to tissues. Ultraviolet radiation induces the formation of pyrimidine dimers (usually between thymines), which brings about alterations in the normal sequence of bases in DNA. Individuals with red hair and with certain genetic conditions, such as xeroderma pigmentosum, are less able to repair ultraviolet-induced damage and are more susceptible to Sun-induced skin cancers than the general population. Protection against ultraviolet radiation is offered by the ozone layer of the atmosphere and by pigmentation of the skin. Ultraviolet radiation also exerts systemic effects by altering the immune function. *See* RADIATION BIOLOGY; RADIATION INJURY (BIOLOGY).

Foreign bodies are only an occasional cause of tumors, and they rarely induce sarcomas and neoplasms in the site into which they are introduced; and in animal models the composition of the foreign body is less important than its size and shape. The same materials are more carcinogenic if they are fibrous than if they are powdered, porous, or perforated. The malignant transformation is probably related to a derangement during the connective tissue reaction to the foreign body. The foreign body that is most carcinogenic in humans is asbestos, a natural mineral fiber that is inhaled and is associated with lung cancer and mesotheliomas. Other potentially carcinogenic fibers include synthetic vitreous and crystalline fibers. *See* ASBESTOS.

*Biological carcinogens.* Although viral carcinogenesis is widespread in nature, examples in humans are not numerous, and knowledge of oncogenes has relied on the prevalence of tumors induced by retro-

viruses in animals; but the precise mechanism of carcinogenesis is not yet clear. There are four documented oncogenic virus types in humans: the hepadnavirus family (specifically hepatitis B virus), associated with hepatocellular carcinoma; the herpesvirus family (specifically Epstein-Barr virus), associated with Burkitt's lymphoma, other lymphomas in the setting of immunodeficiency, nasopharyngeal carcinoma; some papilloma viruses, associated with cervical cancer and skin cancer; and the human retrovirus family (specifically HTLV-I), associated with adult T-cell leukemia/lymphoma. Herpesvirus 6 has been implicated in some lymphomas. *See* ANIMAL VIRUS; EPSTEIN-BARR VIRUS; TUMOR VIRUSES.

Chronic bacterial infections may lead to cancers in certain sites. For example, *Helicobacter pylori* gastritis can lead to gastric MALT (mucosa-associated lymphatic tissue) lymphoma and to gastric adenocarcinoma. Infections with certain parasites can also lead to carcinogenesis, perhaps much like foreign bodies do. An example is bilharzial squamous-cell bladder cancer, which follows chronic bladder infection with *Schistosoma haematobium*.

Hormones may also be considered biological carcinogens. Estrogens can cause endometrial cancer. The synthetic estrogen diethylstilbestrol (DES), administered to women in the 1950s to prevent spontaneous abortion, was found to result in clear-cell carcinoma of the vagina in some of their female offspring between the ages of 15 and 30. Transplacental carcinogenesis in this setting is not understood. *See* ESTROGEN; HORMONE; MUTAGENS AND CARCINOGENS.

### Cancer Prevention

An obvious starting point for cancer prevention is avoidance of environmental agents that contribute to carcinogenesis. Eliminating use of tobacco and alcohol would reduce cancer mortality by more than 30%. Skin cancer prevention is a simple matter of using the available tools (clothing and sunscreens) to protect the skin from Sun damage. Maintaining a dark and even tan can result in premature aging of the skin as well as cancer. Thinning of the ozone layer may one day necessitate routine application of sunscreen products before an individual spends any time outdoors.

The role of diet in cancer prevention is controversial. Epidemiologic evidence suggests a particularly strong link between a high-fat, high-calorie, low-fiber diet and an increased risk of colon cancer. But a change to a low-fat, low-calorie, high-fiber diet may not alter the risk. Calorie restriction and exercise may reduce the carcinogenicity of carcinogenic agents. The addition to the diet of carotenoids, selenium, vitamins A, D, and E, and some short-chain fatty acids may prevent cancers in high-risk populations, but there is no evidence that any dietary supplement will prevent cancer. *See* NUTRITION.

There are a variety of clinical settings in which surgery may prevent cancer. Repositioning of undescended testes may prevent testicular cancer, and resection of the large intestine may prevent colon

cancer in individuals with familial polyposis, familial colon cancer, or ulcerative colitis. Surgical removal of the thyroid will prevent medullary carcinoma in individuals with certain types of multiple endocrine neoplasia, breast removal can be preventive in familial breast cancer, and removal of the ovaries can prevent cancer in familial ovarian cancer. Preventive medical interventions are being evaluated, including the nonsteroidal anti-inflammatory agent sulindac to block the development of neoplastic polyps in familial polyposis, and tamoxifen and/or aromatase inhibitors in familial breast cancer.

### Cancer Biology

The study of cancer biology picks up where cancer etiology leaves off, namely, at the point where the tumor has developed into a clonal cluster of autonomously proliferating cells. The pathological correlate of this stage of tumor development is carcinoma in situ; a condition in which no tissue destruction is evident, but atypical-appearing cancer cells are present at their site of origin. The transition from carcinoma in situ to locally invasive cancer is accompanied by dissolution of the basement membrane, penetration of tumor cells through the membrane and into the supportive tissues, and disruption of the supportive tissues. The tumor cell is thought to bind to the basement membrane and tissue matrix through cell surface receptors for laminin and fibronectin. The tumor cell secretes enzymes that dissolve the membrane or matrix and also secretes an autocrine motility factor that helps the cell move through the membrane and matrix. The tumor may also be affected by chemotactic factors in the host tissues. Expansion of the primary tumor in locally invasive cancer is always accompanied by the development of blood vessels, which are often defective and easily invaded by individual and clumped tumor cells. The tumor cells can also invade regional blood vessels and lymphatics and circulate throughout the body, attaching to endothelium in a distant organ site, inducing retraction of the endothelium, and becoming attached to the endothelial basement membrane. Once attached to the basement membrane, the tumor cells are covered over by the endothelial cells and effectively separated from the flow of blood. Local dissolution of the basement membrane then occurs, allowing the tumor to completely spread into the tissue and reestablish a blood flow in the breached vessel. As it grows, more blood vessel development nourishes the enlarging tumor.

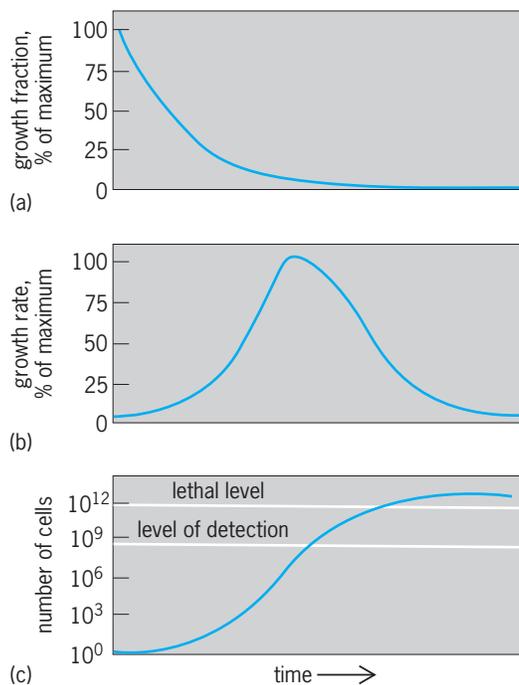
During metastasis, tumor cells must overcome host defenses. They have various mechanisms to do so. For example, they produce new cell surface receptors to facilitate basement membrane and matrix binding; make new enzymes such as collagenases, serine proteases, metalloproteinases, cysteine proteinases, and endoglycosidases to facilitate their invasiveness; and secrete motility factors to enable them to move through the holes and pathways created by their enzymes. They avoid detection by the immune system through a variety of techniques. Unlike animal tumors, most human tumors are poorly immuno-

genic, and when they do express a tumor-related antigen, its expression is often heterogeneous or it is shed by the tumor cell. Tumor cells may fail to express class I major histocompatibility (MHC) antigens. Since cytotoxic T cells recognize tumors only through MHC antigens, this makes the T cells blind to the presence of the tumor. Tumor cells often produce factors that are immunosuppressive. *See* CELLULAR IMMUNOLOGY; HISTOCOMPATIBILITY.

An unexplained feature of metastasis is the propensity of certain tumor types to spread to specific organs. There must be one or more mechanisms by which some metastatic tumor cells distinguish lung from liver, for example, but the nature of the mechanism is not known.

### Cancer Detection, Diagnosis, and Treatment

The curability of a cancer is inversely proportional to its size at diagnosis, and so early diagnosis is critical. However, at the earliest point that a tumor is likely to be detected on physical exam (about 0.4 in. or 1 cm) it already contains  $10^8$ – $10^9$  tumor cells and has undergone about 30 doublings. A lethal tumor burden occurs at around 40 doublings (about  $10^{12}$  cells); thus, most of a tumor's natural history occurs before detection. Most tumors do not grow exponentially. Exponential growth, which is typical of bacteria, occurs when a cell divides to produce two cells capable of division. Tumors follow Gompertzian growth kinetics, in which not all progeny of a dividing cell are capable of division. Generally the growth fraction of a solid tumor decreases with the size of the tumor (see **illustration**), and contrary to common belief, tumors generally do not proliferate more actively than normal tissues. Not all the factors that



**Gompertzian growth kinetics.** (a) Growth fraction declines exponentially. (b) Growth rate of a tumor, a measure of the absolute number of dividing cells plotted as a function of time. (c) Tumor size as a function of time.

**TABLE 2. Guidelines for cancer screening**

Location	Procedures	Baseline age	Screening frequency
Breast	Mammography	35–40	Every 1–2 years between ages 40 and 49; annually thereafter
Cervix	Papanicolaou (Pap) smear + pelvic examination	18 (or earlier if sexually active)	Annually; every 3 years after three consecutive normal results
Colon	Fecal occult blood test + sigmoidoscopy	40	Every 3–5 years following normal baseline examination

regulate tumor cell growth are defined. However, at least some of the tumor cells die or are hypoxic, and others have sustained too much genetic damage to divide. Tumor doubling times range from 3 to 90 weeks in human tumors and, in general, metastases double more quickly than primary lesions, largely because of the selection metastases have undergone to become established.

**Tumor detection.** There are two major strategies to detect tumors at the earliest possible stage in their history: responding to the seven warning signals of cancer and screening populations at high risk. The seven danger signals of cancer are (1) unusual bleeding or discharge, (2) a lump or thickening in the breast or elsewhere, (3) a sore that does not heal, (4) change in bowel or bladder habits, (5) persistent hoarseness or cough, (6) persistent indigestion or difficulty in swallowing, and (7) change in a wart or mole.

Screening has been enormously successful in reducing mortality from cancer of the cervix, but it has taken many years to develop appropriate guidelines for screening. Large-scale screening can be costly, and the method used must be safe, inexpensive, and accurate. For example, the Papanicolaou (Pap) smear for detecting cervical cancer relies heavily on the interpretive abilities of the technician reading the cytology slide. Screening guidelines for other types of cancer are considerably more controversial (Table 2). False positive screening tests increase patient risk and expense associated with follow-up testing. Certain techniques may reduce the size of the population that needs screening, and the screening tests may be able to detect cancers at an earlier stage. The detection of genetic lesions in patient samples promises to further enhance the specificity and sensitivity of screening. For example, screening for colorectal cancer currently depends heavily on finding occult blood in the stool. Detection of mutated genes in stool specimens may improve the detection of colorectal cancer; abnormal cells are continuously shed from tumors, but bleeding in such cancers is only intermittent. See MAMMOGRAPHY; RADIOGRAPHY.

**Diagnosis.** The diagnosis of cancer depends on the careful examination of biopsy material. Light-microscopic analysis of hematoxylin- and eosin-stained material is the primary technique. Most cancers are readily diagnosed in this fashion; the grading of histologic (tissue) atypia can be an important predictor of outcome. However, other techniques can assist the pathologist in defining the diagnosis and in conveying useful prognostic information; special stains can detect intracellular markers of particular

tumors, such as melanin and nerve-specific enolase; electron microscopy can reveal pathognomonic cellular organelles; and immunohistochemistry can define cell surface characteristics and cytoplasmic contents of prognostic importance. The detection of genetic lesions characteristic of certain cancers also aids in diagnosis. See CLINICAL PATHOLOGY.

Cancers arising in tissues having ectodermal or endodermal origins are generally called carcinomas; those derived from glands are called adenocarcinomas. Cancers arising in tissues derived from mesoderm are called sarcomas; those of lymphohematopoietic origin are lymphomas and leukemias. The cardinal microscopic features of malignancy are anaplasia, invasion, and metastasis.

Once a diagnosis of cancer is made, it is critical to determine the extent to which the disease has spread. This is called staging. It is distinct from grading, which is an assessment of histologic atypia performed with a microscope. Staging entails performing a careful physical examination, various radiographic studies (computed tomography, lymphography, nuclear medicine scanning), and perhaps surgical procedures (biopsies, endoscopies) to examine those sites to which a particular tumor type is most likely to spread. For example, patients with breast cancer often undergo evaluation of the liver, brain, and bones to search for metastatic disease, whereas patients with lymphoma generally require assessment of lymph node groups, bone marrow, and liver. Often the results of such staging tests determine the nature and extent of therapy.

For some tumors, assessment of serum levels of proteins or hormones can act as a measure of total body tumor burden. Such tumor markers can complement but not replace other staging techniques. In some instances they can be used to follow the progress of treatment and detect early relapse (Table 3).

**Treatment.** There are four major approaches to cancer treatment: surgery, radiation therapy, chemotherapy, and biological therapy. These modalities are often used together with additive or synergistic effects. Surgery and radiation therapy are most effective in curing localized tumors and together result in the cure of about 50% of all newly diagnosed cases. Once the cancer has spread to regional nodes or distant sites, it is generally incurable with the use of local therapies alone. Systemic administration of a combination of chemotherapeutic agents may cure another 15–18% of all patients.

*Surgery.* This treatment modality is used for the following purposes: (1) definitive treatment of primary

TABLE 3. Selected serum tumor markers useful in assessing disease status

Marker	Disease
Human chorionic gonadotrophin (HCG)	Choriocarcinoma
Alpha fetoprotein (AFP)	Testicular cancer
Immunoglobulin	Testicular cancer
CA-125	Hepatocellular carcinoma
Carcinoembryonic antigen (CEA)	Multiple myeloma
	Ovarian cancer
	Colon cancer
	Breast cancer
Prostate-specific antigen	Prostate cancer
Calcitonin	Medullary carcinoma of the thyroid
CA 19-9	Gastrointestinal malignancies
Interleukin-2 receptors	Hairy cell leukemia
	Adult T-cell leukemia/lymphoma
Neuron-specific enolase	Neuroblastoma
Lactate dehydrogenase (LDH)	Lymphoma
	Wilms' tumor
	Ewing's sarcoma
	Testicular cancer

cancer; (2) reduction of tumor bulk to increase the efficacy of subsequent treatment by other modalities; (3) removal of metastases; (4) treatment of oncologic emergencies (spinal cord compression); (5) palliation (relief of biliary or gastrointestinal obstruction); and (6) reconstruction and rehabilitation.

**Radiation therapy.** This physical method of treating cancer is analogous in some ways to surgery. Radiation will kill only those cells that receive sufficient linear energy transfer. There are two major methods of delivering radiation: teletherapy, where an external beam of radiation is aimed at the tumor site, and brachytherapy, where the radiation source is placed within or near the target. Electromagnetic and particulate forms of radiation can be used to facilitate local control of tumor growth, each of which has advantages for specific clinical situations. Normal tissues vary a great deal in the amount of radiation they can safely tolerate, and it is this tolerance that limits the total acceptable dose of radiation. For radiation to kill a cell, oxygen must be present in the tissue. Many tumors have regions of tissue that are poorly oxygenated, and such tissue is relatively radiation-resistant. Hypoxic cell sensitizers are being tested, and other compounds have been developed in an effort to improve the therapeutic ratio by protecting normal tissues from radiation damage. Radiation therapy is a component critical to the primary treatment of a number of tumor types, including breast cancer, head and neck cancer, cervical cancer, brain tumors, lung cancer, and certain stages of lymphoma and Hodgkin's disease. See RADIOLOGY.

**Chemotherapy.** Chemotherapeutic agents are toxic compounds that exert their maximum antitumor effects when employed at the maximum tolerated dose. Different classes of drugs are used to kill tumor cells by different mechanisms, and are employed mainly for the treatment of metastatic disease. Efficacy is usually directly proportional to dose and inversely proportional to tumor burden. With a drug, as with radiation therapy, toxicity to normal tissues (especially the highly proliferative tissues such as bone marrow and gastrointestinal tract mucosa) limits the

amount that can be safely administered. Available agents are capable of curing at least a fraction of patients with 16 different advanced-stage malignancies, and both breast cancer and colorectal cancer appear curable by adjuvant chemotherapy (administered in patients who appear to be free of disease after primary surgery but who have a high likelihood of relapsing if treated with surgery alone). A list of cancers responsive to chemotherapy follows.

#### Curable in advanced stages

- Choriocarcinoma
- Acute lymphocytic leukemia
- Hodgkin's disease
- Diffuse aggressive lymphoma
- Follicular mixed lymphoma
- Testicular cancer
- Acute myelogenous leukemia
- Ovarian cancer
- Lymphoblastic lymphoma
- Burkitt's lymphoma
- Embryonal rhabdomyosarcoma
- Wilms' tumor
- Peripheral neuroepithelioma
- Neuroblastoma
- Ewing's sarcoma
- Small cell carcinoma of the lung

#### Curable in the adjuvant setting

- Breast cancer
- Colorectal cancer
- Osteogenic sarcoma
- Soft tissue sarcoma

#### Responsive but not yet curable in advanced stage

- Head and neck cancer
- Breast cancer
- Colorectal cancer
- Cervical cancer
- Multiple myeloma
- Chronic lymphocytic leukemia
- Chronic myelogenous leukemia
- Bladder cancer
- Adrenal cancer

Prostate cancer  
 Medulloblastoma  
 Follicular small-cleaved-cell lymphoma  
 Melanoma  
 Endometrial cancer

Two main factors limit the success of chemotherapy: inability to deliver the agents with adequate dose intensity and the development of drug resistance. Theoretically, small increases in the amount of drug delivered could markedly improve the outcome; similarly, small decreases in the amount of drug delivered could drastically reduce the efficacy of the treatment. Analysis of clinical trial results confirms that delivery of agents with less dose intensity results in the cure of fewer patients; however, it has not yet been shown definitively that increasing dose intensity cures more patients. A major emphasis in cancer treatment research is on the development of techniques to deliver chemotherapeutic agents with a higher dose intensity (that is, higher doses of drug per week). Preliminary evidence suggests that the use of high-dose therapy along with bone marrow transplantation to reverse marrow toxicity dramatically improved the outcome in lymphoma and leukemia.

There are diverse mechanisms of drug resistance. The multidrug resistance gene produces a membrane protein that acts as an efflux pump to excrete a variety of structurally unrelated chemotherapeutic agents from the cell. Also, sometimes the tumor cell amplifies the gene whose product is antagonized by the drug; for example, dihydrofolate reductase is amplified to overcome the toxic effects of methotrexate. Drug resistance can result from poor transport into the cell, poor activation of the drug, inactivation of the drug, and altered pools of competing biochemical substrates. Often the use of agents in combinations (combination chemotherapy) overcomes the capacity of the tumor cells to adapt to the treatment. See CHEMOTHERAPY.

**Biological therapy.** This therapy uses an agent that is capable of altering the host-tumor relationship in favor of the host. Biological therapy aims at the rational design of therapeutic interventions to boost host defense or attack the biology of the tumor cell; to inhibit its growth, invasiveness, or metastatic potential; or to promote its differentiation. Biological agents are diverse in constitution and mechanism of action. Cytokines such as interferon, tumor necrosis factor, and interleukin-2 are products of the host and have a variety of mechanisms of action. Monoclonal antibodies are tailored to specifically attack a tumor cell, block critical functional receptors, call forth a cellular response to the tumor, or deliver to it a fatal dose of radiation, toxin, or drug. Adoptive cellular therapy uses expanded host effector cells capable of killing tumor cells in laboratory cultures. Tumor-nonspecific immune stimulants such as BCG (Bacille Calmette-Guérin) are effective in bladder carcinoma in situ. Vaccines against tumors and tumor viruses are showing promising effects in early clinical testing. Differentiating agents such as retinoids

are active in acute promyelocytic leukemia. Novel agents targeting oncogene products (such as imatinib mesylate in chronic myeloid leukemia) or signal transduction pathways (gefitinib against epidermal growth factor receptor) are emerging as effective new therapies. Colony-stimulating factors ameliorate the marrow-destroying tendencies of radiation therapy and chemotherapy. Agents aimed at interfering with the metastatic process, such as laminin fragments to prevent tumor cells from adhering to basement membranes, have also been under investigation. A large number of biological therapies are being tested. The use of biological agents in combination and with the other treatment modalities promises to further increase the fraction of patients with disseminated cancer who respond to therapy. See MONOCLONAL ANTIBODIES.

**Relief of symptoms.** Relieving the symptoms of cancer and alleviating the side effects of agents used to treat it is another important aspect of treatment. Many agents and interventions are available for these purposes. Pharmacologic agents can control nausea and vomiting. Various strategies are available to control pain, improve appetite, and combat insomnia and mood changes. Surgical procedures and radiological techniques can palliate many of the complications of cancer that formerly were incapacitating. Even when the hope for a cure has dwindled, the oncologist can relieve much suffering. See CANCER (MEDICINE); TUMOR. Dan L. Longo

**Bibliography.** M. D. Abeloff et al., (eds.), *Clinical Oncology*, 2004; V. T. DeVita, Jr., S. Hellman, and S. A. Rosenberg (eds.), *Cancer: Principles and Practice of Oncology*, 7th ed., 2004; A. T. Skarin, *Atlas of Diagnostic Oncology*, 2002; I. F. Tannock and R. P. Hill (eds.), *The Basic Science of Oncology*, 3d ed., 1998.

## Onion

A cool-season biennial, *Allium cepa*, of Asiatic origin and belonging to the plant order Liliales. The onion is grown for its edible bulbs (**Fig. 1**). See LILIALES.

Related species are leek (*A. porrum*), garlic (*A. sativum*), Welsh onion (*A. fistulosum*), shallot (*A. ascalonicum*), and chive (*A. schoenoprasum*). See GARLIC.

**Propagation.** The common onion is grown as an annual and is propagated most frequently by seed sown directly in the field. Onions may also be grown from transplants started in greenhouses or outdoor seedbeds or from small bulbs, called sets, grown the previous year. Field spacing varies; plants are generally grown 1–4 in. (2.5–10 cm) apart in 14–18-in. (36–46-cm) rows. The Egyptian tree or top onion (*A. cepa* var. *vivaparum*) produces little bulbs or topsets in the flower cluster, and the multiplier or potato onion (*A. cepa* var. *aggregatum*) multiplies by branching at the base.

**Varieties.** Onion varieties (cultivars) are classified mainly according to pungency (mild or pungent) and use (dry bulbs or green bunching). Bulbs may be

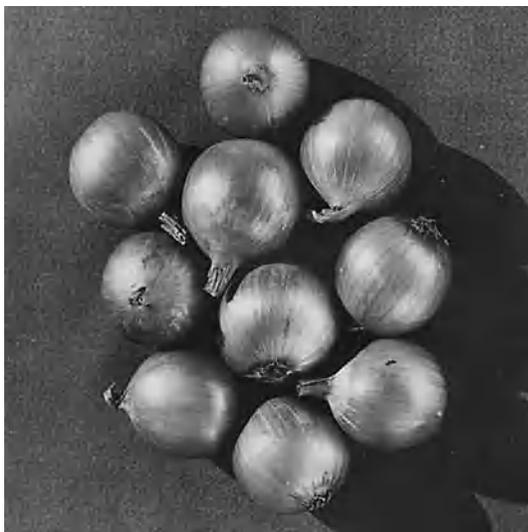


Fig. 1. Onion bulbs. (Asgrow Seed Company, Subsidiary of The Upjohn Company)

white, red, or yellow. Varieties differ markedly in their keeping quality and in their response to length of day. Hybrid varieties, produced with male-sterile breeding lines, and with increased disease resistance, longer storage life, and improved quality, are rapidly displacing older varieties.

**Harvesting.** The harvesting of dry-bulb varieties usually starts after the leaves begin to turn yellow and fall over, generally 3–4 months after planting. Bulbs to be stored are cured by exposure to warm dry air. Bunching onions are ordinarily harvested when the bulbs are  $\frac{1}{4}$  in. (0.6 cm) or larger in diameter.

Texas, New York, and California are important producing states. H. John Carew

**Diseases.** The onion and its relatives (chives, garlic, leek, shallot) are subject to many diseases, the most serious of which are caused by fungi and bacteria. Some diseases that occur on the plant in the field affect the yield, quality, or both; others are important in storage and transit, and some diseases are important both in the field and during marketing. Onion diseases cause millions of dollars in damages each year throughout the world. Weather and other environmental conditions, cultivar susceptibility, cultural practices, existence of primary sources of inoculum, and other parameters determine which diseases become limiting economic factors in onion production.

**Fungal diseases.** White rot, caused by *Sclerotium cepivorum*, is one of the most important onion maladies in the world. White rot usually appears first during the cool, moist weather in the fall and spring. The leaves begin to yellow and wilt and the plants eventually collapse. The soil-borne pathogen invades the roots and the basal parts of the bulb scales, and often develops a white fluffy mass of fungus threads and later black spherical bodies (sclerotia) visible to the naked eye (Fig. 2). Tolerant onion cultivars have not yet been tested adequately in the field. Fungicide seed treatments and soil applications often reduce white rot, but are not always reliable. Biological

control methods have not yet been applied commercially.

Neck rot and blighting of onion flowers and seed capsules, caused by *Botrytis* spp., are destructive and widespread diseases of onion. Sometimes as much as 50% of the crop is lost. Leaf blight, mainly caused by *B. squamosa*, is a very serious disease during wet growing seasons; it begins in early summer and continues until harvest. Elimination of cull piles and other infested debris will reduce *Botrytis*-induced diseases. Isolation of seed production fields from commercial fields will also reduce these diseases.

Downy mildew, caused by *Peronospora destructor*, varies in destructiveness with locality and season and is enhanced by cool, moist weather. Fungicidal sprays may help if the foliage is thoroughly covered.

At high soil temperatures, a basal and bulb rot, caused by the soil-borne *Fusarium oxysporum*, may cause serious losses to onions, in western valleys and in Northern states after midseason. The roots rot and the bulbs become soft. The disease may then become important in transit.

Black mold, *Aspergillus niger*, is a major cause of bulb losses of yellow onion varieties. The disease is readily identified by observation of black, powdery spores of the mold on the scales at the neck and between the outer scales of the bulb. Drying onion bulbs at or below 3% relative humidity and at 98.5–100°F (37–38°C) may reduce black mold

*Colletotrichum circinans*, the fungus that causes onion smudge, is confined largely to white cultivars. It appears in the field before harvest as small, dark-green to black dots on the outer scales. Prompt harvest of the crop and avoiding exposure to rain usually reduce this disease.

Smut, caused by *Urocystis cepulae*, is a serious onion disease in upland and muck soils in which the pathogen survives for many years. The disease appears on seedlings as black elongated blisters formed within the scales or leaves. Smut does not affect onion crops starting with sets or transplants. Seed treatments are effective in direct seeded crops.

Pink root, incited by the soil-borne fungus *Pyrenochaeta terrestris*, is occasionally serious in



Fig. 2. White rot (caused by *Sclerotium cepivorum*) on bunching onions. (P. B. Adams, U.S. Department of Agriculture)

many states in the United States. Infected roots turn yellow, shrivel, and die; they take on a distinct pink color. The disease becomes most apparent at harvest. Crop rotation and use of resistant varieties reduce the disease.

**Bacterial disease.** Bacterial soft rot, caused by *Erwinia carotovora*, the most destructive post-harvest disease, starts at the neck of the bulb and progresses downward. The organism may enter the bulbs through wounds such as those caused by onion maggots. Removal of diseased bulbs at harvest time, avoiding bruising during harvest and packing, and storing bulbs in dry, well-ventilated areas will reduce soft rot and other storage diseases.

**Other diseases.** Other diseases, usually of minor importance, are caused by species of *Pythium*, *Stemphylium botryosum*, *Alternaria porri*, and *Puccinia porri*, and by virus, nematodes, and *Cuscuta*. Chemical injuries and physiological disorders also occur. See PLANT PATHOLOGY. George C. Papavizas

## Ontogeny

The developmental history of an organism from its origin to maturity. It starts with fertilization and ends with the attainment of an adult state, usually expressed in terms of both maximal body size and sexual maturity. Fertilization is the joining of haploid gametes (a spermatozoon and an ovum, each bearing half the number of chromosomes typical for the species) to form a diploid zygote (with a full chromosome number), a new unicellular living being. The gametes are the link between one generation and the next: the fusion of male and female gametes is the onset of a new ontogenetic cycle. Many organisms die shortly after sexual reproduction, whereas others live longer and generations are overlapped. Species are usually perceived as consisting mostly of adults, but in most cases the majority of their representation in the environment is as intermediate ontogenetic stages. See FERTILIZATION; REPRODUCTION (ANIMAL).

**Disciplines.** The study of ontogeny covers most aspects of biology and can be split into distinct disciplines. Embryology, often referred to as developmental biology, is the study of embryonic development, usually from fertilization to the beginning of independent life of the new individual. Larval biology is the study of postembryonic and prejuvenile or preadult organisms. Developmental genetics is the study of genetic regulation and specification of development. Life cycle biology involves the study of all the qualitative aspects dealing with organismal growth, from the zygote to the adult stage, identifying developmental stages and their sequences. Life history studies are of ecological interest and consider ontogeny from a quantitative point of view; they examine information such as the number of eggs produced by each individual, the number of viable zygotes produced by each spawning event, and the number of all subsequent stages of the life cycle, having recruitment evaluation [assessment of the numerical value of larval/juvenile recruitment (enrollment in new co-

horts)]; it indicates the number of larvae/juveniles that become adults] as their final result. Evolutionary biology and phylogenetics encompass the patterns and processes of evolution, often inferred from ontogenetic patterns and processes.

**Development.** Zygotes undergo a series of asexual reproductions (via mitoses). In unicellular organisms, mitoses usually lead to the formation of new independent cells (or sometimes to polynucleated cells, or to colonies of identical cells), deriving from a first sexually derived individual, so forming a clone of genetically identical individuals. In multicellular organisms, the products of the asexual reproductions starting with the first division of the zygote are initially clones of one another but remain connected, and eventually differentiate to form an individual. Clonation of individuals occurs even in humans, when the first results of asexual reproduction of the zygote separate from each other, leading to twin formation. See MITOSIS.

The ontogeny of a multicellular organism involves segmentation (or cleavage): the zygote divides into two, four, etc., cells which continue to divide. These cells are initially similar to the zygote, although smaller in size. They soon start to differentiate from their ancestors, acquiring special features, and forming specific tissue layers and, eventually, organs. These processes lead to the formation and growth of an embryo. Embryos can develop either freely, or within egg shells, or within the body of one parent; they can grow directly into juveniles (as in humans) or into larvae (with an indirect development, as in insects). Ctenophores, commonly known as comb jellies, are marine organisms that are paradoxical in having sexually mature larvae, which transform into sexually mature adults. See ANIMAL GROWTH; EMBRYOLOGY.

Juveniles are similar to adults but are smaller and not sexually mature. Their ontogeny continues until they reach a maximal size and reproductive ability. Usually ontogeny is interrupted at adulthood, but some organisms can grow throughout their life, so that ontogeny ends with death.

Indirect developers undergo rapid but severe morphological and genetic transitions between larval and juvenile/adult bauplans (body plans), referred to as larval metamorphosis. Detection of exogenous stimuli (such as bacterial signals) or internal clocks activate developmental programs. Embryo development and larval metamorphosis usually proceed by a network of signals driving multipotent (stem) cell proliferation and differentiation, apoptosis (programmed cell death), and sometimes cell transdifferentiation.

Stem cells are poorly differentiated or undifferentiated cells that retain the potential to form either one or more tissues, or a number of differentiated cell types, by renewing themselves through mitotic division. Stem cells are located in embryos, larvae, and adult organisms. Usually, embryonic stem cells generate multiple cell types of several tissues, whereas adult stem cells are thought to produce cell types of a single tissue or organ. Multipotent stem cells are known in adults of several invertebrate phyla, such as

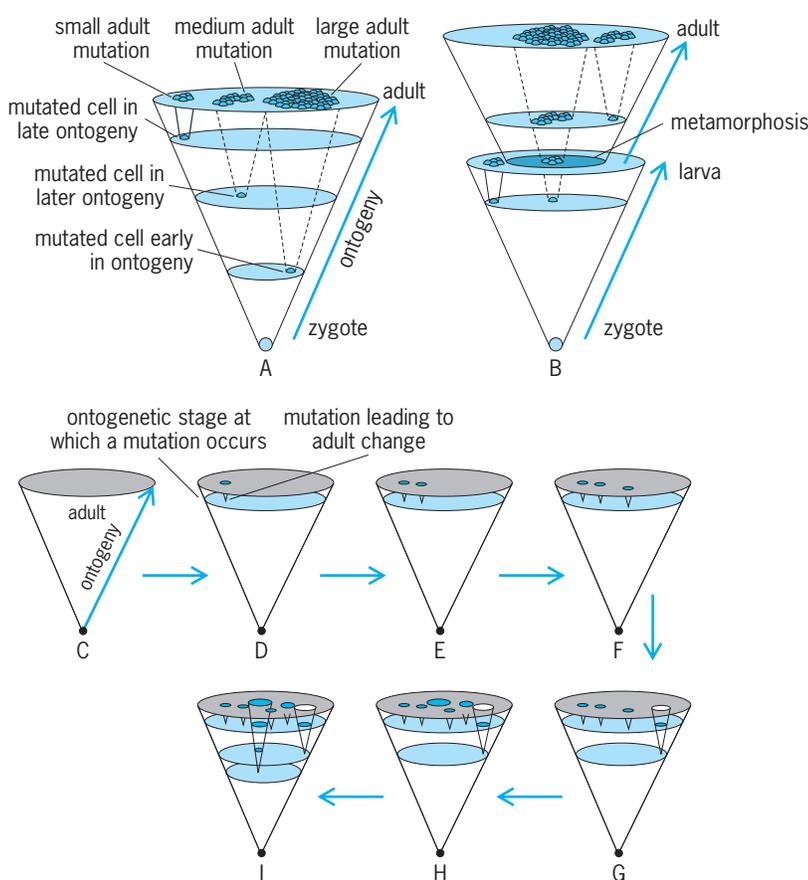
sponges (archeocytes), cnidarians (interstitial cells), and turbellarians (neoblasts), where they are also involved in cell replacement and regeneration of body parts that are lost by either injury or physiological causes. See STEM CELLS.

**Inverted cone theory of development.** The inverted cone theory states that ontogeny starts with a cell (the zygote) which has the potential to produce a whole adult organism. Further developmental steps—from cleavage to tissue and organ formation—restrict the possibilities of expression of the new cells. A viable mutation in the first ontogenetic steps (at the base of the inverted cone) will thus affect the rest of development, with a sharp modification of the mutated organism, whereas a change in late ontogeny produces a slight modification in adult expression. The higher the mutation in the inverted cone (that is, nearer the end of ontogeny), the smaller the effect; the lower the mutation (nearer the beginning of ontogeny), the greater the effect will be (Fig. 1).

When applied to speciation, this theory means that a mutation affecting the base of the inverted cone will lead to a new species similar to the ancestral one (that is, through a process of gradual evolution), whereas a mutation at a lower level of the inverted cone (at the beginning of ontogeny) will affect the rest of development, leading to a new species very different from the ancestral one through a process of saltational (discontinuous) evolution (Fig. 1). Organisms with direct development have a single inverted cone, whereas those with complex life cycles, involving indirect development, have an inverted cone for every developmental stage. See ORGANIC EVOLUTION; SPECIATION.

**Heterochrony.** Heterochrony is a change in the sequence of ontogenetic events within a lineage and is a major evolutionary mechanism, leading to recombination of preexisting structures. The evolution of new body plans requires subtraction or reassembly of old structures (pedomorphosis) as well as the introduction of novel structures added to the ancestral ones (peramorphosis). Ontogenetic changes, through pedomorphosis and peramorphosis, are at the base of evolutionary innovations. Fossil records show that early organisms were simple and that complex organisms are the products of the addition of structures. Since almost no new body plans are known to have evolved after the Cambrian radiation, it is reasonable to presume that most of post-Cambrian evolution occurred via modification of preexisting structures. See HETEROCHRONY.

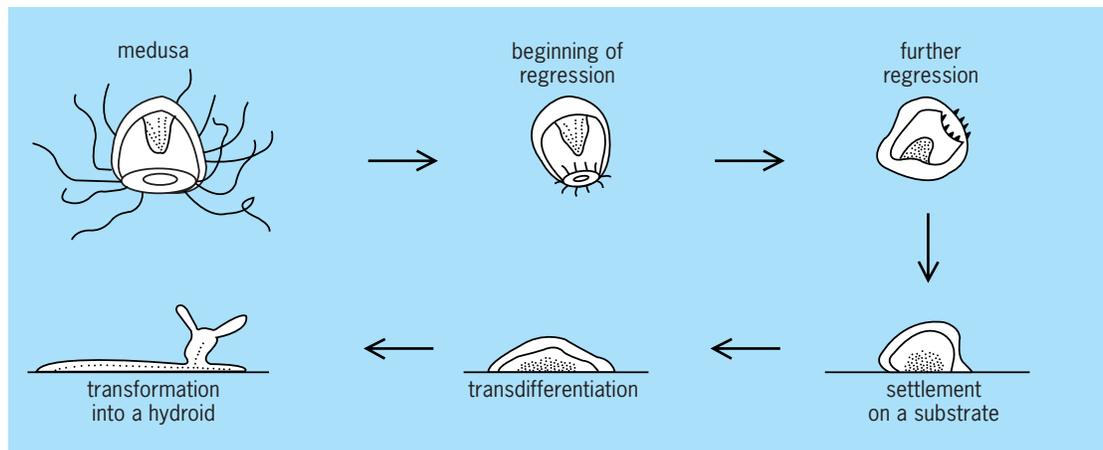
**Developmental genetics.** The discovery of homeobox, or *Hox*, genes in almost all animals investigated shows that the genes of development are conserved, even though their expression in different lineages leads to different body plans. In the past, embryology had a separate evolution from genetics, but this gap is being bridged by the understanding of ontogenetic processes through an integrated approach. Every cell contains all the necessary information to specify a whole organism (the very basis of clonation) but, during ontogeny, cells differentiate from each other despite their uniform genetic informa-



**Fig. 1.** (A, B) Inverted cone theory of development of an individual organism. (A) The ontogeny of an organism with direct development is seen as an increase in cell number, from the unicellular zygote to the pluricellular adult. Mutations occurring at different ontogenetic stages have a different impact on the final product of ontogeny, that is, the adult. (After W. Arthur, *The Origin of Animal Body Plans*, Cambridge University Press, 1998). (B) The ontogeny of organisms with complex life cycles (indirect development) can be depicted as an inverted cone, with the apex representing the zygote, the area closer to the base representing the (first) larval stage (and more truncated cones, not shown, if there is more than one larval stage), and the base of the distal truncated inverted cone representing the adult. Mutations occurring during ontogeny of larval stages might affect adult structures; however, mutations occurring after metamorphosis affect adult ontogeny only. (C–I) Inverted cone theory of development applied to the evolution of species and higher taxa. (C) Ancestral species, founder of a new lineage. (D) A mutation occurring late in ontogeny (that is, near the end of the inverted cone that represents the adult) leads to a slightly modified new species, sharing most of the features of the ancestral one. (E, F) Additional mutations lead to further anagenesis (or to cladogenesis, not shown), which is the alteration of the genetic properties of a lineage over time. However, the general architecture of the ancestral species is only slightly changed. (G) Speciation due to mutation at an earlier ontogenetic stage modifies the ontogenetic pathway, leading the formation of a new higher taxon (for example, a genus). (H) Gradual evolution continues with a load of mutations affecting late ontogeny. (I) After another load of small changes (not shown), another mutation occurring even earlier in ontogeny causes a further shift from the original body plan, and the founder of a new higher taxon (for example, a family) evolves.

tion. It is for this reason that embryology flourished separately from genetics. Apparently, there is a reciprocal information flux through genotype and phenotype, and ontogeny occurs through both genetic and epigenetic processes. See DEVELOPMENTAL GENETICS.

Each individual trait has alternative developmental trajectories, that is, potentially realizable phenotypes. These can be defined genetically at birth, but environmental thresholds may drive changes in the ontogenetic history. The shape of trajectories is dynamically modeled by mutations. The interplay of genomic and environmental pressures produces



**Fig. 2.** Ontogeny reversal in the hydromedusa *Turritopsis nutricula*. (After S. Piraino et al., *Reversing the life cycle: Medusae transforming into polyps and cell transdifferentiation in Turritopsis nutricula* (Cnidaria, Hydrozoa), *Biol. Bull.*, 190(3):302–312, 1996)

variations both in phenotypical traits and in their quantitative dimension, which will be screened by selection. Changes in development ultimately lead to evolution. See GENE ACTION; MUTATION.

**Ontogeny reversal.** Ontogeny is considered a one-way chain of events from fertilization to death. The conversion of an adult into a larva (ontogeny reversal) has been documented in species of the hydrozoan genus *Turritopsis*, whose adult medusae can become reduced to a mass of dedifferentiated cells which transdifferentiate into other cell types and form a hydroid colony. This colony is the developmental stage preceding the medusa (Fig. 2). Hydroids are normally formed by medusae through sexual reproduction, involving fertilization and planula formation, that is, through embryonic development. Most medusae die after sexual reproduction, but those of *Turritopsis* are able to reverse their ontogeny and, if under stress or after spawning, go back to a hydroid stage, escaping death. These medusae transformed into hydroids are able to produce new medusae, redirecting ontogeny in its normal flow.

Transdifferentiation demonstrates that a differentiated cell can still use the whole potential of its genome. This usually occurs at a cellular level and has almost no bearing on the architecture of the individual organism. In *Turritopsis*, it is the whole adult organism which reassembles its transdifferentiated cells, metamorphosing into an earlier ontogenetic stage.

**Ecology of ontogeny.** Ontogeny can be put into an ecological framework (and ecology into a development framework). Life cycles allow the persistence of species from one generation to the other. The study of ecology has been centered mainly on both interspecific (food webs) and inter-extraspecific (biogeochemical cycles) fluxes. The importance of intraspecific fluxes (life cycles and histories) has been neglected in the explanation of ecosystem functioning, having been restricted to populational approaches. The widespread occurrence of ontogenetic patterns involving the presence of resting stages leading to the formation of seed banks in terres-

trial ecosystems and of resting egg-embryo and cyst banks in aquatic ecosystems is a way of tackling the problem of how matter is maintained in a living state. Ontogeny interruption, usually caused by the onset of adverse conditions, with the formation of resting-stage banks, leads to the formation of a potential biodiversity which is then realized at the onset of a new favorable period when the resting organisms become active again to complete their ontogeny. This process explains sharp discontinuities in the occurrence patterns of many organisms in their adult stage, from terrestrial plants and insects to planktonic protists and metazoans. See ECOLOGY.

Ferdinando V. Boero; Jean Bouillon; Stefano Piraino

**Bibliography.** W. Arthur, *The Origin of Animal Body Plans*, Cambridge University Press, 1998; F. Boero et al., The continuity of living matter and the discontinuities of its constituents: Do plankton and benthos really exist?, *Trends Ecol. Evol.*, 11(4):177–180, 1996; A. Minelli, *The Development of Animal Form. Ontogeny, Morphology, and Evolution*, Cambridge University Press, 2003; S. Piraino et al., Reverse development in Cnidaria, *Can. J. Zool.*, 82 (11):1748–1754, 2004; R. Raff, *The Shape of Life*, The University of Chicago Press, 1996; M. J. West-Eberhard, *Developmental Plasticity and Evolution*, Oxford University Press, 2003.

## Onychophora

The only living animal phylum with true lobopods (annulate, saclike legs with internal musculature). There are about 70 known living species in two families, Peripatopsidae and Peripatidae. These terrestrial animals are frequently referred to as Peripatus. Onychophora comprise a single class or order of the same name. They were once considered a missing link between annelid worms and arthropods, but are best considered to be aligned with the arthropods.

They have a cylindrical body, 0.5–6 in. (1.4–15 cm) long, with one antennal pair, an anterior ventral mouth, and 14–43 pairs of stubby, unsegmented legs

ending in walking pads and paired claws. Mandibles are present as modified tips of the first appendage pair. The body surface has a flexible chitinous cuticle; the skin is tuberculate, with numerous transverse folds. The body wall has three layers of smooth muscle, as in annelids, but the coelom is reduced to gonadal and nephridial cavities; the body cavity has an arthropodlike partitioned hemocoel; the heart is tubular with metameric ostia; and the nephridia are segmental with pores on legs or their bases. Gas exchange takes place by means of tracheae; spiracles are minute and numerous, located between skin folds. The brain is bilobed; paired ventral nerve cords are connected by commissures. Slow locomotion is effected by legs and body contractions; the animals can squeeze into very tight spaces. The eyes, located at the antennal base, are the direct type with a chitinous lens and retinal layer. The sexes are separate; the testes and ovaries are paired; and the genital tracts open through the posterior ventral pore. Onychophora are oviparous, ovoviviparous, or viviparous.

The Onychophora are predatory, feeding on small invertebrates. They spurt an adhesive material from modified nephridial glands in papillae at sides of the mouth for food capture and defense. They are largely nocturnal, occurring in humid habitats in forests, being found under logs or in litter.

This is a group of great antiquity. The order Protonychophora was proposed to contain the fossil *Aysbeaia pedunculata*, a marine lobopodial onychophorelike animal known from mid-Cambrian shales of British Columbia (over 500 million years old).

Stewart B. Peck

Bibliography. A. P. Gupta (ed.), *Arthropod Phylogeny*, 1979; S. M. Manton, *The Arthropoda: Habits, Functional Morphology, and Evolution*, 1977.

## Onychopoda

A specialized order of branchiopod crustaceans formerly included in the order Cladocera. The body is up to about 12 mm (0.5 in.) in length, but much of the length of the longest species is made up by a caudal process.

The head and thorax are short, as is the abdomen in some species, but in others it is drawn out into a long caudal process. A carapace is present but is reduced to a dorsal brood pouch, leaving the body naked. A large median compound eye occupies much of the head.

Onychopods swim actively by means of their antennae—the antennules are small and sensory—and seize their food with their four pairs of grasping trunk limbs. Most are predators, but detritus is also eaten by some species. The mandibles are stoutly denticulate.

Reproduction is mostly by parthenogenesis (males are unknown in some species from the Caspian Sea); eggs and young are carried in the brood pouch. Sexual reproduction gives rise to freely shed resistant

eggs that overwinter in temperate zone species. Onychopods occur in the sea and fresh water and are worldwide in distribution, but fresh-water species occur only in the Holarctic temperate zone. There is a remarkable group of endemic species in the Ponto-Caspian region. See BRANCHIOPODA. Geoffrey Fryer

## Onyx

The name onyx is applied correctly to banded chalcidonic quartz, in which the bands are straight and parallel, rather than curved, as in agate. Unfortunately, in the colored-stone trade, gray chalcedony dyed in various solid colors such as black, blue, and green is called onyx, with the color used as a prefix. Because the color is permanent, the fact that it is the result of dyeing is seldom mentioned.

The natural colors of true onyx are usually red or brown with white, although black is occasionally encountered as one of the colors. When the colors are red-brown with white or black, the material is known as sardonyx; this is the only kind commonly used as a gemstone. Its most familiar gem use is in cameos and intaglios, in which the figure is carved from one colored layer and the background in another. See CAMEO; CHALCEDONY; GEM; INTAGLIO (GEMOLOGY); QUARTZ.

Richard T. Liddicoat, Jr.

## Oogenesis

The generation of ova or eggs, the female gametes. Primordial germ cells, once they have populated the gonads, proliferate and differentiate into either sperm in the testis or ova in the ovary. The decision to produce either motile spermatocytes or more sedentary oocytes is based primarily on the genotype of the embryo. In rare cases, this decision can be reversed by the hormonal environment of the embryo, so that the sexual phenotype may differ from the genotype. Formation of the ovum most often involves substantial increases in cell volume as well as the acquisition of organellar structures that adapt the egg for reception of the sperm nucleus, and support of the early embryo. In histological sections, the structure of the oocyte often appears somewhat random, even chaotic, but as the understanding of its chemical and structural organization increases, an elegant but still somewhat cryptic order begins to emerge. See OVUM; SPERMATOGENESIS.

Among lower vertebrates and invertebrates, mitotic divisions of the precursor cells, the oogonia, continue throughout the reproductive life of the adult; thus extremely large numbers of ova are produced. In the fetal ovary of mammals, the oogonia undergo mitotic divisions until the birth of the fetus, but a process involving the destruction of the majority of the developing ova by the seventh month of gestation reduces the number of oocytes from millions to a few hundred. Around the time of birth, the mitotic divisions cease altogether, and the infant

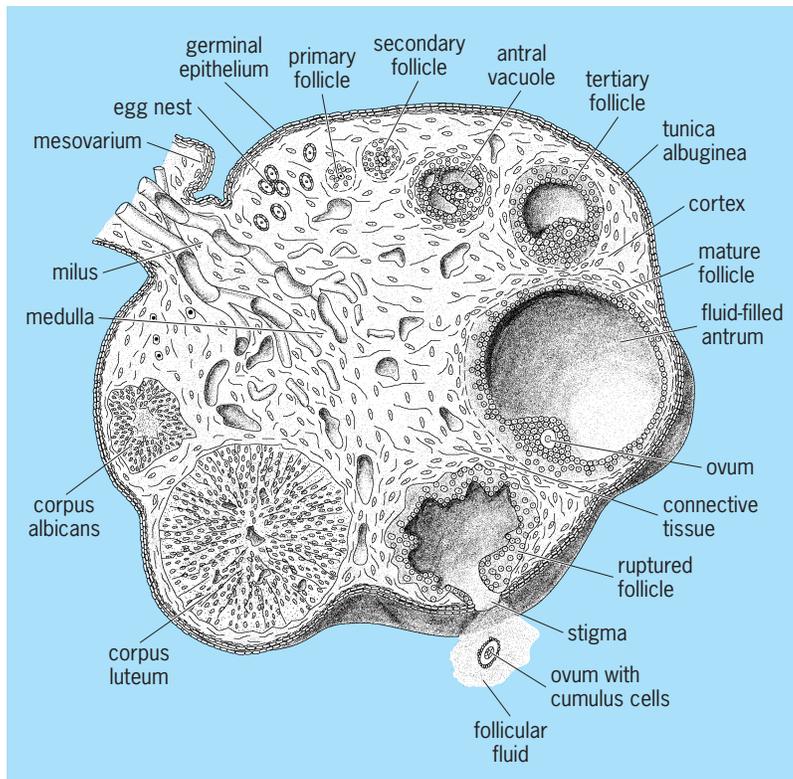


Fig. 1. Three-dimensional view of the cyclic changes in the mammalian ovary.

female ovary contains its full complement of potential ova. At puberty, the pituitary hormones, follicle stimulating hormone (FSH), and luteinizing hormone (LH) stimulate the growth and differentiation of the ova and surrounding cells (Fig. 1). See MITOSIS.

One important feature of oocyte differentiation is the reduction of the chromosome complement from the diploid state of the somatic cells to the haploid state of gametes. Fusion with the haploid genome of the sperm will restore the normal diploid number of chromosomes to the zygote. The meiotic divisions which reduce the chromosome content of the oocyte occur after the structural differentiation of the oocyte is complete, often only after fertilization. Unlike the formation of sperm, in which the two divisions of meiosis produce four equivalent daughter cells, the cytoplasm of the oocyte is divided unequally, so that three polar bodies with reduced cytoplasm and one oocyte are the final products. Generally, each fertilized oocyte produces a single embryo, but there are exceptions. Identical twins arise from the same fertilized egg; the egg of the armadillo normally gives rise to four embryos; and a single egg of a parasitic wasp may give rise to a thousand embryos. See FERTILIZATION; MEIOSIS.

**Yolk.** The provision of nutrients for the embryo is a major function of the egg, and this is accomplished by the storage of yolk in the cytoplasm. Yolk consists of complex mixtures of proteins (vitellins), lipids, and carbohydrates in platelets, which are membrane-surrounded packets dispersed throughout the egg cytoplasm (ooplasm). The amount of yolk in an

egg correlates with the nutritional needs of the embryo. If, as in the case of many invertebrates, the embryo hatches and becomes free-living as a very small larva, the amount of yolk required is small. Sea urchin eggs are a good example of those having a modest amount of evenly distributed yolk, and are classified as oligolecithal. Birds, fishes, and amphibians produce large eggs which sustain considerable growth of the embryo prior to hatching, and are classified as macrolecithal. In addition, the yolk in bird and fish eggs is concentrated at one end of the egg, so they are also described as telolecithal. Although the eggs of mammals are extremely small as compared to the fetus, the bulk of the nutrition is supplied by the placenta; yolk is required only until implantation in the uterine wall.

The contents of the yolk platelets are generally not produced in the oocyte, but are taken up from the blood and surrounding cells. The yolk of vertebrate eggs is thought to originate in the liver and to be carried to the ovary in the blood plasma. Insect yolk proteins are synthesized in the fat body cells and pass between the follicle cells (Fig. 2) to the surface of the oocyte, where they are taken in by pinocytosis. Often the transported form of the yolk protein (vitellogenin) is partially degraded upon arrival in the egg cytoplasm to form vitellins. During the formation of the embryo, the vitellins are degraded to amino acids to support the synthesis of new proteins.

**Ovary types.** Egg cytoplasm also contains large stores of ribonucleic acid (RNA) in the form of ribosomal, messenger, and transfer RNA. These RNAs direct the synthesis of proteins in the early embryo, and may have a decisive influence on the course of development. The mechanism by which the RNA is supplied to the egg is the basis for a major classification of ovary types. Panoistic ovaries, in which the egg nucleus is responsible for the production of all the stored RNA in the ooplasm, are typical of vertebrates,

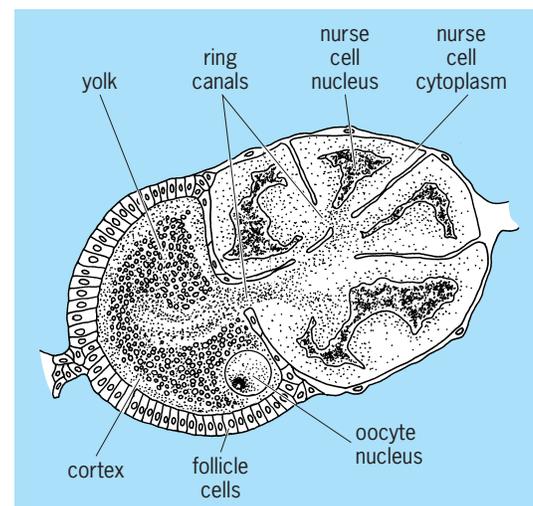


Fig. 2. Developing follicle of the cecropia silkworm. The nurse cells are sister cells to the oocyte. The ring canals arise from incomplete cytoplasmic divisions and provide direct cytoplasmic connections between oocyte and nurse cells.

primitive insects, and a number of invertebrates. The amounts of RNA produced during the meiotic prophase in such ovaries are much larger than those produced by a somatic cell, and thus special mechanisms seem to be involved in the synthetic process. In the case of ribosomal RNA, extra copies of the ribosomal RNA gene are produced which separate from the chromosomes. The deoxyribonucleic acid (DNA) template for messenger RNAs may not be amplified, but the "lampbrush" configuration often associated with oocyte chromosomes may reflect much more efficient use of the DNA as template for messenger RNA. *See* DEOXYRIBONUCLEIC ACID (DNA); GENE AMPLIFICATION; RIBONUCLEIC ACID (RNA).

More advanced insects and many other invertebrate species have adopted a different strategy for increasing the amount of template DNA required for production of oocyte RNA. Once the oogonial mitoses have been completed, the preoocytes begin a series of divisions in which the nuclei, but not the cytoplasm, divide leaving a cluster of connected cells (cystocytes). Only one cell of each cluster is destined to become an oocyte; the rest will become trophic cells, in some cases called nurse cells. These trophic cells often replicate their chromosomes many times, and thus supply template for oocyte RNA. Ovaries of this type are classified as meroistic, and are further subdivided into polytrophic, where the nurse cells are included in the follicle (Fig. 2), and telotrophic, where the forming oocyte is attached to a mass of trophic cells by a long cord. In panoistic follicles, the RNA is discharged directly from the nucleus into the oocyte cytoplasm. By contrast, the RNA often must be transported for considerable distances from the trophic cells to the ooplasm in meroistic follicles. The transport mechanism is not completely understood, but it is thought that an electrical current flowing from the trophic cells through the oocytes may at least assist in the transport.

**Specialized structures of ova.** Because eggs are large, sedentary cells, they face difficult physiological stresses, and have developed various specialized structures to cope with these problems. The plasma membrane (oolemma) of developing oocytes is the site of uptake of various compounds from the blood, and to facilitate this activity, numerous fingerlike (microvillar) or platelike projections are found on the surface during oocyte development. These extensions of the surface membrane are often coated with polysaccharides which may assist in the removal of substances from the blood. Numerous coated vesicles, which consist of bubbles of membrane containing concentrated plasma proteins, are found at the base of the microvilli and are the product of this uptake mechanism. The small vesicles near the surface migrate through a yolk-free area, the cortex, and then fuse to form the large yolk platelets.

The cortex is a poorly understood region just under the oolemma, which may be very significant in the formation of the embryo. Many classical experiments in embryology have involved centrifuging fertilized eggs to displace various cytoplasmic elements. One interpretation of the results of these

treatments is that unless the cortex is disturbed, the manipulations have little effect on the formation of the embryo, suggesting that the cortex may be of critical importance in determining the structure of the embryo. A special microarchitecture consisting of a dense matrix of fibers has been detected in the cortex of the eggs of many different species. In the silk moth, this structure has been demonstrated to consist of a meshwork of actin fibers. The cortex of eggs is also very rich in messenger RNA (mRNA), which may be held in place by being anchored to this cortical actin meshwork. It is possible that the latent image depends upon maternal mRNA held in a specific spatial configuration by its association with the fibrous cortical meshwork. At least one enzyme crucial to later development, RNA polymerase II, has been found to be stored in the cortex of silk-moth eggs. The ooplasm also contains mitochondria, pigment granules, occasional Golgi apparatus, and of course, the nucleus. *See* CENTRIFUGATION.

In some amphibians and insects a unique area in the ooplasm, germ plasm, has been identified. It is of particular interest since it is the site of formation of the primordial germ cells of the next generation. Germ plasm can often be distinguished by light microscopy of stained sections of oocytes because of its greater affinity for basic dyes. Electron microscopy of these areas reveals large numbers of granules, which are dense concentrations of ribonucleoprotein often intimately associated with mitochondria. The exact nature of the process is not understood, but the blastomeres which come to contain the granules are thereby determined to become the precursors of eggs or sperm. The primordial germ cells can often be identified in later embryos by structure, location, and staining properties, and if they are removed the resulting adult is always sterile.

The activity of the nucleus during oogenesis depends upon the presence or absence of trophic cells. Nuclei of panoistic oocytes, as mentioned earlier, are responsible for the entire RNA contents of the oocyte. Nuclei of meroistic oocytes are relatively "silent" during oogenesis, and often the chromosomes are gathered in a dense body called the karyosome or karyosphere. In the early stages of oogenesis in some insects, a large, temporary, DNA-containing structure, the Giardina body, appears in the nucleus, and then breaks down when yolk begins to appear in the ooplasm (vitellogenesis). The Giardina body contains ribosomal genes, and probably other genes as well.

Oocytes are almost always covered by some type of protective coating which is noncellular and varies in complexity according to the mode of fertilization and the environment into which it is shed. Because they are fertilized internally and maintained in the protection of the uterus, mammalian eggs are simply coated with a noncellular vitelline membrane (area pellucida). A single layer of follicle cells (the corona radiata) remains attached to the outer surface of the area pellucida. Bird eggs, in which the embryos develop outside the body of the female in a dry environment, are coated with a heavy layer of albumen

(egg white) and then a complex shell of mineral and protein which provides support and a barrier to insult and desiccation, but also allows gas exchange with the environment. See EGG (FOWL).

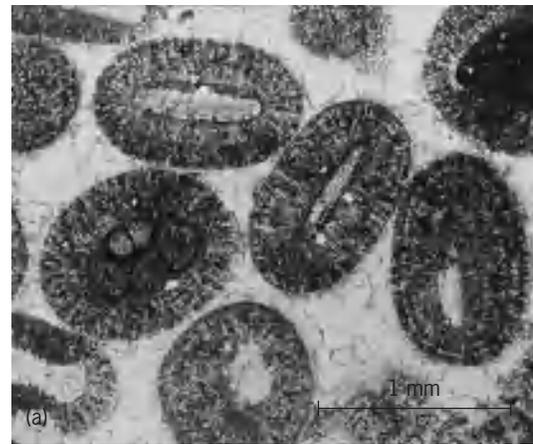
**Fertilization.** Fusion of egg and sperm occur at the top of the reproductive tract in birds, so the albumen and shell are added by oviductal glands after fertilization. Many invertebrates produce eggs with a tough, impervious shell (chorion) secreted by follicle cells before fertilization can occur. A special opening, the micropyle, is incorporated in the structure of the chorion to allow sperm access to the oocyte. Eggs of many lower vertebrates and invertebrates are shed directly into the water and then fertilized by sperm released into the immediate vicinity. Protection in this case is often afforded by jelly coats which expand slowly on contact with water. Although polyspermy, the entry of multiple sperm into an oocyte, is common in some animals, other species have developed mechanisms to prevent multiple sperm entry. The sea urchin egg, for instance, is provided with numerous vesicles just under the oolemma. The first penetration of the oolemma by a sperm causes the eversion of these vesicles, and the release of enzymes that cause the vitelline membrane to pull away from the surface and to harden. This raising and toughening of the vitelline membrane carries away any supernumerary sperm, and affords some protection for the developing embryo. See DEVELOPMENTAL BIOLOGY; GAMETOGENESIS. Spencer J. Berry

Bibliography. L. Browder, *Developmental Biology*, 3d ed., 1997; L. Browder (ed.), *Developmental Biology: A Comprehensive Synthesis*, 1985; A. DeLoof, The meroistic insect ovary as a miniature electrophoresis chamber, *Comp. Biochem., Physiol.*, 74A:3-9, 1983; P. D. Nieukoop and L. A. Sutasurya, *Primordial Germ Cells in the Chordates*, 1970.

## Oolite

A deposit containing spheroidal grains with a mineral cortex, most commonly calcite or aragonite, accreted around a nucleus formed primarily of shell fragments or quartz grains. The term ooid is applied to grains less than 0.08 in. (2 mm) in diameter, and the term pisoid to those greater than 0.08 in. (2 mm). Accretionary layering (growth banding) is usually developed clearly. A flattened or elongate shape may occur if the nucleus shows that form. Ooids formed on nuclei of shells and shell fragments and composed of fine, radial calcite are shown in *illus. a*. These ooids are cemented by coarse, clear calcite. Growth banding is visible in most ooids. The pisoids shown in *illus. b* are composed of many thin layers of very small, tangential (lighter layers) and radial (darker layers) aragonite crystals. These pisoids are cemented with fibrous aragonite.

Ooids are primarily marine, forming in agitated shallow, warm waters such as the Bahama Platform, the Persian Gulf, and off Yucatan. Under those conditions, the ooids are kept intermittently moving, so accretion occurs on all sides. Also, the agitation and



Plain light micrographs of thin sections of ooids and pisoids. (a) Oolite of Mississippian age, southwestern Indiana. (b) Pisoids from the hot springs of Karlovy Vary (Carlsbad), Czech Republic. The lower right pisoid shows a rock fragment nucleus and dark radial rays, possibly of bacterial origin.

warming leads to loss of  $\text{CO}_2$  from the water, thereby enhancing  $\text{CaCO}_3$  precipitation. Some ooids and most pisoids form in nonmarine environments, such as hypersaline and fresh-water lakes, hot springs, caves, caliche soils, and some rivers. Modern marine ooids are composed commonly of minute needles of aragonite arranged approximately tangentially. Many ancient ooids are composed of radial calcite, and it has long been dogma that those ooids formed by calcite replacement of originally tangential aragonite ooids. However, that widespread misconception is refuted by observations of the calcitization behavior of known aragonites, including ooids in the Pleistocene and Pennsylvanian, which produces irregular, coarser calcite. Hence those radial calcite ooids were never aragonite and may preserve textures close to the original ones.

Primary mineralogy of ooids (aragonite or calcite) appears to have varied through geologic time, probably in response to some factor which is influenced by plate-tectonically mediated global changes in  $\text{CO}_2$  abundance.

Although unequivocally secondary replacement of  $\text{CaCO}_3$  ooids, particularly by such minerals as dolomite, silica, or hematite, is relatively common, some ooids composed of phosphatic (collophane) or ferruginous (chamosite, hematite) minerals appear to be primary, rather than replacement of carbonate

oids. Even ooids of gypsum and halite have been reported. See ARAGONITE; CALCITE.

Philip A. Sandberg

Bibliography. D. P. Bhattacharyya and P. K. Karimoto, Origin of ferriferous ooids: An SEM study of ironstone ooids and bauxite pisoids, *J. Sediment. Petrol.*, 52:849-857, 1982; H. Blatt, *Sedimentary Petrology*, 2d ed., 1991; B. D. Keith and C. W. Zuppann (eds.), *Mississippian Oolites and Modern Analogs*, 1993; W. C. Krumbein and F. J. Pettijohn, *Manual of Sedimentary Petrography*, 1938, reprint 1988; T. Peryt (ed.), *Coated Grains*, 1983; P. A. Sandberg, An oscillating trend in Phanerozoic non-skeletal carbonate mineralogy, *Nature*, 305:19-22, 1983; L. Simone, Ooids: A review, *Earth Sci. Rev.*, 16:319-355, 1981.

## Oomycota

A class of fungi in the subdivision Mastigomycotina. They comprise a group of heterotropic, fungallike organisms that are classified with the zoosporic fungi (Mastigomycotina) but in reality are related to the heterokont algae. They are distinguished from other zoosporic fungi by the presence of biflagellate zoospores: the shorter, anteriorly directed flagellum contains two rows of tripartite hairs; the longer, posteriorly directed flagellum is naked and whiplash. Some taxa are nonzoosporic. The thallus may be unicellular, or mycelial and composed of filamentous hyphae that are coenocytic (without cross walls), at least in the early stages of development. The main component of the cell walls is cellulose, or cellulose and chitin. Asexual reproduction involves the release of zoospores from sporangia; in some taxa the sporangium germinates with outgrowth of a germ tube. Sexual reproduction occurs when an oogonial cell is fertilized by contact with an antheridium, resulting in one or more oospores. At least in some, possibly all, species, the thallus is diploid.

There are five orders: Saprolegniales, Rhipidiales, Leptomitales, Lagenidiales, and Peronosporales. The Oomycete systematics is in a state of flux. They have been placed under the Kingdom Protista, Kingdom Chromista, Kingdom Heterokonta, Kingdom Protocista, and most recently Kingdom Stramenopila.

Oomycetes are cosmopolitan, occurring in fresh and salt water, soil, and as terrestrial parasites of plants. Many species can be grown in pure culture on defined media. The Saprolegniales and Leptomitales are popularly known as water molds; some species are destructive fish parasites, especially in salmon hatcheries, where they may reach epidemic levels. Species of Rhipidiales, which once were placed in Leptomitales, grow in stagnant or polluted waters. Many Lagenidiales are parasites of invertebrates and algae. The Peronosporales are primarily plant parasites attacking the root, stem, or leaf, and include some of the more destructive plant pathogens known. One example is *Phytophthora infestans*, the cause of late blight of potato that was responsible for the Irish famine in the midnineteenth century. White

rust of crucifers, blue mold of tobacco, and downy mildews are other examples of crop diseases caused by Oomycetes. See EUMYCOTA; FUNGI; MASTIGOMYCOTINA.

Donald J. S. Barr

Bibliography. G. C. Ainsworth, K. F. Sparrow, and A. S. Sussman, (eds.), *The Fungi: An Advanced Treatise*, Vol. 4. *A taxonomic review with keys*, Academic Press, New York, 1973; C. J. Alexopoulos, C. W. Mims and M. Blackwell, *Introductory Mycology*, 4th ed., John Wiley, New York, 1996; D. J. S. Barr, Evolution and kingdoms of organisms from the perspective of a mycologist, *Mycologia*, 84:1-11, 1992.; M. W. Dick, Oomycota, in L. Margulis et al. (eds.), *Handbook of Protoctista*, Jones and Bartlett, Boston, 1990.

## Opal

A natural hydrated form of silica. Opal is a relatively common mineral in its nongem form, which is known as common opal and lacks the play of color for which gem, or precious, opal is known. All opal is of relatively simple chemical composition,  $\text{SiO}_2 \cdot n\text{H}_2\text{O}$ , in which  $\text{SiO}_2$  represents silicon dioxide, and  $n$  represents a variable amount of water ( $\text{H}_2\text{O}$ ) that is housed in the internal structural arrangement of the mineral. The water content generally ranges from about 4 to 10 wt %, but it may be as high as 20 wt %. The hardness of opal on the Mohs hardness scale ranges from 5 to 6, the specific gravity from 2.25 to 1.99, and the refractive index from 1.455 to 1.435. Both specific gravity and refractive index decrease as a function of increasing water content. See HARDNESS SCALES.

Opal ranges from dense and glassy with a conchoidal fracture to quite porous such as siliceous geyserite (with a specific gravity of 1.8), which occurs at the site of thermal springs and geysers. In addition to being formed at the site of hot springs, opal is deposited by meteoric waters or by low-temperature hydrothermal solutions. It is found lining and filling cavities in rocks, and may replace wood buried in volcanic tuff. It commonly occurs as crusts with botryoidal, globular, and ropy surfaces in stalactitic forms, or as concretionary masses and cavity and vein fillings. Precipitation may result from evaporation or by organic action of organisms such as siliceous tests and diatoms; this form is known as diatomaceous earth (infusorial earth). Most opal is amorphous, that is, apparently lacking an ordered internal, atomic structure. See CRYSTAL STRUCTURE; GEYSER; ORE AND MINERAL DEPOSITS; SPRING (HYDROLOGY).

The color of common opal ranges from transparent, glassy, and colorless to white and bluish white. Common pigmentsing agents, such as iron, produce yellow, brown, red, and green colors, and frequently several colors in a single specimen. Precious opal has a play of color that is the result of white light being diffracted by the relatively regular internal array of silica spheres. That is, when white light passes through an essentially colorless opal, it strikes the planes of spheres and voids between them, and at

these interfaces the white light is diffracted such that certain wavelengths flash out of the stone as nearly pure spectral colors. Some wavelengths may be completely internally reflected, so the full color spectrum may not escape from a flat surface; however, the curved surface of a cabochon-cut opal allows most or all colors to escape. Because opal is a hydrous mineral, certain opals from specific geologic occurrences may crack because of water loss. Therefore, considerable care is required in the polishing and handling of opal.

**Gem classification.** Several trade terms are used to describe the appearance of precious opal based on transparency, body color, and the type of play of color. Some of these terms are black opal, which is translucent to almost opaque, with dark gray to black body color, with play of color; fire opal, which is transparent to semitransparent, with yellow, orange, red, or brown body color and with or without play of color; harlequin or mosaic opal, in which the play of color occurs in distinct, broad, angular patches; and matrix opal, which consists of thin seams of high-quality gem opal in a matrix.

Synthetic gem opal has been produced in Switzerland since 1970. This product is similar to natural material in its chemical and physical properties, including a beautiful play of color. A lay person is unable to distinguish this material from natural precious opal. Simulants such as colored synthetic glass with some internal color reflections, and some plastics, imitate some aspects of precious opal. *See GEM.*

**Sources.** The largest sources of precious opal are in Australia, where it occurs in various districts as vein or void fillings in sedimentary rocks. Important districts are located in Queensland, New South Wales, and South Australia. The opal accumulates in spaces that may be formed by the leaching of fossils, wood, and minerals such as gypsum and glauberite, or in spaces in poorly sorted coarse sediments. In the Lightning Ridge area of New South Wales, the opal occurs as nobbies, which are subrounded shapes up to roughly 40 mm (1.6 in.) in diameter. The origin of the nobbies, not all of which show a play of color, is not understood. *See SILICATE MINERALS; SILICEOUS SINTER.*

Peter J. Darragh; Cornelis Klein

**Bibliography.** J. E. Arem, *Color Encyclopedia of Gemstones*, 2d ed., 1994; C. S. Hurlbut, Jr., and R. C. Kammerling, *Gemology*, 2d ed., 1993.

## Opalescence

The milky iridescent appearance of a dense transparent medium when the medium (or system) is illuminated by polychromatic radiation in the visible range, such as sunlight. Slight changes in the rainbowlike color of the system can occur, depending on the scattering angle, that is, the angle between the directions of incident radiation and of observation. All dense transparent fluids have local density fluctuations due to the thermal motions of molecules, or concentration fluctuations due to the presence of

a second component, such as colloidal suspensions or macromolecules in solution. Local fluctuations in density (or concentration) are accompanied by local fluctuations in the refractive index. Since the fluid is optically inhomogeneous, some of the light is scattered to the side. Normally, the amount of light scattered is very small, perhaps of the order of magnitude of  $10^{-4}$  or less of its incident radiation. Whenever the amplitude of fluctuation becomes large, a significant portion of the incident light may be scattered. The transmitted light is then visibly weakened, and the fluid looks turbid. The local optical inhomogeneities may be frozen-in in solids, making the system turbid (opalescent).

Opalescence is a general term which applies to the optical phenomenon of intense scattering in the visible range of the electromagnetic radiation by a system with strong local optical inhomogeneities. The iridescence, or rainbowlike display of interference of colors, arises because the intensity of scattered light is approximately proportional to the reciprocal fourth power of the wavelength of incident light (Rayleigh's law). *See SCATTERING OF ELECTROMAGNETIC RADIATION.*

**Critical opalescence.** A classical view of the critical point in gas-liquid phase transitions is that it is the state at which the densities of the coexisting gas-liquid phases are equal, and also that it is represented by a characteristic critical temperature above which gas cannot be liquefied, no matter how great the applied pressure is. The corresponding pressure required to liquefy the gas at the critical temperature is the critical pressure. For a one-component fluid the compressibility becomes very large in the neighborhood of the critical point and infinite at the critical point itself. Thus the energy required in the compression of a gas to a given amplitude of fluctuations becomes smaller the closer one approaches the critical point. There the thermal motions of molecules can produce strong density fluctuations, resulting in a very impressive scattering, the so-called critical opalescence. Apart from the critical points of gas-liquid transitions, several other types of second-order phase transitions at which the second derivatives of the free energy are discontinuous, such as critical mixing (consolute) points of binary liquid mixtures and of polymer blends, exhibit critical opalescent behavior. *See CRITICAL PHENOMENA.*

The nature of phase transitions can be studied by observing the size and shape of local fluctuations and their time-dependent changes. By using statistical models, the fluctuations can then be related to thermodynamic and transport properties of the system, such as isothermal compressibility and thermal conductivity. In the critical region, any such properties become very difficult to measure by conventional means. The changes at the critical region in thermodynamic and transport properties are very dramatic; for example, both isothermal compressibility and heat capacity diverge at the critical point. While the scattered intensity is related to the amplitude of local fluctuations in the refractive index, the angular dependence of scattered light reveals the extent of

such fluctuations. As the critical point is approached, the scattered intensity increases because the isothermal compressibility becomes larger; the system consequently looks very turbid. Further analysis shows that the scattered light is concentrated more and more in the forward direction. This indicates that the extensions of fluctuations approach the wavelength of the incident radiation, which is a few hundred nanometers for light in the visible region. These large extensions of local fluctuations result from long-range molecular interactions in the system. A suitable approach is to consider the incident electromagnetic radiation as a measuring scale. Thus, for smaller fluctuations or fluctuations in metal alloys near the consolute point, x-rays or thermal neutrons with wavelengths of the order of 0.5 nanometer or less, instead of visible light with a wavelength of several hundred nanometers, can be used to investigate fluctuation sizes ranging from several to tens (or occasionally hundreds) of nanometers, even when the system is barely opalescent to the naked eye or when it does not transmit visible light.

**Time dependency.** The density (or concentration) fluctuations that are produced by thermal motions of molecules are time-dependent, so that light is quasi-elastically or inelastically scattered. The spectral distribution of the scattered light characterizes the time dependence of such fluctuations and can be resolved by means of interferometric and optical beat frequency techniques, using a laser as a light source. Changes in the structure of local inhomogeneities during phase transitions, including spinodal decomposition and nucleation processes, can be studied by means of time-resolved (laser) light scattering or time-resolved small-angle x-ray scattering, including the application of intense synchrotron x-rays. However, the divergence in thermodynamic and transport properties results in a slowdown in the relaxation times of thermal fluctuations and an increasing difficulty in reaching thermal equilibrium for the system. Refined experiments to test the very successful applications of the renormalization group theory to critical phenomena become difficult to perform on Earth, and they have begun to be carried out in space under microgravity conditions. Analogy to the critical-point behavior has also been extended to semidilute polymer solutions in general and extremely dilute polymer solutions below the Flory theta temperature, that is, the temperature at which the polymer intermolecular interactions become negligibly small as the second virial coefficient vanishes. Polymer coil collapse could occur below the upper Flory theta temperature or above the lower Flory theta temperature. Care should therefore be exercised when studying the phenomenon of critical opalescence. See ABSORPTION OF ELECTROMAGNETIC RADIATION; LASER; SYNCHROTRON RADIATION.

Benjamin Chu

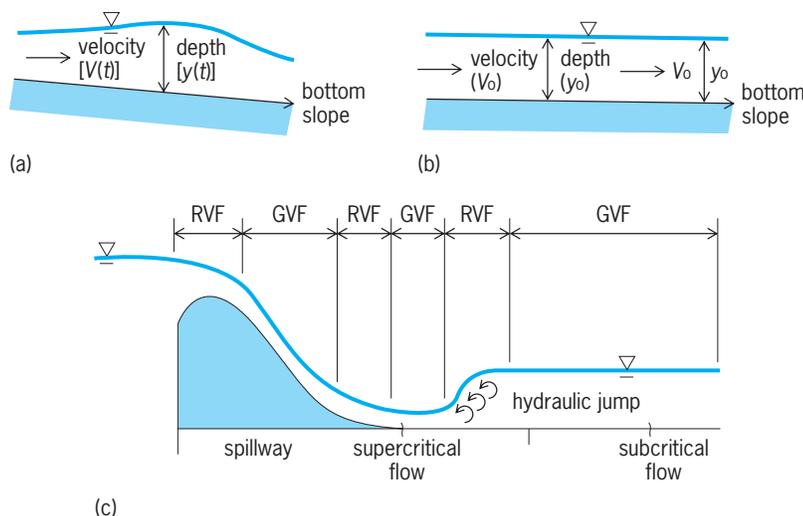
Bibliography. S. H. Chen, B. Chu, and R. Nossal (eds.), *Scattering Techniques Applied to Supramolecular and Nonequilibrium Systems*, 1981; B. Chu, *Laser Light Scattering: Basic Principles and Practice*, 2d ed., 1991; B. Chu et al.,

Critical phenomena and polymer coil-to-globule transition, *J. Appl. Crystal.*, 21:707, 1988.

## Open channel

A natural or artificial conveyance through which liquid (typically water) having a free surface moves. The free surface is the interface with a gas (usually the atmosphere), along which the pressure is constant. The liquid is accelerated or decelerated in the flow direction due to an imbalance between the driving gravity force and the viscous boundary resistance force (friction). Such flows occur naturally in rivers, streams, and estuaries as a part of the hydrologic process of surface runoff and artificially in free-surface conduits for the transport of water for irrigation, water supply, drainage, flood control, and other useful purposes. In contrast to full pipe flow, the free surface introduces an additional freedom into the description of open-channel flow, which is the position of the free surface itself as it adjusts to the imposed flow conditions. See CANAL; PIPE FLOW; RIVER; VISCOSITY.

**Effect of gravity.** The effect of gravity as the driving force in open-channel flows is manifested first as a body force (weight) with a component in the direction of motion, parallel to the sloping channel bottom, and second as a gradient (spatial variation) in the pressure with flow depth. Open-channel flows exhibiting changes in depth or velocity at a fixed point over time are classified as unsteady (illus. a). Flows exhibiting spatial changes in depth or velocity in the flow direction, or the lack thereof, result in a classification as nonuniform or uniform flow, respectively. As a special case, steady uniform flow (illus. b) does not change with time and has the property of constant depth and velocity in the flow direction so that the streamwise gravity force is exactly balanced by the boundary resistance force exerted



Types of open-channel flow. The inverted triangles denote the free surface of the flow. (a) Unsteady flow. (b) Steady, uniform flow. (c) Steady, gradually varied flow (GVF) and steady, rapidly varied flow (RVF). (After T. W. Sturm, *Open Channel Hydraulics*, McGraw-Hill, 2001)

on the flowing fluid. Artificial channels are often designed for this condition.

**Depth of flow.** The design of open channels such as storm sewers, roadside drainage ditches, irrigation canals, or water supply aqueducts proceeds from a specification of design flow rate, channel roughness, longitudinal channel slope, and cross-sectional geometry. The depth of flow must be determined from these variables using an empirical formula for steady uniform flow that relates them. An early formula of this type, which determines the channel flow rate  $Q$  under such conditions, is Eq. (1), attributed to the

$$Q = (1/n) [A]^{2/3} S^{1/2} \quad (1)$$

drainage engineer Robert Manning. Here,  $A$  is the channel flow area,  $P$  is the wetted perimeter of the channel, and  $S$  is the bottom slope. The roughness of the channel is characterized by Manning's  $n$  factor, which depends partly on the rugosity (wrinkledness) of the roughness elements—from brush and trees in the floodplain to sand, gravel, and cobbles on the bed of an alluvial channel. Some typical values of  $n$  are 0.01 for glass, 0.04 for a natural river channel, and 0.10 for a wooded floodplain. The larger the value of  $n$ , the greater the uniform-flow depth if all other variables are the same. In the case of sand-bed channels, the bed itself is deformed by the flow into ripples and dunes, which produce variable-form roughness as the river depth rises and falls. Dunes at the beginning of a flood, for example, may be washed out during the peak flow, a process which tends to reduce roughness and buffer depth changes due to floods.

**Steady, nonuniform flow.** In steady, nonuniform flow, the variation in the streamwise position of the free surface is either gradually varied or rapidly varied (illus. *c*). In the gradually varied case, the streamline curvature is gentle enough that the transverse pressure distribution over the flow depth can be considered to be hydrostatic, and the flow velocity is approximately uniform over the cross section. The shape of the free surface in the streamwise direction is called a water surface profile, and is computed in rivers for the purpose of floodplain mapping and in storm sewers to discover points of surcharging and overflow. Rapidly varied flow, as exemplified by storm surges, dam-break waves, hydraulic jumps, or other abrupt changes in depth, requires analysis of sudden changes in streamwise momentum in the flow region for its prediction. See HYDROSTATICS.

**Unsteady flow.** Unsteady flow is predicted by the Saint-Venant equations, which consist of the general momentum and mass conservation equations for one-dimensional, unsteady, turbulent flow in differential form. In the momentum equation, local and convective acceleration results from the net imbalance in the gravity force component down the slope, the net pressure force arising from a streamwise gradient in depth, and the resisting viscous boundary force. Application of the Saint-Venant equations is made to flood waves in rivers in order to track stage (depth) changes in time and space to provide for warnings and evacuations to prevent loss of life. This

general process is referred to as flow or flood routing. See FLUID-FLOW PRINCIPLES.

**Froude number.** Because of the dominance of the gravity force in shaping the free surface of an open-channel flow, the most important dimensionless parameter in open-channel flow is the Froude number,  $F$ , named after the engineer William Froude, who vastly improved the flow resistance and stability characteristics of ships by building and testing scale models in large towing tanks. He developed the modeling principles required to achieve dynamic similarity between the model and prototype in flows having free liquid surfaces. The primary parameter of this scaling is the Froude number, which represents the ratio of inertial forces to gravity forces and is defined by Eq. (2), in which  $V$  is the mean velocity

$$F = \frac{V}{\sqrt{gD}} \quad (2)$$

of flow,  $g$  is gravitational acceleration, and  $D$  is the length of the ship. The Froude number is also a primary parameter in open channel flow, for which  $D$  is some measure of depth that varies with the cross-sectional shape of the open channel. See DYNAMIC SIMILARITY; FROUDE NUMBER; SHIPPOWERING, MANEUVERING, AND SEAKEEPING; TOWING TANK.

**Flow regimes.** It can be shown from linear wave theory that the speed, or celerity  $c$ , of a low-amplitude wave is given by  $(gD)^{1/2}$  for the case of shallow water, that is, waves for which the wavelength is large in comparison to the flow depth. Because waves in open-channel flow are generally of this type, it can be observed that the Froude number represents the ratio of flow velocity to wave celerity. For Froude numbers less than one, the velocity is less than the wave celerity; this is called subcritical flow and is analogous to subsonic flow of air with respect to the speed of sound. Similarly, supercritical flow is defined as a flow with velocity greater than the wave celerity so that wave disturbances are swept downstream, analogous to supersonic flow in air. In contrast, waves move both upstream and downstream in subcritical flow. Supercritical flow is more descriptively called rapid flow to reflect the relatively small depths and large velocities at which it occurs, while subcritical flow is referred to as tranquil flow, with relatively large depths and small velocities. See WAVE MOTION IN LIQUIDS.

A flow can move from subcritical to supercritical as a result of sudden reductions in channel cross-sectional area or increases in bottom slope. The transition from supercritical to subcritical occurs due to increases in channel cross-sectional area or reduction in bottom slope and often takes the form of a hydraulic jump. In order to illustrate these changes in flow regimes, Boris Bakhmeteff introduced the concept of specific energy, which is the depth or pressure head relative to the channel bottom plus the kinetic energy head. Both these quantities represent energy components per unit weight of fluid; they have the dimensions of length and are referred to as head. The resulting

specific energy diagram, which is a plot of depth versus specific energy, clarifies the existence of critical depth and critical sections in an open channel, where the Froude number is unity and the specific energy is minimum. These sections exert control on the relationship between depth and flow rate, such as might occur in the flow over spillways or in flow-measuring devices such as weirs. See FLOW MEASUREMENT; FLUID FLOW; HYDRAULIC JUMP; HYDRAULICS.

Terry W. Sturm

**Bibliography.** E. F. Brater et al., *Handbook of Hydraulics*, 7th ed., McGraw-Hill, 1996; J. A. Cunge, F. M. Holly, Jr., and A. Verwey, *Practical Aspects of Computational River Hydraulics*, reprinted by Iowa Institute of Hydraulic Research, Iowa City, 1994; P. Y. Julien, *River Mechanics*, Cambridge University Press, 2002; I. Nezu and H. Nakagawa, *Turbulence in Open-Channel Flows*, A. A. Balkema, Rotterdam, 1993; T. W. Sturm, *Open Channel Hydraulics*, McGraw-Hill, 2001; B. C. Yen (ed.), *Channel Flow Resistance: Centennial of Manning's Formula*, Water Resources Publications, Littleton, CO, 1991.

## Open circuit

A condition in an electric circuit in which there is no path for current between two points; examples are a broken wire and a switch in the open, or off, position. See CIRCUIT (ELECTRICITY).

Open-circuit voltage is the potential difference between two points in a circuit when a branch (current path) between the points is open-circuited. Open-circuit voltage is measured by a voltmeter which has a very high resistance (theoretically infinite), such as a vacuum-tube voltmeter.

Clarence F. Goodheart

## Open-pit mining

The process of extracting beneficial minerals by surface excavations. Open-pit mining is a type of surface excavation which often takes the shape of an inverted cone (Fig. 1); the shape of the mine opening varies with the shape of the mineral deposit. Other types of surface mining are specific to the type and shape of the mineral deposit. See COAL MINING; PLACER MINING; SURFACE MINING.

The open-pit mine, like any other mining operation, must extract the product minerals at a positive economic benefit. All costs of producing the product, including excavation, beneficiation, processing, reclamation, environmental, and social costs, must be paid for by the sales of the mineral product. A mineral that is in sufficient concentration to meet or exceed these economic constraints is called ore. The terms ore body and ore deposit are used to refer to the natural occurrence of an economic mineral deposit. See ORE AND MINERAL DEPOSITS.

Ore bodies occur as the result of natural geologic occurrences. The geologic events that lead to the concentration of a mineral into an ore deposit are generally complex and rare. If those events placed

the deposit sufficiently near the surface, open-pit mining may be viable.

Material encountered during the mining process that has little or no economic value is called waste or overburden. One important economic criterion for open-pit mining is the amount of overlying waste which must be removed to extract the ore. The ratio of the amount of waste to the amount of ore is referred to as the strip ratio. In general, the lower the strip ratio, the more likely an ore body is to be mined by open-pit methods.

**Mining operations.** Modern open-pit mining utilizes large mechanical equipment to remove the ore and waste from the open-pit excavation. The amount of equipment and its type and size depend on the characteristics of the ore and waste and the required production capacity. In general, there are four basic unit operations common to most open-pit mining operations. These are drilling, blasting, loading, and hauling. If the rock types to be excavated are soft enough to permit excavation without blasting, the first two unit operations may not be performed.

The equipment used to accomplish the unit operations works most effectively when applied to a specific set of geometries. These geometric constraints result in the rock being excavated in horizontal slices of equal thickness. A slice is called a bench and may range in height from 10 to 50 ft (3 to 15 m). The specific height of the bench is set, based on the ore body geometry, the distribution of ore and waste, production requirements, and the equipment selected to remove the material. Large base metal mines tend to have benches of 40 or 50 ft (12 or 15 m). Many open-pit gold and silver mines have bench heights of 20 ft (6 m) due to the fact that ore and waste are intermixed, requiring better selectivity (less mixing of the ore and waste), and have lower production rates. Mining bench by bench results in the terraced geometry of the pit wall (Fig. 2); this technique aids in the containment of rock falls.

The total depth of the pit is a function of ore geometry, allowable pit slope angle, and the economic constraints of the strip ratio and the ore quality or grade. Large base metal operations can often exceed 1500 ft (460 m) in depth or may be as shallow as 50 ft (15 m). Slope angles depend on the strength of the rock and the natural fractures or joints in the rock. Overall slope angles may vary from 28 to 55° with 35 to 40° being the more common angles.

**Drilling.** The primary purpose of drilling is to provide an opening in the rock mass for explosives. Explosives work most efficiently when confined within the rock mass to be blasted. The drilling machine generally sits on top of a bench and bores a vertical hole downward into the bench area to be blasted. The drill should be able to penetrate the full bench height plus 2–5 ft (1–2 m) additional depth. The upper limits of the bench height are often set by the effectiveness of the drill. The holes provide a secondary purpose in most metal mining operations in that the cuttings extracted from the hole during the drilling process provide a sample for assaying the grade of the zone near the hole. The results of the assayed



Fig. 1. Aerial view of a mature open-pit mine. (Kennecott Copper Corp.)

drill hole cuttings are used to delineate the boundaries between ore and waste and between multiple mineral products, if applicable (for example, copper and molybdenum).

The most common form of drilling is rotary drilling where downward pressure is applied to a rotating tricone rotary bit, a type of bit originally developed by the petroleum industry for oil well drilling. The cuttings are removed from the hole by compressed air forced down the hollow drill pipe, past the rotating bit, and up the annulus between the drill pipe and the blast hole wall. The compressed air serves a second function of cooling the bit and bearings. For rotary drilling of harder formations, the tricone surfaces support buttons made of tungsten carbide, an extremely hard, wear-resistant metal.

When hard and abrasive rocks are encountered, rotary percussion drilling is performed with down-hole hammers. This type of blast-hole drilling uses impact by a tungsten carbide bit. After impact, the bit is mechanically retracted and rotated for another impact. This process occurs very rapidly and is generally powered by compressed air. A downhole

hammer is a rotary percussion tool that places the reciprocating hammer or impact mechanism immediately behind the bit. In this way, the drill pipe does not incur the stresses of impact. Compressed air forced down the hollow drill pipe powers and lubricates the tool, cools the bit, and removes the cuttings.

The drill equipment and air compressor are generally mounted for transport on a truck-type, rubber-tired carrier or a track-type carrier. The track-type carrier is generally more rugged and able to negotiate steeper grades. Rubber-tired carriers are generally used with smaller drills and where more mobility is required in the pit.

Drill hole diameters vary from roughly 3 to 15 in. (7 to 38 cm). The larger mines generally use holes from  $6\frac{3}{4}$  in. (17 cm) upward in size. Holes of about  $12\frac{1}{4}$  in. (31 cm) and larger are generally drilled by track-mounted rotary equipment (Fig. 2). Blast-hole drills can be powered by electric or diesel motors. See DRILLING, GEOTECHNICAL.

*Blasting.* The process of blasting breaks the rock into manageable sizes so that it can be loaded and



Fig. 2. A track-mounted, large-diameter, electrically powered rotary drill. (Kennecott Copper Corp.)

moved. While many blasting agents are available, the most common in open-pit mining is ANFO (ammonium nitrate and fuel oil). By itself ammonium nitrate is fairly insensitive and can be handled easily without many of the precautions required for other high explosives. When mixed with roughly 5% diesel fuel, the mixture is a powerful and inexpensive explosive compound. At large mining operations, ammonium nitrate is delivered in bulk in the form of small pellets known as prills, which are stored in bins or silos. A special truck transports the ammonium nitrate to the drill holes, where it is mixed with diesel fuel immediately before being loaded into the drill hole. For small operations, ANFO can be purchased in bags or cartridges. *See* PRILLING.

ANFO cannot be used if the drill holes contain water because the prills are highly soluble and will not detonate if water-soaked. A slurry, which is a mixture of ANFO and one or more additives, can be used to improve water resistance and detonation characteristics. A variety of slurry mixtures can be made to fit specific blasting requirements. Other blasting agents, for example, water gel explosives and gelatin dynamites, are available and are used in specialized

applications. Aluminum is a common additive; it increases the energy released by the explosion.

Any high explosive must be detonated to achieve the desired explosive results. Detonation is provided by caps and boosters. Caps are either electric or nonelectric. Nonelectric caps are preferred in open-pit mines as they are not sensitive to electric discharges such as lightning. Propagation of the initial detonation from one drill hole to the next is accomplished with a cord or cable known as det cord or primacord, a hollow plastic tubing filled with high explosive. Det cord is used to connect all drill holes which are filled with explosive in a given blast.

It is not efficient to detonate all holes in a blast at once. A timed pattern of drill hole detonation will result in the most efficient breakage of the rock per given weight of explosive. To accomplish this, delays ranging from milliseconds to seconds are used to arrest the detonation of the primacord momentarily.

After blasting is complete, the muck pile or shot rock is surveyed to set ore and waste boundaries based on the assays from the drill hole cuttings. The material is then ready for loading into trucks and haulage out of the mine. *See* EXPLOSIVE.



Fig. 3. A large (27-yd<sup>3</sup> or 21-m<sup>3</sup>) electric track-mounted cable shovel loading a 170-ton (154-metric-ton) haulage truck. (Kennecott Copper Corp.)

*Loading.* The blasted rock is loaded onto some conveyance, such as a truck or a train, for transport. Shovels and rubber-tired front end loaders are the most common pieces of loading equipment in open-pit mines. Track-mounted cable shovels powered by electricity are widely used in large mining operations (Fig. 3). These shovels are classified by the volumetric capacity of the dipper (or loading scoop), with common sizes ranging from 22 to 70 yd<sup>3</sup> (17 to 54 m<sup>3</sup>). While smaller sizes are available, they are generally being replaced by front end loaders. Electric cable shovels are high-capital-cost items, but they are extremely reliable and can last 15 to 20 years with proper maintenance. Electric shovels are highly productive and consequently have the lowest operating cost of all loading equipment types.

Rubber-tired wheel loaders are commonly used in precious metal mines throughout the world. Wheel loaders for mining applications generally range in size from 5 to 23 yd<sup>3</sup> (4 to 18 m<sup>3</sup>) for their loading bucket size (Fig. 4). Although both larger and smaller units are available, they are not usually used in open-pit mines. Direct diesel power is common for sizes up to 23 yd<sup>3</sup> (18 m<sup>3</sup>), with diesel electric more common for larger units. Wheel loaders do not require trailing power cables and can move from one working area to another rapidly. This mobility is a great benefit where working places are widely separated and the production rate does not require an

operating loading unit continuously at each loading area.

Compared to cable shovels, front loaders have lower capital costs but require higher operating costs than shovels of equal capacity. Mining operations with a short mine life or high-mobility requirements are candidates for the application of this type of equipment.

Hydraulic shovels have become more widely applied because of improvements in hydraulic systems



Fig. 4. A 13.5-yd<sup>3</sup> (10-m<sup>3</sup>) front end loader putting waste rock onto a 50-ton (45-metric-ton) haul truck. (Caterpillar Tractor Co.)

and shovel reliability. Hydraulic front shovels generally range in size from 7 to 45 yd<sup>3</sup> (5 to 35 m<sup>3</sup>). The hydraulic machines are generally diesel-powered, but electric-powered models are also available. Although not as mobile as the rubber-tired wheel loader, the hydraulic shovel is usually more productive than the former in terms of tonnage loaded per operating hour.

Two other types of loading equipment used for specific applications in mining are bucketwheel excavators and draglines. Both are used in soft materials that do not require blasting and can be excavated easily, though there are some operators that use draglines to excavate blasted material. Coal stripping is the most common application of both types of equipment; they are used to remove the overburden above the coal. *See* POWER SHOVEL.

*Hauling.* The broken material is loaded into trucks for transport out of the pit to its destination at the processing plant or the various waste dumps and stockpiles. Truck haulage is the most widely used. Direct conveyor haulage from a loading unit is limited to continuous loading systems such as bucketwheel excavators where the material is small and uniform in size so it may be loaded directly onto a belt. Most open-pit applications of conveyor haulage require the material to be hauled by truck from the loading unit to an in-pit crusher. The crushed material is placed on a belt for transport out of the mine. *See* CRUSHING AND PULVERIZING.

Haul trucks are classified by their payload capacity and vary in size from 20 to 350 tons (18 to 318 metric tons) capacity, with trucks in the 35 to 240 tons (32 to 218 metric tons) capacity range being the most common in open-pits. The trucks with capacity less than 35 tons (32 metric tons) are used in very small mines or quarries. All trucks are powered by diesel engines. The power supplied by the diesel engine is transferred to the wheels by two methods. The first method is by a mechanical transmission, and in the second method electric power is generated and used to drive electric motors located in the truck wheels. Trucks of 85 tons (77 metric tons) capacity and larger often use the diesel-electric method, and smaller trucks typically use mechanical drive. Haul trucks in sizes between 85 and 240 tons (77 and 216 metric tons) can be mechanical or electric drive.

Truck haulage ramps in the pit commonly range between 7 and 10% grade, depending on equipment size, weather conditions, and location in the pit. Haul roads are kept well graded to assure safe and efficient truck operation at speeds up to 35 mi/h (56 km/h). All trucks require braking systems that can stop the vehicle safely when descending these grades.

To augment truck haulage and help reduce fuel costs, a few large mines have installed a trolley assist system to be used with trucks that have electric wheel motors. These are modified trucks that are equipped with pantographs to take electric power from overhead trolley wires instead of from the diesel-alternator system. Such systems are used primarily on relatively permanent haul routes, often

on uphill climbs where trucks generally slow down because the diesel engines cannot supply enough power. When not in use, the pantographs are lowered and the trucks operate on power generated by the diesel engine. The system allows the flexibility of truck haulage with the benefit of direct electric power on the more energy-intensive segments of the haul. *See* PANTOGRAPH.

Another modification to material haulage has been the placement of a crusher in or near the pit to shorten the truck haul distance. These crushers are either permanent or movable and feed onto a conveyor system that transports the crushed material to its destination. The system is designed so that the conveyors replace the most costly segment of the truck haul, usually an uphill segment. By doing this, electrical energy replaces diesel energy and reduces the cost of transport. For ore, which must be crushed anyway, crushing in the pit represents no increase in cost. Waste material is not normally crushed, and the cost for this portion of the operation must be offset by the reduction in haul costs.

The implementation of a crush-convey system does have some negative aspects which must be balanced against the reduced haul costs. These include the increase in capital costs early in the mine life derived from installation of the system, and the reduction in the flexibility of the mining operation with semipermanent conveyors and crushers installed in the pit (**Fig. 5**).

In addition to the equipment described above, the open-pit mine has a fleet of auxiliary equipment, including track dozers, rubber-tired dozers, graders, and water trucks. Their function is to maintain the mine working areas, haulage roads, and dump and stockpile areas. *See* BULK-HANDLING MACHINES; CONVEYOR; MATERIALS-HANDLING EQUIPMENT.

**Waste disposal.** Waste material that is generated during the course of mining at most mines must be



**Fig. 5.** An in-pit feeder (left), crusher (center), and conveyor system (right and background). (Cyprus Sierrita Mine)

discarded as economically as possible without jeopardizing future mining activities but while respecting environmental regulations. Two types of waste material are generated at most mining operations: waste rock and overburden from the mine, and tailings—the waste material from the processing plant after treatment of the ore.

The rock, overburden, and soils removed from the pit and not processed are hauled to dumps or storage piles that can be from tens of feet to several hundred feet high, depending on their location and time of existence. These dumps are located as close to the pit as possible without jeopardizing future pit expansion. Lower-grade material that is not being processed but may become valuable is stockpiled for possible processing at some time in the future. These stockpiles must be located so that the material can be conveniently remined and hauled to the processing plant. Topsoil is often stockpiled for later use in reclaiming the dump and tailings areas. In some operations this is stored in shallow stockpiles to maintain the organic characteristics of the soils.

Considerable attention is being paid to the long-term effects of the disposal of mine waste material, particularly material that contains sulfide minerals or other potentially hazardous materials such as mercury or arsenic. Sulfide minerals are of particular concern because over time the sulfides break down chemically and mix with water to form sulfuric acid. Precautionary planning and monitoring are actively undertaken at all mining operations in the United States. Precautions include lining of tailings and storage ponds, catchment and control of water runoff from dumps and stockpiles due to rain and snowfall, diverting streams and small waterways around mining properties to avoid contamination of their waters, and treatment of contaminated water. Air and water are monitored at most mines to comply with government regulations and to protect the air as well as surface and underground water supplies. These environmental practices are being adopted in mining projects around the world.

Once the reserves have been exhausted and mining activity ceases, reclamation of a mining area is required by many states in the United States. Most large coal surface mines have active land reclamation programs. Many planned mining ventures have budgets that include complete or partial land reclamation. This could involve recontouring and revegetating, covering and planting tailings areas, and filling or partial filling of old pits. These measures add to the cost of mining the ore body and will make the lower-grade mines uneconomic. See LAND RECLAMATION.

**Computer technology.** Computer software is available to assist the mining engineer in ore reserve estimation with the application of geostatistics, mine planning and design, and production and maintenance monitoring and reporting. With the help of high-speed computers the engineering and production staff can evaluate aspects of the mining activities, which allows a more efficient and economical extraction of the mineral commodity. Comput-

erized truck dispatch systems, using global positioning systems (GPS), have been installed at many large mines with fleets greater than 30 trucks. These systems monitor all truck movements, shovel productions and locations, and truck destinations. The computer can instruct the truck drivers to go to specific shovels for loading and to use specific routes to their destinations. Once a particular trip is complete, the computer directs the driver to a specific shovel for the next load. This provides for the best deployment of the truck fleet, minimizing truck-waiting and shovel-waiting time. See COMPUTER; DIGITAL COMPUTER; MINING; OPERATIONS RESEARCH; OPTIMIZATION; SATELLITE NAVIGATION SYSTEMS.

Herb Welhener; John M. Marek

**Bibliography.** S. Bandopadhyay (ed.), *Application of Computers and Operations Research in the Mineral Industry—Proceedings of the 30th International Symposium*, Society for Mining, Metallurgy, and Exploration, Inc. (SME), 2002; D. W. Gentry and T. J. O'Neil, *Mine Investment Analysis*, Society for Mining, Metallurgy, and Exploration, Inc. (SME), 1984; H. L. Hartman (ed.), *SME Mining Engineering Handbook*, 2d ed., 2 vols., Society for Mining, Metallurgy, and Exploration, Inc. (SME), 1992; W. Hustrulid and M. Kutcha, *Open-Pit Mine Planning and Design*, 2 vols., A. A. Balkema, 1998; B. A. Kennedy (ed.), *Surface Mining*, 2d ed., Society for Mining, Metallurgy, and Exploration (SME), 1990.

## Open-systems thermodynamics (biology)

The application of the physical properties of energy transformations to biological systems.

Thermodynamic reasoning historically evolved along lines which focused on equilibrium states and fictitious “reversible” transitions between them. These transitions have the characteristic that they are frictionless or dissipationless, exemplified by the dissipationless charging and discharging of an ideal electrical capacitor in electronics. This reasoning led to three postulates, commonly called the laws of thermodynamics. These are often augmented by a zeroth law which establishes the notion of thermal equilibrium. Thermodynamics is generally applied to so-called simple systems, namely those which are large enough that certain average properties can be expected to be measured under conditions where fluctuations can be neglected, and which are also homogeneous or at least small collections of homogeneous regions. Thus each system and subsystem becomes defined by a real or imaginary boundary which has well-defined properties with regard to exchange of energy and matter with the environment through the boundary.

The most restrictive condition results from an isolated system, one with a boundary that passes neither energy nor matter. Isolated systems are compelled to achieve equilibrium, and these end points are the focus of the subject matter of classical thermodynamics. Relaxing the constraint against energy flow

through the boundary produces the next level in the hierarchy, the closed system. “Closed” now refers to matter flow since heat does pass into or out of the system. These systems are capable of achieving a class of stationary states away from equilibrium, all involving heat flow through the system due to maintaining a temperature difference across certain boundaries through which the heat flow occurs. If all constraints are removed, an open system results in which boundaries will pass both matter and energy and a broader class of stationary states away from equilibrium can exist. A simple example of this type of open system is an electrical resistor in series with an ideal voltage or current source. The generalization of this idea is the realm of nonequilibrium or stationary-state thermodynamics.

In both open and closed systems, the possibility exists that transients may occur as the system goes from one state to another after an external perturbation at the boundary. The field of network thermodynamics is an extension of thermodynamic reasoning that uses the generalization of classical thermodynamics as the study of all capacitive energy storage, the generalization of nonequilibrium thermodynamics as the study of all dissipative processes without energy storage, and the generalization of inertial energy storage (generalized inductance) to produce a new form of analysis completely isomorphic with modern dynamic systems theory. The network approach to thermodynamics is completed by adding a method for dealing with the connectedness of topology of the system, in the same manner as in electronic networks.

**Biological systems.** Biological systems obey the laws of physics and chemistry and, in particular, the thermodynamics of open systems. These are also highly organized, hierarchical systems. By using the network approach to their energetics, it is possible to see the flow of matter and energy through them in a particular pattern of connections or topology exactly like the analysis of an electronic network. Since electronic networks have had some powerful methods for their analysis, biological networks are capable of being analyzed by these same, powerful methods. In particular, certain computer simulation techniques developed for the simulation of nonlinear, highly organized, dynamic electrical networks are now being used to simulate some very complex aspects of living networks.

**Principles.** The principles of the network thermodynamic analysis of biological systems are fairly simple. The system is first seen as a network of discrete elements (simple thermodynamic systems) categorized by the way in which the state variables are related in the elements’ description. The state variables are always in conjugate pairs whose product is a power or pseudopower.  $P = ef$ , where  $P$  is the power, and  $e$  is an effort or force across the element that arises from a difference in a potentiallike quantity. The flow  $f$  is the change in some amount (charge, mass, volume, moles) per unit time and occurs by that amount flowing through the element in a unit of time. Two other state variables are obtained by in-

tegration, a displacement [Eq. (1)] and a momentum [Eq. (2)].

$$q = q(0) + \int_0^t f(t) dt \quad (1)$$

$$p = p(0) + \int_0^t e(t) dt \quad (2)$$

Now four kinds of elements are possible if the elements are to define binary relations among the four state variables: resistance [ $R(e, f)$ ], capacitance [ $C(q, e)$ ], inductance [ $L(p, f)$ ], and memristance (membrane resistance) [ $M(p, q)$ ]. Examples of resistances in this general sense are defined by Fick’s law for diffusion (permeability<sup>-1</sup>), chemical reaction kinetics, Poiseuille’s law for bulk flow, and Ohm’s law. The capacitors are a bit more subtle at first but also are binary relations like resistance. For example, in reactions or mass transfer, the volume  $v$  is the capacitance because it relates the displacement, the amount  $n$  to the effort’s potential, namely concentration  $c$ , through a simple binary relation,  $c = n/v$  or  $n = vc$ . The dynamic version of this (easily seen if volume is constant) is expressed by Eq. (3).

$$\frac{dn}{dt} = f = v \frac{dc}{dt} \quad (3)$$

Thus, although resistance yields an algebraic relation between efforts and flows, capacitances (and inductors) yield first-order differential equations. In the steady state, the capacitors are “dormant” since the efforts are no longer changing in time, and they can be omitted from the system’s description. Thus arises the earlier statement that the stationary states of nonequilibrium thermodynamics are determined solely by the dissipators (resistances) in the system and also yield purely algebraic relations between the state variables’ effort and flow. This is how the physics of the discrete elements is systematically built into the formalism, but the connectedness must also be systematically introduced and combined with the element’s description.

**Topology of biological networks.** Although the method for dealing with a biological network’s topology is very general, it is best seen through an example. One such example is shown in **Fig. 1**, which depicts an epithelial membrane such as those lining the gut, kidney tubule, gallbladder, or tongue. For a single, uncharged substance, there are four resistive branches and four compartments with volume for storage, so the network is that shown in **Fig. 1** superimposed on the tissue’s diagram. The rectangular elements are the resistive diffusion barriers, and the triangles represent the capacitive volume storage elements (essentially capacitors to a “ground” node which is some reference concentration, usually zero). The final step in codifying the system’s connectedness is to abstract the network to a linear graph as shown in **Fig. 2**. In this graph, the broken lines are called links or chords and represent the resistive branches, and the solid lines represent the graph’s “tree” which is a structure linking all the nodes without any loops being formed. In the

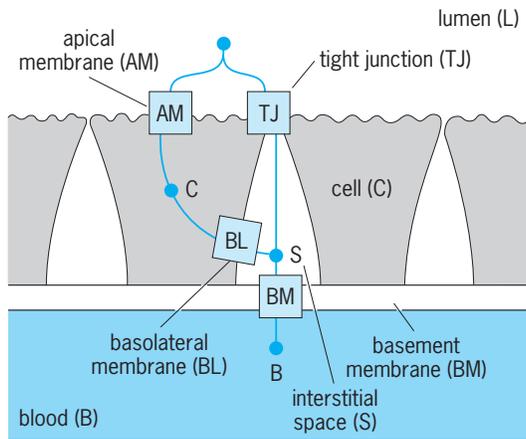


Fig. 1. An epithelial membrane with the network describing the flow through it superimposed on the tissue's morphology. See text for further explanation.

tree, branches are the capacitor-to-ground branches. Now a circle around each node in the graph defines a fundamental cut-set. Each of these circles isolates the node from the rest of the graph by cutting one tree branch and one or more links. This defines a simple code for the graph using the fact that the arrows drawn on each branch define a positive for flow through it and a positive direction for potential drop across it. The code resides in the cut-set-branch incidence matrix  $Q$ , having its elements defined as  $q_{ij} = +1$  if cut-set  $i$  cuts branch  $j$  and is oriented in the same direction (cut-sets are directed by the tree branch they cut),  $q_{ij} = -1$  if the orientation is opposite, and  $q_{ij} = 0$  if they are not incident. In the example above, Notice that  $Q$  partitions nicely into

$$Q = [F: I]$$

$$= \begin{bmatrix} 1 & 0 & 1 & 0 & \cdot & 1 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 & \cdot & 0 & 1 & 0 & 0 \\ 0 & -1 & -1 & 0 & \cdot & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & -1 & \cdot & 0 & 0 & 0 & 1 \end{bmatrix}$$

two distinct submatrices  $F$  and  $I$ , the latter being the identity matrix. This is due to the care taken to arrange the elements so that the links were listed first, then the tree branches in the same order as the cut-sets that cut them. A procedure which is available in any text on network solutions to electronic networks leads to the state-vector equation(4), where

$$\frac{d\bar{x}}{dt} = A\bar{x} + B\bar{u} \quad (4)$$

$\bar{x}$  is the vector of tree branch efforts,  $\bar{u}$  is the vector of sources or inputs energizing the network, and  $A$  and  $B$  are determined by  $Q$ , and the branch resistance and capacitance values. With some modification, the method extends to nonlinear network elements which are characteristic of biochemical reactions. See TOPOLOGY.

**Kirchhoff's laws and Tellegen's theorem.** Another set of relations is easily expressed in this language, namely the conservation of flows at a node, Kirchhoff's flow law, and the closure of the potential drops around any closed path, Kirchhoff's effort law. In the

topological code of the incidence matrix, these are  $Qf = \bar{O}$ , Kirchhoff's flow law, and  $X_l = F^T \bar{X}_t$ , Kirchhoff's effort law, where  $l$  denotes link efforts (resistive),  $t$  denotes tree efforts (capacitive), and  $F^T$  means the transpose of matrix  $F$ . It is easily shown using these relations or variations of them that Tellegen's theorem holds, namely that  $\bar{e} \cdot \bar{f} = 0$ , or that the vector of efforts is orthogonal to the vector of flows. Since the proof of this depends only on the choice of efforts and flows such that they obey Kirchhoff's laws, the quasipower theorem of Tellegen follows as well:  $\bar{e}' \cdot \bar{f} = 0$ . In this version of the theorem,  $\bar{e}'$  is an effort vector from an entirely different system with the same incidence matrix (topology) as that used to define the flow vector  $\bar{f}$ . Equally, these could be an effort and flow vector from the same system at different times. This leads to Onsager's reciprocity theorem for coupled systems.

**Coupled linear resistive multiport.** Onsager's non-equilibrium thermodynamics relates the efforts and flows in a single simple thermodynamic system (membrane, for example) as a set of linear effort flow relations :

$$\begin{aligned} e_1 &= R_{11}f_1 + R_{12}f_2 + \dots + R_{2n}f_n \\ e_2 &= R_{21}f_1 + R_{22}f_2 + \dots + R_{2n}f_n \\ &\vdots \\ e_n &= R_{n1}f_1 + R_{n2}f_2 + \dots + R_{nn}f_n \end{aligned}$$

This set of equations is an extension of the defining relation for a one-port resistor to multiport resistors (a port is a pair of input-output terminals through which a given flow passes and across which a given effort occurs). Onsager's reciprocity law equates all the coupling resistances between pairs of efforts and

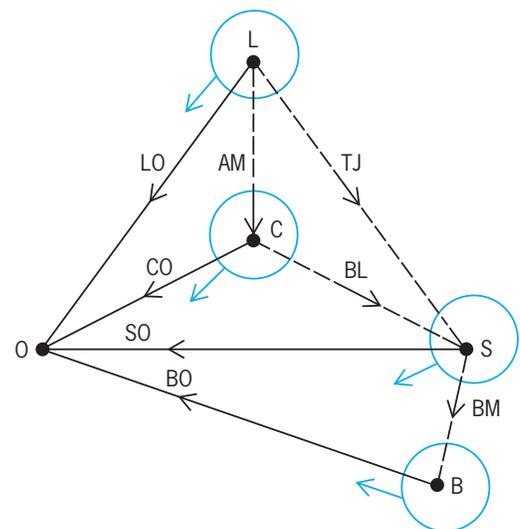


Fig. 2. Graph of the network in Fig. 1, using the same abbreviations. The link branches (AM, TJ, BL, and BM) are the membrane resistances; the tree branches (BO, SO, CO, and LO) are the capacitive storage branches representing accumulation or depletion of material in the compartments. O is a reference or "ground" concentration and is arbitrarily set at zero.

flows  $R_{ij} = R_{ji}$  and is a direct result of Tellegen's theorem.

**Simulation of living systems.** A variety of living systems have now been successfully simulated by using network thermodynamics. Among these are coupled epithelial transport of salt, water, and current; glomerular filtration; microcirculatory absorption and filtration; whole-body pharmacokinetics; cellular pharmacokinetics of anticancer drugs; cellular uptake and metabolism of glucose and its modification by insulin; ion transport systems such as channels and carriers; ecosystems; nervous transmission; synaptic function; and reaction-diffusion systems exhibiting multiple stationary states and hysteresis. See THERMODYNAMIC PRINCIPLES; THERMODYNAMIC PROCESSES.

D. C. Mikulecky

**Bibliography.** M. Athans et al., *Systems, Networks, and Computation: Multivariable Methods*, 1974; L. P. Huelsman, *Basic Circuit Theory*, 3d ed., 1991; D. C. Mikulecky, Network thermodynamics: A candidate for a common language for theoretical and experimental biology, *Amer. J. Physiol.*, 245: R1-R9, 1983; G. R. Oster, R. Perelson, and A. Katchalsky, Network thermodynamics: Dynamic modeling of biophysical systems, *Quart. Rev. Biophys.*, 6:1-134, 1973; L. Peusner, *Studies in Network Thermodynamics*, 1986; J. C. White and D. C. Mikulecky, Application of network thermodynamics to the computer modeling of the pharmacology of anticancer agents: A network model for methotrexate action as a comprehensive example, *Pharmacol. Ther.*, 15:251-291, 1982.

## Operating system

The software component of a computer system that is responsible for managing and coordinating activities and sharing the resources of the computer. The operating system (OS) acts as a host for application programs that are run on the machine. (In computer jargon, application programs are said to run on top of the operating system.) As a host, one of the functions of an operating system is to handle the details of the hardware operation. This relieves application programs from having to manage these details and makes it easier to write applications. Almost all computers use an operating system, such as handheld computers (including personal data assistants or PDAs), laptop and desktop computers, supercomputers, and even modern video game consoles. See COMPUTER SYSTEMS ARCHITECTURE.

Users may interact with the operating system by typing commands or using a graphical user interface (GUI, commonly pronounced "gooey"). The purpose of a GUI is to make it easier to interact with objects in the operating system, such as files, folders, and application programs. The GUI represents these objects as icons and provides easy methods for moving and copying objects, launching and switching between applications, scrolling through windows, and so on. The GUI is generally thought of as part of the operating system; however, most operating systems

implement it as a specialized application program or service that runs on top of the operating system. The Microsoft Windows<sup>®</sup> operating system and Apple's MacOS<sup>®</sup> each have their own distinctive GUIs that run for the most part outside the operating system. Other operating systems, such as Linux, use GUIs based on the X Window System, which similarly runs on top of the OS. See COMPUTER PROGRAMMING; HUMAN-COMPUTER INTERACTION.

Operating systems offer a number of services to application programs and users. Applications access these services through application programming interfaces (APIs) or system calls. By invoking these interfaces (which is usually done by making a special subroutine call in the program), the application can request a service from the operating system, pass parameters, and receive the results of the operation.

Different operating systems exist because no one operating system is capable of being efficient on all types of hardware and for all types of applications. For example, the needs of supercomputer applications are completely different from those of a video game running on a low-cost home game console. Therefore, each operating system provides its own set of APIs that the designers feel are best suited for the intended hardware and application environment. These may be quite different from the APIs of other operating systems, which can make it difficult to take an application written for one operating system and rewrite it to run on another. This is one reason why applications tend to be specific to a certain operating system. Nevertheless, a number of attempts have been made to standardize various subsets of an operating system's APIs so that programs that adhere to the standardized subset can be adapted to run on other operating systems that support the same standard with little effort. One such standard is POSIX, which standardizes a range of APIs for the different variations of the UNIX operating system. Programs that make use of standardized interfaces are said to be "portable" between different operating systems.

Although APIs and services vary from one operating system to another, there are several broad categories of services that most operating systems provide.

**Process management.** Modern operating systems provide the capability of running multiple application programs simultaneously, which is referred to as multiprogramming or multitasking. Each program running is represented by a process in the operating system. The operating system provides an execution environment for each process by sharing the hardware resources so that each application does not need to be aware of the execution of other processes. For example, since the central processing unit (CPU) of the computer can be used by only one program at a time, the operating system is responsible for sharing the CPU among the processes. This may be done using a technique known as time slicing, whereby each process is given a brief amount of time on the CPU (usually measured in milliseconds) before the CPU is passed to the next process. In this manner,

the processes take turns using the CPU. An operating system may also assign a priority to each process and grant use of the CPU to higher-priority processes ahead of lower-priority processes. This allows the operating system to meet scheduling constraints and deliver deterministic response time for high-priority processes. Operating systems that must meet critical scheduling constraints are referred to as real-time operating systems.

**Memory management.** The main memory of a computer (known as random access memory, or RAM) is a finite resource. All running applications require the use of memory to store the program's instructions and data while the application is executing. The operating system is responsible for sharing the memory among the currently running processes. When a user initiates an application, the operating system decides where to place it in memory and may allocate additional memory to the application if it requests it. Most operating systems use capabilities in the hardware to prevent one application from overwriting the memory of another. This provides security and prevents applications from interfering with one another. The operating system can also allow portions of memory to be shared among processes. This provides a way for processes on the computer to communicate efficiently. Shared memory is an example of an interprocess communication (IPC) facility.

Many modern operating systems employ a technique called virtual memory (VM). Here, the operating system utilizes special features in the hardware to provide each process with the illusion that it has its own dedicated RAM. This virtual memory may be larger than the actual RAM in the computer. The operating system stores the portions of the virtual memory (referred to as pages) that exceed the available RAM on disk in a paging file or swap area and transparently moves these portions back and forth between memory and disk as needed. This provides a way to share the RAM between multiple processes and to allow processes with very large memory requirements to run on computers that lack large amounts of RAM. *See* COMPUTER STORAGE TECHNOLOGY.

**Input/output and device management.** Computers can have a variety of input/output (I/O) devices attached to them. Personal computers may include a hard disk, network interface, CD/DVD drive, keyboard, mouse, graphical display, sound, camera, printer, scanner, modem, or other devices. The commands and methods for controlling an I/O device vary from device to device and may even vary for different manufacturers of the same type of device. Because of the great variety of devices available, it would be impossible to imbed the techniques for controlling all possible devices in every application. Therefore, the details of device management are left to the operating system. The operating system provides a set of APIs to the applications for accessing I/O devices in a consistent and relatively simple manner regardless of the specifics of the underlying hardware. The operating system will generally use a software component called a device driver to control an

I/O device. A device driver is typically provided by the manufacturer when the I/O device is purchased and is configured into the operating system when the device is first installed. This allows the operating system to be upgraded to support new devices as they become available. *See* COMPUTER PERIPHERAL DEVICES.

**File system management.** A file system is employed to manage and share the disk space on a computer. A file system allows users and applications to organize their data into files and assign mnemonic names to these files. Modern file systems allow files to be stored in a hierarchical structure for further convenience. File systems also maintain file ownership and permissions, and enforce security by controlling which users can read, write, or execute a given file.

Each file is generally viewed by the application or user as a simple sequence of bytes without regard to the physical location of the data on the disk. Allocating space on the disk is the responsibility of the file system, which may break a given file into pieces and store them in different locations on the disk. Among other benefits, this allows files to be easily extended after they are created. In this manner, the available space on the disk is shared among the different files that are created. This service is completely transparent to applications. However, it can lead to fragmentation of the disk. Some operating systems require that the system administrator periodically run a disk defragmentation utility. Others use a file system designed so that it does not require defragmentation.

So that the file system can be used on many different types of disks, the file system will use the device driver for the disk to manage and control the actual drive that the file system resides on. The device driver used by the file system can also be a virtual driver that controls a set of disks that have been organized to form one large virtual disk. These are generally called redundant arrays of inexpensive disks (RAID) and can allow the file system or even single files to use more space than is available on any single disk drive. RAID can also be used to provide redundancy in various forms so that data are not lost, even if a drive in the array fails.

**Network management.** A network allows multiple computers to be connected for exchanging data. There is a wide variety of hardware that can be used to create a network. Networks can also be wired or wireless. As with other types of I/O devices, the operating system uses a device driver for the network I/O device. In addition, the operating system includes software known as a network protocol and makes various network utilities available to the user. The protocol provides a standardized method for two computers to communicate and specifies how the data are transferred and how transmission errors are detected. The protocol then provides an API for its services, which allows utility programs and applications to more easily transfer files and data. Applications such as e-mail and Web browsing are implemented outside the operating system by using the APIs provided by the network management portion

of the operating system, while services such as file sharing are generally implemented in the operating system.

Probably the most commonly used protocol suite in use is TCP/IP (transmission control protocol/internet protocol) by which most computers communicate with each other over the Internet. By implementing the network protocol and device driver in the operating system, application programs are relieved of the burden of the communication details. This allows applications to communicate with another computer located around the world as easily as it can with one locally. *See* LOCAL-AREA NETWORKS; WIDE-AREA NETWORKS.

**Security.** Operating systems provide security by preventing unauthorized access to the computer's resources. Many operating systems also prevent computer users from accidentally or intentionally interfering with each other. The security policies that an operating system enforces range from none in the case of a video game console, to simple password protection for hand-held and desktop computers, to very elaborate schemes for use in high-security environments. The security policies of an operating system can control who can log into the machine, which users can read or modify which files, which users can install software or reconfigure the operating system, which applications can use the network, which users can terminate applications, and so forth.

With networked computers, a large focus of computer security is in preventing outside users from gaining unauthorized access to a computer's data over the network. This begins in the operating system, which initially determines what outside connections will be accepted. Security continues by having the higher-level networking services and applications (such as e-mail and Web servers) further authenticate outside users before access is granted. Security is most frequently compromised on networked computers due to improper configuration, bugs in the networking or operating system software, or user error that might cause a virus to be installed. Any of these can allow an outside user to gain unauthorized access to the machine. Antivirus software and network firewalls, which further restrict external access to the computer, can be used to help thwart security attacks. *See* COMPUTER SECURITY.

Curt Schimmel

Bibliography. C. Schimmel, *UNIX Systems For Modern Architectures*, Addison-Wesley, 1994; W. Stallings, *Operating Systems: Internals and Design Principles*, 4th ed., Prentice Hall, 2000; A. S. Tanenbaum, *Modern Operating Systems*, 2d ed., Prentice Hall, 2001.

## Operational amplifier

A voltage amplifier that amplifies the differential voltage between a pair of input nodes. For an ideal operational amplifier (also called an op amp), the amplification or gain is infinite.

Most existing operational amplifiers are produced on a single semiconductor substrate as an integrated circuit. These integrated circuits are used as building blocks in a wide variety of applications. *See* INTEGRATED CIRCUITS.

Discrete-component and vacuum-tube operational amplifiers preceded the integrated operational amplifiers. The cost of these operational amplifiers was very high, and the performance was modest. The first integrated operational amplifiers appeared in the mid-1960s. By the mid-1970s, inexpensive, general-purpose integrated operational amplifiers with excellent characteristics were widely available. A wide variety of operational amplifiers are now available. Inexpensive general-purpose operational amplifiers are adequate for most applications. Special-purpose operational amplifiers have performance characteristics specifically tailored for niche applications and must be used when the required performance cannot be attained from the general-purpose devices.

**Operation.** Although an operational amplifier is actually a differential-input voltage amplifier with a very high gain, it is almost never used directly as an open-loop voltage amplifier in linear applications for several reasons. First, the gain variation from one operational amplifier to another is quite high and may vary by  $\pm 50\%$  or more from the value specified by the manufacturer. Second, other nonidealities such as the offset voltage make it impractical to stabilize the dc operating point. Finally, performance characteristics such as linearity and bandwidth of the open-loop operational amplifier are poor. In linear applications, the operational amplifier is almost always used in a feedback mode.

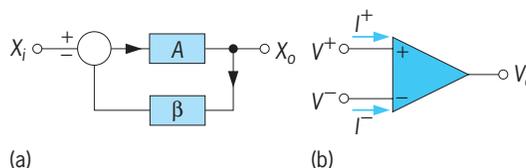
A block diagram of a classical feedback circuit is shown in **Fig 1a**. The transfer characteristic, often termed the feedback gain  $A_f$  of this circuit, is given by Eq. (1). In the limiting case, as  $A$  becomes very large, the feedback gain is approximated by Eq. (2).

$$\frac{X_o}{X_i} = A_f = \frac{A}{1 + A\beta} \quad (1)$$

$$A_f \simeq \frac{1}{\beta} \quad (2)$$

*See* FEEDBACK CIRCUIT.

An operational amplifier is often used for the amplifier designated  $A$  in this block diagram. Since  $A_f$  in the limiting case is independent of  $A$ , the exact gain characteristics of the operational amplifier become unimportant provided the gain is large. Although linear applications of the operational amplifier extend



**Fig. 1. Basic circuits. (a) Classical feedback circuit. (b) Operational amplifier symbol typically used in circuit diagrams.**

well beyond the simple feedback block diagram of Fig. 1a, the applications invariably involve circuit structures with feedback that make the characteristics of the circuit nearly independent of the exact characteristics of the operational amplifier. Such circuits are often termed active circuits.

The commonly used operational amplifier symbol is shown in Fig. 1b. In this circuit, the output voltage is related to the gain  $A$  of the operational amplifier by Eq. (3), where  $A$  is very large and the input currents

$$V_o = A(V^+ - V^-) \quad (3)$$

$I^+$  and  $I^-$  are nearly zero. As a consequence of the high voltage gain, Eq. (4) follows from Eq. (3). This

$$V^+ \simeq V^- \quad (4)$$

observation is often used in the analysis of active circuits.

Several typical operational-amplifier applications are shown in Fig. 2 along with the relationship between the input and output. The circuits in Fig. 2a-d are linear or active circuits using feedback. The circuit of Fig. 2e does not use feedback to realize

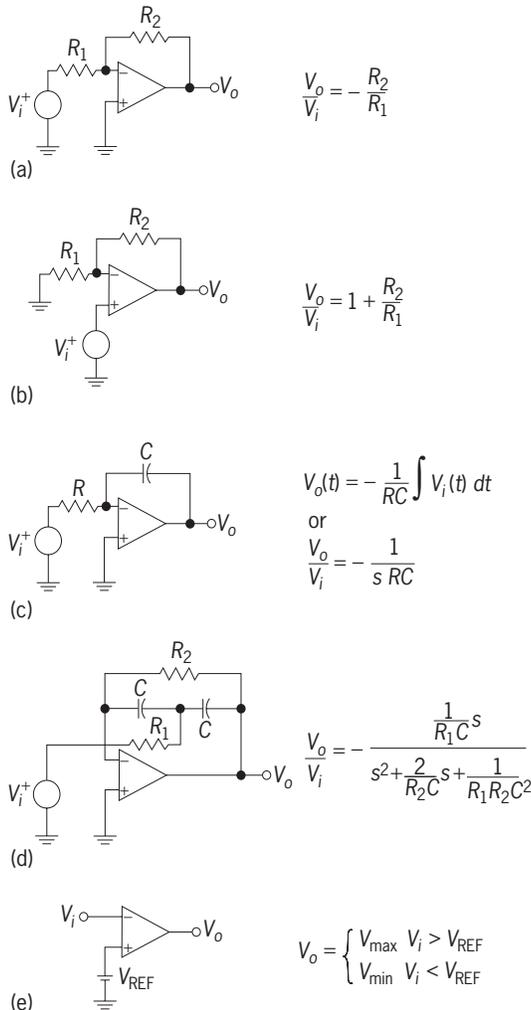


Fig. 2. Typical operational amplifier applications. (a) Noninverting amplifier. (b) Inverting amplifier. (c) Integrator. (d) Band-pass filter. (e) Voltage comparator.

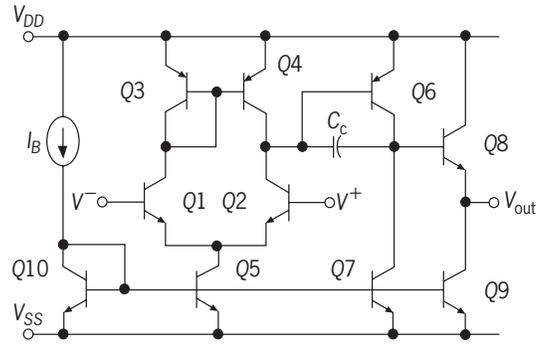


Fig. 3. Simplified circuit diagram of a bipolar operational amplifier.

the nonlinear comparator transfer characteristic. The use of Eq. (4) in the analysis of linear operational-amplifier circuits can be demonstrated easily with the circuit of Fig. 2a. Since the voltage at the plus terminal is at 0 V, the potential at the minus terminal must also be at 0 V. Since  $I^- \simeq 0$ , application of Kirchoff's current law to the node connecting  $R_1$  and  $R_2$  yields Eq. (5), which can be solved to obtain the gain expression, Eq. (6).

$$\frac{V_i}{R_1} + \frac{V_o}{R_2} = 0 \quad (5)$$

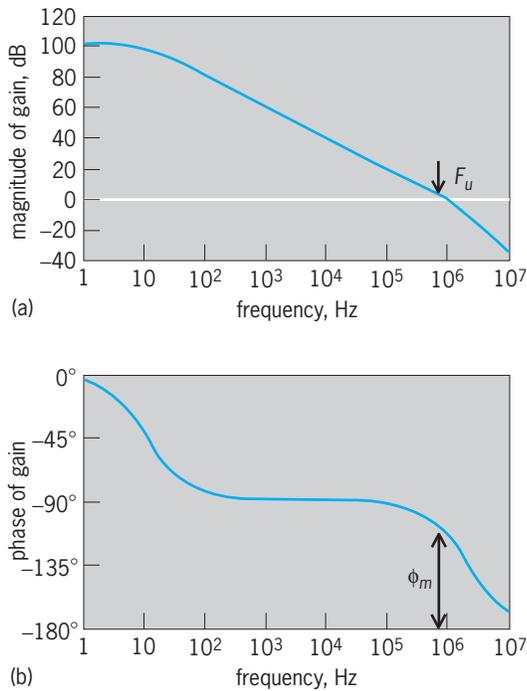
$$\frac{V_o}{V_i} = -\frac{R_2}{R_1} \quad (6)$$

**Design.** Figure 3 is a simplified circuit diagram of a typical bipolar operational amplifier. Every operational amplifier has a differential input stage (here, transistors  $Q_1$  to  $Q_5$ ), which amplifies the difference between the input voltages,  $V^+$  and  $V^-$ , and rejects the sum or common-mode component, followed by one or more secondary amplification stages (here one stage, transistors  $Q_6$  and  $Q_7$ ). To reduce the output impedance, a low-impedance output stage is added, such as the emitter follower ( $Q_8$  transistors and  $Q_9$ ) in Fig. 3. A pole-splitting compensation capacitor  $C_c$  is added over the secondary gain stages, to push one of the poles in the open-loop gain characteristic to very low frequencies, and at least one of the other poles to higher frequencies, thus improving the stability of the operational amplifier in a feedback configuration. Most commercial operational amplifiers are stabilized for the unity-gain follower configuration, which is the worst case for stability considerations. See CONTROL SYSTEMS; EMITTER FOLLOWER.

**Specifications.** Several specifications are used to characterize the performance of an internally compensated operational amplifier.

The open-loop gain  $A_{OL}$  is the gain between the differential input and the output (Fig. 4a) at zero frequency (dc). It is typically high enough to be safely ignored as a source of errors. See GAIN.

The gain-bandwidth product (GB) is the product of the open-loop gain  $A_{OL}$  and the open-loop 3-decibel bandwidth. For frequencies beyond this 3-dB bandwidth, the gain characteristic falls off at



**Fig. 4.** Open-loop gain characteristics of a typical commercial operational amplifier. (a) Magnitude of open-loop gain. (b) Open-loop phase characteristic.

20 dB per decade, since a dominant pole is created through pole splitting. The unity-gain frequency  $F_u$ , which is the frequency at which the gain of the operational amplifier drops to 0 dB (Fig. 4a), is close to the gain-bandwidth product. The gain-bandwidth product is an indication of the bandwidth achievable when feedback is applied around this operational amplifier.

The phase margin  $\phi_m$  is the phase difference between the actual phase of the open-loop gain at the unity-gain frequency  $F_u$  and  $-180^\circ$  (Fig. 4b). Positive values of  $\phi_m$  guarantee stability of a feedback circuit built with that amplifier, with larger values of  $\phi_m$  yielding better stability.

The offset voltage  $V_{os}$  is the differential input voltage that must be applied to return the output voltage to 0 V. It is due to device mismatching in the input stage.

The input bias current  $I_B$  is the constant (dc) current that is drawn by the input nodes to bias the input devices of the operational amplifier. Related to the input bias current  $I_B$  is the small-signal parameter input impedance  $Z_{in}$ . The input offset current  $I_{os}$  is the difference between the input bias currents drawn by the two input nodes, caused by mismatch in the devices in the input stage. These parameters are primarily important for operational amplifiers with bipolar input transistors, since these have a finite current gain  $h_{FE}$ . In operational amplifiers whose input stages are made up of junction field-effect transistors (JFETs),  $I_B$ ,  $I_{os}$ , and  $Z_{in}$  can usually be ignored.

The output impedance is specified in the open-loop configuration. Typically, it will exhibit itself as an apparent reduction in the open-loop gain  $A_{OL}$  for low-impedance amplifier loads.

The settling time  $T_{settle}$  is the time it takes the voltage  $V_o$  at the output node of an operational amplifier (in feedback configuration) to settle to within a specific error band. For small signal swings, the settling process is linear. The slew rate (SR) is the maximal variation of the output voltage per unit time ( $dV_o/dt$ ), which originates from the fact that internal stages are usually biased by constant-current sources. Slewing (operating at maximal output variation) will cause nonlinear distortion in large-swing output signals.

The common-mode rejection ratio (CMRR) is the ratio between the gain of the operational amplifier for difference signals between the input terminals, and the gain for the average or common-mode signal component. The common-mode rejection ratio is important whenever the differential signal is superimposed on a much larger common-mode signal. See DIFFERENTIAL AMPLIFIER.

The power-supply rejection ratio (PSRR) is the ratio between the gain of the operational amplifier for difference signals between the input terminals, and the gain for variations of the power-supply voltages. A finite value of the power-supply rejection ratio affects the effective signal-to-noise ratio and the stability.

The input referred noise voltage  $e_N$  and the input referred noise current  $i_N$  are two noise sources (specified as noise spectral densities) that model the noise performance of a complete operational

Operational amplifier specifications			
Parameter	Ideal value	Typical value*	Optimized value†
Open-loop gain ( $A_{OL}$ )	$\infty$	100 dB	160 dB
Gain-bandwidth product (GB)	$\infty$	1 MHz	1 GHz
Phase margin ( $\phi_m$ )	$90^\circ$	$70^\circ$	—
Offset voltage ( $V_{os}$ )	0	5 mV	20 $\mu$ V
Input bias current ( $I_B$ )	0	200 nA	0.1 pA
Input offset current ( $I_{os}$ )	0	10 nA	20 fA
Slew rate (SR)	$\infty$	0.7 V/ $\mu$ s	5 V/ $\mu$ s
Common-mode rejection ratio (CMRR)	$\infty$	90 dB	140 dB
Power-supply rejection ratio (PSRR)	$\infty$	80 dB	140 dB
Input referred noise voltage ( $e_N$ )	0	20 nV/ $\sqrt{\text{Hz}}$	2 nV/ $\sqrt{\text{Hz}}$
Input referred noise current ( $i_N$ )	0	0.2 pA/ $\sqrt{\text{Hz}}$	1 fA/ $\sqrt{\text{Hz}}$

\*Close to that of a 741 circuit.  
†Best value found in commercial operational amplifiers. (These specifications cannot be found in one single operational amplifier.)

amplifier. Two independent sources are required to predict correctly the noise performance of an active circuit. For a low-impedance feedback network around the operational amplifier,  $e_N$  dominates the noise performance; for a high-impedance feedback network,  $i_N$  dominates. JFET-input operational amplifiers are used primarily for high-impedance signal processing (such as the amplification of capacitive sensor signals), since they have very low values of  $i_N$ .

The most important specifications of operational amplifiers are given in the **table**. See AMPLIFIER; CIRCUIT (ELECTRONICS). Peter M. VanPeteghem

Bibliography. A. Barna and D. I. Porat, *Operational Amplifiers*, 2d ed., 1989; D. J. Dailey, *Operational Amplifiers and Linear Integrated Circuits: Theory and Applications*, 1989; P. R. Gray and R. G. Meyer, *Analysis and Design of Analog Integrated Circuits*, 3d ed., 1993; R. Gregorian and G. C. Temes, *Analog MOS Integrated Circuits for Signal Processing*, 1986; J. Wait, L. Huelson, and G. Korn, *Introduction to Operational Amplifiers: Theory and Applications*, 2d ed., 1992.

### Operations research

The application of scientific methods and techniques to decision-making problems. A decision-making problem occurs where there are two or more alternative courses of action, each of which leads to a different and sometimes unknown end result (Fig. 1). Operations research is also used to maximize the utility of limited resources. The objective is to select the best alternative, that is, the one leading to the best result. Often, however, it is not simply a matter of searching a table, as there may literally be an infinity of outcomes. More intelligent means are needed to seek out the prime result.

To put these definitions into perspective, the following analogy might be used. In mathematics, when solving a set of simultaneous linear equations, one states that if there are seven unknowns, there must be seven equations. If they are independent and consistent and if it exists, a unique solution to the prob-

		resulting end states				
		$\theta_1$	$\theta_2$	$\theta_3$	$\theta_4$	$\theta_5$
alternatives	$a_1$	$r_{11}$	$r_{12}$	$r_{13}$	$r_{14}$	$r_{15}$
	$a_2$	$r_{21}$	$r_{22}$	$r_{23}$	$r_{24}$	$r_{25}$
	$a_3$	$r_{31}$	$r_{32}$	$r_{33}$	$r_{34}$	$r_{35}$
	$a_4$	$r_{41}$	$r_{42}$	$r_{43}$	$r_{44}$	$r_{45}$

Fig. 1. End states for alternatives in a decision-making process.

lem is found. In operations research there may be figuratively “seven unknowns and four equations.” There may exist a solution space with many feasible solutions which satisfy the equations. Operations research is concerned with establishing the best solution. To do so, some measure of merit, some objective function, must be prescribed.

There are several terms associated with the subject matter of this program: operations research, management science, systems analysis, operations analysis, and so forth. While there are subtle differences and distinctions, the terms can be considered nearly synonymous. See SYSTEMS ENGINEERING.

The field can be divided into two general areas with regard to methods. These are those that can be termed mathematical programming and those associated with stochastic processes. While computers are heavily used to solve problems, the term programming should not be considered in that sense, but rather in the general sense of organizing and planning. Also, the tools of probability and statistics are used to a considerable extent in working with stochastic processes. These areas will be explored in greater detail in a later section. With regard to areas of applications, there are very few fields where the methods of operations research have not been tried and proved successful. Following is a brief history of the field, and then the general approach to solving problems.

### History

While almost every art and every science can reach back into antiquity for its roots, operations research can reach back less than a century to find its beginnings. During World War I, F. W. Lancaster developed models of combat superiority and victory based on relative and effective firepower. Thomas Edison studied antisubmarine warfare, and in 1915 F. Harris derived the first economic order quantity (EOQ) equation for inventory. Starting in 1905 and continuing into the 1920s, A. Erlang studied the flow of calls into a switchboard and formed the basis of what is now known as queueing theory.

**Empiricists.** The formal beginning of operations research was in England during World War II, where the term was and still is operational research. Early work concerned air and coastal defense—the coordination of fighter aircraft, antiaircraft guns, barrage balloons, and radar. Typical of the research groups formed was “Blackett’s Circus.” This interdisciplinary group consisted of three physiologists, two math physicists, an astrophysicist, an army officer, one surveyor, one general physicist, and two mathematicians. The basic mode of operation was to observe the problem area, and then call on the expertise of the various disciplines to apply methods from other sciences to solve the particular problem. In retrospect, this was an era of “applied common sense,” and yet it was novel and highly effective.

**Pragmatists.** Following World War II, operations research continued to exist mainly in the military area. The operations research groups formed during the war stayed together, and a number of

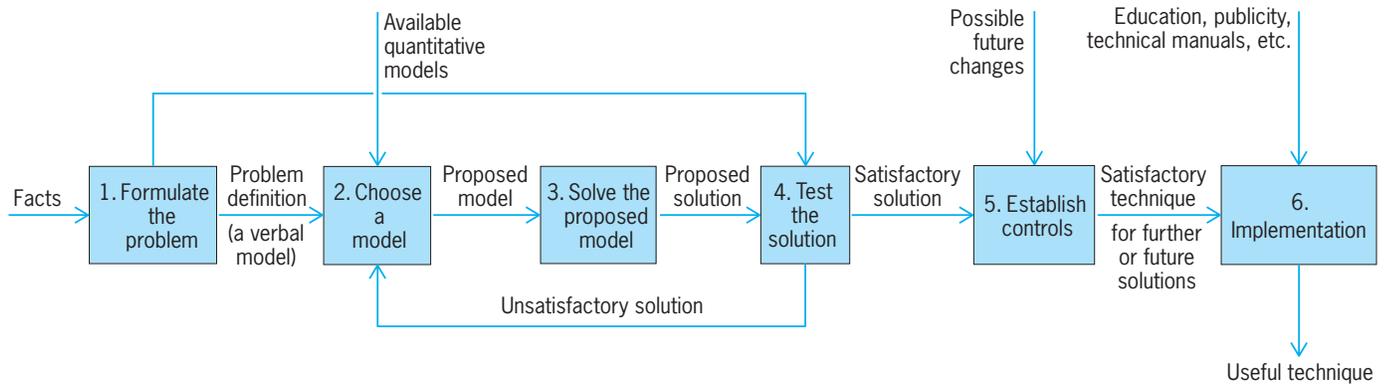


Fig. 2. Operations research approach: the six basic rules of success.

civilian-staffed organizations were established—RAND (1946), ORO (1948), and WSEG (1948). The real impetus to this era and to the whole field was the work done by George Dantzig and colleagues at RAND on Project Scope, undertaken for the U.S. Air Force in 1948. In attacking the problem of assigning limited resources to almost limitless demands, they developed the techniques of linear programming. Perhaps no other method is more closely associated with the field. Its use quickly spread from the military to the industrial area, and a new dimension was added to operations research. No longer was it simply “observe, analyze, and try.” For the first time the field became “scientific.” It could now “optimize” the solutions to problems. See LINEAR PROGRAMMING.

It was during this time that formal courses in operations research were first offered. It was also during this time that operations research suffered its first lapse. Linear programming soon was looked to as the cure for too many of industry’s ills. Unfortunately, industry’s problems were not all linear, and “straightening them out” to fit resulted in many aborted projects and reports that were simply shelved instead of implemented. The early 1950s saw a growth in the number of industrial operations research groups; by the end of the decade, many had disappeared.

**Theorists.** Toward the end of the 1950s a number of highly skilled scientists emerged who made some substantial contributions to the field. Operations research came of age and matured. In fact, the movement was so far advanced toward developing a sound theoretical base that a new problem arose—the practicality gap.

### Methodology

Operations research today is a maturing science rather than an art—but it has outrun many of the decision makers it purports to assist. The success of operations research, where there has been success, has been the result of the following six simply stated rules:

- 1 Formulate the problem.
- 2 Construct a model of the system.
- 3 Select a solution technique.

- 4 Obtain a solution to the problem.
- 5 Establish controls over the system.
- 6 Implement the solution (Fig. 2).

The first statement of the problem is usually vague, inaccurate, and sometimes not a statement of the problem at all. Rather, it may be a cataloging of observable effects. It is necessary to identify the decision maker, the alternatives, goals, and constraints, and the parameters of the system. At times the goals may be many and conflicting; for example: Our goal is to market a high-quality product for the lowest cost yielding the maximum profit while maintaining or increasing our share of the market through diversifications and acquisitions; yielding a high dividend to our stockholders while maintaining high worker morale through extensive benefits, without the Justice Department suing us for constraint of trade, competition, price fixing, or just being too big.

More properly, a statement of the problem contains four basic elements that, if correctly identified and articulated, greatly eases the model formulation. These elements can be combined in the following general form: “Given (the system description), the problem is to optimize (the objective function), by choice of the (decision variable), subject to a set of (constraints and restrictions).”

In modeling the system, one usually relies on mathematics, although graphical and analog models are also useful. It is important, however, that the model suggest the solution technique, and not the other way around. Forcing a model to fit a preferred technique led to some of the bad operations research of the past.

With the first solution obtained, it is often evident that the model and the problem statement must be modified, and the sequence of problem-model-technique-solution-problem may have to be repeated several times. The controls are established by performing sensitivity analysis on the parameters. This also indicates the areas in which the data-collecting effort should be made.

Implementation is perhaps of least interest to the theorists, but in reality it is the most important step. If direct action is not taken to implement the solution, the whole effort may end as a dust-collecting

report on a shelf. Given the natural inclination to resist change, it is necessary to win the support of the people who will use the new system. To do this, several ploys may be used. Make a member of the using group also a member of the research team. (This provides liaison and access to needed data.) Educate the users about what the system does—not to the extent of making them experts, but to alleviate any fear of the unknown. Perhaps the major limitation to successful use of operations research lies in this phase.

**Mathematical Programming**

Probably the one technique most associated with operations research is linear programming. The basic problem that can be modeled by linear programming is the use of limited resources to meet demands for the output of these resources. This type of problem is found mainly in production systems, but is not limited to this area. Since this method is so basic to the operations research approach, its use will be illustrated.

**Linear program model.** Consider a company that produces two main products—X and Y. For every unit of X it sells, it gets a \$10 contribution to profit (selling price minus direct, variable costs), and for Y it gets a \$15 contribution. How many should they sell of each? Obviously there must be some limitations—on demand and on productive capacity. Suppose they must, by contract, sell 50 of X and 10 of Y, while at the other end the maximum sales are 120 of X and 90 of Y. Unfortunately they have only 40 total hours of productive capacity per period, and it takes 0.25 h to make an X and 0.4 h to make a Y. What is their optimal strategy? First, the problem must be formally stated: “Given a production system making two products, the problem is to maximize the contribution to profit, by the choice of how many of each product to make, subject to limits on demand (upper and lower) and available production hours.”

If X equals the number of first products made and Y the number of the second products made, the system can be modeled as follows:

Maximize				
	Profit contrib. =	10X + 15Y		<i>Constraints</i>
				<i>line in</i>
				<i>Fig. 3</i>
Subject to	X	≥ 50	1	
	X	≤ 120	2	
	Y	≥ 10	3	
	Y	≤ 90	4	
	.25X + .40Y	≤ 40	5	

The first two constraints are the lower and upper bounds of demand on X, the next two are for Y, and the last constraint refers to the productive capacity.

The problem is illustrated in Fig. 3. Any point in the feasible region of solution and the edges will satisfy all the demand and capacity constraints. To pick the best, the objective—“maximize contribution to profit”—is used. Several isoprofit lines, that is, lines where the profit is a constant value, have been added to Fig. 3. Note that the \$600 line is below the feasible region, the \$1800 line is above it, and the \$1200

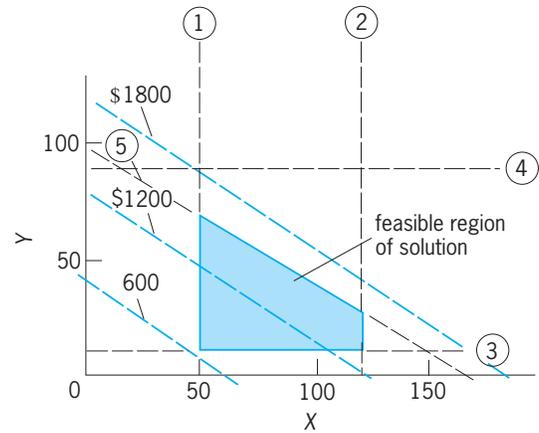


Fig. 3. Graphical presentation of production problem; circled numbers refer to the constraints.

line runs through it. If other lines were formed between \$1200 and \$1800, one could graphically find the maximum-profit-level line that still had one point in the feasible region. This would occur at the intersection of the second and fifth constraints. At this point, the solution is as follows:

$$\begin{aligned} \text{Profit contrib.} &= \$1575 \\ X &= 120 \\ Y &= 25 \end{aligned}$$

It is fairly easy to verify that this is the optimal solution, that all the constraints are satisfied, and that no larger profit can be found.

While this simple example did not pose much of a computational problem, actual cases do, in that there may be thousands of variables and thousands of constraints. Problems of scheduling the product mix from a refinery where a blend of different crudes can be used (each with different costs and properties) are in this category. Efficient algorithms and computer codes have been developed to solve these problems.

**Other models.** There are a number of other linear programming models that relate to specific types of problems. However, advantage is taken of the special structure of the model to develop special and more efficient methods of solution.

**Transportation problem.** While the transportation problem can be modeled as a linear programming model, it is not efficient to solve it in that manner. Consider the tableau shown in Fig. 4. This shows the unit profit of shipping a common item from four plants to six regional distribution points. Supply refers to the number of units that each source can provide. Demand refers to the number of units that each destination requires. The “dummy” column is added to the supply-demand matrix so as to make the supply equal the demand. In Fig. 4, a dummy demand of 300 units is added to make it equal the total 3600 units of supplies available from four sources. Most transportation problems simply minimize the cost of transportation. This example includes variable selling price, variable production cost, and transportation costs that depend on the distance between the plant and destination.

If the problem were to be modeled as a linear programming model, there would be a decision variable for each combination of shipment. For the example problem there are 4 sources and 6 possible destinations, resulting in  $24$  decision variables: the amount shipped from each source to each destination. There would be 10 constraints—4 indicating the capacity limits on the plants and 6 reflecting the demand constraints. In general, there will be  $S \times D$  variables and  $S + D$  constraints, where  $S$  is the number of sources and  $D$  is the number of constraints. For a more realistic problem, there might be a dozen sources and five dozen destinations. Then the linear programming model would have 720 variables and 72 constraints.

Another factor is that the linear programming model matrix will be composed entirely of 0s and 1s. Also these 1s will be regularly placed. Because of this regular structure, more efficient methods have been found to solve the problem. These algorithms typically have two phases, just as in linear programming.

In the first phase, the objective is to find a feasible solution. One such heuristic is called the North-West Corner rule. Simply stated, the solution process starts in the upper left-hand corner of the tableau, and either the minimum of supply or demand is assigned to the first cell. The next step is to move either horizontally or vertically, depending upon where the demand has been satisfied or the capacity has been all used. The result is a set of assignments that generally move down the main diagonal of the tableau.

A more efficient algorithm is known as Vogel's approximation method (VAM). Again using the transportation tableau, the procedure is to examine each row and column to determine the difference between the highest unit profit and the next highest unit profit. Where only costs are considered, the procedure would be to determine the difference between the lowest cost and the next lowest cost.

These differences are examined to determine the largest difference, and a transportation assignment is made to the highest unit profit that caused this largest difference. The assignment is to ship either the remaining capacity of the plant or the remaining unsatisfied demand of the destination—whichever is smaller.

The basic philosophy of Vogel's approximation method is that these differences represent a measure of "regret"—that is, how much would be lost (or how much more would have to be paid) if the row or column representing highest unit profit was not used. That is, if it is not used, then it might be necessary to settle for "second best." Choosing the largest difference is an attempt to minimize regret.

Vogel's approximation method then proceeds to recalculate these differences for the remaining row and columns (one row or column is eliminated at each iteration) and to repeat the procedure. When the method is completed, the result will not only be a feasible solution but an extremely good one—and possibly the optimal.

INCREMENTAL PROFIT MATRIX								
SOURCES	DESTINATIONS							source supply
	St. Louis	Kansas City	Little Rock	Dallas	Houston	Montgomery	dummy	
Atlanta	25.00	27.60	40.50	22.50	32.00	48.50	0.00	1200
Austin	12.00	18.50	30.00	18.00	28.50	32.00	0.00	500
Memphis	32.60	35.20	48.70	30.80	39.80	52.00	0.00	1200
Tulsa	21.40	28.00	37.70	22.50	30.50	38.00	0.00	700
destination demand	750	500	150	750	1000	150	300	3600

Fig. 4. Transportation problem tableau, showing the incremental profit and the supply-demand matrices of sources and destinations.

After either approach, the result is an initial basic feasible solution. The next steps are to determine if the solution is optimal. The specific algorithm is too lengthy to present here.

*Assignment problem.* The assignment problem is also a linear programming model and a variation of the transportation problem. It can be stated as: Given a set of  $N$  tasks to be performed by  $N$  people and a table of values showing the cost or time it takes each person to perform each task, the problem is to minimize the total time or cost it takes to perform the  $N$  tasks by choice of which task is assigned to each person, with the restrictions that each task must be done and each person is assigned to only one task.

This can be considered a special form of the transportation problem where there are  $N$  sources (tasks) and  $N$  destinations (people). Also, the capacities are each 1. The linear programming simplex method to solve the problem can be used. There would be  $N \times N$  variables and  $2N$  constraints. The matrix also would be all 0's and 1's. The transportation algorithm could also be used, but would run into many degenerate solution points.

The special structure of the model is used to form a special algorithm to efficiently solve this model. It is known as the Hungarian method and consists of the following steps:

1. The smallest element in each row of the cost-time matrix is subtracted from each element in that row. (This will create at least one element that is 0.)

2. The same procedure is repeated for each problem.

3. An attempt is made to make one assignment to each row and column using the elements that are 0 as candidates. (The recommended procedure is to start with those rows and columns with only one 0.)

4. If a complete set of assignments has been made, then the optimal solution has been found; if not, the next step in the procedure is followed.

5. Using a minimum number of lines either horizontal or vertical, the 0's in each row and column are crossed out. (This can always be done, as many lines are assignments in the current solution.)

6. The smallest element not crossed out is selected, and it is subtracted from each element not crossed out and added to each element at the intersection of a vertical and horizontal line. (This will result in the creation of at least one new 0 as a candidate for assignment and may result in the loss of one or more 0's if they happened to be at the intersection of two lines.)

7. The entire process is repeated (steps 1-6) until the optimal solution is found.

The solution process is reasonably simple, and even modestly large problems can be solved by hand. Besides the generic assignment problem, this approach has been used for such diverse applications as the machine change-over problem (the sequence of jobs to be run on a machine to minimize setup time; assigning students to projects; allocating parking spaces to employees; and so forth).

**Network models.** There is another set of linear programming models that fall under the general category of network models:

The shortest-path problem determines the shortest path from one node to another. (The reverse of this is the critical path method used in project management.)

The max flow problem determines the capacity of a network such as a pipeline or highway system.

The min cost problem is a variation of the transportation problem.

*Integer linear programming.* There is a class of problems that have the linear programming form, but in which the variables are limited to being integer values. These represent the most difficult set to solve, especially where the variables take on values of 0 or 1. (This problem occurs, for instance, in project selection—the project is either funded, 1, or not funded, 0.) While a number of algorithms have been developed, there is no assurance that a given problem will be solved in a reasonable amount of computer time. These algorithms fall into two general categories—the cutting plane method, which adds constraints that cut off noninteger solution points, and the branch and bound method, which examines a tree network of solution points. See ALGORITHM.

*Nonlinear programming.* Another category of problems arises where either the objective function or one or more of the constraints, or both, are nonlinear in form. A series of methods have been developed that have varying degrees of success, depending on the problem structure.

*Geometric programming.* One of the techniques developed relates to a certain class of nonlinear models that use the arithmetic-geometric mean inequality relationship between sums and products of positive numbers. Such models result from modeling engineering design problems, and at times can be solved almost by inspection. Because of the ease of solution, a considerable effort has been made to identify the various problems that can be structured as a geometric programming model.

*Dynamic programming.* This technique is not as structured as linear programming, and more properly should be referred to as a solution philosophy rather

than a solution technique. Actually, predecessors of this philosophy have been known for some time under the general classification of calculus of variation methods. Dynamic programming is based on the principle of optimality as expressed by Richard Bellman: “An optimal policy has the property that whatever the initial state and initial decision are, the remaining decision must constitute an optimal policy with regard to the state resulting from the first decision.” See CALCULUS OF VARIATIONS; OPTIMIZATION.

The operative result of this principle is to start at the “end” of the problem—the last stage of the decision-making process—and “chain back” to the beginning of the problem, making decisions at each stage that are optimal from that point to the end. To illustrate this process, the fly-away-kit problem (also known as the knapsack problem) will be briefly described: Given a set of components, each with a unit weight and volume, the problem is to maximize the value of units carried to another location, by choice of the number of each to be taken, subject to limitations on volume and weight that can be carried.

In this problem each component will represent a stage at which a decision is to be made (the number to be included in the kit), and the amount of volume and weight left are defined as state variables. The end of the problem, then, is where there is only one more component to consider (Fig. 5a).

Usually at this stage the solution is almost trivial. As many of the last components ( $X_1$ ) are selected as can be within the limits of available volume ( $V_1$ ) and weight ( $W_1$ ). In practice, all possible values of these two state variables are solved for, since it is not known how much will be left when the last stage is reached. To this solution is now “chained” the problem of how to select the second-last component (Fig. 5b).

Here is considered the problem that, for each level of volume and weight, a choice must be made between taking one or more of component 2 or passing it on to the last stage. Combinations of components 1 and 2 must now be looked at to find the best mix, considering the available resources. In like manner, the solution is “chained back” to the beginning of the problem (Fig. 5c).

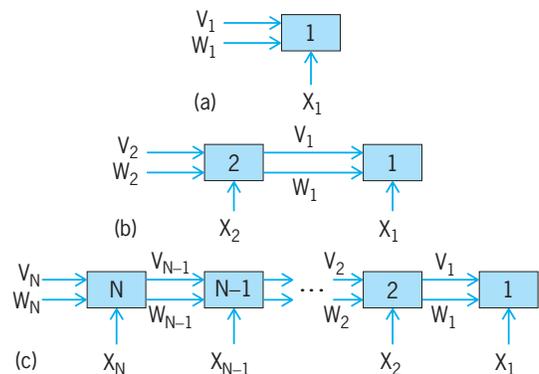


Fig. 5. Dynamic programming problems. (a) Final decision stage. (b) Two-stage problem. (c) “N”-stage problem.

The measures of value can be in any terms. If the components are simple cargo, they could have a monetary value. If they are spare parts for some system, a reliability measure could be used.

**Stochastic Processes**

A large class of operations research methods and applications deals with stochastic processes. These can be defined as processes in which one or more of the variables take on values according to some, perhaps unknown, probability distribution. It takes only one of these random variables to make the process stochastic.

In contrast to the mathematical programming methods and applications, there are not many optimization techniques. The techniques used tend to be more diagnostic than prognostic; that is, they can be used to describe the “health” of a system, but not necessarily how to “cure” it. This capacity is still very valuable. See STOCHASTIC PROCESS.

**Queueing theory.** Probably the most studied stochastic process is queueing. A queue or waiting line develops whenever some customer seeks some service that has limited capacity. This occurs in banks, post offices, doctors’ offices, supermarkets, airline check-in counters, and so on. But queues can also exist in computer centers, repair garages, planes waiting to land at a busy airport, or at a traffic light.

Queueing theory is the prime example of what can be said about the state of the system but not how to improve it. Fortunately, the improvements can be made by increasing the resources available. For example, another teller opens up a window in a bank, or another check-out stand is opened in the supermarket. Other possible changes are more subtle. For example, the “eight items or less” express lane in a supermarket minimizes the frustration of a customer in line behind another with two full shopping carts. Some post offices and banks have gone to a single queue where the person at the head of the line goes to the next available window.

While it is not possible to generalize the results of queueing analysis, it is possible to provide some general measures. One is the load or traffic factor. This is the ratio of arrival rate to combined service rate, considering the number of service centers in operation. It is possible to plot queueing statistics against this factor as shown in Fig. 6.

When the load factor ( $\rho$ ) is low, there is excess serving capacity. There may be occasional lines, but not often. As the load factor rises, a fairly linear rise is obtained in any queueing statistics until approximately the point where  $\rho = 0.75$ . After this there is a sharp and continuous rise in the lengths of line and waits. The system is becoming saturated. If  $\rho = 1.00$ , the best that can be said is, “the larger the line, the larger the wait.”

The following problem is an example of the type of analysis that can be done on a simple queueing system. A check-out counter at a small market can handle customers at a rate of 20 per hour (3 min per customer). Since the number of customers will vary considerably throughout the day and throughout the

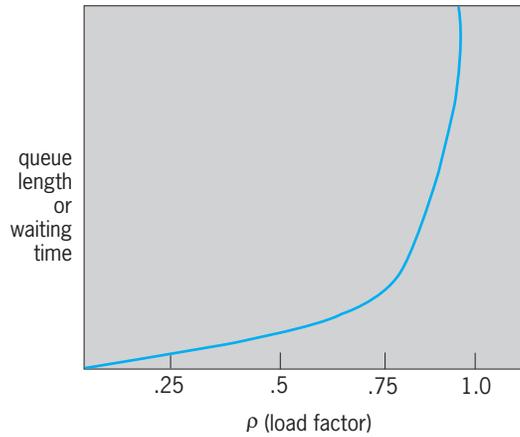


Fig. 6. Plot for general queues relating load factor to serving capacity.

days of the week, the arrival rate of customers at the check-out counter will also vary considerably. The manager wishes to know when to open up a second check-out counter.

The basic set of queueing formulas are given in Eqs. (1)–(5), where  $\lambda$  = the arrival rate;  $\mu$  the service

$$\rho = \frac{\lambda}{\mu} \tag{1}$$

$$L = \frac{\lambda}{\mu - \lambda} \tag{2}$$

$$L_q = \frac{\rho^2}{1 - \rho} \tag{3}$$

$$W_q = \frac{\lambda}{\mu(\mu - \lambda)} \tag{4}$$

$$W = \frac{W_q + 1}{\mu} \tag{5}$$

rate;  $\rho$  = the traffic factor;  $L$  = the average number of customers in the system;  $L_q$  = the average number of customers in the check-out queue;  $W_q$  = the average waiting time for service; and  $W$  the average waiting time in the check-out system. If the arrival rate varies from a low value to one approaching the service rate, the effects on the system can be examined. These results are presented in the table. From this table, it is evident that the queue builds up very rapidly as the traffic factor increases beyond 0.75. The store manager would need to monitor the actual number of customers in the queue and open another check-out counter when this number exceeds 4 or 5.

The mathematics of the probability distributions of arrival rates and service times often define closed-form solutions, that is, being able to directly solve for an answer. In these cases another technique has proved very useful. See QUEUEING THEORY.

**Simulation.** Simulation is defined as the essence of reality without reality itself. With stochastic processes, values are simulated for the random variables from their known or assumed probability distributions, a simpler model is solved, and the process is

Analysis of a simple queueing system					
Arrival ( $\lambda$ )	Traffic factor ( $\rho$ )	Average number of customers in the system ( $L$ )	Average number of customers in check-out queue ( $L_q$ )	Average waiting time for service ( $W_q$ )	Average time in check-out system ( $W$ )
1	.05	.053	.0026	.0026	.0526
5	.25	.333	.0833	.0166	.0666
10	.50	1.000	.5000	.0500	.1000
15	.75	3.000	2.2500	.1500	.2000
16	.80	4.000	3.2000	.2000	.2500
17	.85	5.667	4.8161	.2833	.3333
18	.90	9.000	8.1000	.4500	.5000
19	.95	19.000	18.0500	.9500	1.0000

repeated a sufficient number of times to obtain statistical confidence in the results.

As a simple example of this, assume that someone makes the statement that the average female student at a university is of a certain height. To verify this assumption the actual heights of all female students could be measured, but this may be a long process. As an alternative, a hopefully random sample may be taken, the heights measured, the average determined, and an inference made as to the whole female population. See CHEMOMETRICS.

Many stochastic processes do not lend themselves to such a direct approach. Instead an assumption is made as to what the underlying probability distributions of the random variables are, and these are sampled. This sampling is done with random numbers. True random numbers are difficult to obtain, so pseudorandom numbers are generated by some mathematical relationship. This apparent paradox is

justified by the fact that the numbers appear to be random and pass most tests for randomness. As an example, a queueing system could be modeled by having two wheels of fortune—one with numbers in random order from 1 to 15, and the other from 1 to 20. The first could represent the number of arrivals into the system per 10-min period, and the other the number served during the same period. In this manner a system with a load factor of 0.75 could be simulated. Many simulation computer languages have been developed to ease the work of analyzing queueing systems. See SIMULATION.

**Markov processes.** This class of stochastic processes is characterized by a matrix of transition probabilities. These measure the probability of moving from one state to another. Such methods have been used to determine the reliability of a system, movement of a stock in the stock market, aging of doubtful accounts in credit analysis, and brand switching. This last application will be illustrated.

A company is considering marketing a new product. A preliminary market survey indicates that if a customer uses the product, there is a 60% probability that he or she will continue to do so. Likewise, the advertising campaign is such that 70% of those using all other products will be tempted to switch. The question is, what market share will the product be expected to finally obtain? The probability matrix is shown in Fig. 7.

		brand X	all others
now	brand X	.60	.40
	all others	.30	.70

Fig. 7. Probability matrix for brand-switching model.

Initially the market share of brand X is 0%, but after one period this will rise to 30%; that is, the advertising campaign will induce 30% to try brand X. In the next period, only 60% of these 30% will continue to purchase brand X, but 30% of the other 70% will switch, for a total market share of 39%.

Period by period, there will be transitions from state to state, but soon the variations will dampen out and a steady state will be achieved. This can be found for this problem by matrix value, and the final market share for brand X is 42.8%. See PROBABILITY.

**Decision trees.** While not originated by the field of operations research and not strictly in its domain, decision trees are an important tool in the analysis of some stochastic processes. More properly, they are a part of what may be considered statistical decision making. Their use can be illustrated with an example from the oil industry.

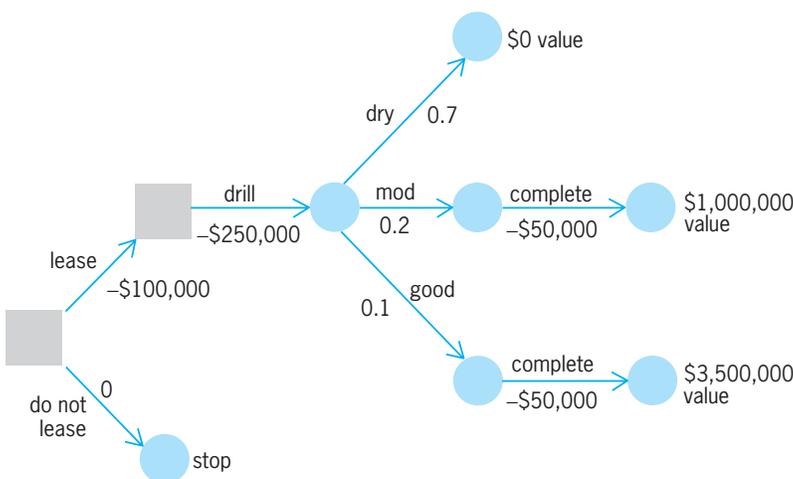


Fig. 8. Decision tree for an oil well problem.

An oil company is developing an oil field and has

the option to lease the mineral rights on an adjacent block of land for \$100,000. An exploratory well can be drilled at a cost of \$250,000. If the well is considered moderate or good, it will cost an additional \$50,000 to complete before production can start.

For simplicity, it can be assumed that the well, if good, has either moderate or good production. Also, the present worth of the net returns on all the future production is \$1,000,000 and \$3,500,000 for these two states. The problem facing the oil company is what they should do.

To analyze this problem, some subjective probabilities must be estimated. It is assumed that the probability that the well is dry is 70%, that the flow is moderate is 20%, and that it is good is 10%. With these percentages, a decision tree can be constructed (Fig. 8).

By a process of "folding back" the values and probabilities, the expected value at each decision point is seen to be represented by a square in Fig. 8. Actually, while there are three sequential decisions—lease, drill, and complete the well—once the decision to lease is made, the sequence is fixed unless the well is discovered to be dry. The expected value of this decision can be calculated as follows:

$$\begin{aligned} E[\text{lease} - \text{drill}] &= -100,000 - 250,000 + 0.70(0) \\ &\quad + 0.20(-50,000 + 1,000,000) \\ &\quad + 0.10(-50,000 + 3,500,000) \\ &= \$185,000 \end{aligned}$$

This expected value is slightly positive. In reality one of three things will happen—the well is dry, and the loss is \$350,000; the well is moderate, with a net gain of \$600,000; or the well is good, with a net gain of \$3,100,000. This expected value can be interpreted as follows: If a large number of wells are drilled, 70% will be dry at a loss of \$350,000 each, 20% will be moderate, and 10% will be good. The average gain will be \$185,000.

### Scope of Application

There are numerous areas where operations research has been applied. The following list is not intended to be all-inclusive, but is mainly to illustrate the scope of applications: optimal depreciation strategies; communication network design; computer network design; simulation of computer time-sharing systems; water resource project selection; demand forecasting; bidding models for offshore oil leases; production planning; assembly line balancing; job shop scheduling; optimal location of offshore drilling platforms; optimal allocation of crude oil using input-output models; classroom size mix to meet student demand; optimizing waste treatment plants; risk analysis in capital budgeting; electric utility fuel management; public utility rate determination; location of ambulances; optimal staffing of medical facilities; feedlot optimization; minimizing waste in the steel industry; optimal design of natural-gas pipelines; economic inventory levels; random jury selection; optimal marketing-price strategies; project management with CPM/PERT/GERT; air-traffic-control simu-

lations; optimal strategies in sports; system availability/reliability/maintainability; optimal testing plans for reliability; optimal space trajectories.

It can be seen from this list that there are few facets of society that do not have an application for operations research. See DECISION THEORY; GERT; PERT.

William G. Lesso

**Bibliography.** R. E. Bellman and S. E. Dreyfus, *Applied Dynamic Programming*, 1962; V. N. Bhat, *Elements of Applied Stochastic Processes*, 2d ed., 1984; F. Budnick, D. McLeavey, and R. Mojena, *Principles of Operations Research for Management*, 1988; B. Cornet and H. Tulkens (eds.), *Contributions to Operations Research and Econometrics*, 1989; F. S. Hillier and G. Lieberman, *Introduction to Operations Research*, 7th ed., 2000; J. W. Sutherland, *Towards a Strategic Management and Decision Technology*, 1989; H. A. Taha, *Operations Research*, 6th ed., 1996.

## Operator theory

At one level of abstraction an operator is simply a function whose arguments and values are real- (or complex-) valued functions of one or more real variables; in more naive terms an operator is a rule for converting such real- (or complex-) valued functions into others. The following are simple examples: (i) the operator which takes each differentiable real-valued function of one variable into its derivative; (ii) the operator which takes each twice-differentiable function  $f$  of one variable into expression (1); (iii) the operator which takes each

$$\left(\frac{df}{dx}\right)^2 + x^2 \frac{d^2f}{dx^2} \quad (1)$$

twice-differentiable function  $f$  of three variables into expression (2); and (iv) the operator which takes

$$\frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} + \frac{\partial^2 f}{\partial z^2} \quad (2)$$

the continuous function  $f$  of one real variable into the function  $g$  where relation (3) holds.

$$g(x) \equiv \int_0^1 \sqrt{x+y} f(y) dy \quad (3)$$

Since an operator is a function, the usual functional notation is applicable.  $L(f)$  may be used to denote the result of operating on  $f$  with the operator  $L$ . The set of all functions  $f$  for which  $L(f)$  is defined is called the domain of  $L$ , and the set of all functions  $g$  such that  $L(f) = g$  for some  $f$  in the domain of  $L$  is called the range of  $L$ . It is obvious that solving a differential or integral equation is equivalent (in many ways) to solving an operator equation  $L(f) = g$ , where  $g$  and  $L$  are given and it is required to find  $f$ . Moreover, the operator concept can be very useful both in theory and practice, producing a great variety of illuminating insights.

In large part the fruitfulness of the operator concept can be traced to two sources. One of these is

the possibility of adding and multiplying operators in such a way that many, though not all, of the laws of ordinary algebra hold. The other is the fact that the ranges and domains of operators behave in many respects like ordinary space and, indeed, may be regarded as contained in infinite dimensional generalizations of the familiar three-dimensional space of solid geometry. This makes it possible to think of an operator as a geometrical transformation and to exploit one's spatial intuition.

Let  $D$  be a family of real-valued functions such that  $\lambda f + \mu g$  is in  $D$  whenever  $\lambda$  and  $\mu$  are real numbers and  $f$  and  $g$  are in  $D$ . Let  $L$  and  $M$  be operators with domain  $D$  and range included in  $D$ . Then the operator which takes each  $f$  in  $D$  into  $L[M(f)]$  is called the product of  $L$  and  $M$  and is denoted by  $LM$ . Moreover, the operator which takes each  $f$  in  $D$  into  $L(f) + M(f)$  is called the sum of  $L$  and  $M$  and denoted by  $L + M$ . In particular, one may form powers  $L^2, L^3, \dots$ , polynomials  $a_0 + a_1L + a_2L^2 + \dots + a_nL^n$ , and in suitably restricted contexts, power series. It is important to note that, while it is always true that  $L + M = M + L$ , it is not always true that  $LM = ML$ . On the other hand, it is possible to show that  $(LM)N = L(MN)$  and that  $(L + M) + N = L + (M + N)$  for all  $L, M$ , and  $N$  so that parentheses may be omitted just as in ordinary algebra.

In many but not all cases the sets of functions with which one deals derive their spacelike properties from the possibility of assigning a distance  $\rho(f, g)$  to each pair  $f$  and  $g$  of members of the set  $D$  under consideration. This is done in such a manner that  $\rho(f, g) = \rho(g, f)$ ,  $\rho(f, g) > 0$  if  $f \neq g$ ;  $\rho(f, f) = 0$  and  $\rho(f, g) \leq \rho(f, h) + \rho(h, g)$  for all  $f, g$ , and  $h$  in  $D$ . When  $D$  is as described in the preceding paragraph,  $\rho$  is often chosen so that  $\rho(f, g) = \|f - g\|$  where  $\|f\| = \rho(f, 0)$ . If  $\|\lambda f\| = |\lambda| \|f\|$  for all real numbers  $\lambda$ , then  $\|f\|$  is said to be a norm for  $D$ . There will usually be more than one way of norming a given  $D$ . For example, if  $D$  is the set of all real-valued continuous functions defined on the interval  $0 \leq x \leq 1$ , Eq. (4) gives one value for  $D$ , and Eq. (5) gives another value for  $D$ .

$$\|f\| = \max_{0 \leq x \leq 1} |f(x)| \tag{4}$$

$$\|f\|_1 = \sqrt{\int_0^1 |f(x)|^2 dx} \tag{5}$$

The analogy with the familiar space of experience is closest when the second norm is used, but the first is useful also.

The operator  $L$  is said to be linear if  $L(\lambda f + \mu g)$  is defined and equal to  $\lambda L(f) + \mu L(g)$  whenever  $f$  and  $g$  are in the domain of  $L$ , and  $\lambda$  and  $\mu$  are numbers. Insofar as there is a general theory of operators, it is largely concerned with linear operators, and this article discusses linear operators exclusively. It is useful to develop this theory from axioms.

**Axioms.** Let  $F$  denote either the field of all real numbers or the field of all complex numbers. A vector space over  $F$  is a set or collection  $X$  whose members are of an unspecified character except that they

may be added together and multiplied by the members of  $F$  in such a way that the following formal laws are satisfied:

1.  $(f + g) + b = f + (g + b)$  and  $f + g = g + f$  for all  $f, g$ , and  $b$  in  $X$ .
2. There is a unique zero vector  $0$  in  $X$  such that  $f + 0 = f$  for all  $f$  in  $X$ .
3.  $\lambda(\mu f) = (\lambda\mu)f$ ,  $(\lambda + \mu)f = \lambda f + \mu f$ , and  $\lambda(f + g) = \lambda f + \lambda g$  for all  $f$  and  $g$  in  $X$  and all  $\lambda$  and  $\mu$  in  $F$ .
4.  $f = f$  for all  $f$  in  $X$ .

By generalizing the more special and concrete definition in the obvious fashion, a linear operator is defined to be a function  $L$  whose domain is a vector space  $X$ , whose range is in a vector space  $Y$ , and for which it is true that  $L(\lambda f + \mu g) = \lambda L(f) + \mu L(g)$  whenever  $f$  and  $g$  are in  $X$  and  $\lambda$  and  $\mu$  are in  $F$ .

**Finite-dimensional case.** A vector space  $X$  is said to be finite dimensional if it contains a finite subset  $v_1, v_2, \dots, v_n$  spanning the space in the sense that every element in the space may be written in the form  $\lambda v_1 + \lambda_2 v_2 + \dots + \lambda_n v_n$  where the  $\lambda_j$  are in  $F$ . The representation  $f = \lambda_1 v_1 + \lambda_2 v_2 + \dots + \lambda_n v_n$  is unique if, and only if, no  $v_j$  is in the span of the rest. In this case  $v_1, v_2, \dots, v_n$  is said to form a basis for  $X$ , and the  $\lambda_j$  are said to be the coordinates of  $f$  with respect to this basis. It is not hard to show that any two bases for the same space have the same number of elements. This number is called the dimension of the space.

Let  $L$  be a linear operator whose domain  $X$  is finite dimensional. It follows immediately that the range is also finite dimensional and that the dimension  $d_R$  of the range is less than, or equal to, the dimension  $d_X$  of the domain. Let  $Y$  be the vector space containing the range of  $L$ , and let  $d_Y$  denote the dimension of  $Y$ . This gives  $d_R \leq d_Y$  and  $d_R \leq d_X$ . The differences  $d_Y - d_R$  and  $d_X - d_R$  measure the extent to which the operator equation  $L(f) = g$  fails to have a unique solution for all  $g$ . In fact if  $X_1$  and  $X_2$  are finite-dimensional sub-vector spaces of a vector space and  $X_1 \supseteq X_2$ , then  $X_1 = X_2$  if, and only if,  $X_1$  and  $X_2$  have the same dimension. Thus  $L(f) = g$  is always solvable if, and only if,  $d_Y = d_R$  and  $d_Y - d_R$  is a measure of the size of the set of  $g$ 's for which no solution exists. On the other hand, it is easily seen that if  $f_0$  is a particular solution of  $L(f) = g$ , then the general solution is  $f_0 + b$  where  $b$  is any element of  $X$  such that  $L(b) = 0$ . The set of all such  $b$  is a vector space  $N$ , called the null space of  $L$ , whose dimension  $d_N$  measures the extent to which the equation  $L(f) = g$  has multiple solutions. It is not hard to show that  $d_N + d_R = d_X$  so that  $d_N = d_X - d_R$ . In the special but important case in which  $X = Y$ ,  $d_X - d_R = d_Y - d_R$ . Thus either  $L(f) = g$  has a unique solution for all  $g$  (nonsingular case) or else for many values of  $g$ ,  $L(f) = g$  has no solutions, and whenever it has any nonzero solutions, it has many (singular case).

There is a certain sense in which most operators are nonsingular. Let  $I$  denote the identity operator which takes every vector into itself. Then it can be shown that  $L - \lambda I$  is nonsingular for all but a finite number of values of  $\lambda$ . Indeed  $L - \lambda I$  will be singular if, and only if, there exists a nonzero vector  $f$  such that

$L(f) - \lambda f = 0$ ; that is,  $L(f) = \lambda f$ . Such an  $f$  is called a proper vector (or eigenvector) belonging to the proper value (or eigenvalue)  $\lambda$ . It is easy to show that proper vectors belonging to distinct proper values are linearly independent in the sense that no one is in the span of the rest. Thus the number of distinct proper values cannot exceed the dimension of the space. See DIFFERENTIAL EQUATION.

Knowledge of the proper values of an operator  $L$  yields a great deal of information about the nature of  $L$ , especially in the important case in which the domain  $X$  admits a basis made up of proper vectors. Let  $v_1, v_2, \dots, v_n$  be a basis for  $X$ , and let  $L(v_j) = \lambda_j v_j$  where the  $\lambda_j$  are (not necessarily distinct) members of  $F$ . Very simple computations lead to the following observations.

1. Every proper value of  $L$  is equal to some  $\lambda_j$ .
2.  $L$  is nonsingular if, and only if, no  $\lambda_j$  is zero.
3. If  $L$  is nonsingular, the inverse operator carries  $\mu_1 v_1 + \mu_2 v_2 + \dots + \mu_n v_n$  into  $(\mu_1/\lambda_1)v_1 + (\mu_2/\lambda_2)v_2 + \dots + (\mu_n/\lambda_n)v_n$ .
4. If  $P$  is any polynomial with coefficients in  $F$ , then  $P(L)$  carries  $\mu_1 v_1 + \mu_2 v_2 + \dots + \mu_n v_n$  into  $P(\mu_1)v_1 + P(\mu_2)v_2 + \dots + P(\mu_n)v_n$ .

The structure of  $P(L)$  revealed by observation (4) suggests a definition of  $F(L)$  where  $F$ , instead of being a polynomial, is an arbitrary function with domain and range in  $F$ . To wit:  $F(L)(\mu_1 v_1 + \mu_2 v_2 + \dots + \mu_n v_n) = F(\mu_1)v_1 + F(\mu_2)v_2 + \dots + F(\mu_n)v_n$ . A similar, but of course more subtle, definition in certain infinite-dimensional cases is the source of the modern rigorization of the celebrated operational calculus of O. Heaviside.

Returning to the case in which  $X$  need not equal  $Y$ , let  $v_1, v_2, \dots, v_n$  and  $w_1, w_2, \dots, w_n$  be bases for  $X$  and  $Y$ , respectively. Let  $L(v_j) = \alpha_{j1}w_1 + \alpha_{j2}w_2 + \dots + \alpha_{jn}w_n$  where each  $\alpha_{ji}$  is in  $F$ . Then  $L(x_1 v_1 + \dots + x_n v_n) = x_1 L(v_1) + x_2 L(v_2) + \dots + x_n L(v_n) = y_1 w_1 + y_2 w_2 + \dots + y_n w_n$  where  $y_i = a_{1i}x_1 + a_{2i}x_2 + \dots + a_{ni}x_n$ . The rectangular array in which  $\alpha_{ji}$  is in the  $i$ th row and  $j$ th column is called the matrix of  $L$  with respect to the basis in question. It is clear that solving the operator equation  $L(f) = g$  in the finite-dimensional case is equivalent to solving  $m$  linear algebraic equations in  $n$  unknowns. The theory sketched above is the theory of such equations couched in the language of operator theory. When  $X = Y$  and  $v_j = w_j$  for all  $j$ , then the condition  $\alpha_{ji} = \bar{a}_{ij}$  (where the overbar denotes complex conjugate) implies that there exists a basis for  $X$  made up of proper vectors of  $L$ . However, this condition is by no means a necessary one.

**Infinite-dimensional case.** When  $X$  is not assumed to be finite dimensional, such simple and general theorems as those described above are no longer available. In certain contexts, however, more complicated and less complete analogs of them may be proved. It is with these that the general theory of linear operators is mainly concerned.

Let  $X$  be a vector space which is not necessarily finite dimensional but instead is equipped with a norm—defined in the abstract case as suggested by the definition given above for real function spaces.

Such a normed vector space is said to be complete if each sequence  $f_1, f_2, \dots$  of members of  $X$  which is convergent in the sense defined by Eq. (6) is also convergent in the sense defined by Eq. (7) for some

$$\lim_{\substack{n \rightarrow \infty \\ m \rightarrow \infty}} \|f_n - f_m\| = 0 \tag{6}$$

$$\lim_{n \rightarrow \infty} \|f_n - f\| = 0 \tag{7}$$

$f$  in  $X$ . This  $f$  is easily seen to be unique and is called the limit of the sequence  $f_n$ . A complete normed vector space is called a Banach space. A normed vector space  $X$  is said to be separable if there exists a sequence  $f_1, f_2, \dots, f_n$  of elements of  $X$  such that every element of  $X$  is the limit of some subsequence. The space of continuous functions defined earlier is a separable Banach space. Now let  $X = Y$  and let  $X$  be a Banach space. Let  $L$  be completely continuous in the sense that whenever  $f_1, f_2, \dots$  is a sequence of elements such that  $\|f_n\| \leq 1$  for all  $n$ , there is a subsequence  $f_{n_1}, f_{n_2}, \dots$  such that  $L(f_{n_1}), L(f_{n_2}), \dots$  is convergent in the sense defined in Eq. (6). Then the following theorem can be proved. For each  $\lambda$  in  $F$  with  $\lambda \neq 0$  there is a pair of vector subspaces  $M_\lambda$  and  $N_\lambda$  of  $X$  such that (1)  $L(f)$  is in  $M_\lambda$  for all  $f$  in  $M_\lambda$ , and  $L(f)$  is in  $N_\lambda$  for all  $f$  in  $N_\lambda$ ; (2) every  $f$  in  $X$  can be written uniquely in the form  $f_1 + f_2$  where  $f_1 \in M_\lambda$  and  $f_2 \in N_\lambda$ ; (3) for each  $g_1 \in M_\lambda$  there is one and only one element  $f_1 \in M_\lambda$  such that  $L(f_1) - \lambda f_1 = g_1$ ; and (4)  $N_\lambda$  is finite dimensional. It follows easily that existence and uniqueness questions for the operator equation  $L(f) - \lambda f = g$  reduce to the corresponding questions for the restriction of  $L$  to the finite-dimensional space  $N_\lambda$  and, hence, that the simple analysis given earlier applies. It may be proved further that  $M_\lambda$  coincides with  $X$  for all values of  $\lambda$  except those in a sequence  $\lambda_1, \lambda_2, \dots$  such that Eq. (8) applies.

$$\lim_{n \rightarrow \infty} \lambda_n = 0 \tag{8}$$

Let  $K$  be a continuous real-valued function defined on the unit square  $0 \leq x \leq 1, 0 \leq y \leq 1$ . Let  $X$  be the Banach space of all continuous real-valued functions defined on the interval  $0 \leq x \leq 1$  with Eq. (9)

$$\|f\| = \max_{0 \leq x \leq 1} |f(x)| \tag{9}$$

applying. Then it can be proved that the operator  $L_K$  which takes  $f$  into  $g$ , where Eq. (10) holds, is

$$g(x) = \int_0^1 K(x, y)f(y) dy \tag{10}$$

a completely continuous linear operator. Application of the theorems quoted above yields most of the results of the Fredholm theory of integral equations. Such integral operators occur in inverting the members of a large class of linear differential operators. There are similar results for integral operators in  $2n$  variables, it being possible to establish complete continuity whenever the region of integration is bounded.

Infinite-dimensional versions of theorems about bases of proper vectors and functions of operators take their simplest and most complete form when the underlying Banach space is a Hilbert space; that is, when there is defined an  $F$ -valued "inner product"  $f \cdot g$  for each  $f$  and  $g$  in  $X$  which satisfies the following conditions: (1)  $(\lambda f + \mu g) \cdot b = \lambda(f \cdot b) + \mu(g \cdot b)$ ; (2)  $(f \cdot g) = (g \cdot f)$ ; and (3)  $(f \cdot f) = \|f\|^2$  for all  $f, g$ , and  $b$  in  $X$  and all  $\lambda$  and  $\mu$  in  $F$ . The simplest (and original) example of a Hilbert space is the vector space of all sequences  $c_1, c_2, \dots$  of complex numbers such that  $|c_1|^2 + |c_2|^2 + \dots < \infty$ , based on the definition  $(c_1, c_2, \dots) \cdot (c'_1, c'_2, \dots) = c_1 \bar{c}'_1 + c_2 \bar{c}'_2 + \dots$ .

Now let  $L$  be a completely continuous linear operator, the domain of which is a separable Hilbert space  $H$  and whose range is contained in  $H$ . Let  $L$  be self-adjoint in the sense that  $L(f) \cdot g = f \cdot L(g)$  for all  $f$  and  $g$  in  $H$ . Then it is a theorem that there exists a sequence  $v_1, v_2, \dots$  of members of  $H$  which has the following properties: (1) Each  $v_j$  is a proper vector for  $L$ ; (2) if  $f$  is any element in  $H$  and  $c_j = f \cdot v_j$ , then  $f = c_1 v_1 + c_2 v_2 + \dots$  in the sense that the partial sums of this series have  $f$  as a limit; and (3)  $v_i \cdot v_j = 0$  if  $i \neq j$  and  $v_i \cdot v_i = 1$ . Such a sequence is said to be an orthonormal basis for  $H$ . Just as in the finite-dimensional case, one can form more or less arbitrary functions of the operator  $L$ . In rough terms the celebrated spectral theorem is a generalization of the preceding theorem in which the operator  $L$ , though self-adjoint, is not required to be completely continuous. Instead of a discrete basis of proper vectors, one finds a sort of continuous basis. More precisely it is possible to map  $H$  onto a Hilbert space  $H'$  whose elements are complex-valued functions in such a manner that the norms and the vector space operations are preserved and so that  $L$  becomes the operator of multiplying the elements of  $H'$  by a fixed real-valued function.

In addition to the abstract theory there are many detailed studies of particular operators of importance, such as the Laplace and Fourier transforms, and numerous applications to differential and integral equations. Moreover, in addition to the theory of single operators there are extensive theories of certain kinds of collections of operators. A ring of operators contains the sum, product, and difference of any two of its members, and a group of operators contains the inverse of every operator in it and the product of each two operators in it. The theory of groups and rings of operators is related to the algebraic theory of groups and rings much as the theory of a single operator is related to systems of linear algebraic equations. It has applications to harmonic analysis and to the conceptual foundations of quantum mechanics. See COMPLEX NUMBERS AND COMPLEX VARIABLES; GROUP THEORY; INTEGRAL TRANSFORM; REAL VARIABLE; RING THEORY; SET THEORY; TOPOLOGY.

George W. Mackey

Bibliography. N. I. Akheizer and I. M. Glazman, *Theory of Linear Operators in Hilbert Space*, 1961, reprint 1983; J. B. Conway, *A Course in Functional Analysis*, 1994; N. Dunford and J. T. Schwartz, *Lin-*

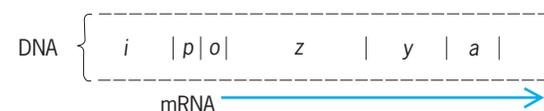
*ear Operators*, 3 vols., 1958–1971, reprint 1988; L. Hoermander, *The Analysis of Linear Partial Differential Operators*, 2d ed., 1990; A. Taylor and D. Lay, *Introduction to Functional Analysis*, 2d ed., 1980, reprint 1986; T. Yoshino, *Introduction to Operator Theory*, 1994.

## Operon

A group of distinct genes that are expressed and regulated as a unit. Each operon is a deoxyribonucleic acid (DNA) sequence that contains at least two regulatory sites, the promoter and the operator, and the structural genes that code for specific proteins (see **illustration**). The promoter ( $p$ ) site is the location at which ribonucleic acid (RNA) polymerase binds to the operon. RNA polymerase moves down the operon catalyzing the synthesis of a messenger RNA (mRNA) molecule with a sequence that is complementary to DNA. This process is called transcription. The mRNA is used as a template by ribosomes to synthesize the proteins coded for by the structural genes (in the original DNA) in a process called translation. This mRNA is referred to as polycistronic because its sequence directs the synthesis of more than one protein. The operator ( $o$ ) site is located between the  $p$  site and the beginning of the coding region for the first structural gene. It is at this site that molecules called repressors can bind to the DNA and block RNA polymerase from transcribing the DNA, thus shutting off the operon. Some systems can be derepressed by the addition of small molecules called effectors, which bind to the repressor protein and cause a conformational (shape) change that makes it no longer able to bind to the DNA at the operator site. See DEOXYRIBONUCLEIC ACID (DNA); RIBONUCLEIC ACID (RNA).

**Promoters.** Promoters are located just before the start of the first gene in an operon, and have two conserved sequences separated by 16–19 base pairs. The  $-10$  region has the consensus sequence of TATAAT (T = thymine; A = adenine). Although most promoters differ to some degree from the consensus sequence, little variability in the separation of two regions, is tolerated. The reason for this sensitivity is DNA's helical structure, where binding sites spiral around the DNA molecule. The required separation of the  $-10$  and  $-35$  regions determines whether they are on the same side of DNA. See MOLECULAR BIOLOGY.

**Repressors.** There are two types of repressors that block or repress the production of a protein. The



The lactose (*lac*) operon from *Escherichia coli*:  $z$ ,  $y$ , and  $a$  are structural genes;  $i$  is the *lac* repressor gene;  $p$  is the promoter site; and  $o$  is the operator site. The arrow indicates length and direction of mRNA synthesis.

better understood are the transcriptional repressors. These proteins bind to the *o* site of the operon. They repress transcription of the gene through blocking binding of RNA polymerase at the promoter or by blocking the movement of RNA polymerase through the gene. In the simplest case, this blockage can be direct by having the operator overlap the promoter. This blockage can also be indirect through the formation of loops of DNA which hide the promoter. Such is the case with the C protein in the arabinose operon from *Escherichia coli*. When the levels of C protein is high, it binds to two sites that are 160 base pairs apart. Portions of the C protein self-associate, forming a loop that inhibits transcription. In this manner, sites that are quite distant from the promoter and the first gene in the operon can influence transcription.

A second type of repression is translational. In order for mRNA to be translated, ribosomes must first bind to the message and then read through the codes adding amino acids to the growing protein chain. In translational repression the repressor binds to mRNA so that the ribosome binding site is blocked.

**Activators.** At a number of locations within the *E. coli* genome, a regulator called the catabolite gene activator protein acts as a repressor, but there are other sites at which it stimulates the expression of genes within an operon.

Activation is believed to arise from the binding of a protein immediately adjacent to the promoter. The protein provides additional locations with which RNA polymerase can interact; the extra interactions result in an increased amount of polymerase binding to the promoter. Activators are more frequently involved in the regulation of genes in eukaryotes than in prokaryotes.

**Antiterminators.** Once RNA polymerase begins transcribing a gene, it continues making RNA until a termination site is reached. Antiterminators are proteins that prevent termination at certain sites either by preventing mRNA from forming a hairpin or by preventing the mRNA from falling off if a hairpin is formed. In the presence of these antiterminators, RNA polymerase continues along the genome and transcribes the genes following the termination site until a different class of termination site is encountered.

**Attenuation.** Attenuation is the premature termination of the mRNA translation. Although the exact mechanism of attenuation has not been unambiguously determined, it is thought that attenuation is due to the formation of a translation termination site in mRNA. As soon as an mRNA strand is synthesized, while it is still associated with DNA, ribosomes begin to bind and produce protein. If the level of tryptophan in the cell is high, the ribosome speeds through this region; the mRNA folds as shown and produces a translation termination site. Since no mRNA for the remaining structural genes is made, the corresponding proteins are not synthesized. If the level of tryptophan is low, the ribosome slows when it comes to that region of mRNA coding for tryptophan. This is believed to give the mRNA time to refold into a structure which has no termination site.

**Antirepressor.** A protein called an antirepressor does not appear to bind DNA but rather repressors. The production of small amounts of this protein is sufficient to bind a substantial amount of repressors. Transcription is activated as the repressor is removed.

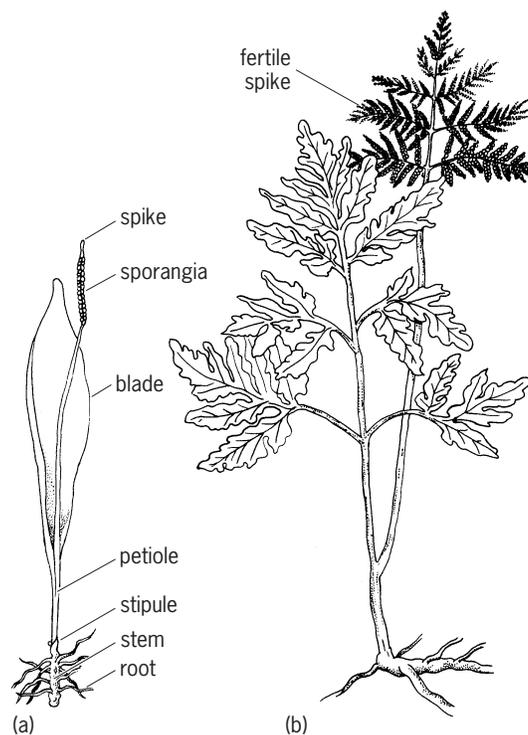
**Methylation of DNA.** The encoding of a protein that adds a methyl group to the adenine in a particular genetic sequence can occur. When the GATC sequence is found in the promoter of a gene, such methylation might be expected to influence transcription of that gene. In eukaryotic cells, approximately 5% of the cytosines in DNA are methylated. During some viral infections, the expression of the viral genes appears to be repressed by methylation of the viral DNA. It is possible that during development from egg to embryo the expression of some genes may be regulated through methylation of the DNA. The mechanism that causes the level of methylation of the DNA to change is not known. See GENE; GENE ACTION; IMMUNOLOGY.

Douglas H. Ohlendorf

Bibliography. B. Lewin, *Genes*, 1994; J. Miller and W. Reznikoff (eds.), *The Operon*, 1978.

## Ophioglossales

An order of the class Polypodiopsida known as the adder's-tongue ferns. It is a small group with only 3 genera and about 80 species. Two genera, *Ophioglossum* and *Botrychium* (see **illustration**), are widely distributed in tropical and temperate regions and



Ophioglossales. (a) Adder's-tongue fern (*Ophioglossum*) with a single leaf. (b) Grape fern (*Botrychium*), showing the division of the leaf into a sterile foliage part and a fertile spike. (After H. J. Fuller and O. Tippo, *College Botany*, Holt, rev. ed., 1954)

have about the same number of species; the third genus, *Helminthostachys*, is represented by a single species confined to southeastern Asia and Polynesia. These are considered the most primitive of the present-day ferns. No fossils have been reported for this group. The plants are homosporous and eusporangiate; that is, spore sacs develop from groups of epidermal cells. Each sporangium produces a large number of spores, as many as 15,000 in some species. The chromosome number is high in the species of *Ophioglossum*. *Ophioglossum petiolatum*, a tropical species, has a chromosome count of over 1000, the largest number observed in a naturally occurring species of vascular plants. This group is distinguished from other ferns by the arrangement of the sporogenous tissue in the characteristic fertile spike (illus. *a*) of the sporophyte (spore-producing generation). The leaves are erect or merely bent over in bud and not circinate. The gametophyte (gamete-producing generation) is a small, nongreen, fleshy, subterranean saprophyte, associated with an endophytic fungus. The group appears to be an evolutionary dead end. See FUNGI; LEAF; POLYPODIALES; POLYPODIOPSIDA; PTEROPSIDA. Paul A. Vestal

## Ophiolite

A distinctive assemblage of mafic plus ultramafic rocks generally considered to be fragments of the oceanic lithosphere that have been tectonically emplaced onto continental margins and island arcs. Ophiolite was named by A. Brongniart, a nineteenth-century French naturalist, who considered its scaly appearance and the greenish color of its main constituent rock, serpentinite. An ophiolite is a formation made up of an association of typical rocks in a clearly defined sequence. As shown in **Fig. 1**, a complete idealized ophiolite sequence from bottom to top includes (1) an ultramafic tectonite complex composed mostly of multilayered, deformed harzburgite, dunite, and minor chromitite; (2) a plutonic complex of layered mafic-ultramafic cumulates at the base, grading upward to massive gabbro, diorite, and possibly plagiogranite; (3) a mafic sheeted-dike complex; (4) an extrusive section of massive and pillow lavas, pillow breccias, and intercalated pelagic sediments; and (5) a top layer of abyssal or bathyal sediments, which may include ribbon chert, red pelagic limestone, metalliferous sediments, volcanic breccias, or pyroclastic deposits. Most ophiolites lack complete sections, and are dismembered and fragmented. Their estimated original thickness is variable, ranging from about 2 km (1.2 mi) to more than 8 km (5 mi). See EARTH CRUST; LITHOSPHERE.

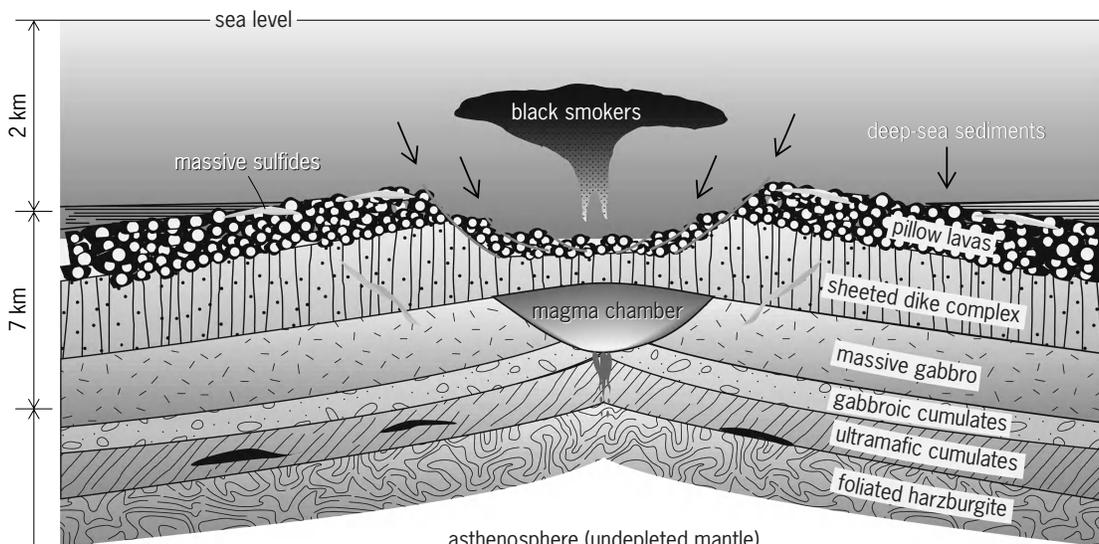
**Occurrence.** Ophiolites typically occur in collisional mountain belts or island arcs and define a suture zone marking the boundary where two plates have welded together. The ophiolite complex is interpreted as evidence for a closed marginal ocean or back-arc basin. Typical examples of such sutures are found along the northern flank of the Himalayas (the Indus suture) and in the central Urals; both extend

for more than 1000 km (600 mi). The occurrence of a suite of deep-sea sediments, pillow basalts, gabbros, and serpentinitized ultramafic rocks within these sutures suggests that they constituted oceanic lithosphere that was subsequently thrust onto the continental margins by a process known as obduction.

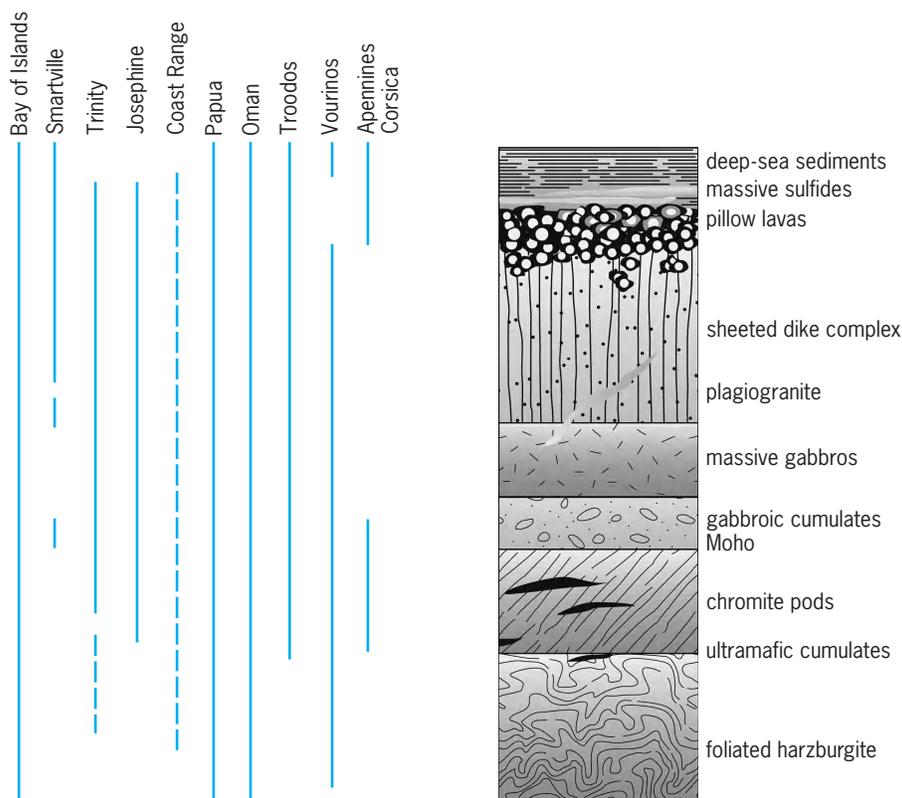
According to plate tectonics, the destiny of the ocean floor is subduction into the mantle. In contrast, geologists have proposed the term "obduction" to describe the particular destiny of ophiolites that, during the advance of oceanic lithosphere against a continent, end up stranded on the edge of a continent instead of disappearing into the subduction zone below it. See PLATE TECTONICS; SUBDUCTION ZONES.

Throughout the world, ophiolites occur as long narrow belts, up to 10 km (6 mi) wide, that can extend more than 1000 km (600 mi) in length, in two distinct geographic settings. (1) Those in the Alpine-Mediterranean region, Tethyan ophiolite, were formed in small ocean basins that were surrounded by older, attenuated continental crust. Many of the classical ophiolites found in Cyprus and Oman belong to this group, which have a nearly complete sequence and were brought above sea level during the collapse of small ocean basins by the convergence of neighboring plates. (2) Those in western North America and the Circum-Pacific (Cordilleran) region seem to have formed in inter-arc basins. The Cordilleran ophiolites, such as the Trinity ophiolite and the Coast Range ophiolite of California, are generally incomplete, metamorphosed, or dismembered, but they commonly form the basement rocks for many North American continental margin terranes. Tethyan ophiolites are characterized by the occurrence of harzburgitic ultramafic rocks and pronounced thick layers of gabbroic rock, whereas Cordilleran ophiolite, such as the Trinity, have undepleted lherzolites and thin layers of recrystallized and deformed gabbros. Such differences have been attributed to the rate of spreading, with fast spreading accounting for more partial melting of the primary mantle rocks, and hence the thicker mantle residue as harzburgite and greater differentiation of gabbroic magma for the mafic-ultramafic rocks of the Oman ophiolite. See BASIN; CONTINENTAL MARGIN; GEODYNAMICS; STRUCTURAL GEOLOGY.

These two types of ophiolites strongly resemble the oceanic lithosphere insofar as the latter is known from dredging, shallow drilling, and geophysical studies. However, the crustal section of most ophiolites is significantly different from the abyssal oceanic crust of the Atlantic and Pacific oceans. Many ophiolitic basalts possess chemical affinities of back-arc basin basalts or island-arc basalts, rather than those of mid-oceanic ridge basalt (MORB). For example, the potassium oxide (K<sub>2</sub>O) content in MORB is usually less than 0.2 wt %, whereas it is higher than 0.2 wt % in back-arc basin basalts and the extrusive rocks of the ophiolite sequence. Many ophiolites carry a weak-to-strong subduction zone geochemical imprint with relative depletions in tantalum and niobium in normalized incompatible-element



(a)



(b)

**Fig. 1.** Idealized cross section of (a) an oceanic ridge showing stratigraphy of the intrusive and extrusive sequences, and circulation of heated seawater at the top of the oceanic crust, and (b) an idealized ophiolite succession compared with various exposed ophiolites.

distribution diagrams compared to MORB. The sedimentary sequence overlying most ophiolites grades from thin layers of pelagic or cherty sediments upward to turbidite or calc-alkaline volcanoclastics, suggesting that the oceanic crust of most ophiolites formed in back-arc basins. This is also supported by the fact that the crustal sequence of most ophiolites is thin (<5 km; 3 mi) in comparison with

average oceanic crust developed in large ocean basins (7–10 km; 4–6 mi). See BASALT; MID-OCEANIC RIDGE; OCEANIC ISLANDS.

**Origins.** Many ophiolites are thought to have formed in submarine extensional tectonic settings with extremely high heat flow, such as the present-day East Pacific Rise and Mid-Atlantic Ridge, where new oceanic crust is being generated. The analogy of

the ophiolite sequence with the oceanic lithosphere (Fig. 1) is supported by the gross similarity in chemistry, metamorphic grades corresponding to temperature gradients existing under spreading centers, the presence of similar ore minerals, and the occurrence of deep-sea sediments. However, in recent years this simple analogy has been challenged. It has been suggested that ophiolites do not represent the typical oceanic lithosphere and do not belong to a unique species; instead they have formed in other extensional regimes, including island-arc or marginal-basin settings, above a subduction zone. Ophiolites are derived from a variety of oceanic sites on which new lithosphere is formed. Despite this ongoing controversy, the general mechanism by which a complete ophiolite succession forms is reasonably well understood and agreed upon.

In extensional oceanic environments, heat flow is high and the mantle asthenosphere is rising. As the pressure decreases in the rising asthenosphere beneath ocean ridges or back-arc basins, undepleted mantle lherzolites partially melt to form basaltic magma according to the general reaction, lherzolite  $\rightarrow$  harzburgite (75%) + basaltic magma (25%). These magmas collect in a chamber (or chambers) at depths about 4–6 km (2.5–4 mi), and 25 km (16 mi) long, and undergo fractional crystallization. Some magma rises as dikes, forming a sheeted-dike complex, and are extruded as pillow lavas in the axial rift on the sea floor. The remaining magma within the chamber fractionates upon cooling to form the layered and massive rocks of the plutonic sequence. Fractional crystallization of the magma gives rise locally to diorite and plagiogranite. The petrologic and chemical data indicate that the lavas, dikes, gabbros, and underlying tectonized harzburgite are all cogenetic; the harzburgite represents a crystalline residue from partial melting in the mantle that produced the overlying igneous rocks. *See* ASTHENOSPHERE; IGNEOUS ROCKS; MAGMA; PLUTON.

**Ages of formation and emplacement.** At the base of most ophiolites, a thin metamorphic sole occurs and records the travel history of the obduction of the oceanic lithosphere on land. The sole ranges in thickness from 10 to 50 m (33 to 165 ft) and can extend laterally for more than 100 km (60 mi). It is composed of highly deformed amphibolites and metasedimentary greenschists, and shows a sharp decrease in metamorphic grade from top to bottom. When young and hot oceanic lithosphere thrusts upon the oceanic crust near a continent margin, it heats the underlying crustal rocks like a flat iron. These rocks start to recrystallize while being deformed under the moving load of the overlying lithosphere, and are subjected to intense metamorphism. During its oceanic travels at a rate of several centimeters per year, the ophiolite nappe loses heat; the temperature at its base will have dropped and the underlying crust will transform into low-temperature metamorphic rocks, such as greenschist. The metamorphic soles formed, apparently, by successive underplating and welding onto the base of an ophiolite as the hot, young oceanic slab migrated toward the continent.

Dating the metamorphic soles' thermally recrystallized minerals provides ages for ophiolite emplacement onto the continental margins. *See* METAMORPHIC ROCKS.

Ages of ophiolite formation can be obtained by direct radioactive lead (U/Pb) dating of zircon from diorite or plagiogranites of the plutonic sequence of an ophiolite. They can also be constrained by determining the ages of radiolarian fossils in pelagic chert overlying these ophiolites. Ophiolites are rather abundant in Phanerozoic orogenic belts; however, several Precambrian ophiolites with ages ranging from 600 million years to about 2 billion years have also been described. Age gaps between the deposition of pelagic sediments and ophiolite emplacement are probably less than 25 million years. Such a short duration is consistent with the life span of a back-arc basin (less than 20 million years), and indicates that obduction of many ophiolites occurred soon after their creation. Ophiolites consequently represent young oceanic lithosphere that was detached while still hot. *See* LEAD ISOTOPES (GEOCHEMISTRY); ROCK AGE DETERMINATION; ZIRCON.

**Hydrothermal alteration and formation of massive sulfide deposits.** Immediately after new oceanic crust forms, seawater percolates downward through fractures and faults from the flanks of the rifted valley. It easily penetrates the crust down to 2–3 km (1–2 mi) at the base of the dike complex. Heated to 400–450°C (752–842°F), the water circulates toward the ridge and starts to ascend, becoming progressively channeled, and finally discharges along the rift axes at temperatures up to 380°C (716°F). During this high-temperature circulation, the hydrothermal solution alters and corrodes the crustal rocks, dissolving metals. When this discharged hydrothermal solution, consisting of metal sulfides leached from the oceanic crust, meets the cold seawater on the ocean floor, a black cloud of metallic sulfides results. Such hot springs are known as black smokers. Subsequent precipitation at, around, and beneath these chimneys can produce substantial amounts of massive sulfide ores in the upper parts of the extrusive sequence (Fig. 1). Such processes for precipitation of massive sulfide deposits were observed on the present-day ocean floor at locations along Pacific, Atlantic, and Indian ocean ridges. *See* HYDROTHERMAL VENT.

**Figure 2** shows a black smoker spewing dark, mineral-rich fluids as observed by scientists during a dive of the deep-sea submersible *Alvin* on the East Pacific Rise (latitude 21°N) in 1979. The dissolved minerals and the heat given off by the black smoker favor the prolific growth of chemoautotrophic bacteria which, in turn, provide nourishment for the surrounding colonies. Around the hot springs, abundant and unusual sea life—giant tubeworms, huge clams, and mussels—has been discovered. *See* MARINE MICROBIOLOGY.

In the upper parts of the extrusive sequences of many ophiolites, thin lenses (up to 50 m or 165 ft thick and ~500 m or 1650 ft across) of sulfide ore occur in some fault-controlled depressions. The



Fig. 2. View of the first high-temperature geothermal vent (380°C or 716°F) ever seen by scientists during a dive of the deep-sea submersible *Alvin* on the East Pacific Rise in 1979. (Photograph by Dudley Foster, RISE Expedition, courtesy of W. R. Normark, U.S. Geological Survey)

ore lenses consist of massive sulfide minerals, including pyrite ( $\text{FeS}_2$ ), chalcopyrite ( $\text{CuFeS}_2$ ), sphalerite ( $\text{ZnS}$ ), pyrrhotite ( $\text{FeS}$ ), and minor molybdenite ( $\text{MoS}_2$ ), and have been mined as major copper deposits. Such copper-zinc deposits are presently exploited in the basalts of ophiolite sequences. The best-known copper deposit is located on Cyprus and is associated with the genesis of the Troodos ophiolites, and it was known to the ancient Greeks, who named the island after the metal (*cupros*, or copper). These Cyprus deposits contain a total reserve of 1.5 megatons of copper. A similar, contemporaneous mineralization of the Kuroko-type black-ore deposits in Japan was generated along embryonic oceanic ridges, which developed in a back-arc setting. See COPPER.

The circulation of heated seawater through cooling, fractured crustal sections also causes significant alteration of the primary minerals. Formation of secondary minerals, including clay minerals and zeolites in basaltic rocks, is selective and incomplete and increases the volatile content of oceanic metabasalts. Hydrothermal metamorphism at depths results in the formation of epidote, chlorite, and amphiboles in the gabbroic sequence, and minor serpentinization of the ultramafic rocks. See SERPENTINITE.

**Summary.** Ophiolites represent new oceanic crust formed in a variety of spreading environments, including oceanic ridge, back-arc basin, and island arcs above a subduction zone, and subsequently emplaced onto the continents. Their occurrence along plate sutures marks the sites of ancient tectonic interaction between oceanic and continental crust. Ophiolites can form in multitude of tectonic settings, and the process of new ocean crust generation in spreading centers can produce different magma types. Depending on the rate of influx of ascending asthenosphere and the spreading rate of the oceanic lithosphere, ophiolites may not consist of plutonic and extrusive sequences resulting from a single magmatic pulse, but several, evidenced by the many in-

trusive relations. For example, the plagiogranites of the best-preserved Oman ophiolite are the product of fractional crystallization of the second magmatic event. Similarly, many Tethyan ophiolites, including the one in Oman, are characterized by thick harzburgites and well-layered gabbros resulting from fast spreading and large magma chambers. Many Circum-Pacific ophiolites, such as the Trinity ophiolite of California, have relatively thin Iherzolitic mantle sequences, which are dismembered and recrystallized as a result of slow spreading at their ridges. In spite of extensive research efforts, it is still difficult to establish what kind of oceanic crust some ophiolites represent, what the actual processes were that formed ophiolites in spreading systems, how particular ophiolites formed in various tectonic settings compared with the oceanic lithosphere, and how they were emplaced onto the continental margins. Nevertheless, ophiolites provide the best opportunity for geologists to walk across the ocean floor on land; they also offer vertical sections in addition to horizontal distributions. Moreover, ophiolite formations record the ages of oceanic fragments that escaped disappearing into subduction zones.

Juhn Liou; Shige Maruyama; Yoshi Ogasawara

Bibliography. American Geological Institute, Penrose Field Conference on Ophiolites, *Geotime*, 17:24–25, 1972; K. C. Condie, *Plate Tectonics and Crustal Evolution*, 1997; R. G. Coleman, The diversity of ophiolites, *Geologic en Mijinhaw*, 16:141–150, 1984; R. G. Coleman, *Ophiolites—Ancient Oceanic Lithosphere?*, 1977; I. G. Gass, Ophiolites, *Sci. Amer.*, pp. 122–131, August 1982; R. A. Kerr, Ophiolites: Windows on which ocean crust?, *Science*, 219:1307–1309, 1983; R. Mason, Ophiolites, *Geol. Today*, 1:36–40, 1985; E. M. Moores, Origin and emplacement of ophiolites, *Rev. Geophys. Space Phys.*, 20:735–760, 1982; A. Nicolas, *The Mid-oceanic Ridges: Mountains below Sea Level*, 1995.

## Ophiuroidea

A class of the Asterozoa, known as the brittle stars, in which the arms are usually clearly demarcated from a central disk and perform whiplike locomotor movements, and the tube feet are nonsuctorial sensory tentacles. In all existing ophiuroids the ambulacral plates fuse together in pairs to form articulating joints termed vertebrae, and the ambulacral groove is converted into an internal epineural canal.

As Fig. 1 shows, typical ophiuroids have a distinctive shape unlike that of the other Asterozoa, and the vigorous lashing movements of the arms distinguish them at once from asteroids. However, fossil ophiuroids are known which approach asteroids in structure, and the two groups are evidently closely related. See ASTEROIDEA.

Ophiuroids are usually five-armed, with four- or six-armed individuals occurring as abnormalities; but a few species are regularly six- or seven-armed. In some Euryalae the arms may branch repeatedly; these are the so-called basket stars. In the smallest

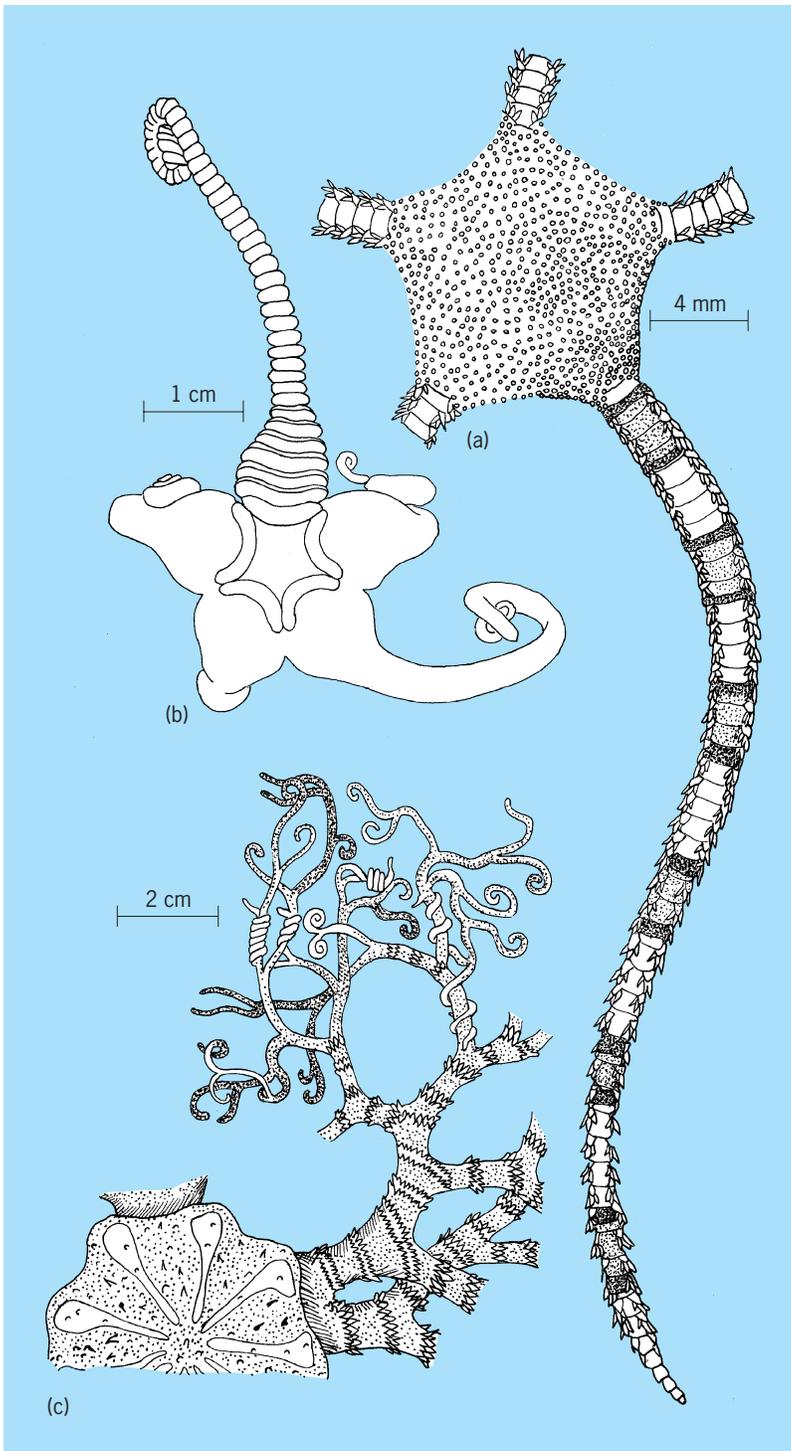


Fig. 1. Representative ophiuroids, which have a quite distinctive shape. (a) *Pectinura cylindrica*. (b) *Astroporpa wilsoni*. (c) *Gorgonocephalus chilensis*.

species the disk and arms together may be no more than a few millimeters across. Large species may have a disk 4 in. (10 cm) across and an arm spread of 20 in. (50 cm). Tropical species are often patterned in contrasting colors, but most ophiuroids tend to match their environment. Their biochromes do not include echinochromes. Some species are luminescent, although not constantly so; the light is emitted

by cells at the bases of the arm spines, usually only on stimulation. Such species are known for the genera *Amphiura*, *Amphipholis*, and *Ophiacantha*.

**Classification.** There are about 1900 extant species referred to 230 genera, arranged to form 3 living orders: Oegophiurida, Phrynophiurida, and Ophiurida. There is also one Paleozoic order, the Stenurida.

*Oegophiurida.* Oegophiurida is the most primitive order of brittlestars (Ophiuroidea), comprising a single living genus, *Ophiocanops*. In oegophiurids, the gonads, together with outpouches of the digestive system (gastric caeca), extend from the central body cavity into the arms. This is a characteristic that they share with asteroids. The digestive system is confined to the disc in all other ophiuroids (order Ophiurida), and in only a small number of primitive taxa do the gonads extend into the arms. In addition, the vertebral ossicles that make up the arms in Oegophiurida, though paired and firmly united, are not fused distally, whereas left and right vertebral ossicles are always fused in other extant ophiuroids. Oegophiurida also primitively lack radial and oral shields in their disc, as well as ventral arm plates. Andrew B. Smith

*Phrynophiurida.* Members of this order of Ophiuroidea are characterized by vertebrae that usually articulate by means of hourglass-shaped surfaces and arms that are able to coil upward or downward in the vertical plane. There is usually a leathery integument, in which calcareous granules or platelets are embedded. Most species are found in deep water, and often the arms are tightly coiled about the branches of black corals, upon which Phrynophiurida feed. Of the families, the Gorgonocephalidae often have branched arms, the Asteronychidae have a large disk and slender arms, and the Asteroschematidae have a small disk and stout arms.

The foregoing families share a number of characteristics and are grouped in one suborder, Euryalina. One remaining family, the Ophiomyxidae, differs in having a soft, unprotected integument, like that of *Ophiocanops*, but lacks the peculiar features of the gut and gonads in oegophiurids. For reasons too specialized to discuss here, it appears best not to associate the Ophiomyxidae with the Euryalina, in the Phrynophiurida, placing the family in a distinct suborder, the Ophiomyxina.

*Ophiurida.* Members of this order of Ophiuroidea possess vertebrae that articulate by means of ball-and-socket joints, and the arms, which do not branch, move mainly from side to side and do not coil in the vertical plane. The disk and arms are usually sheathed in regularly arranged plates. These are disposed in four series on the arms, namely, one dorsal, one ventral, and two lateral. There is a single madreporite. The order embraces most known genera of brittle stars and includes 13 families.

Howard B. Fell

*Stenurida.* Individuals in this Paleozoic (Ordovician–Devonian) order have a double row of plates (ambulacra) that abut across the arm axis either directly opposite one another or slightly offset. In contrast, modern ophiuroids have a single series of axial arm

plates termed vertebrae. In stenurids, as in modern ophiuroids, lateral plates are present at the sides of ambulacrals, and prominent lateral spines are typical. Stenurids lack the dorsal and ventral arm shields that are found in most ophiuroids. Proximal ambulacral pairs can be partially separated, forming a buccal slit, an expansion of the mouth frame. The arms of some stenurids are slender and flexible, but those of others are broad and comparatively stiff. The central disk varies from little larger than the juncture of the arms to an expansion that extends most of the length of the arms. The content of the order is poorly established, and fewer than 10 genera are known.

The relationships among ophiuroids, asteroids, and all other echinoderms provide an enduring problem in invertebrate evolution. Developmental and other studies based on modern organisms imply that asteroids and ophiuroids are not closely related within the echinoderms. Stenurid morphology, in contrast, suggests a close common ancestry for the two; the nature of the ambulacral plates is important, but even their general form is transitional.

Daniel B. Blake

**Skeleton.** The Paleozoic families had open ambulacral grooves on the undersurface of the arms, but in all extant forms the groove has sunk inward, and the ventral midline of the arm usually carries a median row of plates. In five Paleozoic families the ambulacral plates are paired, as in asteroids; in all other ophiuroids these plates fuse in pairs to produce jointed vertebrae, and the corresponding adambulacral plates become the so-called lateral plates on either side of the arm. In most extant forms the arm comprises a series of jointed segments, each one containing a vertebra, and each one covered externally by the ventral plate, right and left lateral plates, and a dorsal plate (Fig. 2a). These features are used in diagnosing the several orders. Spines are often carried by the skeletal plates, especially where the lateral arm plates represent the adambulacral spines of asteroids. In the order Ophiurida they commonly form an erect fringe to the sides of the arms (Fig. 2b); in the suborder Euryalina they commonly are transformed into clubs or hooklets, and hang downward.

**Muscular system.** The arms in extant ophiuroids are provided with well-developed longitudinal muscles linking the successive vertebrae. The two extant suborders are each characterized by a distinctive arm movement, horizontal in the Ophiurida, and vertical in the Euryalina. This distinction is related in either case to the form of the vertebrae, which in turn are defined in the taxonomic diagnoses. Ophiurida move rapidly when disturbed. One of the arms is either trailed or pushed ahead, whereas the other four arms operate as two opposite pairs of levers, thrusting the body forward in a series of rapid jerks. Unlike asteroids, the tube feet play little part in locomotion, or none at all. In very young stages, however, the tube feet may be used as stilts, and seem to be weakly adhesive. Euryalina have bigger vertebrae and smaller muscles, and their movements are less spasmodic;

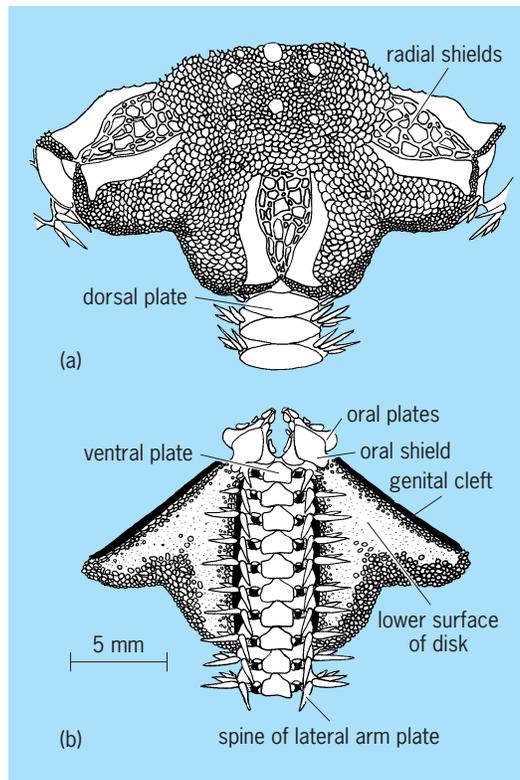


Fig. 2. External features of the ophiuroid *Amphiura*. (a) Aboral view. (b) Adoral view.

but they can coil the arms very firmly around objects, and keep their hold after death.

**Alimentary system.** The mouth, in the middle of the underside of the disk, serves for both ingestion and egestion of food, because there is no anus in existing forms. The gut comprises only the saclike stomach, in the walls of which are glandular hepatic cells. The stomach sends blind ceca into the arms in the Ophiocanopidae, a character recalling the asteroids.

Although many ophiuroids are predatory on small organisms when opportunity offers, most of them evidently spend most of their time scavenging in detritus, eating whatever they find. They are selective, because their gut does not permit the gross mud swallowing practiced by asteroids in similar circumstances. Thus, *Pectinura* will feed selectively on beech pollen in season in the New Zealand fiords, where the trees overhang the water. Among the more consistently predatory ophiuroids are certain Euryalina which cling to the branches of black corals and browse upon the polyps. Ophiuroids which live in large, dense populations evidently rely upon a steady flow of suspended matter, and there is evidence that sea-floor currents supply this. Basket stars sometimes live in these conditions rhythmically sweeping the branched arms toward the mouth.

**Water-vascular system.** This system is typical for the phylum, but the tube feet lack ampullae and

suckers, and can be retracted into so-called tentacle pores. The madreporite normally lies in one interradius of the adoral surface, but some Euryalina have a madreporite in each interradius.

**Nervous system.** This system is also typical for the phylum. No organs of special sense are known. It may be inferred that photoreceptors and chemoreceptors are distributed on the ectoderm of the ambulacral tube feet. See NERVOUS SYSTEM (INVERTEBRATE).

**Reproduction.** Ophiuroids may take up to 2 years to reach sexual maturity, and full growth may require 3 or 4 years. The life span is unknown but may be estimated at about 5 years. Large Euryalina, such as *Gorgonocephalus*, may well live much longer. The sexes are usually separate, but a few species, such as *Amphipholis squamata*, are hermaphroditic. In a few species the female carries a dwarf male clinging to the disk. The ovaries and testes are confined to the disk, and open indirectly to the exterior by way of interradial pouches in the integument, termed genital bursae. In the Ophiocanopidae, however, the gonads are serially paired in the basal arm joints, and do not open into bursae. Those species producing ophiopluteus larvae are apparently fewer than those with direct development. In viviparous species the young are retained in the bursae. Regeneration of lost organs is widespread. Autotomy (shedding of the arms) is practiced by most species when interfered with; the disk alone regenerates the lost members. The Amphiuroidae can also regenerate the gut and gonads, which may be cast off in autotomy. Species of Ophiactidae regularly reproduce by transverse fission, so that the arms (usually six in such forms) occur in two sets, of three large arms and three small arms. In no case have discarded arms been observed to regenerate. See ECHINODERMATA; REGENERATION (BIOLOGY).

**Relation to humans.** No ophiuroids are used as food by humans, and none are venomous. Ophiuroids must have a considerable indirect economic importance in view of their immense numbers and consequent significance in natural food chains involving commercially sought species.

**Ecology.** Ophiuroids occur in all the oceans from low-tide level downward, often in dense populations which number millions to the hectare. Six families range below a depth of 2 mi (3.2 km); the genera *Ophiura*, *Amphiophiura*, and *Ophiacantha* range below 4 mi (6.4 km). The shallow-water forms hide among algae, under stones, within sponges, or bury the disk in sand or mud, leaving only the arms protruding. Deep-water forms lie in or on the bottom material or adhere to corals or cidarids (urchins).

Among their numerous parasites are Protozoa in the stomach or genital organs; nematodes, trematodes, and Crustacea; Myzostomida (polychaete annelids) sometimes occur. Parasitic algae, such as *Coccomyxa ophiurae*, infest the spines and cause malformations; parasitic mollusks are much rarer than is the case with starfishes and sea urchins.

Howard B. Fell

**Bibliography.** R. A. Booloottian (ed.), *The Physiology of Echinodermata*, 1966; H. B. Fell, Phylogeny of sea stars, *Phil. Trans. Roy. Soc. London*, ser. B, 246:381-485, 1963; T. Lyman, Report on the Ophiuroidea, in *Report of the Scientific Results of HMS Challenger*, vol. 5: Zoology, 1882; T. Mortensen, Studies of Indo-Pacific euryalids, *Videnskabelige Meddelelser fra Dansk Naturhistorisk Forening*, 96: 1-75, 1933; J. Nebelsick and T. Heinzeller, *Echinoderms* Taylor & Francis Group, 2004; S. P. Parker (ed.), *Synopsis and Classification of Living Organisms*, 2 vols., 1982; V. B. Pearse, et al., An accessible population of *Ophiocanops* off NE Sulawesi, Indonesia, pp. 413-418 in R. Mooi and M. Telford (eds.), *Echinoderms: San Francisco*, A. A. Balkema, Rotterdam, 1998; A. B. Smith, A. B., G. L. J. Patterson, and B. Lafay, Ophiuroid phylogeny and higher taxonomy: Morphological, molecular and palaeontological perspectives, *Zool. J. Linn. Soc.*, 114: 213-243, 1995; A. B. Smith, et al., From bilateral symmetry to pentaradiality: The phylogeny of hemichordates and echinoderms, pp. 365-383, in J. Cracraft and M. J. Donoghue (eds.), *Assembling the Tree of Life*, Oxford University Press, 2004.

## Opiates

Drugs derived from opium, the dried juice of the oriental poppy seed. The pharmacologically active substances, which constitute approximately 25% of the extract, are the alkaloids morphine, codeine, and papaverine. The newer synthetic compounds, which resemble morphine in their action, are called opioids.

The principal effect of opium and opioids is to relieve pain. Even today morphine remains the best analgesic. It also assuages anxiety and causes slight drowsiness, relaxation, and a euphoric state of mind. These psychic effects are so agreeable that many troubled individuals seek solace by ingesting, smoking, or injecting opiates. Other effects are to suppress cough, allay respiratory distress, and relieve the pain and anxiety of a heart attack. The drugs act on the vital centers in the spinal cord and brainstem (medulla, pons, midbrain, and thalamus). See MORPHINE ALKALOIDS; PAIN.

Codeine has an action similar to morphine, but its analgesic effects are less. Papaverine has almost no analgesic action, and is used as an antispasmodic to relieve vascular spasm and undesirable contraction of smooth muscle.

The most tragic effect of the analgesic opium compounds is their addictive properties. Prolonged use leads to physical dependence, and discontinuance leads to painful and unpleasant withdrawal symptoms. Once addicted, most individuals become incapable of functioning in school and work and will go to extreme measures to obtain the drug and avoid the abstinence symptoms. Few can break the drug habit unless they enter an addiction center. See ADDICTIVE DISORDERS; ANALGESIC; NARCOTIC.

Raymond D. Adams

## Opiliones

The harvestmen or daddy longlegs, an order of the class Arachnida; sometimes known as Phalangida. About 4500 species are known. They are common in temperate and tropical climates. Most are red, brown, or black and 5–20 mm (0.2–0.8 in.) long, although a few are only 1–2 mm (0.04–0.08 in.). Some large tropical species have bright iridescent colors and elaborate spines.

The cephalothorax (prosoma) is broadly attached to the segmented abdomen (opisthosoma). The center of the cephalothoracic shield (carapace) bears a tubercle with a simple eye on each side. Many have scent glands opening on the sides which produce repellent fluids with strong odors, possibly phenols or quinones.

The six pairs of appendages include relatively small chelate chelicerae (jaws), leglike (usually) pedipalps, and four pairs of legs. The legs may be very long and slender, the distal ends being able to wrap around plant stalks. There is a genital opening between the fourth coxae.

Harvestmen are predators of small invertebrates or may scavenge or eat decaying vegetation. Respiration is by means of tracheae (thin tubes), which open through spiracles on the abdomen; there may be additional spiracles on the legs.

When mating, the male faces the female without courtship and projects a tubular penis between the female's jaws and into the female gonopore. The female deposits eggs into soft soil or leaf litter or under stones through an ovipositor. On hatching, the young resemble adults but are less sclerotized. After a series of molts they become sexually mature. Northern species produce eggs in fall, tropical species toward the end of the rainy season. Opilionids live about a year. Some species may occasionally form aggregations of hundreds of individuals with legs intertwined; the reason for such behavior is not known.

The group is divided into three suborders: the mitelike Cyphophthalmi; the tropical Laniatores with strong pedipalps, and some species adapted for cave life; and the long-legged Palpatores, the most common opilionids in temperate areas. See ARACHNIDA.

H. W. Levi

## Opisthobranchia

A subclass in the class Gastropoda containing about 4000 living species, arranged in nine orders, including the herbivorous Aplysiomorpha (sea hares) and Sacoglossa and the carnivorous Thecosomata (sea butterflies); Nudibranchia (sea slugs); Cephalaspidea (or Bullomorpha); Gymnosomata; Pleurobranchomorpha (or Notaspidea); Acochliidae; and Runciniodea. Primitive members of many of the orders show adaptations for burrowing beneath sand or mud; more advanced members are always active surface-living or pelagic forms. This trend is accompanied by a decrease in the importance of the shell and operculum for passive defense. These are re-

placed by more dynamic chemical (some species secrete decinormal sulfuric acid through the skin if annoyed), physical (daggerlike calcareous epidermal spicules), or biological (redirected nematocysts derived from coelenterate prey) defensive mechanisms.

The effects of gastropod torsion have been progressively nullified, then abolished in the opisthobranch line of descent from prosobranch ancestors. Many intermediate stages in this evolutionary loss of visceral torsion have survived and may be studied today. The most advanced members of the subclass show external and internal bilateral symmetry of all systems except the reproductive apparatus (situated on the right side to facilitate reciprocal copulation in these hermaphrodite animals).

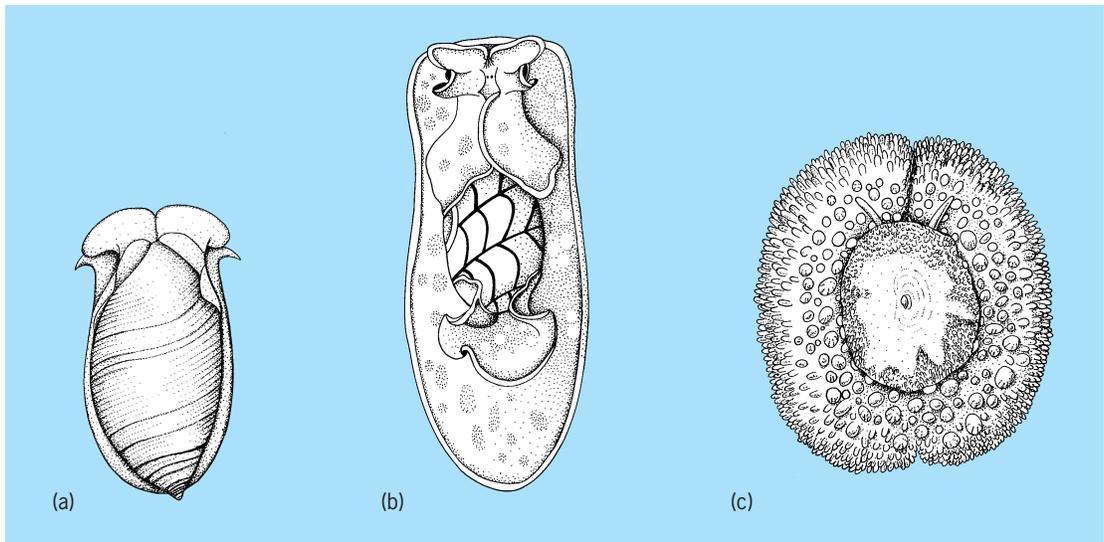
The adult shells of the primitive opisthobranchs living today are often strongly developed and sometimes colorful, as in *Acteon* (illus. *a*) where a horny snug-fitting operculum may be present and functional, but a more typical opisthobranch shell (as in *Micromelo*, illus. *b*, from the Caribbean or *Hydatina* from the Indo-Pacific) is fragile, inflated, and egg-shaped, and has the spire more or less concealed; the operculum is absent after the veliger larval phase. In rather more advanced forms, such as *Umbraculum* (illus. *c*), the external shell has a very wide gape, and in animals like *Berthella* the widely gaping shell is wholly internal, covered by the mantle.

The most varied shells are found in the Sacoglossa, which may have a single capacious coiled shell (*Volvatella*), a flattened open shell situated dorsally (*Lo-biger*), or two lateral shells (the bivalved gastropods, *Berthelinia*).

In the highest sacoglossans, some bullomorphs and aplysiomorphs, and in all the nudibranchs, the true shell is completely lost after larval metamorphosis.

Opisthobranch gastropods vary in size from tiny herbivorous sacoglossans, which live upon and consume delicate marine algae, cell by cell, and the sand-dwelling acochliidaeans, which are sometimes so minute that they can move unharmed between the grains that make up a sandy beach, to the huge nudibranch *Tochiuna* of the Pacific Northwest of America, and the aplysiomorph *Dolabella* of tropical seas, weighing in air as much as 4 lb (2 kg). Some species may be found only by diving, digging, or searching beneath boulders or coral heads; others, like some of the tropical nudibranchs, appear bold and self-advertising.

The only commercial fishery for any opisthobranch is found in coastal regions of China, where dried bodies of *Aplysia* are employed medicinally by physicians. Handling some living opisthobranchs can lead to mild pain and local inflammation. The planktonic nudibranchs *Glaucus* and *Glaucilla*, which feed upon the venomous siphonophore *Physalia* (the Portuguese man-o'-war), utilize for their own defense the nematocysts of this formidable prey. Severe stings by *Glaucus*, resembling those inflicted by *Physalia*, have been reported in eastern Australia by



Representative genera of opisthobranchs. (a) *Acteon*. (b) *Micromelo*. (c) *Umbraculum*.

bathers who misguidedly handled these pretty blue mollusks.

Swimming occurs in many opisthobranchs, sometimes brought about by whole-body undulations, sometimes by vigorous movements of expanded lobes of the foot or mantle. The orders Gymnosomata and Thecosomata contain numerous plankton-feeding species which spend their entire lives swimming near the surface of the oceans. See GASTROPODA; MOLLUSCA; NUDIBRANCHIA; SACOGLOSSA.

T. E. Thompson

Bibliography. S. P. Parker (ed.), *Synopsis and Classification of Living Organisms*, 2 vols., 1982; C. Lydeard, D. R. Lindberg, and G. J. Vermeij, *Molecular Systematics and Phylogeography of Mollusks*, Smithsonian Books, 2003; T. E. Thompson, *Biology of Opisthobranch Molluscs*, II, 1984; T. E. Thompson, *Nudibranchs*, 1976.

## Opisthocomiformes

A small order of birds that contains a single family, the Opisthocomidae. Its one species, the hoatzin, is restricted to South America and has been frequently included in the Galliformes as a suborder; however, little evidence supports this conclusion. More recently, some researchers claimed that it was a typical member of the Cuculidae, the cuckoos, based only on evidence from deoxyribonucleic acid (DNA) hybridization studies. However, important differences in the morphology of the perching foot argue strongly against the inclusion of the hoatzin in Cuculidae. The hoatzin has an anisodactyl perching foot with three anterior toes and a well-developed hallux (first toe) that is completely different from the zygodactyl perching foot of the cuckoos (with the second and third toes anterior, and the fourth toe and hallux posterior). Selective demands for perching could not have resulted in the evolution

of the anisodactyl foot of the hoatzin from the zygodactyl foot of the cuckoos. In the absence of definite evidence of its relationship to other avian orders, the Opisthocomiformes are best considered as a distinct order that evolved in South America since its isolation from the rest of Gondwanaland sometime in the late Mesozoic. See CUCULIFORMES, GALLIFORMES.

The hoatzin is a unique bird, often considered rather reptilian because of its nearly unfeathered young, which leaves the nest soon after birth to crawl around trees by using its clawed wings as well as its feet (see **illustration**). Hoatzins are adept swimmers, capable of surviving a fall into water. They are medium-sized birds of dark brown plumage streaked with white; a distinct crest of reddish-brown feathers tops the head, which is small and has a short stout



A baby hoatzin climbing a tree using its clawed wings as well as its feet. (Reprinted with permission from D. W. Linzey, *Vertebrate Biology*, McGraw-Hill, 2001)

bill. The wings are long and rounded, the legs are short with strong toes, and the long hallux is on the same level as the anterior toes. Hoatzins are weak fliers, but they are almost strictly arboreal, clambering about in trees to feed on leaves and fruit. A large crop, which displaces the sternum in a posterior direction, is used to store food prior to digestion. The nest is placed in trees overhanging water, and both sexes incubate and care for the young. Hoatzins are found in forests of northern South America, especially along rivers and streams.

The hoatzins are known in the fossil record only by the Miocene *Hoazinoides* from Columbia, a typical form which reveals nothing about the evolutionary history of the order. See AVES. Walter J. Bock

Bibliography. B. T. Thomas, Order Opisthocomiformes, pp. 23–32, in J. Del Hoyo et al. (eds.), *Handbook of the Birds of the World*, Lynx Editions, vol. 3, 1996.

## Opossum

A New World marsupial belonging to the family Didelphidae. Didelphidae consists of one North American species, the virginia opossum (*Didelphis virginiana*), and 65 Central and South American species, including the mouse opossums (*Marmosa*, *Gracilinanus*, *Marmosops*, *Micoureus*, *Thylamys*, and *Lestodelphys*); short-tailed opossums (*Monodelphis*); gray and black four-eyed opossums (*Philander*); brown four-eyed opossums (*Metachirus*); water opossums (*Chironectes*); woolly opossums (*Caluromys*); black-shouldered opossums (*Caluromystops*); bushy-tailed opossums (*Glironia*); and thick-tailed opossums (*Lutreolina*). See MARSUPIALIA.

**Morphology.** Opossums are primarily nocturnal and may be either arboreal or terrestrial.

Each front foot has five claw-bearing toes, while each hind foot has four claw-bearing toes and a thumblike, clawless hallux, or big toe, which opposes the other digits. The water opossum, or yapok (*Chironectes*), of Central and South America is semi-aquatic with webbed hind feet. The snout is long and pointed, and the tail is usually prehensile (used for grasping). Body size (excluding tail length) ranges from 75 mm (3 in.) in tiny species (*Monodelphis*) to more than 500 mm (20 in.) in the largest species (*Didelphis*). The abdominal pouch (marsupium), well developed in some genera but rudimentary in others, is lined with soft fur and contains the mammae. The skull has a large preorbital region, large zygomatic arches, a prominent median sagittal crest, and a small braincase. Opossums have heterodont dentition consisting of 50 teeth (dental formula: I 5/4, C 1/1, Pm 3/3, M 4/4 × 2). Females have a well-developed marsupium. The usual number of teats is 13 (12 in a circle and one in the center), although all may not be functional. The male reproductive organ is forked, and the female reproductive system consists of paired vaginae and paired uteri. See DENTITION; REPRODUCTIVE SYSTEM.



Virginia opossum (*Didelphis virginiana*).

**Virginia opossum.** The Virginia opossum *D. virginiana*; (see **illustration**) ranges from southern Canada to Costa Rica. It is a medium-sized mammal with long, rather coarse, grayish-white fur. Individuals have large, thin, leathery, naked ears and a long, scaly, sparsely haired tail. The nose is pink, the eyes are black, and the ears are bluish-black. Adult Virginia opossums are approximately 600–875 mm (23–34 in.) in length, including a 275–350 mm (11–14 in.) tail. Adults usually weigh between 1.3 and 4.5 kg (4 and 13 lb).

**Diet and habitat.** Opossums are omnivorous, eating a wide variety of foods but preferring animal matter during all seasons. They are solitary mammals and inhabit a variety of habitats from grasslands to forests. Home ranges vary considerably and overlap broadly. There is no evidence of territoriality. Opossums are active throughout the year but may retreat to dens during cold periods. Den sites include hollow trees, fallen logs, brush piles and ground burrows. The nest is composed almost entirely of leaves. Opossums readily climb, using their prehensile tail extensively for balance and support. Although most forms are not aquatic, they can swim and dive readily. These mammals are probably best known for “playing possum,” a temporary state of catatonia. This phenomenon, which is a passive defensive tactic that apparently has survival value, seems to be a nonvoluntary (reflex) action and probably is not deliberate or willful.

**Mating and development.** Mating is promiscuous or polygynous with females being seasonally polyestrous. Two litters are usually produced each year. The gestation of 12 to 13 days is the shortest of any

American mammal. Average litter size ranges from 7 to 9, although as many as 18 young have been recorded. A dozen newborn opossums, each weighing approximately 0.16 g, will fit into a teaspoon. At birth, the front limbs are well developed and are used in a swimming motion to help the young ascend through the mother's fur into the marsupium. The young attach to a teat and remain in the marsupium for approximately 60–80 days. Weaning usually occurs between 80 and 100 days. Opossums reach sexual maturity between 6 and 8 months of age, but females do not breed until the first estrus the following year. Average life expectancy is approximately 1.33 years, although captive animals have lived over 7 years. Major predators include great horned owls, carnivorous mammals including dogs, and humans. Automobiles are responsible for the death of many opossums.

Donald W. Linzey

**Bibliography.** G. A. Feldhamer, B. C. Thompson, and J. A. Chapman (eds.), *Wild Mammals of North America: Biology, Management, and Conservation*, 2d ed., Johns Hopkins University Press, 2003; D. W. Linzey, *The Mammals of Virginia*, McDonald & Woodward Publishing Company, 1998; D. Macdonald (ed.), *The Encyclopedia of Mammals*, Andromeda Oxford Limited, 2001; R. M. Nowak, *Walker's Mammals of the World*, 2 vols., Johns Hopkins University Press, 1999; D. E. Wilson and S. Ruff (eds.), *The Smithsonian Book of North American Mammals*, Smithsonian Institution Press, 1999.

## Opportunistic infections

Infections that cause a disease and occur only when the host's ability to fight back (the immune system) is impaired. Some microorganisms can be both opportunistic and nonopportunistic. An example is the bacterium *Mycobacterium tuberculosis*. A healthy individual who is exposed to a sufficient number of *M. tuberculosis* bacteria can develop pulmonary tuberculosis and become quite ill. However, most people with normal immune systems who are exposed to tuberculosis never become clinically ill even though they become infected. In this case, the body's immune defenses are capable of isolating the *M. tuberculosis* and preventing it from spreading throughout the body and causing symptoms or clinical disease. However, if an individual with tuberculosis also has an immune deficiency, such as acquired immunodeficiency syndrome (AIDS), the disease process is much different. Individuals with AIDS who are exposed to an infected person with active tuberculosis are much more likely to become ill, have more widespread disease, infect others, and die from the disease. Fortunately, if the diagnosis is made in time, antituberculous medications work effectively regardless of the underlying condition. See ACQUIRED IMMUNE DEFICIENCY SYNDROME (AIDS); TUBERCULOSIS.

The classic opportunistic infection actually leads to disease only in individuals with abnormal immunity and never in the normal host. The proto-

zoon *Pneumocystis carinii* infects nearly everyone at some point in life but never causes disease unless the immune system is severely depressed. The most common immunologic defect associated with pneumocystosis is AIDS. Other immunologic deficiencies such as hematologic cancers and profound malnutrition can also predispose the individual to pneumocystosis, but only rarely when compared to AIDS, in which it is the most common life-threatening opportunistic infection.

**Compromised host defenses.** A compromised host is an individual with an abnormality or defect in any of the host defense mechanisms that predisposes that person to an infection. A variety of specific host defenses can become altered and may lead to a relatively predictable set of opportunistic infections. The altered defense mechanisms or immunity can be either congenital, that is, occurring at birth and genetically determined, or acquired. Congenital immune deficiencies are relatively rare.

Acquired immunodeficiencies are associated with a wide variety of conditions such as (1) the concomitant presence of certain underlying diseases such as cancer, diabetes, cystic fibrosis, sickle cell anemia, chronic obstructive lung disease, severe burns, and cirrhosis of the liver; (2) side effects of certain medical therapies and drugs such as corticosteroids, prolonged antibiotic usage, anticancer agents, alcohol, and nonprescribed recreational drugs; (3) infection with immunity-destroying microorganisms such as the human immunodeficiency virus that leads to AIDS; (4) age, both old (the older one gets, the more the immune system wanes) and young (the younger one is, particularly under 1 month of age, the more immunologically immature); and (5) foreign-body exposure, such as occurs in individuals with prosthetic heart valves, intravenous catheters, and other indwelling prosthetic devices.

**Granulocytopenia.** Granulocytes are white blood cells that act as a first line of defense against potential pathogens (see **table**). When their numbers are sufficiently depleted, usually as a result of an underlying hematologic cancer, cancer chemotherapy, or alcohol use, infection with aerobic bacteria and fungi that are already present on or in the body becomes more likely. The lack of granulocytes diminishes the overall immune response, making the ability to diagnose an infectious disease even more difficult.

**Cellular immune deficiency.** Cellular immunity refers to the immune defect that results in the inability of monocytes and macrophages, which are essentially scavenger cells, to kill pathogens that are capable of infecting individual host cells. This type of infection includes such diseases as legionnaires' disease, tuberculosis, pneumocystosis, and cryptosporidiosis. These infections rarely cause disease in the normal host. Conditions associated with impaired cellular immunity include AIDS, Hodgkin's disease, chronic corticosteroid use, irradiation therapy, and collagen vascular diseases such as rheumatoid arthritis and lupus. See CELLULAR IMMUNOLOGY.

**Humoral immune dysfunction.** Humoral immunity refers to the immune defect that results from impaired

Opportunistic infections and immunologic deficiencies			
Immunologic deficiency	Associated underlying disease or conditions	Likely opportunistic pathogen	Associated clinical infectious diseases
Granulocytopenia	Acute leukemia Chemotherapy Drugs	<i>Escherichia coli</i>	Bacteremia
		<i>Pseudomonas aeruginosa</i>	Pneumonia
		<i>Candida albicans</i>	Sinusitis
Cell-mediated deficiency	AIDS Hodgkin's disease Solid tumors Sarcoidosis Aging Malnutrition Lupus	<i>Aspergillus fumigatus</i>	
		<i>Pneumocystis carinii</i>	Pneumonia
		<i>Toxoplasma gondii</i>	Encephalitis
		<i>Cryptococcus neoformans</i>	Meningitis
		<i>Mycobacterium tuberculosis</i>	Widely disseminated infection
		<i>Listeria monocytogenes</i>	
		<i>Legionella pneumophila</i>	
Humoral immune deficiency	Congenital splenectomy Chronic lymphocytic leukemia Multiple myeloma Sickle cell anemia	<i>Streptococcus pneumoniae</i>	Bacteremia
		<i>Haemophilus influenzae</i>	Pneumonia
			Meningitis
Foreign body	Mechanical heart valve Artificial joint Intravenous catheter	<i>Staphylococcus aureus</i>	Bacteremia
		<i>Staphylococcus epidermidis</i>	Phlebitis
			Endocarditis
Tumor	Obstructed bronchus	Various anaerobic and aerobic bacteria	Postobstructive pneumonia Enteritis Urinary tract infection

antibody production. Antibodies are important in combining with certain antigens, in this case pathogens, so that the antigen-antibody complex is taken up by host phagocytes and neutralized. The type of infections associated with this deficiency include those by encapsulated bacteria, such as the aerobes *Streptococcus pneumoniae* and *Haemophilus influenzae*. Conditions associated with humoral immune dysfunction include multiple myeloma, congenital deficiencies, chronic lymphocytic leukemia, and splenectomy regardless of the cause. See HAEMOPHILUS; STREPTOCOCCUS.

**Foreign-body exposure.** The presence of any mechanical apparatus, intravascular needle, or prosthetic device such as a joint or heart valve is a risk for infection by organisms that can seed or enter the tissues through the device. The most common organism to do this is a *Staphylococcus* species. It is through this route that many hospital-acquired infections occur. See STAPHYLOCOCCUS.

**Tumors and other medical conditions.** Depending on the nature of the specific malignancy, any number of the host's defense mechanisms may become compromised. In addition, the tumor itself may obstruct or physically invade vital organs, causing yet another form of impaired host defense. An example is the effect of a bronchogenic carcinoma, a type of lung cancer. If the tumor itself obstructs one of the major airways in the lung, a postobstructive pneumonia may result. The usual pathogen is an aerobic or anaerobic bacterium that had formerly colonized the respiratory tract but now is actually causing a pneumonia. Besides tumors, any condition that adversely effects organ function can lead to the development of infection. Examples include asthma, strokes, emphysema, and cystic fibrosis. See TUMOR.

**Opportunistic pathogens.** Virtually any microorganism can become an opportunist. The typical ones fall into

a number of categories and may be more likely to be associated with a specific immunologic defect. Examples include (1) gram-positive bacteria: both *Staphylococcus aureus* and the coagulase-negative *S. epidermidis* have a propensity for invading the skin and as well as catheters and other foreign implanted devices; (2) gram-negative bacteria: the most common is *Escherichia coli* and the most lethal is *Pseudomonas aeruginosa*; these pathogens are more likely to occur in cases of granulocytopenia; (3) acid-fast bacteria: *M. tuberculosis* is more likely to reactivate in the elderly and in those individuals with underlying malignancies and AIDS; (4) protozoa: defects in cell-mediated immunity, such as AIDS, are associated with reactivated infection with *Toxoplasma gondii* and *Cryptosporidium*; in individuals with normal immune defenses, these infections are relatively benign; with AIDS, they are sometimes lethal; (5) fungi: *Cryptococcus neoformans* is a fungus that causes meningitis in individuals with impaired cell-mediated immunity such as AIDS, cancer, and diabetes; *Candida albicans* typically causes blood and organ infection in individuals with granulocytopenia. See ESCHERICHIA; FUNGI; MEDICAL MYCOLOGY.

**Diagnosis and treatment.** The first step in treatment of opportunistic infections involves making the correct diagnosis, which is often difficult as many of the pathogens can mistakenly be thought of as benign. The second step involves administration of appropriate antimicrobial agents. As a third step, the underlying immune defect needs to be corrected. In the case of granulocytopenia, physicians can administer cytokines such as granulocyte colony stimulating factor that actually boosts the white blood cell count considerably. In some cases, indwelling catheters and devices must be removed or replaced. Unfortunately, the biggest limitation lies in the fact that the overall

ability to significantly improve immune function is quite limited. See IMMUNOLOGICAL DEFICIENCY; INFECTION; MEDICAL BACTERIOLOGY.

Robert Murphy Bibliography. M. A. Sande and P. A. Volberding (eds.), *The Medical Management of AIDS*, 6th ed., 1999; S. T. Shulman, J. P. Phair, and H. M. Sommers (eds.), *The Biologic and Clinical Basis of Infectious Diseases*, 1992.

## Opsonin

A term used in serology and immunology to refer to a substance that enhances the phagocytosis of bacteria by leukocytes. As originally proposed by A. E. Wright and E. R. Douglas, the term denoted a thermolabile, relatively nonspecific substance present in normal sera. In modern usages, opsonin is more generally synonymous with the bacteriotropin of F. Neufeld and coworkers, a relatively thermostable antibody, increased in amount during specific immunization, that renders the corresponding bacterium more susceptible to phagocytosis. There is evidence that this action can be promoted to some extent by antibody alone, but that it is substantially increased by the further addition of the thermolabile complement system. Opsonic activity can be displayed by antibodies that also give precipitation, agglutination, lytic, and neutralization reactions.

The opsonic index is a measure of the opsonic activity of sera, found by dividing the average number of bacteria per phagocytic cell, as determined in the presence of an immune serum, by the corresponding value obtained in the presence of normal serum. See AGGLUTINATION REACTION; ANTIBODY; IMMUNOLOGY; LYTIC REACTION; NEUTRALIZATION REACTION (IMMUNOLOGY); PHAGOCYTOSIS; PRECIPITIN TEST; SEROLOGY.

Henry P. Treffers

Bibliography. J. Kuby, *Immunology*, 4th ed., 2000; J. D. Sleight and M. C. Timbury, *Medical Bacteriology*, 3d ed., 1990.

## Optical activity

The effect of asymmetric compounds on polarized light. To exhibit this effect, a molecule must be non-superimposable on its mirror image, that is, must be related to its mirror image as the right hand is to the left hand. An optically active compound and its mirror image are called enantiomers or optical isomers. Enantiomers differ only in their geometric arrangements; they have identical chemical and physical properties. The right-handed and left-handed forms of a molecule can be distinguished only by their optical activity or by their interactions with other asymmetric molecules. Optical activity can be used to probe other aspects of molecular geometry, as well as to identify which enantiomer is present and its purity.

As an example of optical isomers, consider tartaric acid (Fig. 1), which was one of the first synthetic

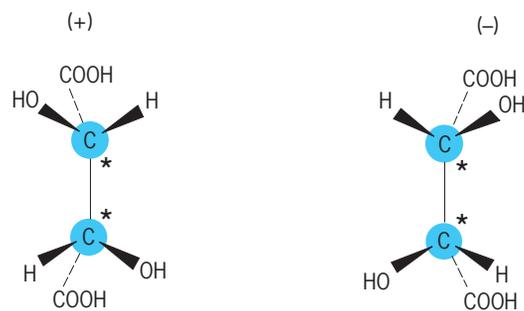


Fig. 1. Enantiomers of tartaric acid.

molecules to be separated into its enantiomers. In this case the asymmetry of each isomer is magnified when trillions of molecules form a crystal; two types of asymmetric crystals are formed.

The physical basis of optical activity is the differential interaction of asymmetric substances with left versus right circularly polarized light. If solids and substances in strong magnetic fields are excluded, optical activity is an intrinsic property of the molecular structure and is one of the best methods of obtaining structural information from a sample in which the molecules are randomly oriented. The relationship between optical activity and molecular structure results from the interaction of polarized light with electrons in the molecule. Thus the molecular groups that contribute most directly to optical activity are those that have mobile electrons which can interact with light. Such groups are called chromophores, since their absorption of light is responsible for the color of objects. For example, the chlorophyll chromophore makes plants green. See FARADAY EFFECT; MAGNETOOPTICS; POLARIZED LIGHT.

**Methods of measurement.** Optical activity is measured by two methods, optical rotation and circular dichroism.

**Optical rotation.** This method depends on the different velocities of left and right circularly polarized light beams in the sample. The velocities are not measured directly, but both beams are passed through the sample simultaneously. This is equivalent to using plane-polarized light. The differing velocities of the left and right circularly polarized components yield a rotation of the plane of polarization. A polarimeter for observing optical rotation consists of a light source, a fixed polarizer, a sample compartment, and a rotatable polarizer. A cell containing solvent is placed between the polarizers, and one of them is adjusted to be perpendicular to the other, excluding the passage of light. The solvent in the cell is then replaced by a solution of the sample, and the polarizer is rotated to again exclude passage of light. The optical rotation  $\alpha$  is the number of degrees the polarizer was rotated. A positive or negative sign indicates the direction of rotation. Enantiomers have rotations of equal magnitude, but opposite signs. The optical rotation depends on the substance, solvent, concentration, cell path length, wavelength of the light, and temperature. Standardized specific

rotations  $[\alpha]$  are reported as defined in Eq. (1), where

$$[\alpha]_{\lambda}^T = \frac{\alpha}{cl} \quad (1)$$

$T$  is the temperature ( $^{\circ}\text{C}$ ),  $\lambda$  the wavelength (often the orange sodium D line),  $l$  the cell path length in decimeters, and  $c$  the concentration in grams per milliliter. Alternatively,  $M_{\phi}$  is defined by normalizing to the rotation for a 1-molar solution, Eq. (2), where

$$M_{\phi} = \frac{[\alpha]_{\lambda}^T \text{MW}}{100} \quad (2)$$

$M_{\phi}$  is the molar rotation and MW the molecular weight. For polymers, the mean residue rotation,  $m_{\phi}$ , may be defined by the right side of Eq. (2) by using the mean residue (monomer unit) weight for MW. The variation of optical rotation with wavelength is known as optical rotatory dispersion (ORD).

**Circular dichroism.** Circular dichroism (CD) is the difference in absorption of left and right circularly polarized light. Since this difference is about a millionth of the absorption of either polarization, special techniques are needed to determine it accurately. Circular dichroism spectrometers consist of a light source, a monochromator to select a single wavelength, a modulator to produce circularly polarized light, a sample compartment, a phototube to detect transmitted light, and associated electronic components. The modulator rapidly switches (typically 50,000 times per second) between left and right circular polarization of the light beam. The absorption of an optically inactive sample is independent of polarization, so that the light intensity at the phototube is constant; thus a constant direct current is generated. The absorption of an optically active sample depends on the polarization, so that the light intensity at the phototube varies at the frequency of the modulator; thus an alternating current is generated. The circular dichroism is proportional to the amplitude of the alternating current. The proportionality constant is determined through calibration by using a compound of known circular dichroism.

Circular dichroism is reported as a difference in absorption, Eq. (3), or as an ellipticity (a measure

$$\Delta\epsilon = \epsilon_L - \epsilon_R = \frac{A_L - A_R}{c'l} \quad (3)$$

of the elliptical polarization of the emergent beam), Eq. (4), for a 1-molar solution, where  $\epsilon$  is the ex-

$$M_{\theta} = 3300\Delta\epsilon \quad (4)$$

inction coefficient,  $A$  is the absorbance [ $\log(I_0/D)$ ], subscripts  $L$  and  $R$  indicate left or right circular polarization,  $c'$  is the concentration in moles per liter,  $l$  is the path length in centimeters,  $I_0$  and  $I$  are the light intensities in the absence and presence of the sample, respectively, and  $M_{\phi}$  is the molar ellipticity. Either  $\Delta\epsilon$  or ellipticity,  $m_{\phi}$ , may be expressed per residue by making  $c'$  the concentration of residues (monomer units). As in the case of optical rotation, enantiomers have circular dichroism spectra of equal magnitude but opposite signs.

**Variation with wavelength.** Optical rotation and circular dichroism are two manifestations of the same interactions between polarized light and molecules. They are related by a mathematical transformation. An important difference between the two measurements is the way in which they vary with wavelength. Optical rotation extends to wavelengths far from any absorption of light. Thus colorless substances still have significant optical rotation at the sodium line. However, all groups which absorb light (chromophores) contribute at all wavelengths, and it can be difficult to extract the contribution of a single group. On the other hand, circular dichroism is confined to the narrow absorption band of each chromophore. Thus it is easier to determine the contribution of individual chromophores, information vital to structural analysis. See COTTON EFFECT; DICHROISM; OPTICAL ROTATORY DISPERSION.

**Correlation with molecular structure.** In synthesizing enantiomers, chemists focus on an asymmetric center, that is, a locus which imparts asymmetry to the whole molecule. A common asymmetric center is a tetrahedral carbon atom with four different groups attached, such as the carbons marked with asterisks in tartaric acid (Fig. 1). However, in correlating optical activity with molecular structure, the focus is on the three-dimensional arrangement of the chromophores which interact most strongly with light.

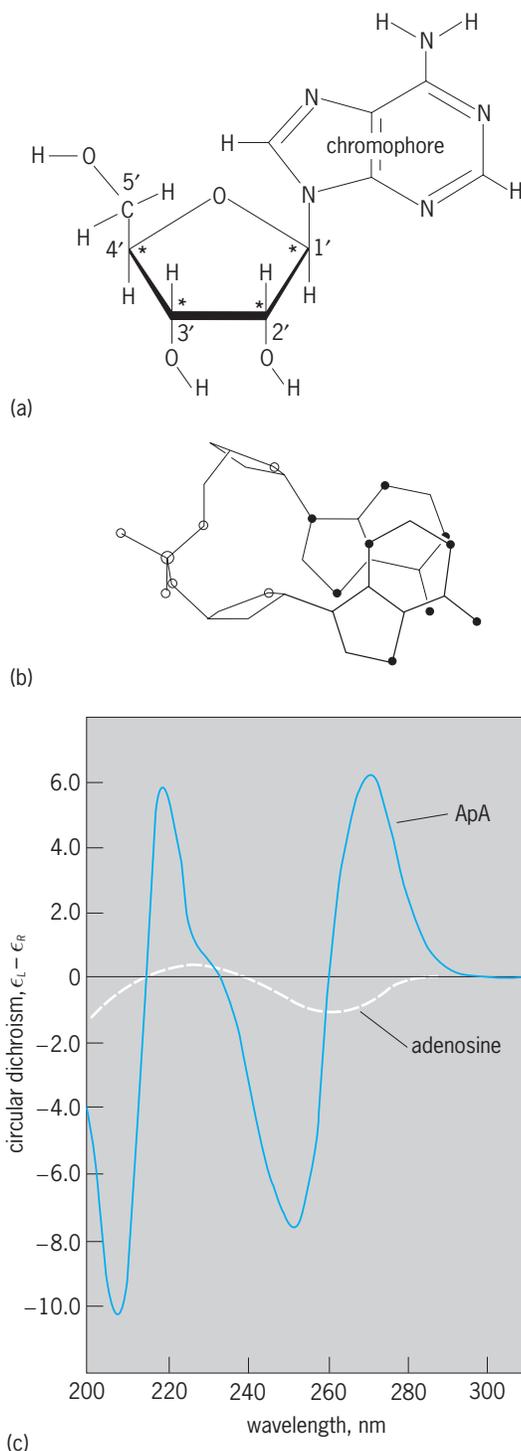
As examples, consider the nucleoside adenosine and its dimer (Fig. 2). The most mobile electrons are in the aromatic ring system, the chromophore (Fig. 2a). The electrons in the sugar ribose are more tightly bound and interact less strongly with visible and ultraviolet light. However, all the asymmetric centers are in the ribose part of the molecule. For adenosine, light interacting with the aromatic chromophore is only weakly influenced by the asymmetric centers in ribose, so that small circular dichroism bands are observed (Fig. 2c).

In the covalently linked dimer of adenosine, the observed circular dichroism bands are about 10 times larger than those of the monomer. In the 240- to 300-nm region of the spectra (Fig. 2c), two bands are observed for the dimer, but only one for the monomer. This indicates strong interaction of the two aromatic chromophores, and hence their close proximity in the dimer. Analysis of the circular dichroism spectra expected for various arrangements of the two chromophores, as well as other types of experimental data, indicates that the aromatic rings are stacked (Fig. 2b). The asymmetric centers in ribose cause the formation of the stacked arrangement shown rather than its mirror image. See ASYMMETRIC SYNTHESIS.

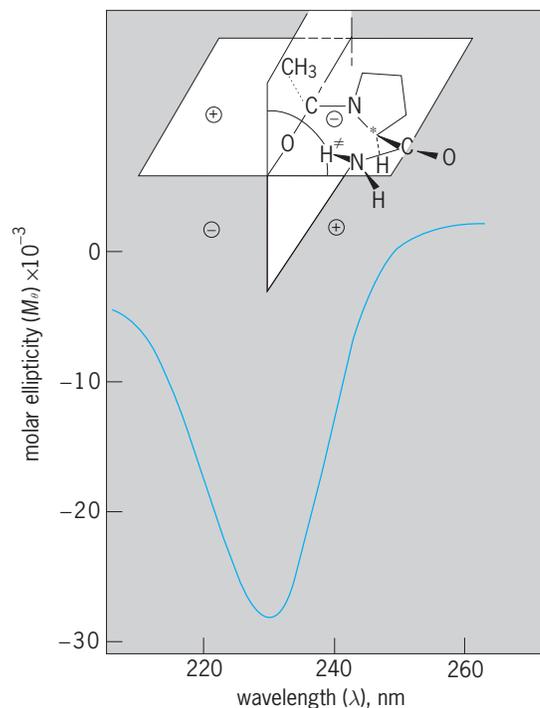
Stacking of aromatic rings, as exemplified by the adenosine dimer, is a common feature of nucleic acid polymers (deoxyribonucleic acid and ribonucleic acid) isolated from biological sources. Slight differences in the stacking geometry gives each of these polymers a characteristic circular dichroism spectrum. Alterations in the stacking arrangement caused by some pharmacologically active agents can be

detected through alterations in the circular dichroism spectra. These structural changes may in turn be related to the pharmacological action.

A derivative of the amino acid proline (**Fig. 3**) can be used to illustrate another way in which opti-



**Fig. 2.** Adenosine and its dimer. (a) Structure of adenosine. Asymmetric centers are marked by an asterisk. (b) Stacked arrangement of adenosine dimer (ApA). The 3' carbon of one adenosine is linked to the 5' carbon of the other by a phosphate group. (c) Circular dichroism spectra of ApA and adenosine at neutral pH in aqueous solution at room temperature.



**Fig. 3.** Folded arrangement of L-proline derivative (*N*-acetyl-L-proline-amide) and its circular dichroism spectrum in *p*-dioxane solution at room temperature.

cal activity depends on molecular structure. In this molecule only the OCN group (amide chromophore) which is in the horizontal plane of the drawing and the hydrogen which is marked H<sup>≠</sup> need be considered. By forming the N-H bond, H<sup>≠</sup> acquires a charge of about  $+1/3$  electron. It has been predicted that such a positive charge will perturb the motion of the electrons in the amide chromophore in a manner which will produce a negative circular dichroism band when the charge is above the plane of the amide group and to the right of the oxygen. Only for the arrangement shown is the magnitude of the circular dichroism band expected to be as large as observed. Furthermore, it has been shown that there will be no circular dichroism if H<sup>≠</sup> is in either of the two planes shown, and that for H<sup>≠</sup> in adjacent quadrants the sign of the circular dichroism band alternates (Fig. 3). For this compound, reflection through the horizontal plane will generate the enantiomer. This would place H<sup>≠</sup> in the lower right quadrant and generate a positive circular dichroism band with magnitude equal to that of Fig. 3. See STEREOCHEMISTRY.

Vincent Madison

**Bibliography.** E. Charney, *The Molecular Basis of Selected Papers on Optical Activity*, 1985; A. Lakhakia (ed.), *Selected Papers on Natural Optical Activity*, 1985; J. A. Schellman, Symmetry rules for optical rotation, *Accounts Chem. Res.*, 1:144-155, 1968; I. Tinoco, Jr., K. H. Sauer, and J. C. Wang, *Physical Chemistry: Principles and Applications in Biological Sciences*, 3d ed., 1995; R. W. Woody, Optical rotatory properties of biopolymers, *J. Polym. Sci. Macromol. Rev.*, 12:181-321, 1977.

## Optical bistability

A phenomenon exhibited by certain resonant optical structures whereby it is possible to have two stable steady transmission states for the device, depending upon the history of the input. Such a bistable device may be useful for optical computing elements because of its memory characteristics. The bistability can result from the intrinsic properties of the optical device or from some external feedback such as an electrical voltage supplied by another device. This second type, extrinsic or hybrid optical bistability, is not true optical bistability.

Optical bistability is an inherently steady-state phenomenon, and typically any cycling of the device through its hysteresis cycle must be done adiabatically; that is, changes in the propagating light amplitude, envelope phase, and profile must occur sufficiently slowly that their impact on the evolution of the system may be neglected. This requirement imposes some rather severe frequency-response limitations on the use of intrinsically bistable devices in optical circuits. The two primary types of intrinsic optical bistability, each arising from a distinct physical mechanism, are absorptive bistability and refractive bistability. *See* ADIABATIC PROCESS; HYSTERESIS.

**Absorptive bistability.** Absorptive optical bistability, discovered in 1969, is based upon coupling the feedback mechanism inherent in an optical cavity with an absorbing nonlinear optical medium in which the absorption coefficient decreases with increasing light intensity (a saturable absorber). The basic theory of operation is as follows: The saturable absorber is placed in the cavity, and the cavity is resonantly pumped. For low light intensities, the transmission coefficient for the cavity is small because of the presence of the highly absorbing medium inside the cavity. As the pump intensity is increased, the absorption of the nonlinear medium decreases. Finally, for some threshold pump intensity, the cavity switches into a high transmission state, because the absorption coefficient is reduced sufficiently that the intrinsic cavity feedback mechanism dominates. The threshold is very sharp because, when the cavity is in a highly transmissive state, the built-up intensity inside the cavity becomes very large compared to the pump intensity (due to the feedback) and effectively bleaches virtually all of the absorption in the nonlinear medium. The intense pump is then largely transmitted, although some energy is stored in the cavity to bleach the absorber. *See* ABSORPTION OF ELECTROMAGNETIC RADIATION; LASER; OPTICAL PUMPING.

When the pump wave is reduced in intensity, the built-up intensity inside the cavity does not drop below the point where it effectively saturates the nonlinear absorption until the pump intensity is well below the switch-up threshold. Thus, the cavity stays in its highly transmissive state for intensities well below the threshold. As a result, the transmission characteristic of the cavity exhibits hysteresis and possesses two steady-state outputs for a range of pump intensities.

This device exhibits two characteristics that constrain its usefulness in particular applications. (1) The device is based on an absorption mechanism, so the energy absorbed from the pump light must be dissipated in the bistable element or heat-sinked elsewhere. (2) It is highly frequency-sensitive because its operation is based on the switching characteristics of a resonant cavity.

**Refractive bistability.** Refractive optical bistability, discovered in 1974, is based on coupling the feedback mechanism inherent in an optical cavity with a nonlinear optical medium that exhibits a change in the refractive index as a function of light intensity. The nonlinear refractive medium is placed inside the optical cavity, and the cavity is pumped slightly off-resonance so that the transmission coefficient is small compared to unity. However, a small amount of light intensity does exist inside the cavity, and changes the effective optical path length inside the cavity by inducing change in the refractive index of the nonlinear medium. As the pump intensity is increased, this change in the effective path length becomes larger, until at some point the cavity switches into, and possibly past, resonance. The transmission coefficient switches abruptly to a value close to unity, and the built-up intensity inside the cavity increases abruptly. If the pump intensity is increased further, it is possible to switch the cavity through a second resonance, with an additional threshold in the transmission coefficient. *See* REFRACTION OF WAVES.

In principle, this process can be repeated for arbitrarily large values of the pump intensity. However, from a practical perspective it is very difficult to obtain more than two or three thresholds because of the intervention of other, higher-order nonlinear processes which arise from the presence of very intense light.

When the pump intensity is reduced, the effective cavity path length returns to its low-intensity value. However, because of the feedback mechanism in the cavity and the presence of the built-up intensity inside the cavity, the switch-down thresholds occur at lower pump intensities than the switch-up thresholds, resulting in hysteresis of the transmission characteristic. In this case, it is possible to observe multiple hysteresis loops, as the cavity is switched through multiple resonances by the built-up intensity within it.

**Implementation.** The most common implementation scheme for a bistable optical device is the nonlinear Fabry-Perot etalon. The device is typically fabricated from a semiconductor such as indium antimonide (InSb), gallium arsenide (GaAs), or zinc selenide (ZnSe), and consists of a slab of material of approximately 1 micrometer thickness. On each surface of the semiconductor, a highly reflective coating may be deposited to increase the bandwidth of the Fabry-Perot cavity. The choice of a proper nonlinear material is based upon the operating wavelength and the temporal response time desired, and possibly other considerations. Typically for applications in the far-infrared, near-infrared, and visible wavelengths, the proper materials are indium

antimonide, gallium arsenide, and zinc selenide, respectively. Although the response time of all bistable optical devices is limited by the quasi-steady-state nature of the phenomenon, it is also subject to the response time of the nonlinear mechanism that exists in the chosen material. Electronic nonlinear mechanisms such as those found in indium antimonide and gallium arsenide typically have response times in the subnanosecond regime, whereas thermal response times such as those used in zinc selenide interference filter-based bistable devices may be slower. Materials engineering holds the potential to modify the associated response times for various materials through bandgap engineering of devices fabricated by multiple-quantum-well technologies. See INTERFEROMETRY; QUANTIZED ELECTRONIC STRUCTURE (QUEST); SEMICONDUCTOR.

**Applications.** A large number of application schemes for bistable optical devices have been proposed. A major goal of optical bistability technology has been the implementation of optical logic gates. In spite of the fact that most basic logic circuits have been demonstrated experimentally by using optical bistability in some form, the technology has not been competitive with electronic semiconductor technologies. The primary reason is that packing densities for bistable logic circuits are limited by the diffraction of light waves. It is not possible to build circuits by using optical devices significantly smaller than the wavelength of the light being used. In gallium arsenide, for example, theoretically the smallest size of a given device is approximately  $0.5 \mu\text{m} \times 0.5 \mu\text{m}$ . In contrast, silicon electronic circuits are fabricated with a feature size of  $0.1\text{--}0.3 \mu\text{m}$ . In addition, a certain amount of energy is used by the device in order to cause it to switch from a low-transmission state to a high one, or vice versa. This energy, called the switching energy, is prohibitively large for the large packing densities that are called for in order to compete successfully with electronic technologies. Various computing architecture schemes have been proposed whereby the parallel nature of light propagation may be exploited to perform, in conjunction with bistable optical devices, highly parallel computations that are difficult to do. If these efforts are successful, they will lead the way for the most probable integration of optical computing technology, namely niche computations where some of the inherent properties of light-based circuits and devices may be exploited. See INTEGRATED CIRCUITS; INTEGRATED OPTICS; LOGIC CIRCUITS; OPTICAL INFORMATION SYSTEMS.

**Active devices.** In addition to passive bistable optical devices, it is possible to observe optical bistability in active devices such as laser systems. Although such systems are much more complicated than the passive devices discussed above, several experiments have been performed demonstrating optical bistability, and applications have been proposed. See NONLINEAR OPTICAL DEVICES; NONLINEAR OPTICS.

David R. Andersen

Bibliography. H. M. Gibbs, *Optical Bistability; Controlling Light with Light*, 1985; H.-Y. Zhang and

K. K. Lee (eds.), *Optical Bistability, Instability, and Optical Computing*, 1988.

## Optical coherence tomography

A noninvasive technique for imaging subsurface tissue structure with micrometer-scale resolution. The principles of time gating, optical sectioning, and optical heterodyning are combined to allow cross-sectional imaging. Depths of 1–2 mm (0.04–0.08 in.) can be imaged in turbid tissues such as skin or arteries; greater depths are possible in transparent tissues such as the eye.

Optical coherence tomography complements other imaging modalities commonly used to image subsurface tissue structure, including ultrasound and confocal microscopy. It has a resolution about an order of magnitude better than ultrasound, although the depth of imaging is less. Unlike ultrasound, it does not require a coupling medium between the instrument probe and tissue, facilitating endoscopic applications and imaging of sensitive structures such as the eye. Optical coherence tomography is a type of confocal system, although device parameters are usually set for a lower resolution than is available in commercial confocal microscopes. However, it has a depth of imaging several times that of confocal microscopy. See BIOMEDICAL ULTRASONICS; CONFOCAL MICROSCOPY; MEDICAL ULTRASONIC TOMOGRAPHY.

**Principles of operation.** In a typical optical coherence tomography system (Fig. 1), light from a broadband, near-infrared source and a visible aiming beam is combined and coupled into one branch of a fiber-optic Michelson interferometer. Broadband sources include superluminescent diodes, fiber amplifiers, and femtosecond pulse lasers in the wavelength range of 800–1550 nanometers. The light is split into two fibers using a  $2 \times 2$  coupler, one leading to a reference mirror and the second focused into the tissue. Light reflects off the reference mirror and is recoupled into the fiber leading to the mirror. Concurrently, light is reflected from index-of-refraction mismatches in the tissue and recoupled into the fiber leading to the tissue. Reflections result from changes in the index of refraction within the structure of the tissue, for instance between intercellular fluid and collagen fibers. Light that has been back-reflected from the tissue and light from the reference arm recombine within the  $2 \times 2$  coupler. See INTERFEROMETRY; OPTICAL FIBERS; OPTICAL PULSES.

Because the broadband source has a short coherence length, only light which has traveled very close to the same time (or optical path length) in the reference and tissue arms will interfere constructively and destructively. By changing the length of the reference arm, reflection sites at various depths in the tissue can be sampled. The depth resolution of the optical coherence tomography system is determined by the effectiveness of this time gating and hence is inversely proportional to the bandwidth of the source. An optical detector in the final arm of the Michelson interferometer detects the interference

between the reference and tissue signals. During optical coherence tomography imaging, the reference-arm mirror is scanned at a constant velocity, allowing depth scans (analogous to ultrasound A-scans) to be made. Either the tissue or the interferometer optics is mounted on a stage so that the beam can be scanned laterally across the tissue to build up two- and three-dimensional images, pixel by pixel. See COHERENCE.

The optical sectioning capability of optical coherence tomography (its ability to image thin slices of tissue) is determined by the tissue-arm optics. Light from the single-mode optical fiber is essentially a point source, which is focused to a small spot in the tissue. The highly turbid nature of many biological tissues tends to scatter light away from this focus, but the scattered light is not efficiently coupled back into the fiber. Therefore the lateral resolution of the optical coherence tomography system is approximately equal to the focused spot size. Often, the focus-spot diameter and lateral resolution are made approximately 10–20 micrometers to match the depth resolution of the system (using superluminescent diodes) and to provide a relatively long depth of focus. This long working distance is desirable for in vivo systems, since it is difficult to control the exact distance between the instrument optics and the tissue.

Optical coherence tomography has extremely high sensitivity; reflections of less than  $10^{-10}$  of the incident optical power can be detected by using the optical heterodyne technique. Movement of the reference arm mirror induces a modulation of the interferometric signal at the Doppler frequency. In the optical heterodyne detection technique, the envelope of the detector current is recorded by demodulating at this frequency, thus rejecting noise outside the signal bandwidth. The magnitude of the detected signal at the modulation frequency is proportional to the reflectivity of the tissue. See DOPPLER EFFECT; HETERODYNE PRINCIPLE.

**Color Doppler OCT.** Several instruments have been built based on variations of the basic optical coherence tomography system. For instance, polarization-sensitive optical coherence tomography uses polarization-altering optics in the arms of the interferometer to determine the sample birefringence from the magnitude of the back-reflected light. Optical coherence microscopy uses a system of high numerical aperture to achieve resolutions comparable to confocal microscopy but with increased depth of penetration.

Color Doppler optical coherence tomography (CDOCT) is an augmentation capable of simultaneous blood flow mapping and spatially resolved imaging. This technique (also called optical Doppler tomography) makes use of the interferometric phase information ignored in conventional optical coherence tomography. Doppler shifts in light backscattered from moving objects (such as red blood cells) add or subtract from the modulation frequency. If the returned signal is coherently demodulated, a short-time Fourier transform can be performed to find the signal strength at various Doppler-shift frequencies.

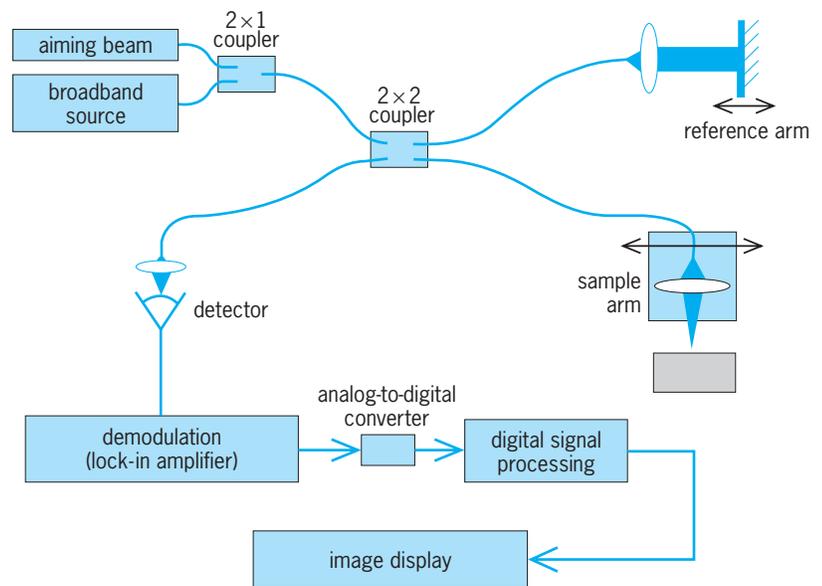


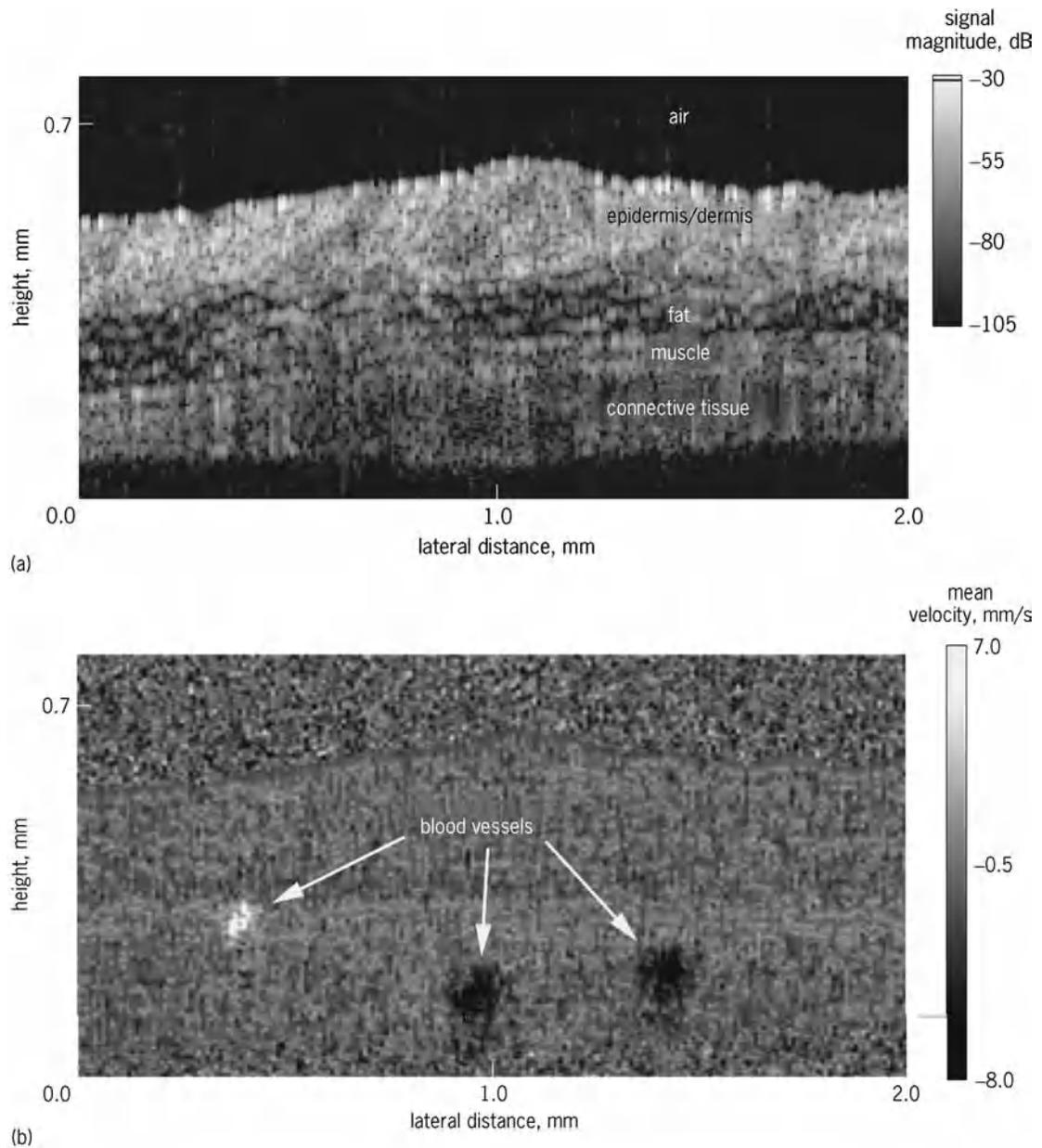
Fig. 1. Typical optical coherence tomography system, based on a broadband source and fiber-optic Michelson interferometer.

The mean velocity of the scatterers,  $\bar{V}_{is}$ , within each Fourier window can be estimated from the centroid,  $\bar{f}_s$ , of the localized Doppler frequency spectrum, according to the equation below.

$$\bar{V}_s = \frac{\bar{f}_s c}{2\nu_o n_t \cos \theta}$$

Here,  $c$  is the speed of light,  $\nu_o$  is the center frequency of the broadband light source,  $n_t$  is the local index of refraction, and  $\cos \theta$  is the angle between the instrument probe and the flow direction. The lateral resolution of a Doppler image remains the same as that of the magnitude image obtained in conventional optical coherence tomography; however, the depth resolution is often limited by the Fourier transform window size. A mean velocity can be calculated for each three-dimensional pixel in the image and combined to create a Doppler image separate and complementary to the magnitude image obtained from the amplitude of the signal alone. Alternatively, a threshold can be placed on the signals contributing to the Doppler image in order to eliminate noise, and the Doppler image can be overlaid on the magnitude image.

**Applications.** Optical coherence tomography can be used to probe the structure of any accessible tissue. Noninvasive studies of the eye and skin are being performed, and a commercial device has been developed for retinal imaging. Using optical coherence tomography, parameters such as eye length can be accurately measured, and the cross-section images of the retina give a clear and quantifiable assessment of retinal separation and macular degeneration, among other pathologies. In skin, the morphology of normal skin layers and components, and disorders such as psoriasis, can be imaged. For example, the expected layers of skin, including epidermis and dermis, fat, muscle, and connective tissue, are seen in a structure



**Fig. 2.** Optical coherence tomography images of healthy hamster skin in vivo. (a) Conventional magnitude image. Various layers of skin can be seen in cross section. (b) Result after Doppler processing of the same signal used to create the magnitude image. Three blood vessels are visible because of the Doppler shift caused by their moving red blood cells.

image of healthy hamster skin in vivo (Fig. 2a). Blood vessels located in the connective tissue and fat layers are not seen in this image, because blood and other skin tissues have similar optical properties in the near infrared. When the same signal is processed to distinguish Doppler shifts, however, the moving blood is readily apparent (Fig. 2b). See EYE (VERTEBRATE); EYE DISORDERS; SKIN; SKIN DISORDERS.

Endoscopes and catheters are used to facilitate imaging of the cardiovascular system and gastrointestinal, respiratory, urinary, and reproductive tracts. Being fiber-based, optical coherence tomography is readily adapted to these applications. Miniature electromechanical and piezoelectric scanning devices have been developed to perform lateral scanning of the interferometer tissue arm optics. In blood ves-

sels, the layered structures of the vessel wall are visible, and distinction can be made between fibrous and calcified plaques. Various mucous membranes of the body have also been successfully imaged. Differences in backscattering properties allow mucosal tissue layers to be differentiated. Cancerous regions are often marked by a disordering of the layered structure and a more homogeneous appearance of the optical coherence tomography image. Because of this distinction, optical coherence tomography has the potential to guide excisional biopsy and improve the random technique frequently used today. Samples of tissue would be taken from areas which appeared suspect on optical coherence tomography images, for standard histological analysis. See FIBER-OPTICS IMAGING.

**Potential.** Research groups are working on improving the performance of this novel imaging modality. Sources with broader bandwidths and higher powers will increase system resolution and depth of imaging. Signal processing techniques that efficiently extract parameters of interest and display data in an easily comprehended fashion are being developed. Efforts are under way to better understand optical coherence tomography images in terms of the morphology of the tissue under study. Jennifer Kehlet Barton

Bibliography. D. Huang et al., Optical coherence tomography, *Science*, 254:1178-1181, 1991; J. A. Izatt et al., Optical coherence tomography and microscopy in gastrointestinal tissues, *IEEE J. Selected Top. Quantum Electr.*, 2:1017-1028, 1997; X. J. Wang, T. E. Milner, and J. S. Nelson, Fluid flow velocity characterization by optical Doppler tomography, *Opt. Lett.*, 20:1337-1339, 1995.

## Optical communications

The transmission of speech, data, video, and other information using light in the frequency range below the infrared limit of the visible spectrum. Optical communications networks now perform this function. An optical communications link consists, at minimum, of a data or information source, a modulated laser transmitter, an optically transparent transmission medium, possibly an amplifier, and a photodetector receiver that recovers the transmitted data (Fig. 1). The predominant medium is optically transparent fiber; in this form, the technology is called fiber-optic communications. In another application, optical communications uses free space: free-space optical (FSO) communications.

Light by nature is an electromagnetic wave and it follows the laws of waves, such as refraction, diffraction, interference, and polarization. It is also a particle and exerts pressure; this is particularly noticeable when a satellite enters the Sun's light and starts drifting due to pressure exerted on it. The smallest quantity of light is known as a photon, which, according to atomic physics, is generated when an atom from an excited state makes a transition to either a less excited state or to its nonexcited (ground) state. The frequencies used in optical communications are much higher than those used for radio, television, and microwaves, and lower than those of visible light (Fig. 2). Thus, light used in optical communications is at frequencies below the infrared limit of visible light, and therefore it is invisible to the human eye. See DIFFRACTION; ELECTROMAGNETIC RADIATION; INFRARED RADIATION; INTERFERENCE OF WAVES; LIGHT; PHOTON; POLARIZATION OF WAVES; RADIATION PRESSURE; REFRACTION OF WAVES; WAVE MOTION.

The amount of information that a wave can carry depends on its frequency; the higher the frequency, the more the information. Thus, frequencies in the optical spectrum can carry orders-of-magnitude more information than radio and microwave frequencies. For example, the frequency of light is of the order of  $3 \times 10^8$  MHz as compared with radio and

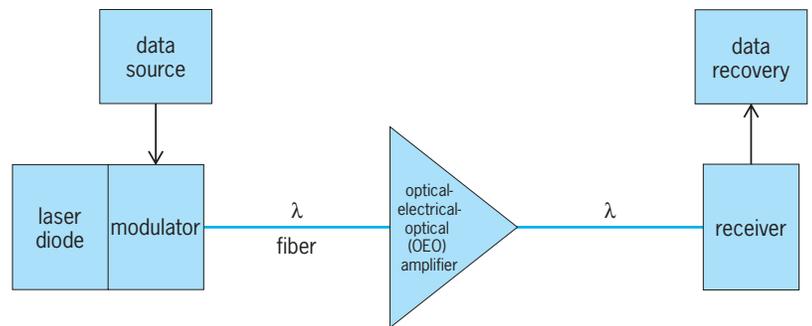


Fig. 1. Block diagram of a simplified optical communications link.

television frequencies that are in the kilohertz-to-megahertz range, and this is what makes optical communications so attractive. Therefore, optical communication is currently the preferred communications technology for very high capacity transmission, and will remain so for many years to come, since it is capable of delivering an unprecedented amount of information per second compared with other communications technologies. See BANDWIDTH REQUIREMENTS (COMMUNICATIONS); INFORMATION THEORY.

**Basic components.** For optical communications to be effective, certain basic components are required to construct the communications link. These, in addition to couplers, are a laser transmitter, a modulator, the optical fiber, and a photodetector receiver (Fig. 1).

**Laser transmitter.** The laser (light amplification by stimulated emission of radiation) used in optical communications is a semiconductor device. When a voltage is applied to it, a strong field is generated within it that excites specific atoms. When excited atoms are stimulated, they emit light at a specific frequency (or wavelength), which is determined by the material, its geometry, and the applied field. Light-emitting diodes (LEDs) are semiconductor devices that emit incoherent light. However, unlike LEDs, a laser radiates stimulated coherent light, that is, light in which all the stimulated emitted wavephotons are in phase. Typical lasers that are used in optical communications generate a continuous optical (yet invisible) beam, and are made with compound materials that consist of the elements indium, gallium, arsenic, and phosphorus. See COHERENCE; LASER; LIGHT-EMITTING DIODE.

**Modulator.** In order to carry information, the continuous optical beam that is generated by the laser must be modulated. That is, the characteristics of the laser light must be altered by impressing on it a stream of digital data or a string of ones and zeros. An optical device that modulates the continuous laser beam is known as a modulator. In optical communications, a modulator may be thought of a fast shutter that acts on the continuous beam. When the shutter opens it allows light to pass through, and when it closes it blocks light; that is, the optical shutter affects the optical power of the beam. This modulation method is known as on-off keying (OOK) and currently is the most used; other modulators affect the frequency or

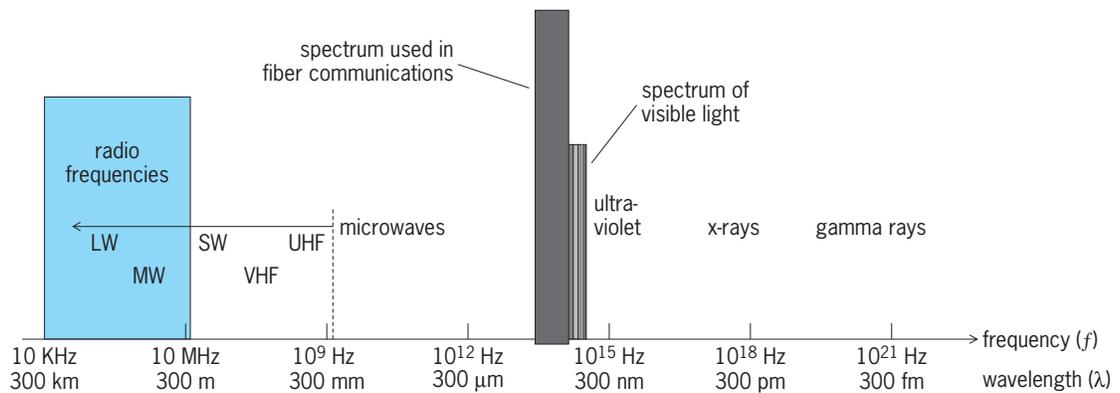


Fig. 2. Optical communications spectrum with respect to the overall electromagnetic spectrum. LW = long-wave, MW = medium-wave, SW = short-wave, VHF = very high frequency, UHF = ultrahigh frequency.

the phase of the laser beam. When laser light passes through the modulator or is blocked for a predetermined period, the period of the optical bit is established. Thus, in 1 second many optical bits may be created by the shutter. Current modulator rates generate up to  $4 \times 10^{10}$  bits per second, or 40 gigabits per second (Gbps). Depending on the technology and modulation rate, the modulator may be external to the laser device or be incorporated internally with it. In low-cost optical communications, it is preferable to incorporate the modulator with the laser device to reduce the number of components and simplify design and manufacturing. See MODULATION; MODULATOR.

**Optical fiber.** The fiber used in optical communications consists of two silica layers, the core and the cladding. What distinguishes the two is the refractive index. The core includes elements, known as dopants, such that its refractive index is higher than the refractive index of the cladding. This refractive index variation helps to keep light within the core in which it is transmitted. The cladding is surrounded by nonsilica layers that are added for strength and for fiber-type identification. Based on geometry, there are two major fiber classifications, the multimode fiber (MMF) and the single-mode fiber (SMF). The multimode fiber has a core diameter of 50 or 62.5 μm surrounded by the cladding layer to a total diameter of 125 μm. The single-mode fiber has a core diameter of only 9 μm surrounded by the cladding layer to a total diameter of 125 μm. In addition to the classification of optical fibers as single-mode and multimode, different types of fiber are produced, such as multimode fiber with a graded refractive index profile in its core, known as GRIN (graded-index), and single-mode fiber with inverted dispersion properties, known as dispersion compensating fiber (DSF). DSF plays a compensating role in the propagation of optical pulses in standard single-mode fiber, since single-mode fiber disperses different frequencies unequally, thus degrading the optical propagation of data at very high rates. See OPTICAL FIBERS.

**Photodetector receiver.** Photodetectors are diode semiconductor devices that generate one or more electron-hole pairs for each photon that enters the biased *p-n* junction. When modulated light impinges

on the photodetector device, the generated electron-hole pairs constitute an electric current known as photocurrent. The generated photocurrent emulates the optical digital signal to an electrical digital signal, which is passed through filters and a resistive load to convert it to voltage pulses. The two photodetector types used predominantly in optical communications are, the *p*-intrinsic-*n* (PIN) diode and the avalanche photodiode (APD). See OPTICAL DETECTORS; PHOTOVOLTAIC CELL; SEMICONDUCTOR DIODE.

PIN photodiodes consist of a lightly doped (intrinsic) region, which is sandwiched between *p*- and *n*-type semiconductor materials. A reverse-biased PIN photodetector exhibits almost infinite internal impedance with an output current proportional to the input optical power. However, capacitance in the reversed-biased PIN limits its switching speed. In addition, as switching speed increases, parasitic inductance becomes significant and causes shot noise, in addition to the dark noise that is generated due to the temperature of the material. Typically, PIN photodetectors are made with indium gallium arsenide. See ELECTRICAL NOISE; MICROWAVE SOLID-STATE DEVICES; PHOTODIODE.

When avalanche photodiodes are reverse-biased, a strong field is developed in the junction region. A photon entering this region generates an electron-hole pair, which, because of the strong field, gains enough energy to generate secondary electron-hole pairs, which generate tertiary pairs, and so on. Thus, this process causes an avalanche, and from a few photons entering the photodetector a substantial photocurrent is generated. However, avalanche photodiodes generate shot noise which is multiplied by the avalanche process as well. Thus, which photodetector type will be used in a particular network application depends on the application requirements that meet the characteristics of the photodetector. Among the key photodetector characteristics, the number of electron-hole pairs produced for each photon determines the efficiency and sensitivity of the photodetector, and this number depends on the material used. The speed of the electron-hole pairs determines how fast a photodetector responds to modulated light. The amount and type of noise

produced determines the expected quality of the signal. Noise is always present and cannot be eliminated or ignored.

The amount of photocurrent due to the optical signal compared with the amount of noise (shot or dark current) generated by the photodetector and associated electronic devices is a measure of the quality of the receiver. This ratio of signal photocurrent to noise, known as the signal-to-noise ratio (SNR), determines whether the original digital signal can be recovered faithfully or not. In practice, the SNR and other factors determine if all the bits that have been received are a faithful reproduction of all the bits transmitted. Calculating or estimating the number of bits in error in 1 billion or trillion bits transmitted for a given data rate provides an important performance parameter in communications, known as the bit error rate (BER). In optical communications, one error bit in  $10^{12}$  bits transmitted is a typical requirement (that is, a BER of  $10^{-12}$ ). See SIGNAL-TO-NOISE RATIO.

**Wavelength-division multiplexing.** Initial optical communications networks used a single wavelength at 1310 nm as the carrier frequency, and at modulation rates 2.5 Gbps or less. Soon thereafter, a second wavelength at 1550 nm was added that was suitable for longer fiber lengths. The protocol used was called synchronous optical network (SONET) in the United States, and synchronous digital hierarchy (SDH) in Europe. SONET and SDH, although very similar, differ enough, as a result of specific local needs for services and demographics, to create two different specification standards. As bandwidth requirements increased, another modulation rate was added at 10 Gbps, and more recently another at 40 Gbps.

Despite this incremental and dramatic increase in data rate, bandwidth demand exceeded bandwidth deliverability by a single fiber; this situation is known as capacity exhaust. To meet the ever-increasing bandwidth demand, a need arose for additional fiber, which is very costly. However, at about the same time (in the early 1990s) optical and photonic developments came to fruition that could be applied cost-effectively to optical communications. These developments allowed more than one optical wavelength to be multiplexed in the same fiber, giving birth to a technology called wavelength-division multiplexing (WDM). Providing a multiplicity of wavelengths, or optical channels, in a single fiber, solved the capacity exhaust problem and alleviated the need for additional fibers and networks. See MULTIPLEXING AND MULTIPLE ACCESS.

Before WDM technology was deployed, the fiber-optic communications frequency spectrum and the optical channels needed to be defined and standardized for interoperability reasons. The low-loss optical spectrum of silica fiber was subdivided into subbands known as O, E, S, C, and L (see table; Fig. 3). Initially, 80 narrow spectral ranges were specified in the C subband and an additional 80 in the L subband, with a separation between center frequencies of adjacent channels, known as channel separation

Subbands of the low-loss optical spectrum of silica fiber			
Label	Range, nm	Fiber type*	Applications*
O	1260–1360	SMF	Future DWDM
E	1360–1460	SMF	Future DWDM
S	1460–1530	SMF	Future DWDM
C	1530–1565	SMF	DWDM
L	1565–1625	DSF	DWDM

\* SMF = single-mode fiber, DSF = dispersion compensating fiber, DWDM = dense wavelength-division multiplexing.

of 50 GHz. The starting reference frequency, from which all other frequencies are calculated, was selected to be a standard frequency at 193.1 terahertz (THz). Since the original recommendation standard, the channel separation has been divided in half to define twice as many optical channels: 160 channels in the C subband and 160 channels in the L. As optical technology evolves and optical communications demand more bandwidth, the number of frequencies may be expanded accordingly.

Figure 3 shows that at wavelengths around 1400 nm the glassy fiber exhibits high attenuation; this is because hydroxyl (O-H) radicals that are trapped in the fiber during the manufacturing process absorb the energy of photons with these wavelengths as they travel in the fiber. However, new manufacturing methods have produced a type of fiber, known as water-free fiber, with low attenuation in this range. If the low-loss spectrum, from below 1300 to over 1600 nm or from O to L, is used, the potential number of channels can be of the order of 1000. If 1000 optical channels are coupled into a single optical fiber, and if each channel is modulated at 10 or 40 Gbps, then the aggregate amount of information transportable in a single fiber is enormous. (For comparison, a digitized uncompressed voice channel requires 64 kilobits per second, and compressed voice less than 4 kbps.) The technology that makes this possible is known as dense wavelength-division multiplexing (DWDM). See HYDROXYL.

Currently, such technology has many manufacturing challenges and is costly; it is used only in high-capacity networks and in long-haul systems. In certain communications applications, a more relaxed and inexpensive optical technology was needed. To address this need, WDM communication standards have defined 18 optical channels with 20-nm

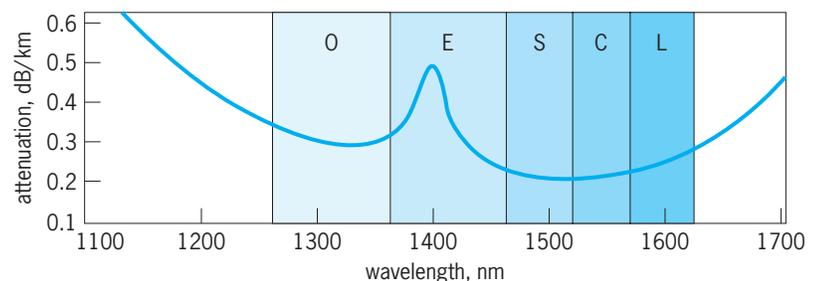


Fig. 3. Low-loss spectrum of silica fiber used in fiber-optic wavelength-division multiplexing communications. The C-band is at a frequency of about 1550 nm (1.55 μm), and the L-band is at about 1600 nm (1.6 μm).

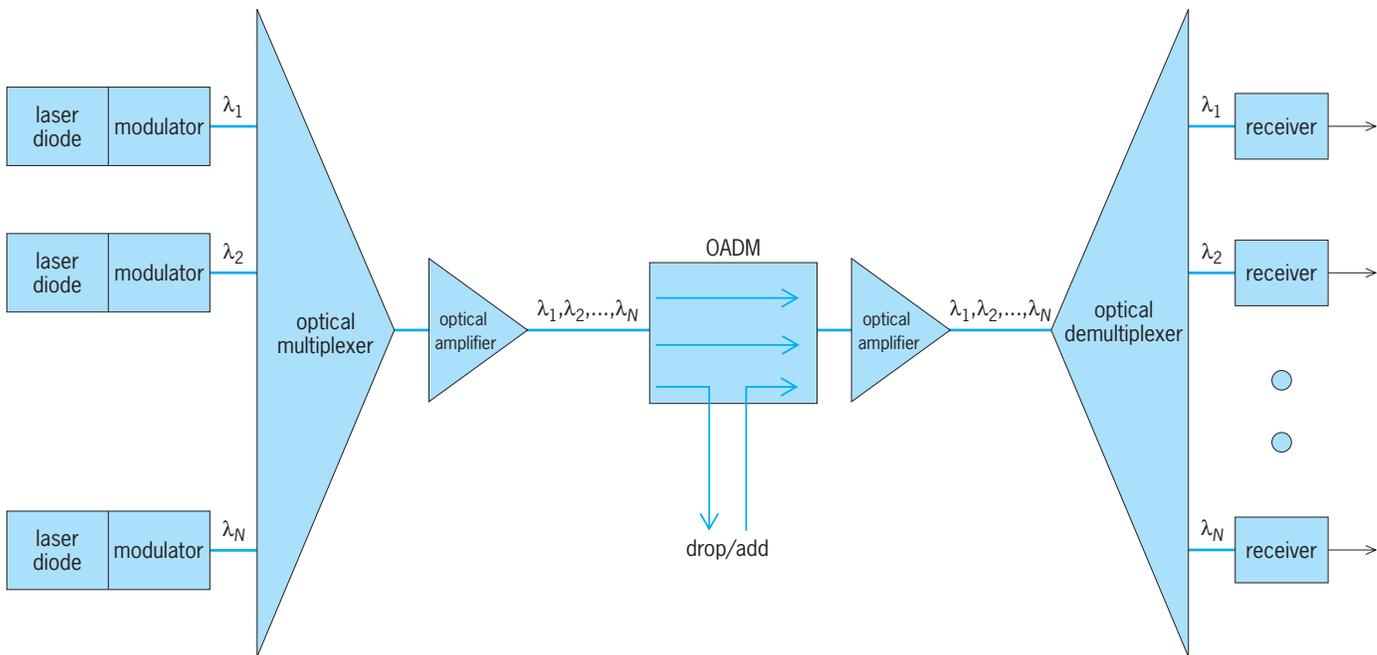


Fig. 4. Point-to-point wavelength-division multiplexing (WDM) link. Shown are laser transmitters, modulators, multiplexers and demultiplexers, and other optical components that make WDM work.

separation (called coarse channels) over the complete low-loss spectrum. This technology, known as coarse wavelength-division multiplexing (CWDM), uses the aforementioned water-free fiber with low loss over the complete spectrum from O to L, including the E subband; CWDM is not shown in the table.

**WDM technology.** The optical technology that was required in basic optical communications (laser, fiber, and photodiode) was now enhanced with more components to make WDM possible. Key WDM components are optical multiplexers, optical demultiplexers, and optical amplifiers. Other components include couplers, splitters, filters, equalizers, compensators, and optical add-drop multiplexers.

*Optical multiplexers and demultiplexers.* An optical multiplexer combines many optical channels of different wavelengths into one beam coupled onto the core of a single fiber; an optical demultiplexer separates the WDM beam into its constituent channels; and the optical add-drop multiplexer (OADM) selectively removes one of many channels from one direction and adds the same channel in the same direction but with different data on it (Fig. 4). The distance between the optical multiplexer and demultiplexer is of the order of tens to hundreds of kilometers. Therefore, in many applications there is a need to add and drop traffic (that is, one or more channels) in between, and the OADM accomplishes this task. For example, if New York City is connected with Washington, DC, via Philadelphia, with a point-to-point DWDM link, it may be necessary to drop some traffic in Philadelphia (for example, servicing calls in which one of the parties is in the Philadelphia area). The one-way communication in Fig. 4 is for illustrative purposes; in reality, the same link exists in the opposite direction. Thus, the OADM must handle adding and dropping traffic

in both directions, and it is therefore more complex than that shown in Fig. 4.

*Optical amplifiers.* The optical amplifier is particularly important in WDM optical communications. It is the optical amplifier that makes WDM optical networks a reality. Optical amplification in a long fiber link is needed to overcome optical signal power loss. Thus, if the signal is amplified every 60–80 km (40–50 mi), it can travel distances that exceed 1000 km (600 mi). In the original optical network with a single channel at 1310 nm, signal amplification was not a major issue because a single regenerator that consisted of a photodetector, an electronic amplifier, electrical filters, and a laser would do; this is known as opaque or optical-electrical-optical (OEO) amplification (Fig. 1). However, in WDM this method is not cost-efficient because many optical channels must be demultiplexed first, then each channel must be amplified by a separate OEO amplifier, and then the channels must be multiplexed and coupled onto the fiber (Fig. 5). Thus, the overall cost of amplification is multiplied by the number of channels in the WDM signal. The amplification stage is too complex to be practical; it requires large amounts of space and power because of the hundreds of amplifiers, and maintenance cost is too high. Fortunately, an optical amplifier was developed that simultaneously could amplify all optical channels in a specific subband directly without electronic amplifiers. This amplifier is known as an erbium-doped fiber amplifier (EDFA).

An EDFA is a specialty fiber whose core contains a high concentration of erbium atoms or erbium dopant. The energy of erbium atoms is excited by absorbing optical radiation at a wavelength of 980 or 1480 nm. The excitation frequency in optical amplification is known as the pump. When a photon with a wavelength in the range from 1520 to

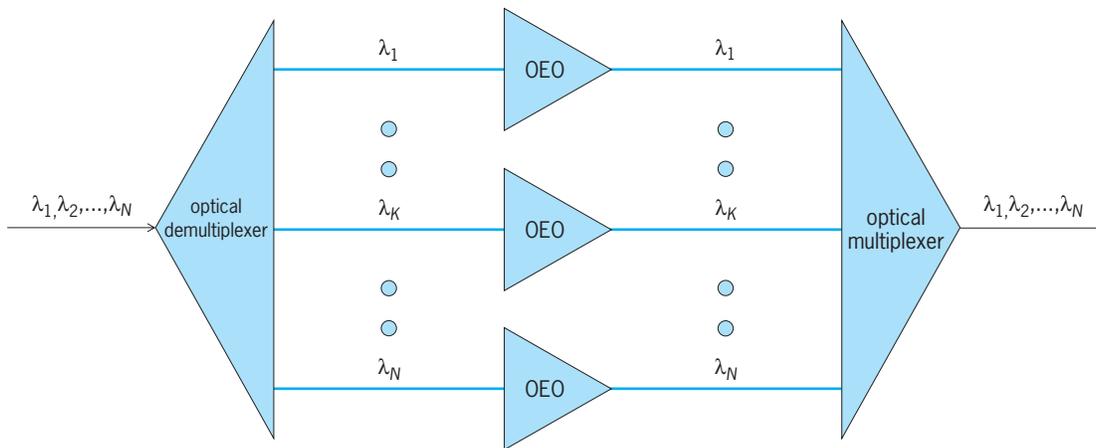


Fig. 5. Use of optical-electrical-optical (OEO) amplifiers in a WDM link. This method is complex and cost-inefficient.

1570 nm (that is, the C-band) passes near an excited erbium atom in the EDFA, it stimulates the atom, and the excited atom emits a photon at the same frequency and phase as that of the stimulating photon; this atom makes a transition to a lower energy level, where it is reexcited by the pump. When the two propagating photons encounter more excited erbium atoms, more photons are emitted. Thus, from one photon many photons are generated at the same frequency and phase, which is manifested as optical amplification.

If two photons with different frequencies A and B, but still in the C-band, enter the excited EDFA, then two excited atoms are stimulated. However, now one atom produces a photon with frequency A and the other atom a photon with frequency B, and this process continues as these photons propagate within the excited EDFA. Thus, two different signals are amplified simultaneously. This example can be extended to more photons of different frequencies, as long as they are within the C-band. When one or more optical signals in the C-band pass through an excited EDFA, they are all amplified. By adding more elements in the core of the fiber, such as aluminum and germanium, the amplification range of EDFAs was extended to include both the C and L subbands. Such fiber-optical amplifiers have been termed extended-range EDFAs.

Thus, an EDFA is a fiber, a few meters long, in which the weak information-carrying WDM signal and the laser pump are coupled and propagate. As they propagate, an evolutionary amplification process takes place (Fig. 6). At the end of the EDFA, an optical filter rejects any of the residual pump light, allowing only the amplified WDM signal to continue its travel over the next single-mode fiber link for another 60–80 km (40–50 mi) to the next EDFA. However, the EDFAs produce noise, known as amplifier stimulated emission (ASE), which blends with the WDM signal and cannot be eliminated. This ASE noise is cumulative, and as the signal is amplified by several EDFAs the noise content in the signal increases and degrades the signal-to-noise ratio. This degraded SNR imposes a limit on the number of concatenated

EDFAs, and thus a limit on the number of fiber links and on overall fiber length. Thus, in paths that are several thousands of kilometers long, such as transoceanic fiber cables, a regenerator amplifier is placed after every 7 to 10 concatenated EDFAs to filter out noise and reconstitute the signal to its original quality.

Another type of fiber amplifier, known as the Raman amplifier, takes advantage of a phenomenon that is based on the excitation of atoms in the single-mode fiber due to the residual nonlinearity of its dielectric material. When a powerful laser light of a specific wavelength (or frequency) propagates in a single-mode fiber, atoms are excited. Then, if photons at a wavelength that are offset by about 80 nm from the pump wavelength enter the excited fiber, they stimulate the excited atoms. The stimulated atoms emit photons that have the same wavelength as the stimulating photons. Thus, Raman amplification supports signals that are offset about 80 nm from the pump wavelength and within a range of about 40 nm (Fig. 7). Unlike EDFAs, which have a fixed usable spectral bandwidth, Raman amplifiers can be used in any spectral subband as long as the Raman pump is offset about 80 nm from the signals to be amplified and the signals are within a 40-nm spectral range. Consequently, several Raman pumps at convenient wavelengths can amplify many signals in a wider spectral band. See NONLINEAR OPTICS; RAMAN EFFECT.

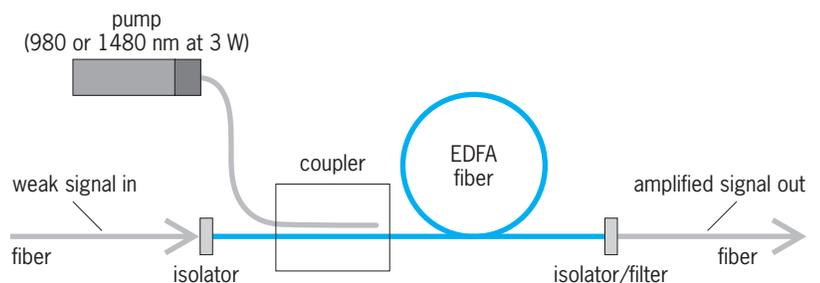


Fig. 6. Fundamental components of an erbium-doped fiber amplifier (EDFA) in a WDM system.

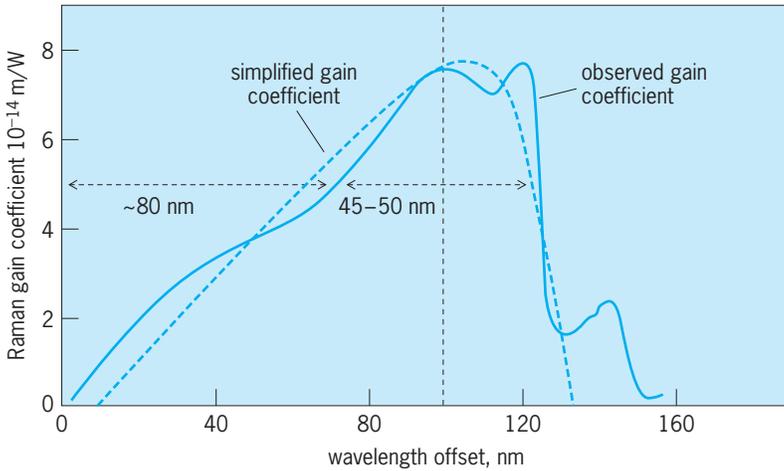


Fig. 7. Typical Raman gain coefficient over the amplification spectrum. In silica fiber at a wavelength of 1550 nm, the coefficient is  $6.6 \times 10^{-14}$  m/W.

In addition to the EDFA, there are other types of fiber amplifiers that contain different dopants, such as tellurium (TDFAs), praseodymium (PDFAs), and yttrium (YDFAs). Each has a useful amplification range which is also fixed but is in a different frequency subband than the EDFA. Because DWDM technology currently uses the C and L subbands, these fiber amplifiers have not been widely used yet. As DWDM evolves to include the remaining frequency subbands, the applicability and popularity of these fiber amplifiers will also increase.

**Optical switching and routing.** Optical networks consist of a mesh of nodes that are interconnected with optical fiber. Nodes receive traffic from one fiber and, based on the predetermined route, they pass it onto another fiber that connects with the next node in the mesh, and so on. This is known as a mesh topology, which is used in the optical networks that connect cities across the United States, also known as the backbone.

To be able to pass traffic from one fiber to another, a node must support an optical switching function. Several technologies provide this function. One such technology uses lithium niobate ( $\text{LiNbO}_3$ ) solid-state devices that, by the application of a voltage, couple an optical signal from one waveguide to another, similar to a train switching tracks. This

technology is fast but attenuates the signal and is not suitable for very large switches. Another technology uses micromirrors that have been manufactured using a standard photolithography and etching process similar to that used to make integrated circuits. This technology is also known as micro-electro-mechanical systems (MEMS). Tilting the mirrors by applying an electrostatic field will redirect the optical beam in the desired direction, and thus a switching function for many optical channels is accomplished (Fig. 8). Multiple micromirrors are manufactured in array or matrix configurations; matrices of up to  $1000 \times 1000$  micromirrors have been demonstrated. See ELECTROOPTICS; MICRO-ELECTRO-MECHANICAL SYSTEMS (MEMS); MICRO-OPTO-ELECTRO-MECHANICAL SYSTEMS (MOEMS); OPTICAL MODULATORS.

In current large-mesh optical networks, an end-to-end path is established by routing traffic from node to node using the same optical channel or wavelength. In many networks, this means that the network management system (NMS), that is, the system that overlooks and manages traffic and the operation of the entire network, finds the best route and establishes the end-to-end path. This is accomplished by executing a routing and wavelength assignment (RWA) algorithm and sending commands to nodes in the networks via a protocol; this is also known as path establishment. In practice, the NMS communicates with the element management system (EMS) of each subnetwork, and each EMS then communicates with the network elements (NE) or nodes that are within its responsibility in the subnetwork (Fig. 9). Because the number of wavelengths per fiber is finite as well as the number of nodes and links in a mesh network, there is a finite number of paths. Thus, during periods of high demand, when all possible paths have been assigned, any additional path connectivity request will be blocked.

Networks with systems that support wavelength conversion reduce the blocking probability. Wavelength converters are devices that shift an optical wavelength from that of one channel to that of another. Thus, if it happens that the same wavelength is not available over an end-to-end path, but different wavelengths are available in segments and links over the intended path, then wavelength converters

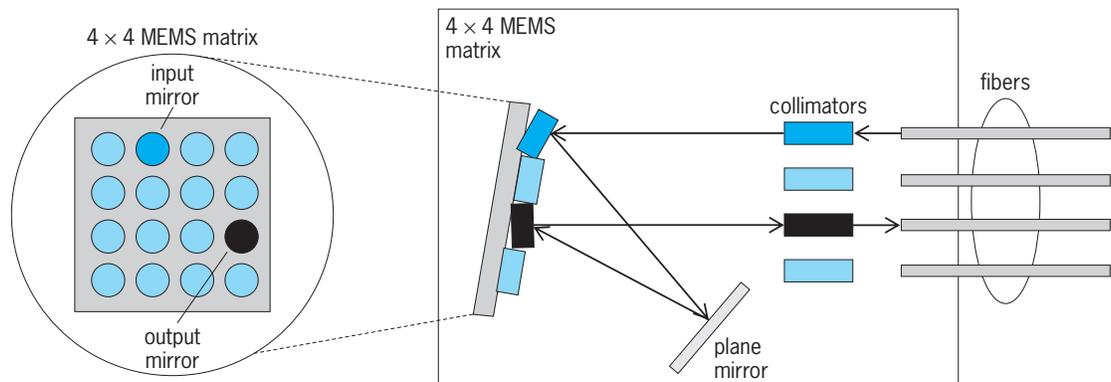


Fig. 8. Optical switch with  $4 \times 4$  MEMS micromirror matrix.

“stitch” different wavelengths together, and the path is established. In this case, the complete path may be envisioned as a colorful array of different wavelengths stitched one after the other; this is known as path establishment with wavelength concatenation (Fig. 9). Although path establishment with wavelength concatenation is more complex and requires protocols and wavelength conversion tables at each node, it greatly decreases the probability for blocking and increases the network bandwidth throughput and bandwidth efficiency of the network.

**Protection.** As in previous communication networks, the network protection strategy in optical networks is very important. The mesh topology has excellent protection because, if one fiber is cut (fiber cuts happen frequently when digging roads and fields through which fiber is laid), then the switching nodes detect the fiber failure and, by using sophisticated algorithms and control messages, find another route on to which to move the affected traffic. This is similar to moving traffic in a city to another road when a particular road is blocked, but in this case the process is more orderly and takes place with minimal congestion. Another topology of interest is the ring, which is used in many metropolitan and large campus applications, and is thus named the Metro ring. The Metro ring consists of a single fiber, a dual fiber, or a quad fiber. Optical add-drop nodes are placed on the ring to remove and add traffic to be sent to its destination. Clearly, a single ring does not exhibit good protection against fiber cuts unless it is a bidirectional ring; that is, traffic flows in both directions of the ring. The dual-ring Metro has much better protection. In this case, when one fiber ring is cut, traffic from the affected ring is passed onto the healthy one and thus service continues. In the case of a quad fiber, a combination of switching and loopback mechanisms provides superb protection (Fig. 10); thus, the quad-ring Metro is preferred

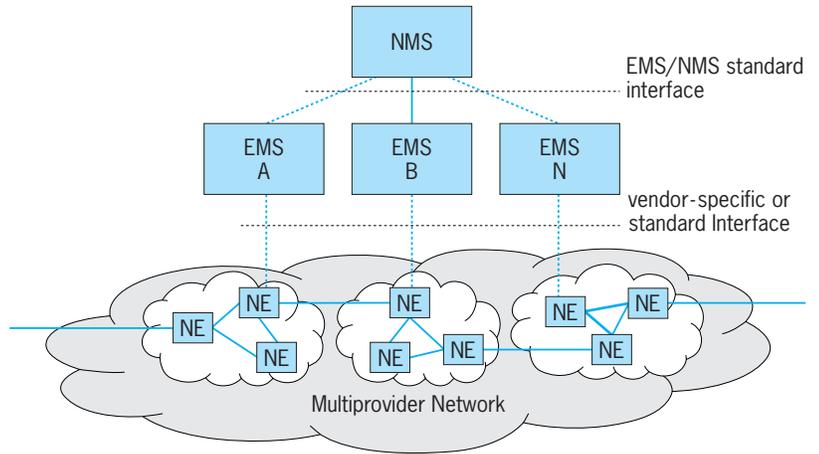


Fig. 9. Large optical network with wavelength converters. Such a network is able to concatenate wavelengths to establish an efficient end-to-end path. NMS = network management system, EMS = element management system, NE = network element.

in large metropolitan applications that support ultrahigh bandwidth transmission of the order of 400 Gbps or more.

**Free-space optical (FSO) communications.** In several applications, a relatively quick point-to-point path must be established in the inner city between two tall buildings. Installing fiber in a busy city is a rather lengthy process because over 1 or 2 km distance many rights of way must be obtained. However, shooting a laser beam from the top of one building to the top of another building that is 1 or 2 km away is much easier, and it does not require rights of way licences. In fact, such a network, using towers and the light of a torch, has been known since antiquity.

Although FSO is quick to deploy, it has a severe drawback. Laser light is severely absorbed by fog. Thus, in cities where fog is frequent, FSO may not be suitable unless an alternative transmission method

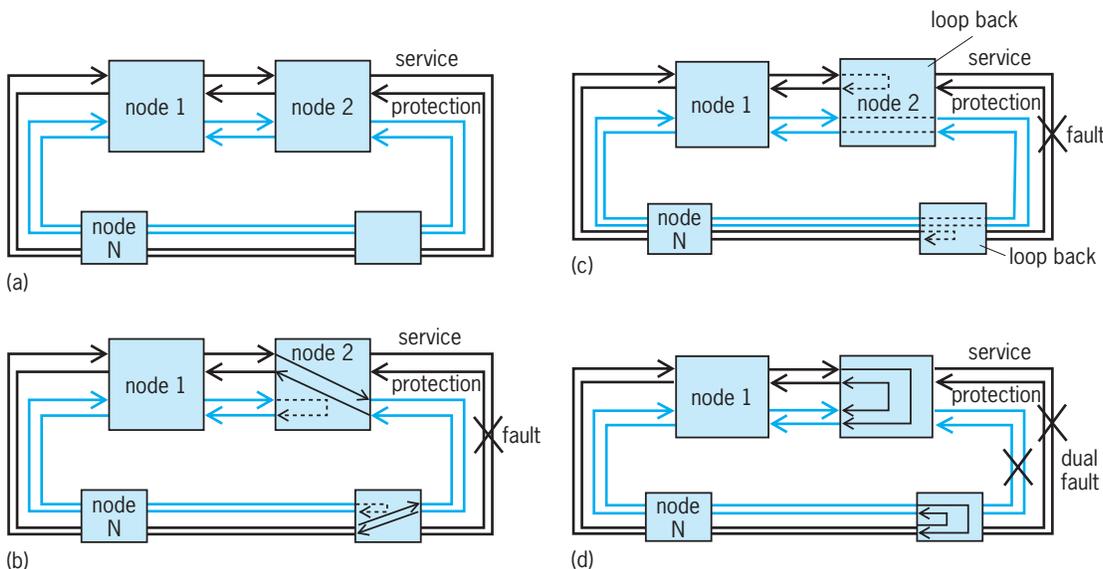


Fig. 10. Large optical four-fiber Metro ring, with protection, loopback, and path switching. (a) System operating with no faults. (b) System using path switching to operate in the presence of a single fault. (c) System using loopback to operate in the presence of a single fault. (d) System using loopback to operate in the presence of a dual fault.

can be used, such as microwaves, which are not substantially affected by fog. On the other hand, some microwave frequencies are affected by rain, which does not affect FSO as much. Thus, the two technologies can work side by side, and in case of fog the microwave link becomes operable, but not at the same data rate. [FSO transmits more than 1 Gbps whereas a microwave link transmits several megabits per second (Mbps).] See MICROWAVE; RADIO-WAVE PROPAGATION.

FSO technology may also include multiple wavelengths, thus utilizing WDM technology. WDM in this case may reduce the data rate per channel, yet transport more aggregate bandwidth over many channels.

In another application, laser beams may be used as intersatellite links. Satellites form a network of their own, communicating with each other via the laser beams. The neodymium yttrium-aluminum-garnet (Nd:YAG) solid-state laser is suited to this application. Satellites are thousands of kilometers apart, but light can travel such distances in space since the attenuation of light is insignificant there. The key issue is to maintain connectivity as the satellites move, which is a tracking problem and not an optical communications one. The optical issue is the divergence of the laser beam. Although the beam is very narrow at the transmitter, it diverges as it propagates for thousands of kilometers. Because of this divergence, a large-aperture telescope is needed at the receiver, to obtain an adequate signal-to-noise ratio, typically about 10 cm (4 in.) in diameter. While millimeter-wave systems compete with laser beams, laser beams can transfer much higher data rates, exceeding gigabits per second. This is what makes them attractive for satellite communications networks. Such intrasatellite optical communications links are used in the Iridium mobile communications satellite system. See COMMUNICATIONS SATELLITE; MOBILE RADIO.

**Next-generation optical networks.** The initial optical network has grown rapidly since the 1980s, responding chiefly to the rapid growth of voice and data services, as was testified by the initial offering of SONET and SDH protocols. During this evolution, however, electronic devices kept shrinking so that the personal computer became a powerful and inexpensive machine that was ubiquitous, as was mobile telephony. The rapid expansion of the Internet and its services, and also the rapid deployment of mobile and other wireless devices, gave rise to a need for more bandwidth. Thus, on the aggregate, the optical communications network created a bottleneck at access points, which for one reason or another did not follow in step with the evolution of the main optical network. See INTERNET; MOBILE RADIO.

However, this situation is changing, and fiber-optic communications are reaching user premises with a passive optical network (PON) technology known as fiber to the premises (FTTP). When this technology is established, the main network is expected to be flooded with all types of traffic, including voice, Ethernet, Internet, video, interactive video, text, games, and music, and all these signals are expected to be integrated over a single optical network. The

next-generation network is thus already at hand, and new efficient protocols, known as next-generation SONET/SDH, have been developed that are able to encapsulate a diverse range of traffic onto a single transmission path. See DATA COMMUNICATIONS.

For very long haul applications over distances of the order of several thousand of kilometers, the optical transport network (OTN) has been defined with superb bit error detection and correction capability. This capability is necessary to reach such long distances, such as in the case of trans-Atlantic links, where there is no land between the east coast of the United States and the west coast of Europe. On such links, the number of amplifiers is limited as is the number of repairs (to 3 per 25 years) since repairing an undersea cable is an extremely expensive and complex operation. Among the protocols that make this capability possible are the generic framing procedure (GFP), the link access procedure SDH (LAPS), and link capacity adjustment scheme (LCAS). Both the next-generation SONET/SDH and OTN are now defined over WDM technology, and thus the next-generation network will be ready to meet the most aggressive bandwidth demands for many years to come in order to provide reliable, cost-effective, scalable, and secure service. See SUBMARINE CABLE.

S. V. Kartalopoulos

**Bibliography.** A. Borella, G. Cancellieri, and F. Chiaraluce, *Wavelength Division Multiple Access Optical Networks*, Artech House, 1998; E. B. Carne, *Telecommunications Primer*, 2d ed., Prentice Hall, 1995; I. P. Kaminow and T. L. Koch (eds.), *Optical Fiber Communications IIIA and Optical Fiber Communications IIIB*, Academic Press, 1997; S. V. Kartalopoulos, *DWDM: Networks, Devices and Technology*, Wiley-IEEE Press, 2003; S. V. Kartalopoulos, *Fault Detectability in DWDM Systems*, Wiley-IEEE Press, 2001; S. V. Kartalopoulos, *Introduction to DWDM Technology*, Wiley-IEEE Press, 2000; S. V. Kartalopoulos, *Next Generation SONET/SDH*, Wiley-IEEE Press, 2004; S. V. Kartalopoulos, *Optical Bit Error Rate*, Wiley-IEEE Press, 2004; S. V. Kartalopoulos, *Understanding SONET/SDH and ATM*, Wiley-IEEE Press, 1999; J. Nellist, *Understanding Telecommunications and Lightwave Systems*, 2d ed., IEEE Press, 1996; J. C. Palais, *Fiber Optic Communications*, 5th ed., Prentice Hall, 2004; R. Ramaswami and K. N. Sivarajan, *Optical Networks*, 2d ed., Academic Press, 2002.

## Optical detectors

Devices that respond to incident ultraviolet, visible, or infrared electromagnetic radiation by giving rise to an output signal, usually electrical. Based upon the manner of their interaction with radiation, they fall into three categories. Photon detectors are those in which incident photons change the number of free carriers (electrons or holes) in a semiconductor (internal photoeffect) or cause the emission of free electrons from the surface of a metal or semiconductor (external photoeffect, photoemission). Thermal

detectors respond to the temperature rise of the detecting material due to the absorption of radiation, by changing some property of the material such as its electrical resistance. Detectors based upon wave-interaction effects exploit the wavelike nature of electromagnetic radiation, for example by mixing the electric-field vectors of two coherent sources of radiation to generate sum and difference optical frequencies. This discussion concentrates on photon and thermal detectors. *See* NONLINEAR OPTICAL DEVICES.

**Photon effects.** The most widely used photon effects are photoconductivity, the photovoltaic effect, and the photoemissive effect. Photoconductivity, an internal photon effect, is the decrease in electrical resistance of a semiconductor caused by the increased numbers of free carriers produced by the absorbed radiation. The change in resistance is measured by passing a bias current through the photoconductor and measuring the change in resistance between the illuminated and dark state. *See* PHOTOCONDUCTIVE CELL.

The photovoltaic effect, also an internal photoeffect, occurs at a *pn* junction in a semiconductor or at a metal-semiconductor interface (Schottky barrier). Absorbed radiation produces free hole-electron pairs which are separated by the potential barrier at the *pn* junction or Schottky barrier, thereby giving rise to a photovoltage. This is the principle employed in a solar cell. *See* PHOTODIODE; PHOTOVOLTAIC CELL; PHOTOVOLTAIC EFFECT; SEMICONDUCTOR DIODE; SOLAR CELL.

The photoemissive effect, also known as the external photoeffect, is the emission of an electron from the surface of a metal or semiconductor (cathode) into a vacuum or gas due to the absorption of a photon by the cathode. The photocurrent is collected by a positively biased anode. Internal amplification of the photoexcited electron current can be achieved by means of secondary electron emission at internal structures (dynodes). Such a vacuum tube is known as a photomultiplier. Internal amplification by means of an avalanche effect in a gas is employed in a Geiger tube. *See* GEIGER-MÜLLER COUNTER; LIGHT AMPLIFIER; PHOTOELECTRIC DEVICES; PHOTOEMISSION; PHOTOMULTIPLIER; PHOTOTUBE.

**Role of materials.** Semiconductors are key to the development of most photon detectors. These materials are characterized by a forbidden energy gap which determines the minimum energy that a photon must have to produce a free hole-electron pair in an intrinsic photoeffect. Since the energy of a photon is inversely proportional to its wavelength, the minimum energy requirement establishes a long-wavelength limit of an intrinsic photoeffect. It is also possible to produce free electrons or free holes by photoexcitation at donor or acceptor sites in the semiconductor; this is known as an extrinsic photoeffect. Here the long-wavelength limit of the photoeffect is determined by the minimum energy (ionization energy) required to photoexcite a free electron from a donor site or a free hole from an acceptor site. In a given semiconductor, the long-wavelength

limit of an extrinsic photoeffect exceeds that of the intrinsic photoeffect. In either the intrinsic or extrinsic photoeffect, the ability to detect a photon is described by the quantum efficiency, which is the ratio of the photoexcited carriers produced to the number of incident photons. Quantum efficiencies of 0.5 are typical of many photon detectors. *See* SEMICONDUCTOR.

*Detector cooling.* Free electrons and holes can also be thermally excited by lattice vibrations (phonons) in the semiconductor, whose magnitude is a function of the detector temperature. Detection of a few photoexcited free carriers in the presence of numerous thermally excited carriers is difficult. The problem is eased by cooling the detector in order to reduce the thermally excited carrier background. Long-wavelength infrared photon detectors having spectral responses extending to 10 micrometers have small forbidden energy gaps or small donor or acceptor ionization energies. Accordingly, cooling the detector becomes more important for long-wavelength detectors. Use of liquid nitrogen, which boils at 77 K, or mechanical refrigerators (cryocoolers) is required for photon detectors whose spectral response extends to 10  $\mu\text{m}$ . Even longer-wavelength spectral responses require cooling below 77 K. *See* CRYOGENICS.

*Synthetic materials.* The usual way to provide a given spectral response for a photon detector is to synthesize a semiconductor whose forbidden energy gap or donor or acceptor ionization energy meets the long-wavelength limit requirement. The semiconductor material most frequently used in long-wavelength military systems is mercury cadmium telluride, wherein the mercury-to-cadmium ratio determines the forbidden energy gap.

An alternative method is to tailor a new material by building up numerous alternating layers of controlled thickness of two semiconductors which have compatible material properties (crystal structure, lattice constant, thermal expansion coefficient). Such structures are known as multiple quantum wells. The most frequently used pair of materials is gallium arsenide and aluminum gallium arsenide. By adjusting the aluminum-to-gallium ratio and the thicknesses of the layers during growth, the spectral response of the photoeffect exhibited by these structures can be tailored within broad limits. When such structures are employed for infrared detection, they are known as quantum-well infrared photodetectors (QWIPs). *See* ARTIFICIALLY LAYERED STRUCTURES; CRYSTAL GROWTH; QUANTIZED ELECTRONIC STRUCTURE (QUEST); SEMICONDUCTOR HETEROSTRUCTURES.

**Thermal effects.** The choice of materials also plays a role in thermal detectors. The most widely used thermal detector is a bolometer, that is, a temperature-sensitive resistor in the form of a thin metallic or semiconductor film (although superconducting films are also used). Incident electromagnetic radiation absorbed by the film causes its temperature to rise, thereby changing its electrical resistance. The change in resistance is measured by

passing a current through the film and measuring the change in voltage. Materials with a high temperature coefficient of resistance are desired for bolometers, a criterion which usually favors semiconductors over metals. *See* BOLOMETER.

Another widely used thermal detector is based upon the pyroelectric effect. Here a change in temperature causes positive and negative charges to appear on opposite faces of thin samples of certain ferroelectric materials such as triglicine sulfate and strontium barium niobate. This charge can be detected by means of electrodes deposited on the faces. In contrast to a bolometer, a pyroelectric detector requires the incident radiation to be temporally modulated, which can be done by means of a radiation chopper. *See* CHOPPING; PYROELECTRICITY.

A third thermal detector is the radiation thermocouple, formed by two junctions between two dissimilar metals or semiconductors. One junction, exposed to the incident radiation, is warmed by it; the other, shielded junction is not. A voltage generated by the temperature difference between the junctions is detected by an external circuit. *See* THERMOCOUPLE; THERMOELECTRICITY.

In contrast to photon detectors, the spectral response of thermal detectors is determined by the wavelength dependence of their absorption of electromagnetic radiation. If the absorption is invariant with wavelength, the spectral response is invariant with wavelength and is said to be flat.

**Elemental and imaging detectors.** Optical detectors can also be classified as elemental or imaging. An elemental detector averages any spatial variation in the incident radiation to produce an output signal characteristic of the average incident intensity. An imaging detector produces a time-varying output signal which bears a one-to-one relationship with the spatial variation of intensity falling on it. An example is a television camera tube known as a vidicon, which contains a broad-area thin photoconductive film upon which the radiant image of a scene is projected by a lens. An internal electron beam scanned across the film in raster fashion generates a time-varying electrical signal representing the spatial variation of film resistance due to the projected image. It is also possible to detect an image by mechanically scanning the scene across an elemental detector by using two moving mirrors, one to scan horizontally and the other vertically. Use of a linear array of detectors simplifies the system complexity by requiring scanning in one direction only. *See* TELEVISION CAMERA TUBE.

**Applications.** The applications of optical detectors can be organized according to the spectral interval within which they respond.

*Ultraviolet detectors.* These find limited applications in laboratory instrumentation such as spectrometers and in astronomy. Industrial applications include the monitoring of gas-burner flames in large furnaces. Ultraviolet detectors are used also in dual-color (infrared and ultraviolet) missile-guidance schemes. The principal ultraviolet detector is the photomultiplier. *See* ULTRAVIOLET ASTRONOMY.

*Visible detectors.* Visible radiation detectors have many applications, for example, in television cameras and camcorders. Solid-state arrays based upon silicon charged-coupled-device technology are widely used in simple, inexpensive high-performance imaging systems. Such arrays also find use in astronomy, where their sensitivity at low light levels is of key importance. They are used in industrial automation systems, including those employing machine vision. Electronic still photography employs imaging arrays to replace photographic film in solid-state cameras. *See* CAMERA; CHARGE-COUPLED DEVICES; COMPUTER VISION.

*Infrared detectors.* Detectors of infrared radiation are widely used in military and, to less extent, commercial night-vision systems. Both image intensifiers, which exploit photoemissive photocathodes responding to the low-level ambient illumination of the night sky at wavelengths out to about 1.1  $\mu\text{m}$ , and thermal imaging systems, employing cryogenic linear and matrix (two-dimensional) arrays of long-wavelength photon detectors, are widely used. Uncooled matrix arrays of bolometers and pyroelectric detectors offer high performance in thermal imaging systems without the complexity and cost of cryogenic systems. *See* INFRARED ASTRONOMY; INFRARED IMAGING DEVICES; INFRARED RADIATION.

Paul W. Kruse

**Bibliography.** E. L. Dereniak and D. G. Crowe, *Optical Radiation Detectors*, 1984; R. J. Keyes (ed.), *Optical and Infrared Detectors*, 2d ed., 1980; R. H. Kingston, *Optical Sources, Detectors and Systems: Fundamentals and Applications*, 1995; Technology guide detector handbook, *Laser Focus World*, pp. A3-A42, March 1991; R. K. Willardson and A. C. Beer (eds.), *Semiconductors and Semimetals*, vol. 5: *Infrared Detectors*, 1970, vol. 12: *Infrared Detectors*, 1976.

## Optical fibers

Transparent strands of glass, plastic, or other flexible material used for light delivery. The light delivery can be for the purpose of signal (data) transmission, light amplification, image transmission, or energy transmission. The light can be confined by total internal reflection, by mirrorlike reflection from an internal coated surface, or by a suitably designed set of interior air holes formed parallel to the axis of the fiber. Optical fibers are, in their most common form, a type of dielectric guide for electromagnetic waves (waveguide). In certain forms, fibers are also termed lightguides. *See* REFLECTION OF ELECTROMAGNETIC RADIATION.

In the most common form, the optical fiber consists of a core of material with a refractive index higher than the surrounding cladding. If a light source such as a laser is directed into the fiber, the light travels (propagates) along the fiber and may be detected at the other end. Information may be encoded on the light by on-off keying (a binary system in which on = 1, off = 0), frequency modulation,

or phase modulation. If the wavelength is chosen correctly, fibers may transmit light over very long distances of 100 km (60 mi) or more. For even longer distances, specialty fibers are fabricated with rare-earth ions, which enable optical amplification. Fiber amplifiers enable both undersea and terrestrial communication over great distances. A fiber amplifier which is equipped with suitable end mirrors may also function as a laser source. *See* FIBER-OPTIC CIRCUIT; FREQUENCY MODULATION; MODULATION; OPTICAL COMMUNICATIONS; PHASE MODULATION; REFRACTION OF WAVES.

Fibers may also be fused together into bundles for image transmission. By assigning one fiber core to each image element (pixel), the fiber provides a flexible device in which the light pattern at the entrance to the bundle is approximately replicated at the output of the bundle. These bundles are often used in medical imaging and inspection. In other cases, optical fibers are a flexible means of laser power delivery. For example, certain medical applications require laser light or similar illumination to a specific point inside the human body. Fibers can be designed to deliver a very high irradiance to places which would otherwise require invasive surgery. Another application for optical fibers is in sensors, where a change in light transmission properties is used to sense or detect a change in some property, such as temperature, pressure, or magnetic field. *See* FIBER-OPTIC SENSOR; FIBER-OPTICS IMAGING.

Fused silica ( $\text{SiO}_2$ ) is the most common fiber material. Plastic fibers are used for low-cost, short-distance optical links. Fluoride-based glasses are required for transmission in the mid-infrared region of the optical spectrum.

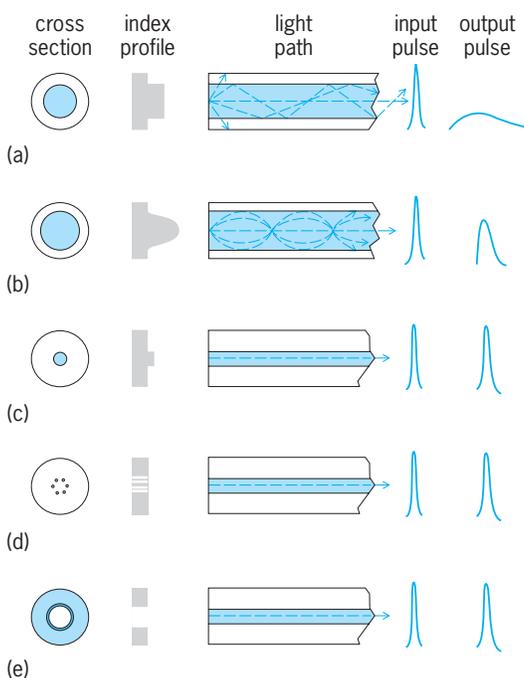
**Fiber designs.** There are five basic types of optical fibers (Fig. 1). Propagation in these lightguides is most easily understood by ray optics, although the wave or modal description must be used for exactness. In a multimode, stepped-refractive-index-profile fiber (Fig. 1a), the number of rays or modes of light which are guided, and thus the amount of light power coupled into the lightguide, is determined by the core size and the core-cladding refractive index difference. Such fibers are limited to short distances for information transmission due to pulse broadening. An initially sharp pulse made up of many modes broadens as it travels long distances in the fiber, since high-angle modes have a longer distance to travel relative to the low-angle modes. This limits the bit rate and distance because it determines how closely input pulses can be spaced without overlap at the output end. At the detector, the presence or absence of a pulse of light in a given time slot determines whether this bit of information is a zero or one.

A graded-index multimode fiber (Fig. 1b), where the core refractive index varies across the core diameter, is used to minimize pulse broadening due to intermodal dispersion. Since light travels more slowly in the high-index region of the fiber relative to the low-index region, significant equalization of the transit time for the various modes can be achieved to reduce pulse broadening. This type of fiber is

suitable for intermediate-distance, intermediate-bit-rate transmission systems. For both fiber types, light from a laser or light-emitting diode can be effectively coupled into the fiber. *See* LASER; LIGHT-EMITTING DIODE.

A single-mode fiber (Fig. 1c) is designed with a core diameter and refractive index distribution such that only one fundamental mode is guided, thus eliminating intermodal pulse-broadening effects. Material and waveguide dispersion effects cause some pulse broadening, which increases with the spectral width of the light source. These fibers are best suited for use with a laser source in order to efficiently couple light into the small core of the waveguide and to enable information transmission over long distances at very high bit rates. The specific fiber design and the ability to manufacture the fiber with controlled refractive index and dimensions determine its ultimate bandwidth or information-carrying capacity. *See* WAVEGUIDE.

A special class of single-mode fibers comprises polarization-preserving fibers. In an ideal, perfectly circular single-mode fiber core, the polarization state of the propagating light is preserved, but in a real fiber various imperfections can cause birefringence; that is, the two orthogonally polarized modes of the fundamental mode travel at different speeds. For applications such as sensors, where controlling the polarization is important, polarization-maintaining fibers can be designed that deliberately introduce a polarization. This is typically accomplished by using noncircular cores (shape birefringence) or by introducing asymmetric stresses (stress-induced birefringence) on the core. In this manner, the polarization



**Fig. 1.** Types of optical fiber designs. (a) Multimode, stepped-refractive-index-profile. (b) Multimode, graded-index-profile. (c) Single-mode, stepped-index. Graded-index is possible. (d) Microstructured optical fiber. (e) Hollow-core fiber for high-power laser delivery.

properties of the input light traveling through the fiber can be controlled. See BIREFRINGENCE; PHOTOELASTICITY; POLARIZED LIGHT.

There are two fiber structures that confine light without total internal reflection. Microstructured optical fibers (Fig. 1*d*), a category which includes photonic bandgap fibers, confine light in a small core region surrounded by micrometer-size air holes. Such fibers provide useful and interesting applications in nonlinear optics, optical switching, and frequency conversion. The confinement of an ultrashort laser pulse within such a waveguide can result in the generation of a supercontinuum, or a very large range of optical frequencies. A hollow-core optical fiber (Fig. 1*e*) confines light through the use of an interior coating (metal or multilayer dielectric) of high reflectance. Such fibers are used in cases of very high laser power or for wavelengths at which transparent fiber materials are not available.

**Attenuation.** The attenuation or loss of light intensity is an important property of the lightguide since it limits the achievable transmission distance, and is caused by light absorption and scattering. Every material has some fundamental absorption due to the atoms or molecules composing it. In addition, the presence of other elements as impurities can cause strong absorption of light at specific wavelengths. Fluctuations in a material on a molecular scale cause intrinsic Rayleigh scattering of light. In actual fiber devices, fiber-core-diameter variations or the presence of defects such as bubbles can cause additional scattering light loss. The light loss of a material, after the light has traveled a length  $L$ , is related to the initial power coupled into the fiber,  $P_0$ , versus the power at the output end,  $P$ , by the equation. See ABSORP-

$$\text{Loss (dB/km)} = \frac{10}{L(\text{km})} \log \left( \frac{P_0}{P} \right)$$

TION OF ELECTROMAGNETIC RADIATION; SCATTERING OF ELECTROMAGNETIC RADIATION.

Optical fibers based on silica glass have an intrinsic transmission window at near-infrared wavelengths with extremely low losses (Fig. 2). Very special glass-making techniques are required to reduce iron and

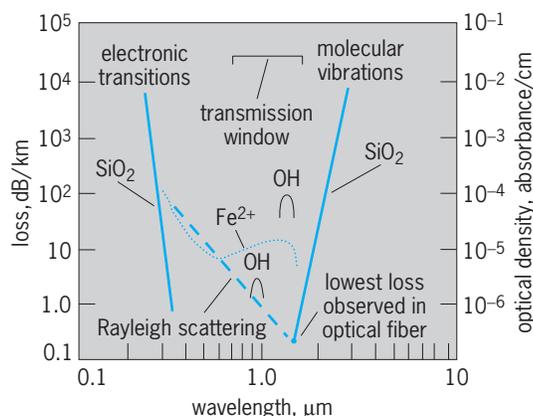


Fig. 2. Loss mechanisms in silica-based fibers.

water (OH) to the parts-per-billion level, and have resulted in losses as low as 0.16 dB/km (0.26 dB/mi). Such fibers are used with solid-state lasers and light-emitting diodes for information transmission, especially for long distances (greater than 1 km or 0.6 mi). Plastic fibers exhibit much higher intrinsic as well as total losses, and are more commonly used for image transmission, illumination, or very short distance data links. See OPTICAL MATERIALS.

Many other fiber properties are also important, and their specification and control are dictated by the particular application. Good mechanical properties are essential for handling; plastic fibers are ductile, while glass fibers, intrinsically brittle, are coated with a protective plastic to preserve their strength. Glass fibers have much better chemical durability and can operate at higher temperatures than plastics. Very tight tolerances on core and outer-diameter control are essential for information transmission fibers, especially to allow long lengths to be assembled with low-loss joining or splicing.

Suzanne R. Nagel; Thomas G. Brown

**Bibliography.** J. A. Buck, *Fundamentals of Optical Fibers*, Wiley, 1995; R. J. Hoss and E. A. Lacy, *Fiber Optics*, 2d ed., Prentice Hall, 1997; S. L. Wymer Meardon, *The Elements of Fiber Optics*, Prentice Hall, 1993; S. E. Miller and I. P. Kaminow (eds.), *Optical Fiber Telecommunications*, vol. 2, Academic Press, 1988; B. P. Pal, *Fundamentals of Fiber Optics in Telecommunication and Sensor Systems*, 1993; J. C. Palais, *Fiber Optic Communications*, 5th ed., Prentice Hall, 2004.

## Optical flat

A disk of high-grade quartz glass approximately 0.75 in. (2 cm) thick, having at least one side ground and polished with a deviation in flatness usually not exceeding 0.000002 in. (50 nanometers) all over, and a surface quality of 5 microfinish or less. When two surfaces of this quality are placed lightly together so that the air is not wrung out from between them, they are separated by a film of air and actually touch at only one point. This point is the vertex of a wedge of air separating the two pieces.

If parallel beams of light pass through the flat, part will be reflected against the surface being inspected, while part will be reflected directly back through the flat. Because the distance between the surfaces is constantly increasing along the angle, the beams reflected from the flat and the beams reflected from the workpiece will alternately reinforce and interfere with each other, producing a pattern of alternate light and dark bands (Fig. 1). Each succeeding full band from a point of contact means the distance between surfaces is one wavelength thicker. If the light is relatively monochromatic, the wavelength is known. Red with a wavelength of 0.000116 in. (295 nm) is commonly used. Thus a definite relationship is established between lineal measurement and light waves. Optical flats are used for two general purposes.



Fig. 1. Optical flat being used to determine flatness of seal ring. Interference bands on the seal ring face show lines of constant depth. (Van Keuren Co.)

**Determination of surface contour.** If the surface is flat, the light bands are parallel. Deviation from flatness shows as curvature of the lines. The principle in interpreting a pattern is almost identical with the principle in interpreting a topographical map in that the bands connect points of equal distance from the master surface of the optical flat. Deviation from flatness can be reduced to rational figures.

**Comparison of lineal measurement.** When an optical flat is placed across a gage block or a buildup of blocks and another object, as an end standard, or a cylinder or sphere, both resting on a precisely flat surface, then the angle between the blocks and the flat can be measured, and the difference in length between the gage blocks and the length or height

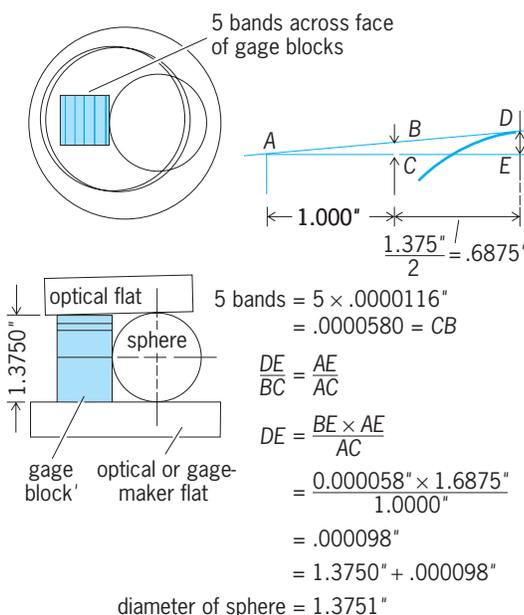


Fig. 2. Measurement of height of sphere determined by means of gage blocks and optical flat. 1" = 25.4 mm.

or diameter of the unknown can be determined (Fig. 2). With the high accuracy available in electrical and pneumatic comparators, optical flats are seldom used in this way except for spheres or irregular surfaces where point contact by the comparator is impractical.

The principle of interferometry is the standard method for measurement of gage blocks. However, an interferometer and not the optical flat itself is used. The wavelength of light is the present standard of all lineal measurements. See GAGES; INTERFEROMETRY.

Rush A. Bowman

## Optical guided waves

Optical-frequency electromagnetic waves confined within an optical waveguide, a structure designed to carry such waves from one place to another somewhat as a pipe carries water. (The terms optical and light are used here in broadest sense to include visible and near-infrared electromagnetic radiation.) The demonstrations of the first semiconductor laser and the first low-loss glass optical fiber initiated a technological revolution. Because of the high data rates that can be achieved, the transmission of information in the form of optical guided waves confined within an optical-fiber waveguide has become the preferred method for the telecommunications industry. That the semiconductor laser, the source that produces the optical signals, also employs an optical waveguide makes clear the pivotal role played by optical guided waves in modern communications technology.

**Total internal reflection.** Optical waveguides confine light by the method of total internal reflection. At a plane interface between two media (Fig. 1a), the upper and lower media are characterized by indices of refraction  $n_1$  and  $n_2$ , respectively. When a beam of light strikes the interface at angle  $\theta_1$ , part of the beam is reflected at the same angle  $\theta_1$ , and the remainder of the beam is transmitted into the second medium at angle  $\theta_2$ . According to Snell's law, Eq. (1),

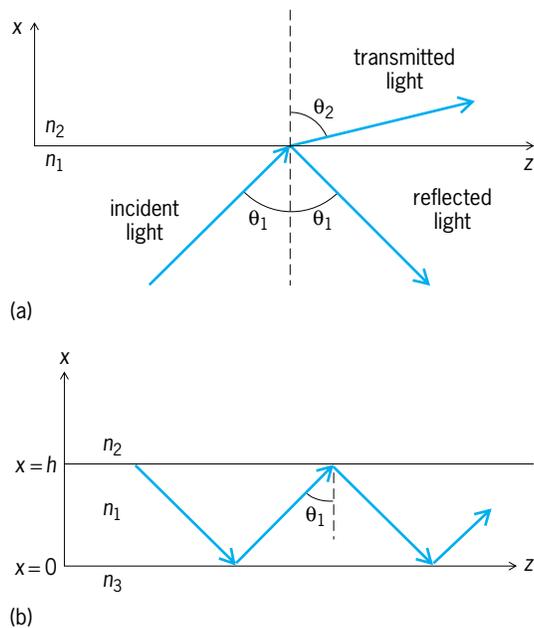
$$n_1 \sin \theta_1 = n_2 \sin \theta_2 \quad (1)$$

the angle  $\theta_2$  exceeds  $\theta_1$  if the refractive index  $n_1$  exceeds  $n_2$ , which means that  $\theta_2$  will reach its physical limit of  $90^\circ$  for an incident angle  $\theta_1 < 90^\circ$ . The incident angle for which  $\theta_2 = 90^\circ$ , called the critical angle ( $\theta_c$ ), is given by Eq. (2).

$$n_1 \sin \theta_c = n_2 \quad (2)$$

For angles of incidence greater than or equal to the critical ( $\theta_1 \geq \theta_c$ ), there is no transmitted beam and, in the absence of absorption, the energy carried by the incident beam is completely reflected into the first medium at angle  $\theta_1$ . See REFLECTION OF ELECTROMAGNETIC RADIATION; REFRACTION OF WAVES.

**Rays or waves.** It is useful to first think of light as propagating as a ray. Total internal reflection can be used to trap a ray of light in a thin layer (Fig. 1b), provided that its refractive index is larger than the



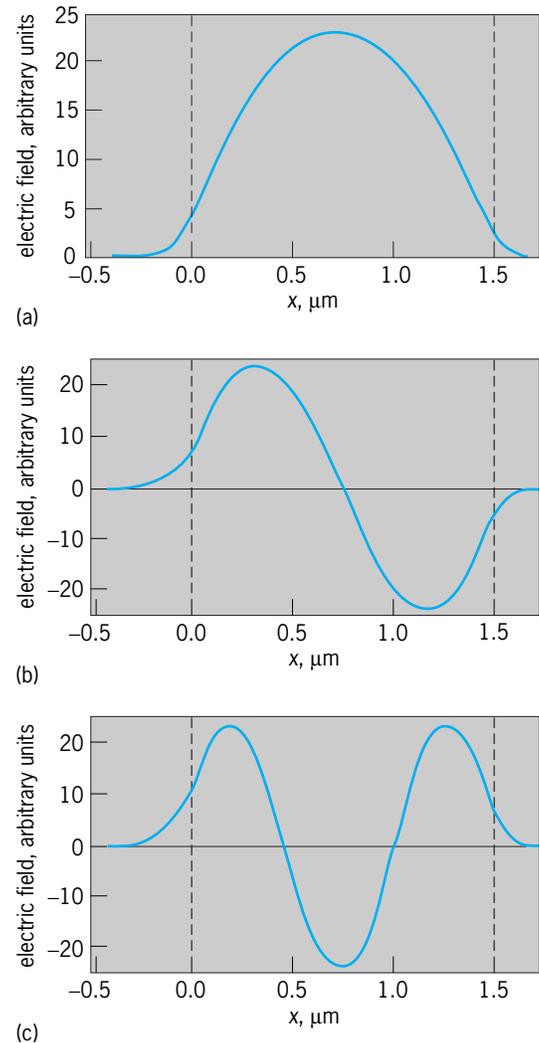
**Fig. 1.** Total internal reflection. (a) Geometry of reflection at the plane interface between two media. (b) Light trapped by total internal reflection in a planar optical waveguide of thickness  $h$  and refractive index  $n_1$ . Refractive indices of boundary media,  $n_2$  and  $n_3$ , are less than  $n_1$ .

refractive indices of the two boundary media. If the internal propagation angle  $\theta_1$  exceeds the critical angle for total internal reflection at both the upper and lower interfaces, the light will be unable to escape the thin layer. Each reflection will be perfect, and the light will travel along a direction parallel to the boundaries of the layer without loss in a zigzag pattern until it reaches the end of the structure. A waveguide of this sort is called a planar (or slab) optical waveguide and is used to confine light, for example, in the active region of a semiconductor laser. Optical guided waves can thus be thought of as rays of light trapped within a structure by total internal reflection.

The above argument suggests that light should be able to follow any zigzag path as long as the angle  $\theta_1$  exceeds the critical angle for total internal reflection at both interfaces. A more detailed analysis based on Maxwell's equations reveals that the allowed angles in any optical waveguide form a discrete set rather than a continuum. For any waveguide, there exists a mathematical expression called the dispersion relation that shows how the allowed propagation angles are related to the optical wavelength, the size of the waveguide, and the refractive indices of the materials that make up the waveguide structure. The dispersion relation also contains at least one integer, say  $m$ , that is used to label a particular solution. The allowed angles are called the mode angles, and each unique trajectory through the waveguide is said to be a mode of the waveguide. The thickness or diameter of an optical waveguide can be made sufficiently small that only a single mode is supported for a given orientation of the electric field (polarization). See POLARIZATION OF WAVES.

The ray picture of the propagation of light is necessarily limited. Light is an electromagnetic wave phenomenon that is better described with Maxwell's equations. A rigorous analysis reveals that an optical guided wave, or mode, is best pictured as a unique electric field distribution that propagates along the waveguide axis with a unique velocity. Similar distributions are found for planar waveguides (Fig. 2) and optical-fiber waveguides. It is significant to note that not all of the guided-wave's electric field, and hence energy, is confined within the layer. The spillover into the adjacent media is characteristic of a proper treatment of total internal reflection using Maxwell's equations. See ELECTROMAGNETIC WAVE TRANSMISSION; MAXWELL'S EQUATIONS.

**Optical waveguides.** Optical guided waves can propagate in a variety of structures. The optical fiber used in the communications industry consists of two concentric glass cylinders. The inner core region is made of a glass that has a slightly higher index of



**Fig. 2.** Electric field distributions for the three lowest-order modes of a planar optical waveguide of thickness  $h = 1.5 \mu\text{m}$ . Electric field magnitudes are plotted as functions of the  $x$  coordinate in Fig. 1b for an arbitrary, fixed value of  $z$ . (a) Mode  $m = 0$ . (b)  $m = 1$ . (c)  $m = 2$ .

refraction than the outer cladding region, as required for total internal reflection to occur. In telecommunications applications, core diameters in the range of 4–50  $\mu\text{m}$  are used routinely to carry near-infrared optical radiation. Pulses of light with wavelengths between 0.8 and 1.55  $\mu\text{m}$  are injected into the fiber on the near end. The presence or absence of a pulse is interpreted as a one or a zero by a receiver on the far end, typically many miles (kilometers) away. These bits of information are carried by the optical fiber from the transmitter to the receiver as optical guided waves. It is a feature of optical waveguides that the light pulses broaden as they travel down the optical fiber. If the pulses broaden too much, they begin to overlap, making it difficult to distinguish one from another. This effect provides the fundamental limitation to the information capacity of a fiber-optic communication system. See OPTICAL COMMUNICATIONS; OPTICAL FIBERS.

Devices such as semiconductor lasers make use of optical guided waves over distances much shorter than those spanned by optical fibers. In a semiconductor laser, a planar waveguide (Fig. 2) also serves as an electrically driven amplifying medium. Gallium arsenide (GaAs), aluminum gallium arsenide (AlGaAs), or indium gallium arsenide phosphide (InGaAsP) are common examples of semiconductor laser materials. When mirrors are formed on the ends of the semiconductor waveguide, typically a few hundred micrometers apart, light travels back and forth between those mirrors just as with any laser, but in the form of optical guided waves. The semiconductor laser is a waveguide laser. See LASER.

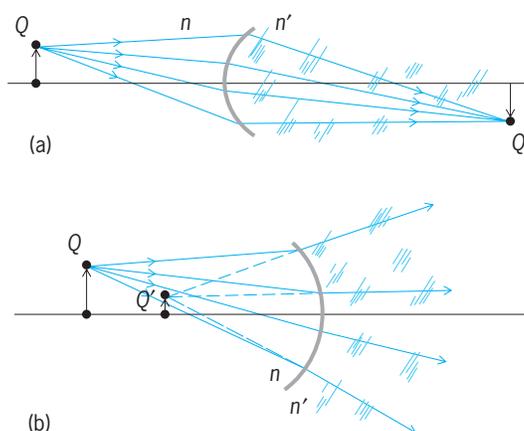
Optical waveguides are the basic constituents of an emerging technology known variously as integrated optics, integrated optoelectronics, or photonic integrated circuits. This technology integrates optical and electronic components into or on an optical waveguide. The aim is to process and manipulate light while it is trapped as optical guided waves within the confines of the optical waveguide. See INTEGRATED OPTICS; WAVEGUIDE. Dennis G. Hall

**Bibliography.** M. J. Adams, *An Introduction to Optical Waveguides*, 1981; M. Koshiha, *Optical Waveguide Analysis*, 1992; D. Marcuse, *Theory of Dielectric Optical Waveguides*, 2d ed., 1991; A. R. Mickelson, *Guided Wave Optics*, 1993; T. Tamir et al. (eds.), *Guided-Wave Optoelectronics*, 2d ed., 1990.

## Optical image

The image formed by the light rays from a self-luminous or an illuminated object that traverse an optical system. The image is said to be real if the light rays converge to a focus on the image side and virtual if the rays seem to come from a point within the instrument (see *illus.*).

The optical image of an object is given by the light distribution coming from each point of the object at the image plane of an optical system. The ideal image of a point according to geometrical optics is obtained when all rays from an object point unite



**Optical images.** (a) Real image. Rays leaving object point  $Q$  and passing through the refracting surface separating media  $n$  and  $n'$  are brought to a focus at the image point  $Q'$ . (b) Virtual image. Rays leaving  $Q$  and refracted by the concave surface separating  $n$  and  $n'$  appear to be coming from the virtual image point  $Q'$ . As the rays are diverging, they cannot be focused at any point. (After F. A. Jenkins and H. E. White, *Fundamentals of Optics*, 4th ed., McGraw-Hill, 1976)

in a single image point. However, diffraction theory teaches that even in this case the image is not a point but a minute disk. The diameter of this disk is about  $1.22\lambda/(NA)$ , where  $\lambda$  is the wavelength of the light considered and  $NA$  is the numerical aperture, the sine of the largest cone angle on the image side multiplied by its refractive index (which is usually equal to unity). See FOCAL LENGTH; GEOMETRICAL OPTICS.

**Aberrations.** From the standpoint of geometrical optics, if this most desirable type of image formation cannot be achieved, the next best objective is to have the image free from all but aperture errors (spherical aberration). In this case the light distribution in the image plane is still circular, resembling the point image; there is a true coordination of object point and image, although the image may be slightly unsharp. If the aperture errors are small, or if the image is viewed from a distance, such an image formation may be very satisfactory.

Asymmetry and deformation errors may be very disturbing if not held in check, because the light distribution of the image of a point in this case has a decidedly undesirable shape.

When the image of an axis point is considered, the rays through a fixed aperture circle converge to an axis point. For this type of imagery, the term half-sharp image will be used. A small object at the object point is then imaged by a circular stop at the focus of the image bundle with a magnification as given by Eq. (1), where  $u$  and  $u'$  are the angles of the imaging

$$m = \frac{n \sin u}{n' \sin u'} \quad (1)$$

cone in object and image space, respectively, and  $n$  and  $n'$  are the corresponding refractive indices.

If the axis point is sharply imaged, an object of finite extent is sharply imaged if, and only if,  $m = m_0$  (the gaussian magnification) for all values of  $u$  (sine condition).

In the case of aperture errors, the most desirable image formation for an axis point is attained when the different images appear under the same angle from the exit pupil. If  $k'$  is the distance of the image point from the exit pupil,  $\Delta s'$  is the aperture aberration, and if Eq. (2) gives the magnification error

$$\Delta m = \frac{n \sin u}{n' \sin u} - m_o \quad (2)$$

compared with the magnification  $m_o$  on the axis, the condition is given by Eq. (3). The fulfillment of this

$$\frac{\Delta s'}{k'} - \frac{\Delta m}{m_o} = \text{constant} \quad (3)$$

condition gives equal quality for an object near the axis of a system with rotation symmetry. Corresponding conditions can be ascertained for the image of an off-axis element if all the asymmetry errors and deformation errors are balanced. See ABERRATION (OPTICS).

**Resolution.** Two points are resolved by an optical system if the two images lie apart. Photometric analysis of an image may indicate the existence of two object points even if their images overlap, but in such an analysis the illumination of the object, as well as the imagery, plays a role.

In interference experiments, it is found that the image of two self-luminous points (that is, two light sources that are sufficiently separated) is incoherent; that is, the intensities of the two beams simply add. If the two object points are illuminated by the same light source, however, the phase relation of the light at the two points has to be taken into consideration. This is of the greatest importance for microscopes and telescopes, which image very small or distant objects. In this case an artificial change of phase by phase plates and apodization may improve resolution. See DIFFRACTION; RESOLVING POWER (OPTICS).

Resolving power is not the only consideration in image formation. The eye recognizes only contrast differences, and therefore objects may not be discerned if the contrast difference is too small. Again, for the image of a point, or an object illuminated by a point light source, means can be found to change the apparent contrast, making it possible to discern biological objects, for example, having small differences of refractive index.

**Image analysis.** Methods have been suggested for obtaining information about optical images by sine-wave analysis. A sinusoidal test object is imaged by an optical system as a sinusoidal image, but altered in phase and amplitude. A large number of sinusoidal test objects with different frequencies (number of maxima per millimeter) are imaged, and the amplitude and phase are measured.

The curve of amplitude versus frequency gives a measure for resolving power and contrast as a function of the frequency of the test object, whereas the adjusted curve of phase versus frequency describes the lack of symmetry in the image. These amplitude-frequency curves can be measured as well as calculated from the spot diagrams, onto which the ef-

fects of diffraction can be superimposed if necessary.

Max Herzberger

Bibliography. B. D. Guenther, *Modern Optics*, 1990; M. Herzberger, *Modern Geometrical Optics*, 1958, reprint 1980; F. A. Jenkins and H. E. White, *Fundamentals of Optics*, 4th ed., 1976; J. Meyer-Arendt, *Introduction to Classical and Modern Optics*, 4th ed., 1995; Optical Society of America, *Handbook of Optics*, 2 vols., 2d ed., 1995.

## Optical information systems

Systems that use light to process information. Optical information systems or processors consist of one or several light sources; one- or two-dimensional planes of data such as film transparencies, various lenses, and other optical components; and detectors. These elements can be arranged in various configurations to achieve different data-processing functions. As light passes through various data planes, the light distribution is spatially modulated proportional to the information present in each plane. This modulation occurs in parallel in one or two dimensions, and the processing is performed at the speed of light. Optical processors offer various advantages compared to other technologies: data travels at the speed of light; all data in one-dimensional and two-dimensional arrays are operated on in parallel; multiple planes of data can be processed in parallel by various multiplexing schemes; it is possible to have large numbers of interconnections with no interaction (which is not possible with electrical connections); and power dissipation is less and size and weight can be less for optical processors than for their electronic counterparts. See CONCURRENT PROCESSING; MULTIPLEXING AND MULTIPLE ACCESS.

In practice, the processing speed is limited by the rate at which data can be introduced into the system and the rate at which processed data (produced on output detector arrays) can be analyzed. The reusable real-time spatial light modulators used to produce new input data, filters, interconnections, and so forth, are the major components required for these optical information-processing systems to realize their full potential. Spatial light modulators convert electrical input data into a form suitable for spatially modulating input light, or react to an optical input and generate a different optical output. The manipulation of the light passing through the system is controlled by spatial light modulators, lenses, holographic optical elements, computer-generated holograms, or fiber optics. Four major application areas are image processing, signal processing, computing and interconnections, and neural networks. See HOLOGRAPHY; OPTICAL FIBERS; OPTICAL MODULATORS.

**Components.** Inexpensive liquid-crystal television displays (when properly modified) can serve as two-dimensional spatial light modulators. Several commercially available two-dimensional spatial light modulators use magneto-optic techniques, deformable mirrors, and ferroelectric liquid crystals.

However, these devices are binary (except for the deformable mirror device). Optically addressed volume-storage optical media and optically addressed bistable devices and structures that function as optical transistors are not yet competitive with digital technology in size and power dissipation, but research advances may yield such competitive elements. Laser-diode array technology for communications and parallel recording of multiple channels of data on disks, plus optical read/write alterable storage media have greatly advanced the possibilities in this area. Acoustooptic devices represent one of the most mature and robust optical transducer technologies, and architectures that utilize these one-dimensional spatial light modulators to implement various two-dimensional processing functions offer the promise of hardware implementation in the near future. See ACOUSTOOPTICS; COMPUTER STORAGE TECHNOLOGY; FERROELECTRICS; LASER; LIQUID CRYSTALS; MAGNETOOPTICS; OPTICAL RECORDING.

**Optical image processing.** The basic optical image-processing functions will be discussed (Fig. 1). An image  $g(x,y)$  is placed in a plane,  $P_1$ , and illuminated with laser light. The light distribution incident on a second plane,  $P_2$ , is the two-dimensional Fourier transform  $G(u,v)$  of the input image  $g(x,y)$ . This Fourier transform distribution is a representation of the spatial frequencies  $(u,v)$  present in the input image. Lower spatial frequencies (corresponding to larger input shapes) lie closer to the center of  $P_2$ , and higher spatial frequencies (corresponding to smaller input image regions) lie further from the center of  $P_2$ . The orientation of each input object is reflected in the angular location of its corresponding spatial frequency distribution in  $P_2$ . Detectors placed in  $P_2$  can feed this Fourier-coefficient information to a digital postprocessor for subsequent analysis and classification of the type of input object and its size and orientation. This is one form of optical feature extraction for pattern recognition and precise measurement of object size and orientation. See FOURIER SERIES AND TRANSFORMS.

Other architectures can optically compute other features of an input object such as its geometrical moments, its chord histogram distribution (the length and angle of all chords that define the object), or wedge and ring samples of the Fourier transform. A digital postprocessor can analyze these various features to determine the final class and orientation of the input object. Such feature extraction systems are the simplest form of optical pattern recognition image processors. Systems to compute optical wedge-ring Fourier transform samples have been tested and demonstrated on a wide variety of product inspection problems and applications. A Hough transform optical processor, which determines the position, orientation, and length of all lines in an input image has also been developed for product inspection. In this system, a computer-generated hologram consisting of  $N$  pairs of orthogonal cylindrical lenses is placed behind an input spatial light modulator on which an image of the product is placed. The

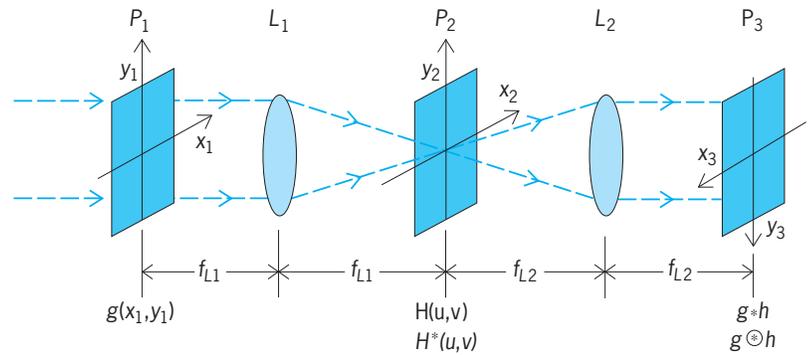


Fig. 1. Optical Fourier transform and correlation image processor. Lenses  $L_1$  and  $L_2$  have focal lengths  $f_{L1}$  and  $f_{L2}$ .

output consists of  $N$  slices of the Hough transform at  $N$  different angles.

However, the full optical system of Fig. 1 can also perform the correlation operation. In this case, the transmittance of  $P_2$  is made proportional to  $H^*(u,v)$ , the conjugate Fourier transform of a reference function  $b(x,y)$ . The  $H^*(u,v)$  filter function at  $P_2$  is formed by holographic techniques. Then, the  $P_3$  pattern is the correlation of the two space functions  $g$  and  $b$ , written  $g \otimes b$ . The correlation of two functions is equivalent to translating  $b$  spatially over  $g$  and, for each translation, forming the product and sum of  $b$  and the associated region of  $g$ . The regions of an input scene  $g$  that most closely match the reference function  $b$  being searched for yield the largest output. This template matching operation can thus locate the presence and positions of reference objects in an input image. This object pattern recognition operation is achieved with no moving parts and is quite effective for extracting objects in clutter or high noise and for locating multiple occurrences of an object. Advanced filter synthesis techniques allow such systems to operate independent of scale, rotation, and other geometrical distortions between the input and reference object. Character recognition, robotics, missile guidance, and industrial inspection are the major applications that have been pursued for such architectures. If  $H(u,v)$  rather than  $H^*(u,v)$  is recorded at  $P_2$ , the  $P_3$  output is an image,  $g^*b$ , that is a filtered version of the input. The filter function of the system can be controlled by spatially varying the contents of  $P_2$  to achieve image enhancement and restoration of blurred and degraded images. See CHARACTER RECOGNITION; COMPUTER VISION; GUIDANCE SYSTEMS; IMAGE PROCESSING; ROBOTICS.

A major shortcoming of prior optical image processors was the need for different architectures or systems to implement different functions and the lack of attention given to the role for optics in low-level iconic (pixel-based) image-processing functions, which are the more intensive image-processing functions required. Techniques exist to optically implement all basic morphological image-processing functions on the optical correlator in Fig. 1 (using a fixed set of simple structuring-element filters at  $P_2$ ) and different thresholding operations at  $P_1$  and

$P_3$ . The  $P_3$  output is the convolution of the input image and the structuring element at  $P_2$ . With a low or high  $P_3$  threshold, the output is the erosion or dilation of the input image. Intermediate thresholds are more useful and produce rank-order filters. Sequences of these operations yield opening or closing functions. These are useful for filling in regions or gaps in an image, and removing peninsulas, salt-and-pepper noise, and other undesirable features. The size of the structuring element and the number of times it is applied determine the size of the image gaps filled in or the size of image regions removed. Alternatively, the skeleton, median filter, or various edge-enhanced versions of the image can be formed. These operations constitute all of the basic morphological image-processing functions. Optical processors that use them can operate at all levels of computer vision: low-level (iconic processors performing morphological functions), medium-level (optical processors for feature extraction), and high-level (correlators and neural networks).

Other new optical filters produce wavelet and Gabor transforms for object detection. Thus, the repertoire of operations achievable on the same optical processor is quite significant.

**Optical signal processing.** Signal processing systems are used in radar, sonar, electronic warfare, and communications to determine the frequency and direction of arrival of input signals, and for correlation applications. These systems generally employ acousto-optic transducers to input electrical signals to the optical processor. The electronic input to an acousto-optic cell is converted to a sound wave which travels the length of the cell. If the cell is illuminated with light, the amplitude of the light leaving the cell will be proportional to the strength of the input signal, and the angle at which the light leaves will be proportional to the frequency of the input signal. Since the device and system are linear, if  $N$  input signals are simultaneously present,  $N$  light waves leave the cell at  $N$  angles and strengths. A lens behind the cell focuses each light wave to a different spatial location in an output detector plane. Thus, an optical spectrum analyzer results (Fig. 2). Compact systems of this type exist, and bandwidths in excess of 8 GHz can be achieved and up to 1000 signals can be simultaneously processed. If a multichannel acousto-optic

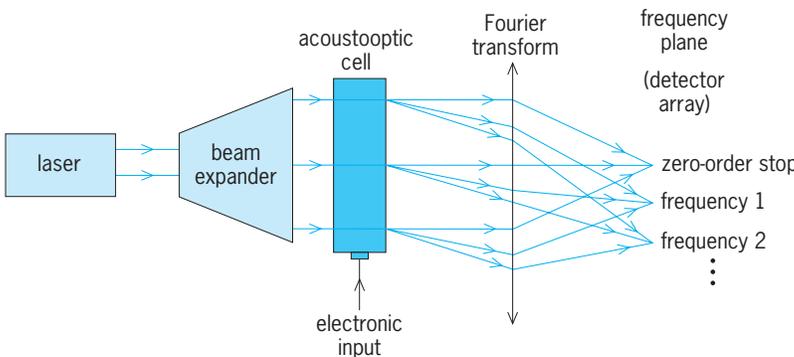


Fig. 2. Acousto-optic signal spectrum analyzer.

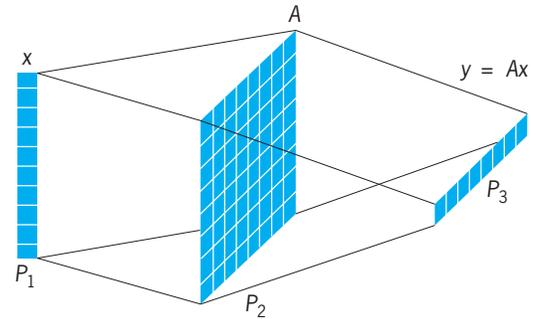


Fig. 3. Optical matrix vector, crossbar, or neural net processor.

cell is used with input signals from different antennas or antenna elements, the two-dimensional Fourier transform of its output provides a two-dimensional display of the simultaneous frequency and direction of arrival of all received signals within the bandwidth of the cell. The angle at which light is deflected can be varied by varying the frequency of the input signal to the acousto-optic cell; a beam deflector or scanner based on this principle is used in some xerographic reproduction systems and laser printers. Other architectures using multiple cells can correlate two or more signals. These systems are used for simultaneous range and Doppler processing, for demodulation and synchronization of signals for communication applications, and for production of adaptive processors that can null interference signals. See ELECTRONIC WARFARE; RADAR; SONAR; SPECTRUM ANALYZER.

**Optical computing and interconnections.** The term optical computing has had different meanings as optical processing has evolved. It has been used to describe optical array processors that operate on one-dimensional or two-dimensional arrays of data (vectors and matrices) to perform various functions in linear algebra. In the basic version of such a system (Fig. 3), the light leaving point modulators (laser diodes or light-emitting diodes) at a vector,  $P_1$ , is expanded to uniformly illuminate a two-dimensional data array (matrix),  $P_2$ . The light leaving  $P_2$  is then integrated onto a linear detector array,  $P_3$ , whose output is the matrix vector product of the  $P_2$  and  $P_1$  data. Implementations of this basic architecture in integrated optics are being studied. See INTEGRATED OPTICS; LIGHT-EMITTING DIODE; LINEAR ALGEBRA; MATRIX THEORY.

With  $N$  inputs at  $P_1$  and  $M$  outputs at  $P_3$ , the mask at  $P_2$  can be used to interconnect any of the  $N$  inputs to  $M$  outputs. One-to-many and many-to-one interconnections are allowed, and a nonblocking optical crossbar switch results. Many optical interconnection systems using fiber optics, holograms, waveguides, and volume materials are rapidly maturing.

Most present optical computing work concerns optical techniques to achieve general-purpose logic and numeric functions. These are generally referred to as optical-digital circuits. There are three basic approaches.

In symbolic substitution, input binary 0 and 1 data are represented by different symbols. Input digit patterns (for example, 00, 01, 10, and 11) are recognized, and different patterns are substituted for each. The pattern substituted determines the operation that the system performs. Substitution rules have been devised to allow all logic and numeric functions to be realized. Different encoding methods allow improved realization with fewer carries required. A variety of optical realizations exist.

A second type of general-purpose optical computer uses acousto-optic devices and optical AND and OR operations. (The AND or OR of multiple optical inputs occurs when several input light beams are incident on a detector or optical device with a high or low threshold.) From a basic parallel optical A-O-I (AND-OR-INVERT) unit, a general optical arithmetic logic unit (ALU) can be fabricated. See COMPUTER SYSTEMS ARCHITECTURE; DIGITAL COMPUTER; LOGIC CIRCUITS.

The third major approach uses optical nonlinear or optical bistable devices to develop optical switches (optical transistors). In one use of the device, the wavelength of the input light is offset slightly from the value that gives maximum transmission. However, when the intensity of the input light increases, the transmittance curve for the device shifts to allow transmission of the input wavelength. A hysteresis effect can occur, and the input wavelength can be shifted by different amounts from resonance to produce a family of optical circuits. See NONLINEAR OPTICAL DEVICES.

**Optical neural networks.** The ability of many optical beams to cross with no interference between them, one beam to be broadcast to several locations, or several beams to combine at one location, is the basis for optical interconnections. These same properties are being pursued extensively to produce optical neural networks. In these cases, input points of light (as in Fig. 3) represent neurons, the amounts of output light represent neurons strengths, matrix masks (such as at  $P_2$  of Fig. 3) represent sets of weights or interconnections between neurons, and the  $P_3$  outputs in Fig. 3 represent a second layer of neurons. A cascade of two (or more) such matrix-vector systems thus represents a neural network. With various nonlinear elements used at the different neuron layers, volume materials used for the matrix connections, and feedback between neuron layers, powerful artificial optical neuron networks are possible. See NEURAL NETWORK.

David Casasent

**Bibliography.** N. J. Berg and J. M. Pellegrino, *Acousto-Optic Signal Processing: Theory and Implementation*, 2d ed., 1994; D. G. Feitelson, *Optical Computing: A Survey for Computer Scientists*, 1988, reprint 1992; H. S. Hinton et al. (eds.), Special issue on optical computing, *Appl. Opt.*, vol. 33, no. 8, March 10, 1994; Special issue on optical interconnections, *Opt. Eng.*, vol. 21, no. 10, October 1986; Special issue on spatial light modulators, *Appl. Opt.*, vol. 28, no. 22, November 15, 1989; H. Szu (ed.), Special issue on adaptive wavelet transforms, *Opt. Eng.*, vol. 21, no. 10, October 1986.

## Optical isolator

A device that is interposed between two systems to prevent one of them from having undesired effects on the other, while transmitting desired signals between the systems by optical means. Optical isolators are used for both electrical systems and optical systems such as lasers.

### Isolators for Electrical Systems

An optical isolator for electrical systems is a very small four-terminal electronic circuit element that includes in an integral package a light emitter, a light detector, and, in some devices, solid-state electronic circuits. The emitting and detecting devices are so positioned that the majority of the emission from the emitter is optically coupled to the light-sensitive area of the detector. The device is also known as an optoisolator, optical-coupled isolator, and optocoupler. The device is housed in an integral opaque package so that the only optical emission impinging on the detector is that produced by the emitter. This configuration of components can perform as a solid-state electronic transformer or relay, since an electronic input signal causes an electronic output signal without any electrical connection between the input and the output terminals.

Optical isolators are used in electrical systems to protect humans or machines when high-voltage or high-power equipment is being controlled. In addition, optical isolators are used in electronic circuit design in situations where two circuits have large voltage differences between them and yet it is necessary to transfer small electrical signals between them without changing the basic voltage level of either.

Before optical isolators were developed, this circuit function was performed by such devices as isolation transformers or signal relays. The optical isolator provides a number of advantages over such devices in that it is much smaller, much faster, has no contact bounce (unlike a relay), has no inductance (unlike a transformer), and provides a very high voltage isolation between the input and output circuits.

In order to discuss applications of the optical isolator, it is necessary to identify the different types of optical isolators that can be produced through the use of different combinations of light emitters and detectors.

**Optical emitters.** The optical emitter most commonly selected for use in optical isolators is the gallium arsenide light-emitting diode (LED), which emits in the infrared or near-infrared regions of the spectrum. The typical wavelength of the emission is 850 nanometers. The LED is extremely small (about the size of a transistor chip) and provides extremely fast infrared light pulses. Photodetectors fabricated from silicon are particularly sensitive to the wavelength of light emitted by the gallium arsenide LED, so that the transfer of a signal from the input to the output of the optical isolator is particularly efficient.

The other input devices commonly used are gas-discharge and incandescent lamps. These emitters tend to be somewhat larger and slower than LEDs,

but are used when their particular electronic characteristics are desired, and in conjunction with detectors that have a spectral peak in the visible portion of the spectrum. See INCANDESCENT LAMP; LIGHT-EMITTING DIODE; SEMICONDUCTOR DIODE; VAPOR LAMP.

**Optical detectors.** The light detectors that are used in the construction of optical isolators include light-dependent resistors (such as photocells), light-sensitive devices that generate a voltage without any electrical input (such as photovoltaic devices), light-sensitive devices that switch from one state to another (such as photothyristors), and light-sensitive devices that modify a voltage or current (such as phototransistors, photodiodes, and photodetector-amplifier combinations). See OPTICAL DETECTORS; PHOTOCONDUCTIVE CELL; PHOTODIODE; PHOTOELECTRIC DEVICES; PHOTOTRANSISTOR; PHOTOVOLTAIC CELL.

**Emitter-detector combinations.** Having such a wide variety of emitter and detector combinations available means that circuitry and system designers can use them over a wide range of applications. For example, the combination of a photocell and an incandescent lamp produces a device with entirely different characteristics from those of the combination of an LED and a phototransistor. Each combination is described below and its primary application identified.

*Incandescent lamp-photocell optical isolators.* In these devices the resistance of the photocell varies from about 10 to 1000 megohms, as the input light is varied from off to maximum brightness. If the input changes instantaneously, the output resistance change occurs in 20 milliseconds (typical). If the input voltage changes gradually, the resistance of the photocell varies somewhat proportionally to the input voltage. As a result of this variable-resistance characteristic, the device whose schematic is shown in Fig. 1 can be used as a remote-controlled rheostat where there is no electrical connection between the control wire and the circuit being controlled. In applications where signals are required for two separate circuits, without an electrical connection between any of the three circuits, devices are used with two photocells built into the case of the optical isolators.

*Neon lamp-photocell combinations.* These devices have output characteristics similar to the incandescent lamp-photocell devices. However, the input char-

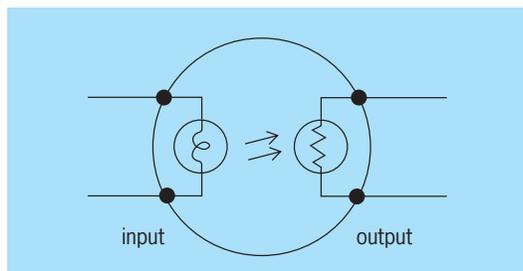


Fig. 1. Schematic diagram of incandescent lamp-photocell optical isolator.

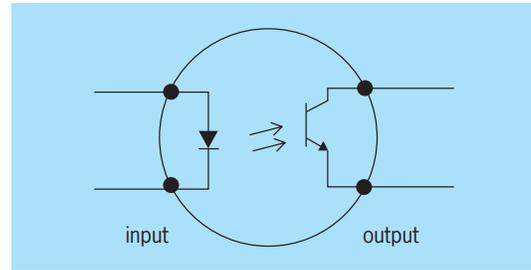


Fig. 2. Schematic diagram of LED-phototransistor optical isolator.

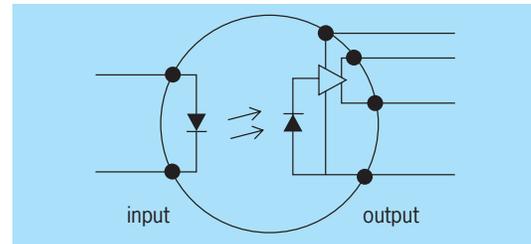


Fig. 3. Schematic diagram of LED-photodiode-integrated-circuit optical isolator.

acteristics vary dramatically, because the neon lamp does not emit light (ignite) until the input voltage reaches about 70 V. Also, the current after ignition is very small. These devices are employed where large input voltage swings are typical and where long life is required. See NEON GLOW LAMP.

*LED-silicon detector combinations.* All of the optical isolators that employ silicon photodetectors have LED emitters. The most common photodetector is the phototransistor. A schematic diagram of a type of LED-phototransistor optical isolator is shown in Fig. 2. In such devices the light from the LED causes the phototransistor current to vary as a function of the amount of light impinging on the photosensitive area of the phototransistor. The devices can isolate circuits that differ by as much as 5000 V. The devices are very fast, with the output current changing in some devices in as little as 10-20 nanoseconds after the occurrence of an input pulse. This fast response time means that these devices are particularly useful in computer circuit applications where such fast pulses are common. Not only do these devices provide high voltage isolation between circuits, but because the phototransistor has built-in gain, the detected signal is amplified. These devices are also immune to certain types of noise and are extremely small, the package being just larger than a typical transistor package. The devices can be used in most circuits that require an isolation transformer, and provide the additional advantages of higher speed and higher voltage ratings. Such devices also can perform as a solid-state relay (no mechanical wear and so no contact bounce), as a high-speed chopper, and as a pulse amplifier.

A similar type of optical isolator that produces a higher output amplitude for the same input employs a photo-Darlington sensor and output device. The photo-Darlington is in reality two phototransistors

that are integrally connected. Although it does have increased amplification, its speed is much lower.

**LED-photodetector-amplifier.** In some optical isolators, integrated circuits are also included in the package to provide specific output characteristics. For example, one optical isolator includes an integrated circuit amplifier so that the optical isolator can perform like a broadband pulse transformer that is compatible with diode-transistor-logic (DTL) computer circuitry as well as provide a frequency response to zero frequency. A schematic diagram of an LED-photodiode-integrated-circuit optical isolator is shown in Fig. 3. See AMPLIFIER; CIRCUIT (ELECTRONICS); INTEGRATED CIRCUITS; LOGIC CIRCUITS.

Robert D. Compton

### Isolators for Optical Systems

The need for optical isolation has broadened considerably since the advent of lasers. It is often necessary to prevent light from reentering the laser, irrespective of any electrical consideration. One example is a small laser followed by high-power laser amplifiers. If the powerful amplified light reenters the small (master oscillator) laser, it can destroy it. Another example is a frequency-stabilized laser, whose oscillation frequency is perturbed by reentering (injected signal) light. The general configuration of an optical isolator for optical systems is shown in Fig. 4.

**Quarter-wave polarizer.** A polarizer-plus-quarter-wave-plate isolator prevents laser light from reentering the laser when the light is scattered back by specular reflectors. This device cannot ensure isolation if

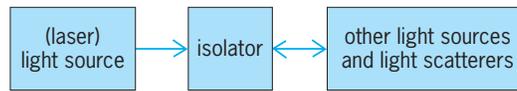


Fig. 4. General configuration of an optical isolator for optical systems.

there is diffuse reflection or if polarization-altering (birefringent) optics are encountered. Another limitation of this isolator is that the transmitted light is circularly polarized. The mechanism of operation is illustrated in Fig. 5, which shows what happens when light is reflected, double-passing the isolator. (Light reflected back through the system shows as if it were traversing a second system whose components are in reverse order.) See BIREFRINGENCE; POLARIZED LIGHT.

**Faraday rotator.** The Faraday effect is found in many solids, liquids, and gases. When plane polarized light is sent through the material in a direction parallel to a strong magnetic field, the plane of polarization is rotated (Fig. 6). The Faraday mechanism differs from that of a wave plate in that the plane of polarization is rotated through an angle  $\theta$  that is proportional to the path length  $l$ . That is, there is no fast or slow axis. This sort of corkscrew effect is called magnetically induced optical activity, in contrast to birefringence in the case of the wave plate. Because the corkscrew rotation has a particular handedness, specularly reflected (double-passed) light receives an additional rotation, which can be stopped by the entrance polarizer. Specifically, if the Faraday rotation

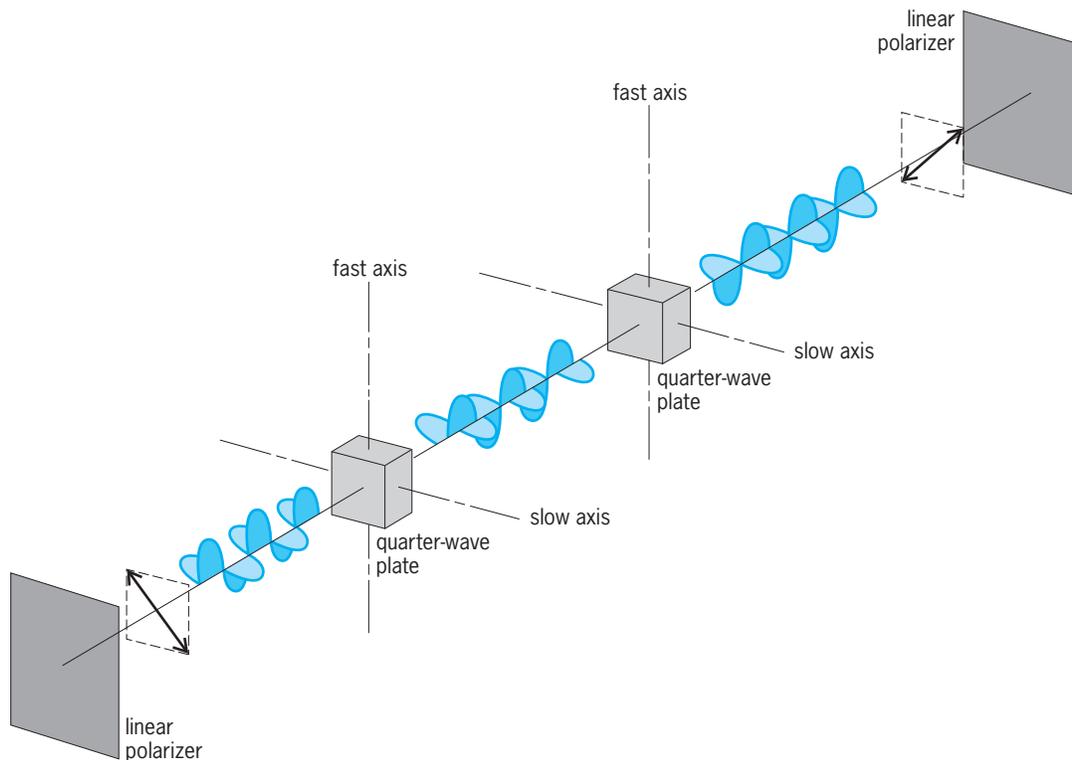


Fig. 5. Device consisting of quarter-wave plate and linear polarizer, showing how it serves as an isolator against specular reflections when they double-pass the system. Light reflected back through the system is shown as if it were traversing a second system whose components are in reverse order.

angle  $\theta$  equals  $45^\circ$ , then such reflected, double-passed light will be rotated  $90^\circ$  with respect to an entrance polarizer and will not be passed. See FARADAY EFFECT.

In contrast to the quarter-wave polarizer isolator, the Faraday isolator can provide truly one-way transmission irrespective of polarization changes from the exit side if an exit polarizer (which passes light that has undergone the Faraday rotation after passing through the entrance polarizer) is used in addition to the entrance polarizer. For example, it isolates against diffuse reflections and any light source on the exit side.

**Acoustooptic deflector.** The isolation properties of an acoustooptic deflector are based on the fact that light deflected by it is shifted in frequency by an amount equal to the acoustic frequency (Fig. 7). The reflected beam, passing through the deflector a second time, is again shifted in frequency by the same amount and in the same sense if the deflector is operated in the Bragg mode. Hence, the reflected light that is returned to the laser is shifted in frequency by an amount  $2f$ , where  $f$  is the frequency of the acoustic wave. Provided the frequency of the light returned to the laser is not close to any resonant frequency of the laser cavity, it will not perturb the laser and will simply be reflected from the output mirror. See ACOUSTOOPTICS.

For general use as isolators, acoustooptic deflectors have the limitation that they are really only frequency shifters. On the other hand, acoustooptic deflectors are quite insensitive to the polarization state of the light, which can be a serious problem with quarter-wave-polarizer and Faraday

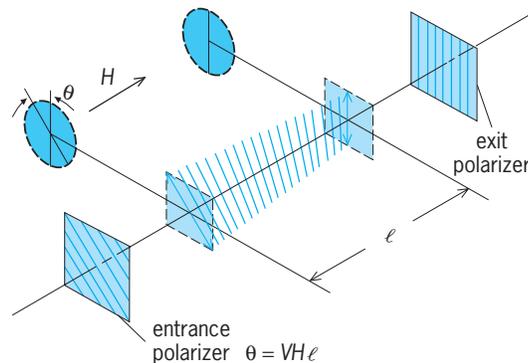


Fig. 6. Faraday effect.  $\theta$  = angle of rotation,  $V$  = Verdet constant,  $H$  = magnetic field strength,  $l$  = length of medium traversed.

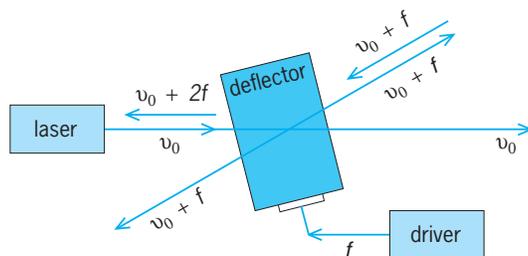


Fig. 7. Acoustooptic isolator, operated in the Bragg mode.

isolators. A combination of acoustooptic and Faraday isolators has been shown to avoid both these limitations. Stephen F. Jacobs

Bibliography. C. J. Georgopoulos, *Fiber Optics and Optical Isolators*, 1982; A. K. Ghatak and K. Thyagarajan, *Optical Electronics*, 1989; H. Lee, Optical isolator using acoustooptic and Faraday effects, *Appl. Opt.*, 26:969-970, 1987; A. Yariv, *Optical Electronics*, 5th ed., 1997.

## Optical materials

All substances used in the construction of devices or instruments whose function is to alter or control electromagnetic radiation in the ultraviolet, visible, or infrared spectral regions. Optical materials are fabricated into optical elements such as lenses, mirrors, windows, prisms, polarizers, detectors, and modulators. These materials serve to refract, reflect, transmit, disperse, polarize, detect, and transform light. The term "light" refers here not only to visible light but also to radiation in the adjoining ultraviolet and infrared spectral regions. At the microscopic level, atoms and their electronic configurations in the material interact with the electromagnetic radiation (photons) to determine the material's macroscopic optical properties such as transmission and refraction. These optical properties are functions of the wavelength of the incident light, the temperature of the material, the applied pressure on the material, and in certain instances the external electric and magnetic fields applied to the material. See ATOMIC STRUCTURE AND SPECTRA; DISPERSION (RADIATION); ELECTROMAGNETIC RADIATION; ELECTROOPTICS; INFRARED RADIATION; LENS (OPTICS); LIGHT; MAGNETOOPTICS; MIRROR OPTICS; OPTICAL DETECTORS; OPTICAL MODULATORS; OPTICAL PRISM; POLARIZED LIGHT; REFLECTION OF ELECTROMAGNETIC RADIATION; REFRACTION OF WAVES; ULTRAVIOLET RADIATION.

There is a wide range of substances that are useful as optical materials. Most optical elements are fabricated from glass, crystalline materials, polymers, or plastic materials. In the choice of a material, the most important properties are often the degree of transparency and the refractive index, along with each property's spectral dependency. The uniformity of the material, the strength and hardness, temperature limits, hygroscopicity, chemical resistivity, and availability of suitable coatings may also need to be considered. See HARDNESS SCALES; STRENGTH OF MATERIALS.

**Transmission.** Throughout the visible region of the spectrum, there are many transparent materials such as glasses, crystals, polymers, and plastics with desirable properties; however, in the ultraviolet and infrared regions the transmission range of the individual material becomes of primary importance. The bars in Fig. 1 represent the wavelength regions in which the optical materials transmit appreciably. More specifically, the bars indicate the wavelength range for which the external transmittance of the

given material is greater than 10% for a sample 1 mm thick. The wavelength scale is logarithmic.

Typical transmission curves (Fig. 2) give more detailed information concerning the wavelength dependence of the transmission of selected optical materials. The external transmittance is defined as the ratio of the intensity of light leaving a sample to the intensity of the light entering it. The external transmittance depends on both internal absorption and the reflection losses at the surfaces.

The need for new materials that will withstand high-power laser beams or transmit signals with minimum loss, as in optical fibers, indicates the significance of the transmission (or absorption) characteristics of a material. The equation below, called

$$I = I_0 e^{-\alpha x}$$

Bouguer's law, gives the attenuation of a light beam propagating through a solid, where  $I_0$  is the beam intensity after entering the solid,  $I$  is the intensity at a distance  $x$  into the solid, and  $\alpha$  is the absorption coefficient of the material. Absorption is the process in which the electromagnetic radiation energy is converted to internal energy (vibrational, rotational, and electronic potential energy) of the electrons, atoms, or ions in the lattice or ensemble of atoms within the material. The absorption loss and thus the transmission vary depending on the individual material, and are characterized by both wide and narrow bands of wavelengths over which the loss of energy occurs (Fig. 2). Reflection at the surface of a material can vary from about 4% (glass) to 36% (germanium). For normally incident light in air, reflection loss is given by  $[(n - 1)/(n + 1)]^2$ , where  $n$  is the refractive index of the material. See ABSORPTION OF ELECTROMAGNETIC RADIATION; CRYSTAL ABSORPTION SPECTRA.

The attenuation of the light beam as it traverses a solid can also result from scattering. Scattering varies with wavelength and results from a variety of sources that must be taken into account when choosing the proper material for fabrication into an optical element. Bouguer's law can be adjusted to include a scattering coefficient analogous to the absorption coefficient. See SCATTERING OF ELECTROMAGNETIC RADIATION.

**Refractive index.** For many optical elements such as lenses and prisms, the most important optical property is the refractive index, which is defined as the ratio of the speed of light in a vacuum to the speed of light in the material. The refractive index varies with wavelength; this variation is referred to as dispersion. The dispersion curves for selected optical materials are illustrated in Fig. 3, where the refractive index is plotted versus wavelength on a logarithmic scale. A wide range of refractive index exist for optical materials (see table) from as low as 1.3 to greater than 4.0, with optical glasses generally ranging from 1.5 to about 1.7.

The choice of an optical material for a particular application can be facilitated by a reference such as the table, which presents relevant physical properties for a variety of optical materials.

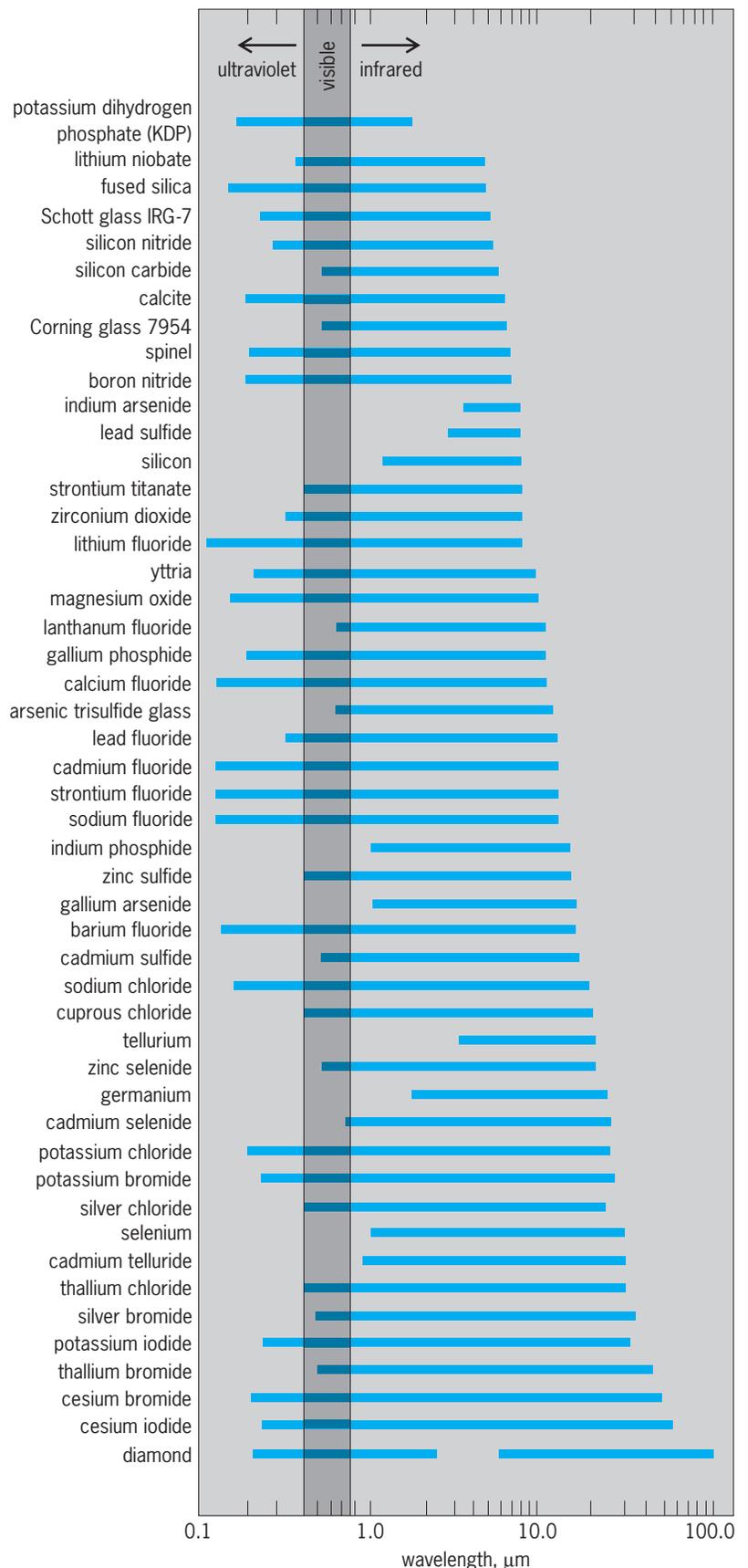


Fig. 1. Transmission range for a variety of optical materials. The bars indicate the wavelength region for which the external transmittance for a 1-mm-thick sample of the material is greater than 10% when measured at room temperature. The wavelength scale is logarithmic.

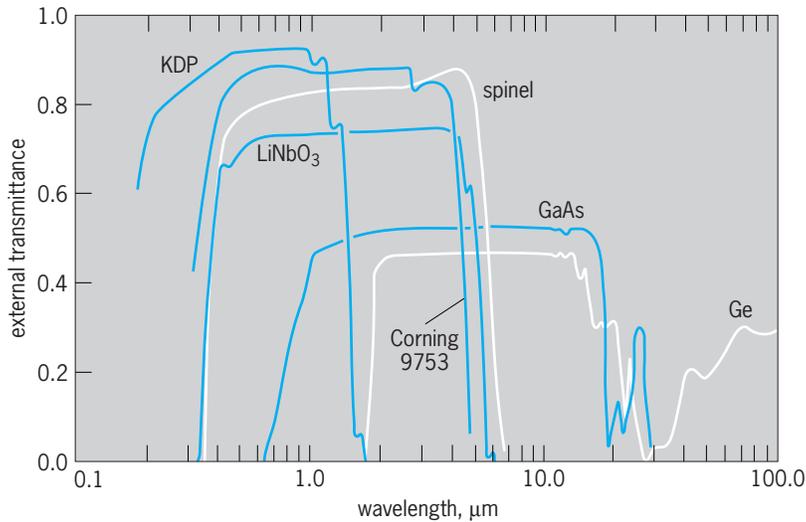


Fig. 2. Transmission versus wavelength for several optical materials. The wavelength scale is logarithmic.

**Glass.** The development of optical glass was due primarily to E. Abbe, C. Zeiss, and O. Schott. Glasses were the vanguard of optical materials because they transmit in the visible, provide a fairly wide range of refractive indices, can be easily shaped, and are relatively inexpensive. Early optical glasses were classified as either crown or flint. Crown glasses are composed mainly of silica, alkaline earth oxides, and alkali metal oxides. They exhibit low refractive index values and low dispersion. Flint glasses are composed typically of silica, alkali metal oxides, and lead oxides. They exhibit high refractive index values and high dispersion. Advances made by Schott in introducing the borosilicate, fluoride, and barium crown and flint glasses, and later by G. Morey in developing the rare-earth glasses, expanded the range of refractive indices and dispersive powers of optical glasses.

Hundreds of glass compositions are manufactured. Glass suppliers publish catalogs with comprehensive

data on each glass. These catalogs contain extensive tables of the refractive index of the glass versus wavelength. The Abbe number indicates the degree of dispersion of the glass. This is the ratio  $(n_d - 1)/(n_f - n_c)$ , where the subscripts on the refractive index  $n$  indicate the wavelength at which it was measured; the subscripts  $d$ ,  $c$ , and  $f$  refer to spectral lines at 589.3, 486.1 and 656.3 nanometers respectively. See CHROMATIC ABERRATION.

Glass technology provided the foundation for classical optical elements, such as lenses, prisms, and filters. Glasses developed for use in the visible region have internal transmittances of over 99% throughout the wavelength range of 380–780 nm. However, the silicate structure in glasses limits their transmission to about 2.5  $\mu\text{m}$  in the infrared. Chalcogenide glasses, heavy-metal fluoride glasses, and heavy-metal oxide glasses extend this transmission to 8–12  $\mu\text{m}$ . See COLOR FILTER.

Beginning with a breakthrough in the development of ultrapure glass in 1970, the process for manufacturing optical fibers changed from melting raw silica to forming the glass from vaporized chemicals. This new process is referred to as outside vapor deposition (OVD) or modified chemical vapor deposition (MOVD). These advances led to the present fiber-optic communication systems that operate in the near-infrared region with windows at wavelengths of 850, 1310, 1550, and 1625 nm. An advanced fiber-optic system, LEAF (Large Effective Area Fiber), was designed to minimize nonlinearities by spreading the optical power over large areas. See OPTICAL COMMUNICATIONS; OPTICAL FIBERS; VAPOR DEPOSITION.

The use of photolithography for printing integrated circuits has necessitated the improvement in the transmission of glasses for the ultraviolet region. Fused silica, which transmits to about 180 nm, is well suited for the lithography in the ultraviolet region. However, the crystalline material calcium fluoride, which transmits into the ultraviolet region to about 140 nm, outperforms any glass in printing microchips using fluorine excimer lasers. Corning 7940 and Suprasil are two commercially available fused-silica glasses designed specifically for deep-ultraviolet applications. They have uses as high-energy laser glasses, spacecraft window glasses, and mirror blanks for large astronomical mirrors. These and other fused-silica glasses are available for optical imaging in the ultraviolet region. Ultraviolet glasses combined with advanced ultraviolet laser sources and matched detectors have improved cancer detection by using laser-induced autofluorescence. See FLUORESCENCE; INTEGRATED CIRCUITS; TELESCOPE.

Glasses also serve an important function as host materials for lasing ions. Lawrence Livermore National Laboratories has created the largest combined system of glass lasers, known as NOVA. These lasers contain a series of Schott (LG-750 and LG-760) phosphate laser glass amplifiers. See GLASS; LASER.

**Plastics.** The need for an inexpensive, unbreakable lens that could be easily mass-produced precipitated the introduction of plastic optics in the mid-1930s. Although the variety of plastics suitable for

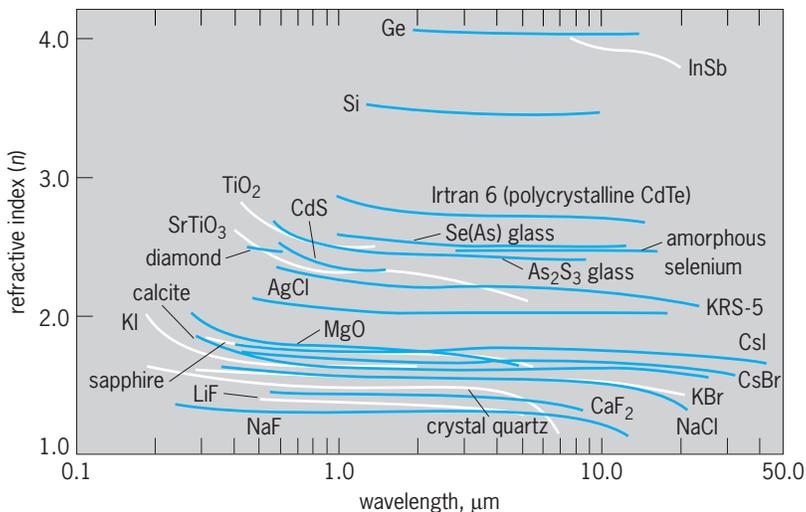


Fig. 3. Refractive index versus wavelength for several optical materials. The wavelength scale is logarithmic.

precision optics is limited compared to glass or crystalline materials, plastics are often preferred when difficult or unusual shapes, lightweight elements, or economical mass-production techniques are required.

The softness, inhomogeneity, and susceptibility to abrasion intrinsic to plastics often restrict their application. Haze (which is the light scattering due to microscopic defects) and birefringence (resulting from stresses) are inherent to plastics. Plastics also exhibit large variations in the refractive index with changes in temperature. Shrinkage resulting during the processing must be considered. *See* BIREFRINGENCE; PHOTOELASTICITY.

The refractive index of most usable plastics is from 1.45 to 1.60. Acrylic (methyl methacrylate) has an index of 1.49 and polystyrene an index of 1.59. An acrylic and polystyrene combination can be used to produce achromatic plastic lenses. Many other polymers, such as polyethylene, can be used in optical elements, along with the optical adhesives and epoxies. *See* POLYACRYLATE RESIN; POLYETHER RESINS; POLYOLEFIN RESINS; POLYSTYRENE RESIN.

Two of the more common plastic optical elements are the Fresnel lenses found in overhead projectors and the dichroic materials used in polarizing sunglasses. The need for mass-produced optical devices such as disposable contact lenses, compact discs, projection television systems, optical fibers, and electrooptical instruments has motivated the search for additional molecularly designed polymers.

**Polymeric materials.** Organic synthetic polymers are emerging as key materials for information technologies. In many cases, organic polymers compete with and even outperform inorganic materials. Polymers often have an advantage over inorganic materials because they can be designed and synthesized into compositions and architectures not possible with crystals, glasses, or plastics. They are manufactured to be durable, optically efficient, reliable, and inexpensive. Many uses of polymers in photonic and optoelectronic devices have emerged, including light-emitting diodes, liquid crystal-polymer photodetectors, polymer-dispersed liquid-crystal devices (for projection television), optical-fiber amplifiers doped with organic dyes (rhodamine), organic thin-film optics, and electrooptic modulators. Transparent optical fibers have been made from poly(methyl methacrylate). A dye-doped fiber of this polymer has been designed as an optical-fiber laser that is much less complicated and toxic than its liquid-dye laser counterpart. A fiber-optic polymer laser made of poly(methyl methacrylate-co-2-hydroxyethyl methacrylate) containing 0.1% of the rhodamine 6G dye has emitted yellow light pulses of about 0.5 millijoule, achieving high efficiency when pumped by a neodymium YAG laser. *See* LIGHT-EMITTING DIODE; LIQUID CRYSTALS.

Polymers can be produced that exhibit strong nonlinear optical properties that were earlier found only in inorganic crystalline materials. An especially promising class of nonlinear photonic materials is the photorefractive polymers. The photorefractive

effect is a nonlinear optical phenomenon in which materials exposed to relatively low-powered laser beams exhibit large refractive index changes. This effect was discovered in 1966 in inorganic crystalline materials. These nonlinear crystals prove more difficult to grow and develop than synthetic organic polymers. In 1991 the photorefractive effect was first observed in an organic material, the epoxy polymer bis- $\alpha$ -NPDA, made photoconductive by doping it with the hole transport agent DEH. The application of the photorefractive effect in polymers is expected to result in a leap forward in high-density information storage and high-speed optical processing. Other promising applications are optical encoding, dynamic real-time holography, image enhancement, phase conjugation, and advanced flat-screen television receivers and computer displays. *See* ELECTRONIC DISPLAY; FLAT-PANEL DISPLAY DEVICE; HOLOGRAPHY; IMAGE PROCESSING; NONLINEAR OPTICS; OPTICAL INFORMATION SYSTEMS; OPTICAL PHASE CONJUGATION; POLYMER.

**Crystalline materials.** Although most of the early improvements in optical devices were due to advancements in the production of glasses, the crystalline state has taken on increasing importance. Historically, the naturally occurring crystals such as rock salt, quartz, and fluorite plus suitable detectors permitted the first extension of visible optical techniques to harness the invisible ultraviolet and infrared rays.

In the 1930s, synthetic crystal-growing techniques made more readily available single crystals such as lithium fluoride (of special value in the ultraviolet region, since it transmits at wavelengths down to about 120 nm), calcium fluoride, and potassium bromide (useful as a prism at wavelengths up to about 25  $\mu\text{m}$  in the infrared). Many alkali-halide crystals (grown synthetically by the Czochralski technique) are important because they transmit into the far-infrared. The cubic symmetry of the alkali halide crystals results in isotropic optical and physical properties. Mixed crystals made of thallium bromide and thallium iodide called KRS-5 were grown in the late 1940s. *See* CRYSTAL GROWTH; CRYSTAL STRUCTURE; SINGLE CRYSTAL.

In the 1950s, following the invention of the transistor, germanium and silicon ushered in the use of semiconductors as infrared optical elements or detectors. Polycrystalline forms of these semiconductors could be fabricated into windows, prisms, lenses, and domes by casting, grinding, and polishing. Compound semiconductors such as gallium arsenide (GaAs, made of elements from the 13 and 15 columns of the periodic table), ternary compounds such as gallium aluminum arsenide ( $\text{Ga}_{1-x}\text{Al}_x\text{As}$ ), and quaternary compounds such as indium gallium arsenide phosphide (InGaAsP) now serve as lasers, light-emitting diodes, and photodetectors. *See* SEMICONDUCTOR.

In the 1960s the so-called polycrystalline compacts were developed. Pure powders of several infrared-transmitting materials were heated and compacted into blanks that could be fabricated into

desired infrared optical elements. An improvement over the hot-pressing techniques is the chemical vapor deposition (CVD) process. Polycrystalline zinc sulfide and zinc selenide manufactured by this process provide superior infrared materials available in diameters greater than 1 ft (0.3 m) and thicknesses of 1 in. (2.5 cm) for large windows in optical systems operating at wavelengths from 8 to 12  $\mu\text{m}$ .

Polycrystalline diamond films, or coatings, are produced by a wide variety of chemical vapor deposition techniques. These techniques require high temperatures that preclude substrate materials with melting points below 1000–1400 K (1340–2060 $^{\circ}\text{F}$ ); thus diamond coatings on plastics are not possible. The most used substrate has been single-crystal silicon. Diamond's extreme properties, high thermal conductivity, low electrical conductivity, low thermal expansion, extreme hardness, and large range of opti-

cal transparency make it suitable for diverse applications. Diamond produced by chemical vapor deposition and single-crystal diamond (produced in the laboratory) share these benefits. See DIAMOND.

Single crystals are indispensable for transforming, amplifying, and modulating light. Birefringent crystals serve as retarders, or wave plates, which are used to convert the polarization state of the light. In many cases, it is desirable that the crystals not only be birefringent, but also behave nonlinearly when exposed to very large fields such as those generated by intense laser beams. A few examples of such nonlinear crystals are ammonium dihydrogen phosphate (ADP), potassium dihydrogen phosphate (KDP), beta barium borate (BBO), lithium borate (LBO), and potassium titanyl phosphate (KTP) [see table]. See CRYSTAL OPTICS.

Nonlinear optics continues at a steady pace with

Physical properties of optical materials*										
Material	Symbol	Refractive index (wavelength = 500 nm)	Density, g/cm <sup>3</sup>	Hardness: Knoop, kg/mm <sup>2</sup> ; or Mohs (M)	Solubility, g/100 g H <sub>2</sub> O	Specific heat, cal/(g·K) <sup>†</sup>	Melting or softening temp., K	Thermal conductivity, W/(m·K) <sup>‡</sup>	Linear expansion coefficient, 10 <sup>-6</sup> /K <sup>‡</sup>	Young's modulus, <sup>‡</sup> GPa
<b>Crystalline materials</b>										
Germanium	Ge	4	5.33	800	Insoluble	0.074	1210	59	6.1	102.66
Lithium fluoride	LiF	1.394	3.5	100	0.27	0.37	1140	11.3	34.4	64.77
Magnesium fluoride	MgF <sub>2</sub>	1.39	3.18	415	Insoluble	0.24	1528	21	14(P), 8.9(S)	138.5
Sodium chloride	NaCl	1.53	2.17	15.2	35.7	0.2	1070	6.5	40	39.96
Zinc sulfide	ZnS	2.42	4.08	230	Insoluble	0.112	2100	17	6.6	74.5
Zinc selenide	ZnSe	2.43	5.42	137	0.001	0.0090	1790	19	7	70.97
Barium titanate	BaTiO <sub>3</sub>	—	5.9	200–580	—	0.103	1870	1.34	19	33.76
Cesium iodide	CsI	1.75	4.51	—	44	0.048	894	1.1	48.3	5.3
Diamond	C	2.4	3.51	5700–10400	Insoluble	0.124	3770	2600	0.8	1050
Lanthanum fluoride	LaF <sub>3</sub>	1.6	5.94	4.5(M)	Insoluble	0.121	1766	5.1	11(P), 17(S)	—
Magnesia	MgO	1.74	3.585	910	Insoluble	0.24	3053	40.6	8	—
Potassium chloride	KCl	1.49	1.98	7.2	34.7	0.162	1050	6.7	36.6	29.63
Sapphire	Al <sub>2</sub> O <sub>3</sub>	1.77	3.98	1370	Insoluble	0.18	2300	33.0	8(P) 5(S)	335
Crystal quartz	SiO <sub>2</sub>	1.55	2.65	741	Insoluble	0.17	1740	10.7(P), 6.2(S)	8.0(P), 13.4(S)	97.2(P), 76.5(S)
Barium fluoride	BaF <sub>2</sub>	1.47	4.89	0.82	0.17	0.096	1550	11.7	19.9	53.05
Calcium fluoride	CaF <sub>2</sub>	1.43	3.18	140	0.0017	0.204	1630	10	18.9	75.79
Cadmium telluride	CdTe	2.69	6.2	56	Insoluble	0.056	1320	6.3	5.9	36.52
Calcite	CaCO <sub>3</sub>	$n_o = 1.665,$ $n_e = 1.490$	2.710	3(M)	0.0014	0.203	1612	5.526(P), 4.646(S)	25(P), -5.8(S)	72.35(P), 88.19(S)
Cuprous chloride	CuCl	2.0	4.14	2–2.5(M)	0.0061	—	695	—	10	—
Gallium phosphide	GaP	3.65	4.13	845	Insoluble	0.2	1623	54	4.7	102.6
Indium arsenide	InAs	4.5	—	330	Insoluble	0.06	1215	50	5.3	—
Lead fluoride	PbF <sub>2</sub>	1.78	8.24	200	0.064	0.085	1100	—	29	—
Lead sulfide	PbS	4.3 at 3 $\mu\text{m}$	7.5	—	—	0.050	1387	0.67	18	—
Silicon carbide	SiC	2.68	3.217	2130–2755	Insoluble	0.165	3000	490	2.8	386
Selenium	Se	2.83	4.82	2.6(M)	Insoluble	0.077	490	1.3	48.7	—
Silicon	Si	3.45	2.329	1100	Insoluble	0.18	1690	163	2.6	130.91
<b>Infrared glasses</b>										
Arsenic trisulfide glass	As <sub>2</sub> S <sub>3</sub>	2.7	3.43	109	Insoluble	0.109	483	0.1674	24.62	15.85
Germanate glass (Schott IRG2)	—	1.9	5.00	481	0.012	0.108	—	0.91	8.8	95.9
Chalcogenide glass (Schott IRG 100)	—	2.73	4.67	150	—	—	624	0.3	15	21
Lead silicate glass (Schott IRG7)	—	1.57	3.06	379	0.171	0.151	—	0.73	9.6	59.7
Ultra-low-expansion (ULE) titanium silicate (Corning 7971)	—	1.484	2.205	459	Insoluble	0.183	1763	1.31	0.015	67.52
Germanate glass (Corning 9754)	—	1.67	3.581	560	—	0.13	1147	1.0	6.2	84.1
Phosphate laser glass (Schott LG750)	—	—	—	—	—	—	—	0.52	13.2	—

Physical properties of optical materials* (cont.)										
Material	Symbol	Refractive index (wavelength = 500 nm)	Density, g/cm <sup>3</sup>	Hardness: Knoop, kg/mm <sup>2</sup> ; or Mohs (M)	Solubility, g/100 g H <sub>2</sub> O	Specific heat, cal/(g·K) <sup>†</sup>	Melting or softening temp., K	Thermal conductivity, W/(m·K) <sup>‡</sup>	Linear expansion coefficient, 10 <sup>-6</sup> /K <sup>‡</sup>	Young's modulus, <sup>‡</sup> GPa
<b>Ultraviolet glasses</b>										
Fused silica	SiO <sub>2</sub>	1.43	2.203	461	Insoluble	0.22	1448	1.38	0.55	73.1
Fused silica (Corning 7940)	—	1.46	2.202	500	—	0.177	1858	1.38	0.52	73
<b>Nonlinear/photorefractive optical materials</b>										
Ammonium dihydrogen phosphate (ADP)	NH <sub>4</sub> H <sub>2</sub> PO <sub>4</sub>	1.51	1.803	—	22.7	—	4.63	.71162(P), 1.2558(S)	39.3	—
Lithium niobate	LiNbO <sub>3</sub>	$n_o = 2.286$ , $n_e = 2.203$ (0.6328 $\mu\text{m}$ )	4.64	5(M)	Insoluble	—	1523	38	2.2(P), 2.0(S)	—
Potassium dihydrogen phosphate (KDP)	KH <sub>2</sub> PO <sub>4</sub>	1.5	2.338	—	33	0.21	525.6	1.3395	21.6	—
Yttrium vanadate	YVO <sub>4</sub>	$n_o = 1.9929$ , $n_e = 2.2154$ (0.6328 $\mu\text{m}$ )	4.22	5(M)	Insoluble	—	—	5.23(P), 5.10(S)	11.37(P), 4.43(S)	—
Iron-doped lithium niobate	Fe:LiNbO <sub>3</sub>	$n_o = 2.286$ , $n_e = 2.203$ (0.6328 $\mu\text{m}$ )	4.64	5(M)	Insoluble	—	1523	38	2.2(P), 2.0(S)	—
Silver gallium sulfide	AgGaS <sub>2</sub>	$n_o = 2.4521$ , $n_e = 2.3990$ (1.064 $\mu\text{m}$ )	4.702	—	—	—	1270	0.015	12.5(P), -13.2(S)	—
Silver gallium selenide	AgGaSe <sub>2</sub>	$n_o = 2.7010$ , $n_e = 2.6792$ (1.064 $\mu\text{m}$ )	5.700	—	—	—	1124	—	16.8(P), -7.8(S)	—
Magnesium oxide-doped lithium niobate	MgO:LiNbO <sub>3</sub>	$n_o = 2.286$ , $n_e = 2.203$ (0.6328 $\mu\text{m}$ )	4.64	5(M)	Insoluble	—	1523	38	2.2(P), 2.0(S)	—
Potassium titanyl phosphate (KTP)	KTiOPO <sub>4</sub>	$n_x = 1.78$ , $n_y = 1.79$ , $n_z = 1.89$	3.01	5(M)	Insoluble	0.1643	1445	13	—	—
Lithium triborate (LBO)	LiB <sub>3</sub> O <sub>5</sub>	$n_x = 1.58$ , $n_y = 1.61$ , $n_z = 1.62$ ,	2.47	6(M)	—	—	1107	—	—	—
Cadmium sulfide	CdS	2.6	4.82	122	Insoluble	0.0882	1773	15.91	2.1(P), 4(S)	—
Gallium arsenide	GaAs	3.35	5.32	731	Insoluble	0.076	1511	55	5.7	82.68
Beta-barium borate (BBO)	$\beta$ -BaB <sub>2</sub> O <sub>4</sub>	$n_o = 1.6749$ , $n_e = 1.5555$	3.85	4.5(M)	—	—	1363	1.6(P), 1.2(S)	36(P), 4(S)	—

\*All values are given for temperatures near room temperature.  
<sup>†</sup>1 cal = 4.18 J.  
<sup>‡</sup>P denotes values measured parallel to the c axis of the crystal, and S denotes values measured perpendicular to the c axis.

new types of crystals becoming available on a regular basis. Important applications are second-harmonic generation (frequency doubling), parametric amplification (using two monochromatic waves to amplify a third), parametric oscillation (feedback is added to the parametric amplifier to create an oscillator), optical amplification, and optical phase conjugation.

A promising nonlinear phenomena is the photorefractive effect. This effect was first observed in 1966 in the crystal lithium niobate, and soon afterward a variety of nonlinear optical materials were discovered (see table), including potassium niobate, strontium barium niobate, bismuth silicon oxide, and barium sodium niobate. Many applications of the photorefractive effect were mentioned earlier in regard to polymers.

**Other materials.** Other optical materials that deserve mention are the liquid crystals used in displays as light valves, materials used in erasable optical discs for computers and in liquid cells (Kerr cells), laser dyes, dielectric multilayer films, filter materials, and

the many metals (aluminum, gold, beryllium, and so forth) and alloys that are important as coating materials. See COMPUTER STORAGE TECHNOLOGY; KERR EFFECT; OPTICAL RECORDING. James Steve Browder

Bibliography. H. Bach and N. Neuroth (eds.), *The Properties of Optical Glass*, 1995; K. S. Hansen (ed.), *Fiber Optic Reference Guide*, 2d ed., 1999; S. A. Jenekhe and K. J. Wynne (eds.), *Photonic and Optoelectronic Polymers*, 1997; I. M. Khan and J. S. Harrison (eds.), *Field Responsive Polymers*, 1999; P. Klocek (ed.), *Handbook of Infrared Optical Materials*, 1991; Optical Society of America, *Handbook of Optics*, vols. 1, 2, 2d ed., 1995, vols. 3, 4, 2000; R. Wood, *Optical Materials*, 1993.

## Optical microscope

An instrument used to obtain an enlarged image of a small object. In general, a compound microscope consists of a light source, a condenser, an objective,

and an ocular or eyepiece, which can be replaced by a recording device such as a photoelectric tube or a photographic plate. The optical microscope is limited by the wavelengths of the light used and by the materials available for manufacturing the lenses.

### Lenses

The quality and design of the lens system determines the magnifying power, details of image formation, and color correcting capabilities of a light microscope.

**Magnifying power.** The magnifying power of a compound microscope is the product of the magnification of the objective and the magnifying power of the eyepiece. The latter is computed like that of any magnifier. The magnification of the objective is equal to the distance from the second focal point to the image formed by the objective, divided by the focal length. An objective of 18-mm (0.7-in.) focal length thus has a power of  $10\times$ . It is customary to specify objectives in terms of magnifying power instead of focal length. The distance mentioned is called the optical tube length (generally 180 mm or 7 in.), and is to be distinguished from the mechanical tube length, which is the length of the mechanical tube itself. See MAGNIFICATION.

**Objectives.** A microscope objective consists of a set of achromatic lenses which are partially or wholly corrected for longitudinal color, aperture errors, and asymmetry errors. The numerical aperture (NA) of the system is given by  $n \sin u$ , where  $n$  is the refractive index of the object space and  $u$  is the angle made by the ray of largest aperture and the axis.

A microscope objective generally consists of a collection of positive lenses comparatively close together (Fig. 1a). Color magnification and curvature of field in the objective are frequently balanced out at least partly by the ocular. See ABERRATION (OPTICS).

For high magnifications, the first lens is planoconvex, with the convex surface either concentric or aplanatic to the aperture rays (Fig. 1b). To increase the aperture, the object can be embedded in an immersion liquid (a special oil) having a refractive index equal or nearly equal to the index of the first lens. Such an arrangement is termed a homogeneous im-

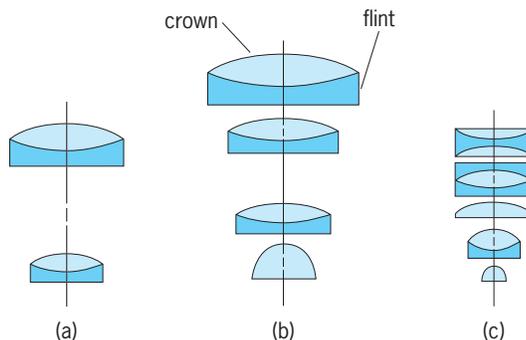


Fig. 1. Microscope objectives: (a) 16-mm (0.63-in.) dry achromat; (b) typical high-power dry achromat; (c) 2-mm (0.08-in.), NA 1.40, oil-immersion apochromat. (Photographic Service Department, Kodak Research Laboratory)

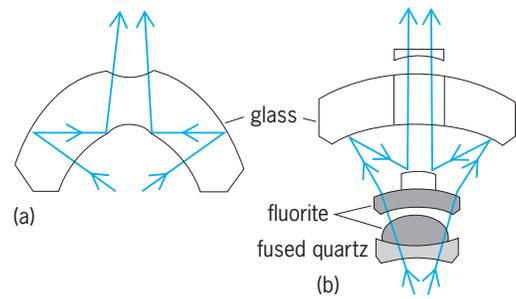


Fig. 2. Two types of catadioptric objective. (a) Maksutov type. (b) 53X, NA 0.72, ultraviolet objective, designed by Gray. Glass elements in the latter serve purely as reflectors. (Photographic Service Department, Kodak Research Laboratory)

mersion system (Fig. 1c). A cover glass of fixed but small thickness separates the object from the immersion liquid for biological investigations.

**Color correction.** Achromatic color correction is not good enough for lenses of high aperture. The use of fluorite led to objectives corrected for more than two colors, and most achromatic lenses of high aperture are now made with crystal fluorite. An objective whose chromatic errors are corrected for three colors and whose aperture and asymmetry errors are also corrected is termed apochromatic. The lateral chromatism of such an objective is great and must be removed by a special compensating eyepiece. See CHROMATIC ABERRATION; EYEPIECE.

High-aperture microscopes for a field larger than the usual  $3^\circ$  have been designed by adding either a negative lens or a very thick positive lens having a negative Petzval sum. Such objectives are termed plane-achromats.

Because resolving power depends upon wavelength, objectives which are corrected for ultraviolet radiation have been designed. Such objectives are generally corrected for a single wavelength and are termed monochromats. They are made of quartz.

**Catadioptric systems.** Catadioptric systems have been developed for microscopes. Their great advantage is their comparatively small chromatic aberration. Pure mirror systems have no color aberrations. In catadioptric systems, therefore, it is customary to assign all the power to the mirror or mirrors, keeping the refracting system nearly afocal (Fig. 2a). The chromatic errors of the entire system remain small, and the refracting part can be used to correct the remaining monochromatic errors. However, in catadioptric systems part of the aperture is obscured by the mirror and the ensuing diffraction may damage the fine detail in the image. All microscopic work in the ultraviolet region is done with catadioptric systems (Fig. 2b). See SCHMIDT CAMERA.

**Image formation.** Geometrical optics is not sufficient to explain all the details of image formation at high magnifications. According to geometrical optics, a point should be imaged by a perfect objective as a point, but there is diffraction at the aperture, and there may also be diffraction at the object. Diffraction theory applied to the aperture shows that, because

of the finite aperture of the optical system, a spherical concentric wave bounded by the exit pupil produces a light pattern in the image plane in which the light is distributed over a disk of diameter  $1.22\lambda/NA$  with faint rings outside the disk. If the magnifying power of the microscope becomes so large that the rings are visible, the image will contain details which are not in the object. This is undesirable. Thus, the aperture determines the useful magnification of the microscope; it is important that the aperture be enlarged when the magnification has to be increased. See DIFFRACTION.

Because the microscope is used to study very small objects at usually high magnifications, diffraction at the aperture is more noticeable in microscopes than in other optical systems. Moreover, most objects viewed with microscopes are so small that there is a significant amount of diffraction at the object, as is seen in the case of dark-field illumination and phase microscopy.

This is of special importance if the object has a periodic structure since even a point light source will then give rise to an imagelike structure. Interference of the rays coming from the light source in phase and being diffracted at the structured object causes this effect and gives rise to sets of images in different planes. This theory (E. Abbe's theory of image formation in the microscope) has been successfully carried further in F. Zernike's theory, which led to the construction of the phase microscope.

In the case of illumination with a large cone of light, these interference patterns fall together at the gaussian focus, so that there is only one image plane. The results in this case, however, can also be derived from Abbe's theory, but only if integration over the aperture is carried out. See INTERFERENCE MICROSCOPE; PHASE-CONTRAST MICROSCOPE.

### Condensers

An external auxiliary lens is used to condense the light from a light source so that the object is brightly and uniformly illuminated. The usual purpose of a condenser system is to make sure that as much light as possible coming from the object goes through an optical system.

Condensers are used in macroscopic projection, in which an illuminated film or slide is imaged with the help of a projection objective or magnifier. In microscope systems, they are used to direct the light from a light source so that the rays from any object point fill most of the entrance pupil.

A condenser system is usually arranged to image the light source onto the entrance pupil of the optical system (Köhler illumination). The condenser is generally corrected for spherical aberration, color, and sine condition, although the requirements are slightly different than in an image-forming system. Condensers frequently consist of a number of planoconvex lenses with the plane side toward the objective. Sometimes one surface is made aspheric to improve the light concentration. Condensers for projection optics are rarely achromatized, but the effect of color magnification is decreased by vignetting

the colored borders. For microscope substage condensers, however, achromatism is a necessary requirement.

The aperture of the condenser must be at least as large as that of the objective with which it is used. Because microscope objectives are generally designed in such a way that they are excellently corrected for color, aperture, and asymmetry errors for only about  $\frac{7}{8}$  of their aperture, the condenser need fill only this much of the entrance pupil. A good test of whether the condenser of a microscope is well adjusted is to remove the object and ocular and see whether the exit pupil of the objective is filled uniformly with light up to  $\frac{7}{8}$  of its aperture. Max Herzberger

### Types of Microscope

The types of microscope discussed in this section are variations of the light- or bright-field microscope.

**Light microscope.** The mirror, condenser, oculars, and body tube of the light microscope are frequently known as the optical train. The stand, stage, and adjustments comprise the mechanical part of the microscope (Fig. 3).

A mirror is usually attached to the substage of the microscope to reflect light along the optic axis of the microscope. When no condenser is used, the concave mirror is used because it concentrates more light on the specimen; a plane mirror is used with a condenser.

Laboratory microscopes are usually supplied with an uncorrected two-lens Abbe condenser with NA 1.25, consisting of a double convex lower and a hyperhemisphere upper lens. A 1.40 NA Abbe is useful for concentrating radiation for fluorescence microscopy. Well-corrected objectives require aplanatic condensers corrected for chromatic aberration for efficient observation.

Objectives vary from a simple doublet lens to complex corrected lens systems. Achromatic objectives are corrected for spherical aberration in one color and for chromatic aberrations in two colors. Apochromatic objectives are corrected to focus three colors together and the spherical aberration is minimized for two colors. Some semiapochromatic and fluorite objectives are of intermediate correction.

With air between the specimen and the objective the maximum NA is about 0.92. Water-immersion objectives have a greater NA, and with immersion oil, an NA of 1.4 is available. Objectives with an NA of 1.6 have been made but require special immersion and mounting media. The resolving power of an objective, the least distance at which two objects can be seen to be separate, is equal to the wavelength of light  $\lambda$  divided by the sum of the numerical apertures of the condenser and objective used. The larger the numerical aperture, the greater is the resolving power. The depth of field seen in focus at one time and the working distance of the objective decrease with an increase in the numerical aperture. The light passed through a microscope is proportional to the square of the numerical aperture and to the inverse of the square of the magnification. Objectives are described also by the equivalent focal length.

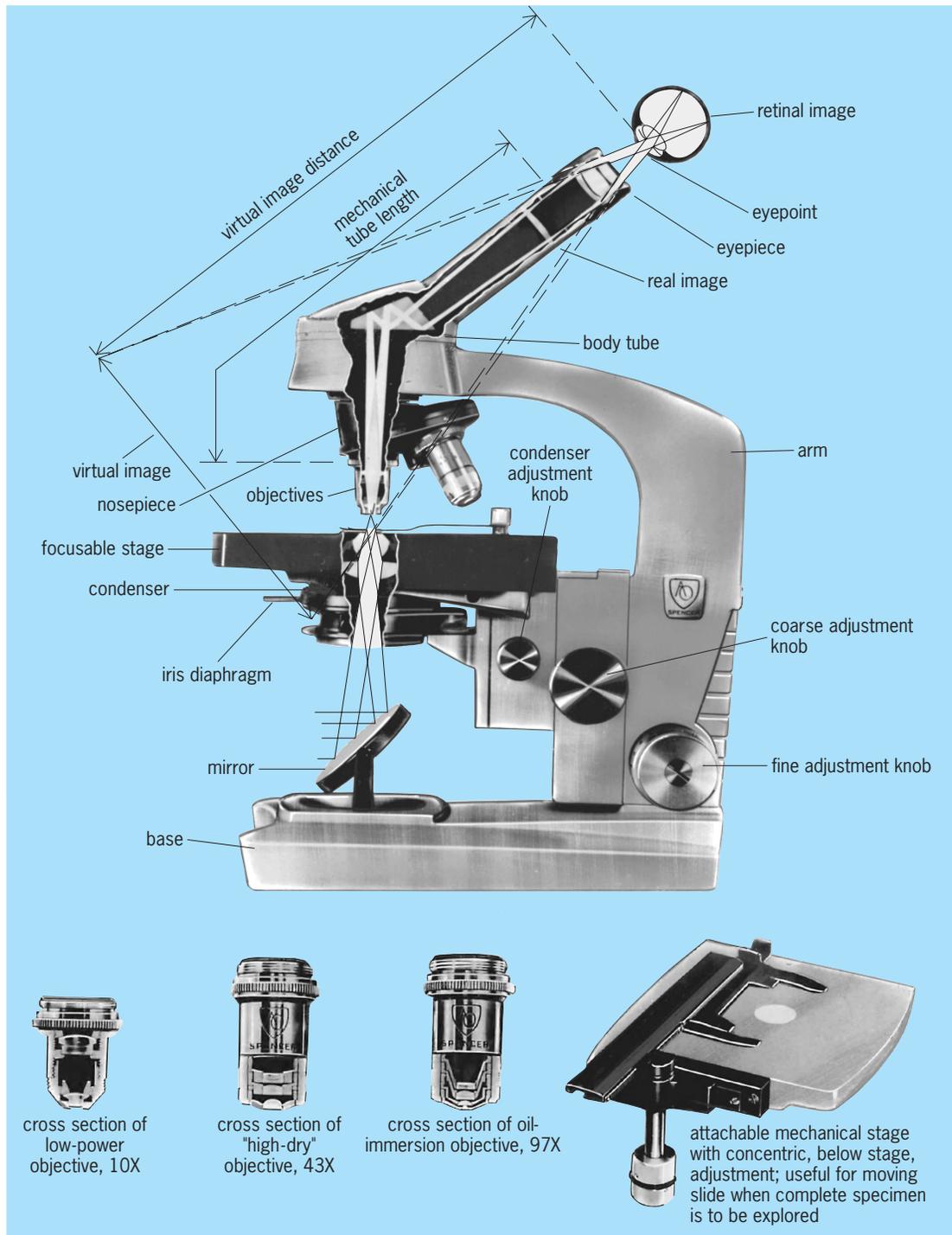


Fig. 3. Diagram of light microscope. (American Optical Corp.)

Objectives of shorter focal length have less depth of field, less working distance, and greater magnification.

Photomicrographic objectives are designed to produce a flat image with little distortion. Some objectives obtain a flatter field by means of a concave rather than a flat front lens. Apochromatic objectives are undercorrected and compensating oculars must be used with them to complete the correction for color and for best resolution.

For convenience, two to five objectives can be mounted on a revolving nosepiece to be parfocal and parcentric, so that the specimen remains almost in focus at the center of the field as the objectives are changed. For more critical work, utilizing interference and polarizing microscopes, individually adjustable quick-change nosepieces are employed.

The commonly used Huygenian ocular has a fairly flat field with marked pincushion distortion. Compensating oculars complete the color correction for

apochromatic objectives and have less distortion, but they do have curvature of field. To obtain a flat field with minimum distortion for microprojection or photomicrography special projection oculars are designed, one type with a color-corrected minus lens called a negative amplifier, and the other a positive-projection ocular with a focusable eye lens. Oculars with a high exit pupil are useful for spectacle wearers. Other oculars are designed to give a wide field of view.

The monocular body tube may be of adjustable length. American microscopes are designed for a mechanical tube length of 160 mm (6.3 in.) and a cover-glass thickness of 0.18 mm (0.007 in.). The draw tube is lengthened for thinner and shortened for thicker cover glasses to correct for the spherical aberrations from cover glasses of incorrect thickness.

Binocular bodies are designed for the use of both eyes. Most binocular bodies use prisms to reflect one-half of the light to each eye. Because each eye sees the same field, these binocular bodies do not give stereoscopic vision. The binocular is often longer than the monocular body and the proper tube length is maintained with a compensating lens. Because the tube length is fixed, it is essential for critical microscopy with binocular bodies to use cover glasses 0.18 mm thick. Binocular bodies can be made with stops, or polarizing materials, so that each eye sees with the corresponding half-aperture of the objective and stereoscopic vision is possible, but at one-half the resolving power that would have been obtained with full aperture.

Trinocular bodies are binocular bodies with a third tube for a camera.

The lowest magnification which will resolve the detail required should be used. High magnification gives images that are less bright and therefore difficult to see. Magnification greater than that required for complete visible detail is called empty magnification and is useless for seeing, although sometimes helpful for measurement. For visual microscopy the total magnification (magnification of the objective times that of the ocular) should be about 1000 times the numerical aperture in use. Apochromatic objectives and compensating oculars are desirable for best vision and color photomicrography.

**Inverted microscope.** The inverted microscope has the body of the microscope, including the objective and the ocular, below the stage and the illumination above the stage for transmitted light. With opaque materials, the vertical illuminator is used under the stage near the objective. The inverted microscope is especially useful for the examination of surfaces (Fig. 4). Specimens placed on the stage can be held substantially in focus. Large and awkward specimens can be moved over the stage more readily than with the usual microscope. The inverted microscope is also useful for microdissection and the observation of hanging-drop preparations and is convenient for observing chemical reactions, melting-point determinations, and photomicrography. The camera can be included in the base, as in the metallograph microscope, for stability. Either monocular or binoc-

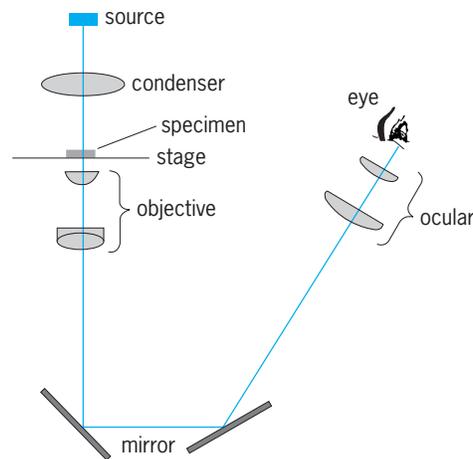


Fig. 4. Optical diagram of inverted microscope.

ular bodies can be used with the inverted microscope.

**Comparison microscope.** The comparison microscope is an arrangement of two microscopes connected by a special viewing ocular so that the field of one microscope is seen at one side of a vertical dividing line and the field of the other microscope on the opposite side of the dividing line; or it may be a projection type of microscope in which the image is compared with a template or known pattern.

When two microscopes and a comparison ocular are used, the magnifications of each must be matched. The specimens are placed on the microscopes and usually require separate lighting. A common application is the examination of bullets. A test bullet fired from the suspected gun is placed under one microscope and the bullet recovered from the scene of the crime is placed under the other microscope. Holders which allow rotation of the bullets are turned, and if the grooves on the two bullets match, it is evident that they came from the same gun. Comparison microscopes can be used to compare grain size, structure, distribution of elements, color, and various other characteristics of any two specimens. Either transmitted light, vertical illumination, or a combination of lighting types can be used.

A projection microscope with a built-in screen can also be used as a comparison microscope. The image of the part is focused on the screen, and the contours of the image are measured with scales or compared with a template of the desired form, for example, a profile of a gear or a screw thread. The stage is usually modified to have a proper holder for the kind of specimen to be examined.

**Dissecting microscope.** Dissecting microscopes are of two types. The simplest is a magnifying glass mounted on a support above a glass plate, used for the dissection of materials. A mirror may be present to reflect light to the specimen.

The more usual dissecting microscope, often called a Greenough microscope, is a stereoscopic microscope composed of two separate microscopes

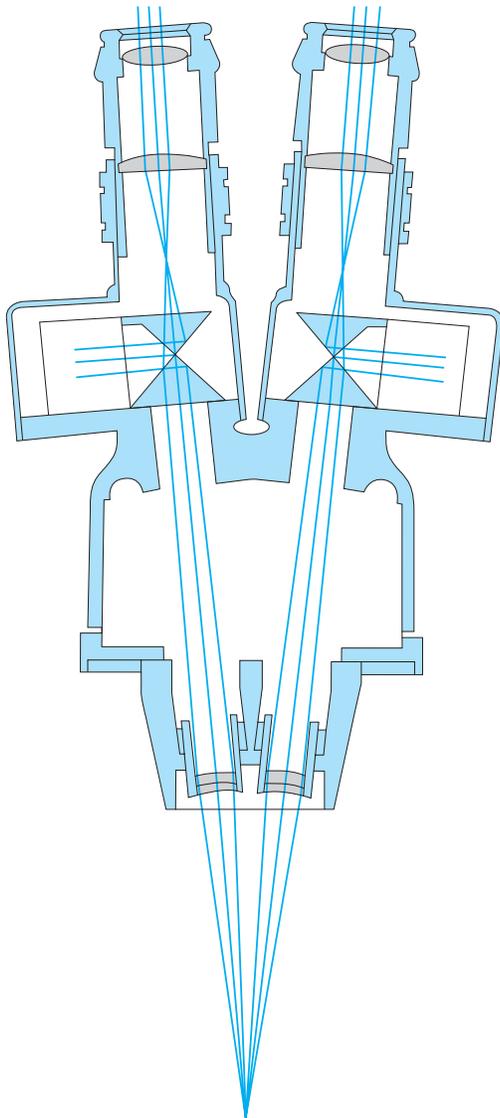


Fig. 5. Diagram of light rays as they pass through a binocular biobjective microscope.

fastened together and used as a single unit on one stand (Fig. 5).

This is a truly stereoscopic instrument because the right eye sees the specimen from the right side and the left eye from the left side. Prisms are usually included in the body tube to erect the image; thus movements of the specimen are direct and are not reversed as with the monobjective microscope. Because two objectives are required, the mechanical difficulty of placing the lenses close together limits the numerical aperture to approximately 0.12. There is no advantage in using more than 120 diameters magnification because further magnification would be empty or useless.

The binocular microscope was developed for biological dissecting, but it is used extensively in industry for the examination and assembly of small parts such as transistors. The body and focusing adjustment can be mounted on lathes and other machinery when the work must be controlled visually to close

tolerances; they can be mounted on a stand with a long arm for examination of large specimens or may be used to assist the surgeon in the operating room.

When the angle of the objectives and oculars is the same, the specimen is seen in true depth. By changing these angles the perception of depth can be increased or decreased. Hyperstereoscopy is useful for biological dissection.

Examination with the stereoscopic microscope is helpful in orienting the microscopist to the specimen before instruments of greater magnification are used. By proper illumination both opaque and transmitted materials can be examined.

**Metallurgical microscope.** The metallurgical microscope is a laboratory microscope with a focusing stage and a vertical illuminator, used primarily for the examination of metal surfaces.

The specimens are usually embedded in molded plastic so that the surface is at a definite position, surfaced with a series of increasingly finer abrasives, and polished. To differentiate the constituents of the metal, the surface is etched lightly by chemical treatment before examination with the microscope.

The metallurgical microscope shows the structure of the metal, including grain size, as well as the nature and distribution of the components. The roughness or polishability of the metal can be studied. Because many of the metals used commercially are mixtures, or alloys, rather than pure metals, the metallurgical microscope is important for analysis and for assessing the effects of heat treatment and surface changes.

### Microscopy

This discussion of microscopy considers the following with relation to the optical microscope: illumination, calibration and measurement, immersion fluids, mountants, and the hanging-drop method.

**Illumination.** Illumination of the microscope is obtained from a bright surface or from a luminous source concentrated with the aid of condensing lenses. Sources commonly used are tungsten-filament lamps and carbon, mercury, and xenon arcs.

With an illuminated-surface light source the lamp is positioned and the microscope condenser focused so that the field and aperture of the microscope are lighted as uniformly as possible.

Illuminators with focusing condensers and iris diaphragms, called research lamps, are used for critical, Köhler, and other more efficient methods.

**Critical illumination.** In critical illumination an image of the light source is focused on the specimen from a uniform source, for example, a ribbon-filament lamp (Fig. 6). Parallel light from an illuminator focused to infinity is directed into the microscope condenser which is then focused so that the image of the source is in the plane of the specimen.

**Köhler's method.** This method is used with coiled filaments or other sources of irregular form or brightness to obtain a uniformly illuminated field. An image of the filament large enough to fill the opening of the iris is focused on the microscope condenser. The microscope condenser then is focused so that the image of the iris diaphragm on the lamp is in

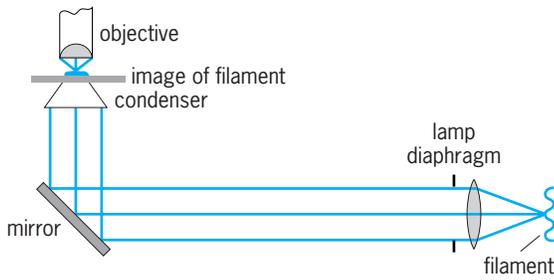


Fig. 6. Light path in critical illumination. (American Optical Corp.)

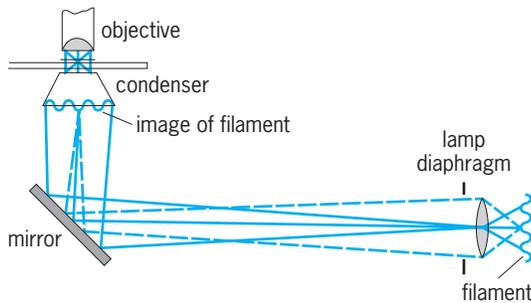


Fig. 7. Light path in Köhler illumination. (American Optical Corp.)

focus with the specimen and the lamp iris is opened only enough to fill the field of view. The iris of the microscope condenser is opened only enough to illuminate the back aperture of the objective. No ground glass is used. The condensing lens of the illuminator becomes the effective source and the field is uniformly illuminated, even though the source itself is not uniform. Loss of contrast from misplaced glare light is minimized (Fig. 7).

*Shillaber's type 3.* This method is useful when the lens systems of the condensers are inadequate to illuminate the field and aperture of the microscope.

A ground glass is placed close to the condensing lens of the illuminator between the lens and the lamp bulb, or the surface of the lamp is ground on the side facing the lens. Because the ground surface does not diffuse all the light, the condensing lens can concentrate more light into the microscope than could be obtained from the same area of a surface-type illuminator. An effectively larger source results which is useful with low-power (searcher) objectives for a large field of view and when the lamp filament is not large enough for the condensing system—for example, some sources built into the base of the microscope. This method is preferable when a diffuser must be used, but more stray or glare light is produced than with the Köhler method.

*Vertical illumination.* Vertical illumination uses a partly silvered mirror, or a prism, to reflect light through the objective onto the specimen. The light reflected back through the objective from the specimen forms the image which is seen.

*Epi-illumination* is vertical illumination with the illuminating light passing around a special objective to the specimen. Only the light reflected from the

specimen passes through the objective. Better definition results from the separation of the paths of the illuminating and viewing light.

Metals, ores, minerals, and opaque materials with adequate reflecting surfaces are examined with vertical illumination.

*Dark-field illumination.* In dark-field illumination the specimen appears bright, or self-luminous, against a black background. The illuminating beam is a hollow cone of light formed by an opaque stop at the center of the condenser, large enough to prevent any direct light from entering the objective. A specimen placed at the concentration of the light cone is seen with the light scattered or diffracted by it, and the smallest particle revealable depends upon the intensity of the available light, even though the particle may be too small for resolution as to size and shape. Size can sometimes be inferred from the number of particles found in a given volume of the specimen (Fig. 8).

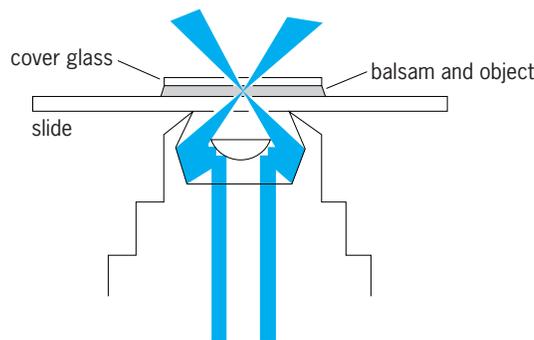


Fig. 8. Dark-field illumination. (American Optical Corp.)

Ultramicroscopy, used for the examination of colloids and smokes, is a dark field obtained with an intense narrow beam of light directed through the specimen at right angles to the optical axis of the microscope.

Rheinberg illumination or optical staining is also a modification of the dark-field method. The central disk is transparent and colored, rather than opaque, and an annulus of a complementary color fills the remaining condenser aperture. The specimen is seen in the color of the annulus against a background of the color of the central disk; for example, when the annulus is yellow and the background blue, the specimen appears yellow against a blue background.

*Modification by filters.* Filters are used between the microscope illuminator and the microscope to control the intensity or quality of illumination. Filters can be liquids in a flat-sided container or solids. Solid filters are made of colored glass, gelatin, or other materials.

Clear filters of water or of heat-absorbing glass are used to remove the excess heat from the lamp beam when delicate specimens must be protected. Blue filters remove excess yellow from tungsten light so that it resembles daylight. Neutral filters of glass or metal (Inconel) deposited on glass are used to reduce the amount of light without altering its color.

Colored filters change the quality of the light by selectively absorbing certain wavelengths and are grouped into broadband and narrow-band types. Broadband filters are used to increase visibility by modifying the color contrast of the specimen; for example, a yellow (minus blue) filter transmits all of the spectrum except the blue. Complementary filters increase contrast, and filters of similar color decrease color contrast. Polychromatic filters, transmitting two or more spectral colors (for example, a minus green filter passing blue and red) are useful with some stained specimens. Smaller regions of the spectrum are isolated with narrow-band-pass filters, interference filters, and by combinations of filters. Monochromatic light usually is obtained from a single spectral line of an arc source rather than with filters. *See* COLOR FILTER; INTERFERENCE FILTER.

**Calibration and measurement.** The size of the object under examination is frequently a desirable criterion for identification. A reference reticule is placed on the diaphragm in the focal plane of the ocular and seen in focus with the specimen from measurement in microscopy. A cross-hair reticule is satisfactory for position; scales are used for measuring distance and a net reticule for counting and drawing. Oculars with a focusing eye lens permit focusing the scale to the microscopist's eye.

Angular measurements are made by orienting one side of the specimen to the cross hair, reading the position of the graduated stage of the microscope, and rotating the stage until the other side is in line with the cross hair. The difference between the first and second readings of the scale gives the angle. Another method uses a goniometer ocular and measures the angle with the scale of the goniometer.

Linear measurements are made by placing one edge of the specimen at the cross hair, reading the position of the graduated mechanical stage, moving the specimen to the other side, and reading the position of the mechanical stage. The difference of the readings is the distance. This method is useful for larger specimens because the mechanical stages are ordinarily not graduated closer than 0.1 mm (0.004 in.). For measurements of smaller specimens a scale is used in the eyepiece and the size of the specimen obtained from the scale. The actual or true distance of the ocular scale depends upon the magnification of the microscope and is obtained by calibration with the aid of a stage micrometer.

More precise measurements can be made with a screw micrometer eyepiece, a filar micrometer, or a step micrometer eyepiece. The movable scales with graduated controls facilitate measurement, but must also be calibrated.

**Immersion fluids.** Air or other low-refractive-index material between the specimen and the objective of the microscope limits the angle of light that can enter the objective, the numerical aperture, and the resolving power of the system. To obtain greater values, a medium of higher refractive index must be placed between the condenser of the microscope and the microscope slide, and between the cover slip of the microscope and the front lens of the objective. Im-

mersion fluids may be aqueous substances (such as water or glycerin), oils, or organic materials.

In addition to the correct refractive index and dispersion, a satisfactory immersion oil must possess the following physical properties: chemical inertness, no tendency to spread or creep, no hardening when exposed to air, and little change with age.

**Mountants.** Mountants for specimens to be observed under the microscope are usually classed as temporary or permanent. Temporary mountants are water, physiological saline, serum, and the like. Such preparations last longer when the cover glass is sealed to the slide with petroleum jelly or paraffin. Semipermanent mounts are made with glycerin, glycerin jelly, or oils, and permanent mountants from gums, resins, or polymers which harden and remain solid at room temperatures. Canada balsam and damar have been largely replaced by synthetic polymers, such as Clarite, and polyvinyl alcohol.

A good mountant should provide enough difference in refractive index for contrast within the specimen, be transparent, not change color or consistency with time, hold the cover glass firmly in place, and not fade stains, or otherwise alter the specimen. *See* MICROTECHNIQUE.

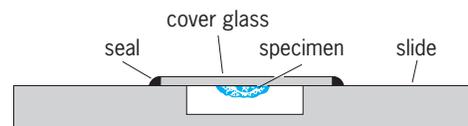


Fig. 9. Hanging-drop preparation.

**Hanging-drop preparations.** The specimen is placed in a drop of a suitable fluid on a cover slip and the cover slip is inverted over a slide which has a hollow ground in it (Fig. 9). Various concavities are used—round, flat, and so on—and the thickness of the slide varies with the preparation. The cover glass is usually sealed to the slide to retard evaporation. More complex preparations have channels so that the environment of the specimen can be changed or kept constant. Some hanging-drop-preparation slides are open at the bottom to make it possible to bring microdissection needles to the specimen for surgical procedures on the specimen while it is observed through the microscope. *See* MICROMANIPULATION. Oscar W. Richards

### Near-Infrared Microscopy

Near-infrared microscopy is an optical method that can be used for studying a variety of materials that are opaque in transmitted visible light (400–700 nm) yet translucent in the near infrared (700–1200 nm). The method utilizes the near-infrared optical microscope, a device for the conversion of the near-infrared image to a visible image.

The heart of the microscope is the image converter tube which is positioned above the microscope objective in the plane of the primary image. The near-infrared to visible conversion is achieved within the

image tube, a cathode tube containing a positive terminal coated with fluorescent material. The impingement of the near-infrared light image onto the negative terminal of the cathode tube releases electrons which are focused by means of an electrostatic lens onto the fluorescent screen.

The image formed on the fluorescent screen either can be viewed directly by means of an eyepiece positioned above the screen or can be photographed with standard cameras using ordinary film. Sharpening of the image is achieved by use of near-infrared filters located in front of the image tube. These filters serve to suppress the visible-light image and limit the wavelength range of the transmitted beam. However, such filters also tend to reduce the percentage of transmitted light so that a need arises for a light source that has a sufficiently intense near-infrared component (such as xenon and tungsten filament lamps).

Lawrence A. Harris

### Near-Field Optical Microscopy

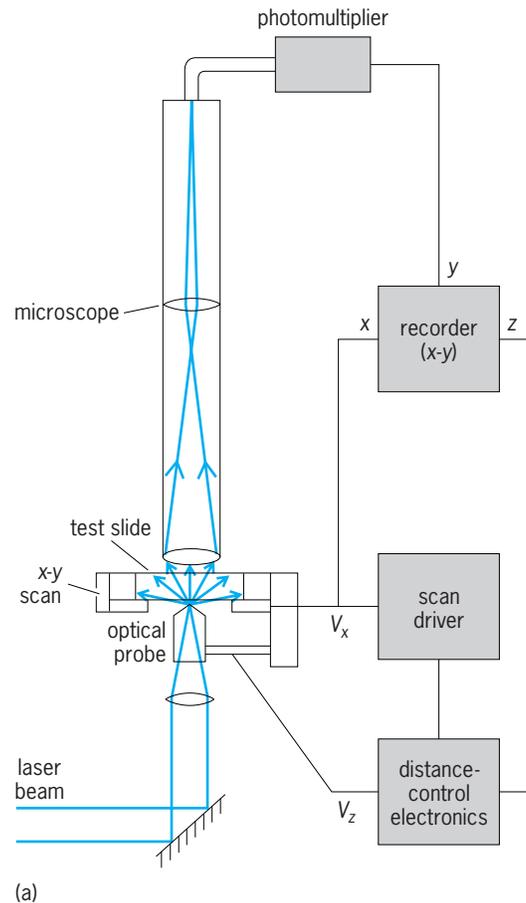
A fundamental law of optics, the so-called diffraction limit, states that two objects can be imaged as separate entities only if their distance is larger by about one-half the wavelength of visible light, which ranges from 400 to 700 nm. As a consequence, conventional optical microscopy is restricted to a resolution of about 200 nm, and this is not enough for many important observations. The scanning near-field optical microscope (NSOM or SNOM) circumvents the diffraction limit. In contrast to nonoptical methods of surpassing this limit, such as electron microscopy, it can provide information on such qualities as color, luster, transmissivity, and birefringence, which are sensitive indicators of material composition and status. Furthermore, it operates at ambient conditions, a prerequisite for observations of living organisms.

The ability to image an object is closely related to the ability to confine radiation to a narrow spot. Lenses and mirrors, the common focusing elements, are subject to the diffraction limit and hence can be ruled out. When illuminated, however, holes in an opaque screen or small scattering particles can also act as confined sources of radiation, and they can be made much smaller than half a wavelength.

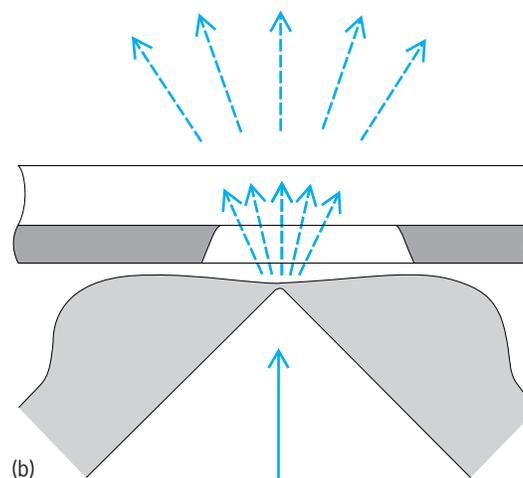
Light emerging from such a small source remains confined only over a distance comparable to its size; this is the so-called near-field zone. Beyond this zone, that is, in the far field, it diverges rapidly. The light source (termed an optical probe hereafter) hence must be brought so close to the object to be investigated that its near field is disturbed by the object. This requirement can best be satisfied by mounting the optical probe at the apex of a sharply pointed tip. The most common optical probe is an aperture in an opaque metallic coating which covers the corner of a transparent crystal or the end of an optical fiber. This probe is prepared so that a sharp tip is formed.

Optical detectors such as photomultipliers are fairly large and therefore cannot be mounted in the near field. The disturbance of the near field by the object, however, also influences the far field. This

property is illustrated by the case in which the optical probe is a small aperture and the object is a strongly absorbing metal film with small holes. If the probe is positioned directly above a hole, the light passing through the aperture can also pass the object without much attenuation; otherwise, the metal film will act like a lid, reducing the light transmission the closer the aperture is to the film surface (Fig. 10).



(a)



(b)

Fig. 10. Scanning near-field optical microscope (SNOM or NSOM). (a) Setup. (b) Optical probe next to a metallic film with small holes. (After U. Dürig, D. W. Pohl, and F. Rohner, *Near-field optical scanning microscopy*, *J. Appl. Phys.*, 59:3318–3327, 1986)

An image of the object surface can be composed by scanning the probe line by line over the object surface and recording the intensity of radiation as a function of position. An image as it would appear to the eye can be easily created by associating the signal intensity with the gray scale on a computer display; color images can also be created in this way.

The scanning motion is conveniently generated by means of piezoelectric elements which can be made to bend sideways or expand in length when voltage is applied in an appropriate way. Such scanners were originally developed for scanning tunneling microscopy. They can be used simultaneously for distance control. See MICROSCOPE; SCANNING TUNNELING MICROSCOPE.

D. W. Pohl

**Bibliography.** R. Barer, A. Bouwers, and D. S. Gray, Reflecting microscope, *Proc. London Conf. Opt. Instrum.*, pp. 43–76, 1951; E. M. Chamot and C. W. Mason, *Handbook of Chemical Microscopy*, vol. 1, 3d ed., reprint 1958; P. J. Duke and A. G. Michette (eds.), *Modern Microscopy: Techniques and Applications*, 1989; P. Gray, *The Microtome's Formulary and Guide*, 1954, reprint 1975; B. Herman and J. J. Lemasters (eds.), *Optical Microscopy: Emerging Methods and Applications*, 1992; R. P. Loveland, *Photomicrography: A Comprehensive Treatise*, 2 vols., 1970, reprint 1981; M. Pluta, *Advanced Light Microscopy*, 3 vols., 1988, 1989, 1993; D. W. Pohl and D. Courjon (eds.), *Near Field Optics*, NATO ASI Ser. E242, 1993; J. Sanderson, *Biological Microtechnique*, 1994; E. M. Slayter and H. S. Slayter, *Light and Electron Microscopy*, 1992; B. H. Walker, *Optical Engineering Fundamentals*, 1994; V. K. Zworykin and G. A. Morton, Applied electron optics, *J. Opt. Soc. Amer.*, 16:181–189, 1936.

## Optical modulators

Devices that serve to vary some property of a light beam. The direction of the beam may be scanned as in an optical deflector, or the phase or frequency of an optical wave may be modulated. Most often, however, the intensity of the light is modulated.

Rotating or oscillating mirrors and mechanical shutters can be used at relatively low frequencies (less than  $10^5$  hertz). However, these devices have too much inertia to operate at much higher frequencies. At higher frequencies it is necessary to take advantage of the motions of the low-mass electrons and atoms in liquids or solids. These motions are controlled by modulating the applied electric fields, magnetic fields, or acoustic waves in phenomena known as the electrooptic, magneto-optic, or acousto-optic effect, respectively. See ACOUSTOOPTICS; ELECTROOPTICS; MAGNETOOPTICS.

For the most part, it will be assumed that the light to be modulated is nearly monochromatic—either a beam from a laser or a narrow-band incoherent source. See LASER.

**Electrooptic effect.** The quadratic or Kerr electrooptic effect is present in all substances and refers to a change in refractive index,  $\Delta n$ , proportional

to the square of the applied electric field  $E$ . The liquids nitrobenzene and carbon disulfide and the solid strontium titanate exhibit a large Kerr effect. See KERR EFFECT.

Much larger index changes can be realized in single crystals that exhibit the linear or Pockels electrooptic effect. In this case,  $\Delta n$  is directly proportional to  $E$ . The effect is present only in noncentrosymmetric single crystals, and the induced index change depends upon the orientations of  $E$  and the polarization of the light beam with respect to the crystal axes. Well-known linear electrooptic materials include potassium dihydrogen phosphate (KDP) and its deuterated isomorph (DKDP or KD\*P), lithium niobate ( $\text{LiNbO}_3$ ) and lithium tantalate ( $\text{LiTaO}_3$ ), and semiconductors such as gallium arsenide (GaAs) and cadmium telluride (CdTe). The last two are useful in the infrared (1–10 micrometers), while the others are used in the near-ultraviolet, visible, and infrared regions (0.3–3  $\mu\text{m}$ ). See CRYSTAL OPTICS; FERROELECTRICS; SEMICONDUCTOR.

The phase increment  $\Phi$  of an optical wave of wavelength  $\lambda$  that passes through a length  $L$  of material with refractive index  $n$  is given by Eq. (1). Thus phase

$$\Phi = \frac{2\pi}{\lambda} nL \quad (1)$$

modulation can be achieved by varying  $n$  electrooptically. Since the optical frequency of the wave is the time derivative of  $\Phi$ , the frequency is also shifted by a time-varying  $\Phi$ , yielding optical frequency modulation. See MODULATION.

The refractive index change is given in terms of an electrooptic coefficient  $r$  by Eq. (2), where typical

$$n(E) = n(0) - \frac{n^3 r E}{2} \quad (2)$$

values for  $n$  and  $r$  are 2 and  $3 \times 10^{-11}$  m/V, respectively.

**Electrooptic intensity modulation.** Intensity modulation can be achieved by interfering two phase-modulated waves as shown in Fig. 1. The electrooptically induced index change is different for light polarized parallel ( $p$ ) and perpendicular ( $s$ ) to the

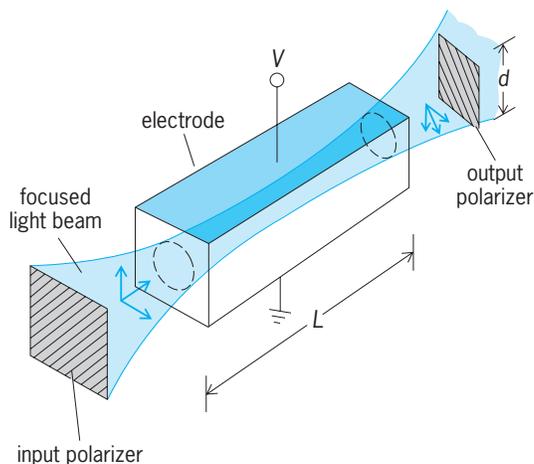


Fig. 1. Electrooptic intensity modulator.

modulating field; that is,  $r_p \neq r_s$ . The phase difference for the two polarizations is the retardation  $\Gamma$  given by Eq. (3), which is proportional to the ap-

$$\Gamma(V) = \Phi_p - \Phi_s = \frac{2\pi}{\lambda}(n_p - n_s)L \quad (3)$$

plied voltage  $V$ . If it is assumed that  $n_p(0) = n_s(0)$  for  $V = 0$ , then  $\Gamma(0) = 0$ . Further assume that  $\Gamma(V') = \pi$ . Then incident light polarized at  $45^\circ$  to the field may be resolved into two equal, in-phase components parallel and perpendicular to the field. For  $V = 0$ , they will still be in phase at the output end of the electrooptic crystal and will recombine to give a polarization at  $+45^\circ$  which will not pass through the output polarizer in Fig. 1. For  $V = V'$ , however, the two components will be out of phase and will recombine to give a polarization angle of  $-45^\circ$  which will pass through the output polarizer. The switching voltage  $V'$  is called the half-wave voltage and is given by Eq. (4), with  $d$  the thickness of the crystal and  $r_c$

$$V' = \frac{\lambda d}{n^3 r_c L} \quad (4)$$

an effective electrooptic coefficient. A typical value for  $n^3 r_c$  is  $2 \times 10^{-10}$  m/V. See POLARIZED LIGHT.

Electrooptic devices can operate at frequencies from dc through hundreds of gigahertz.

**Acoustooptic modulation and deflection.** All transparent substances exhibit an acoustooptic effect—a change in  $n$  proportional to strain. Typical materials with a substantial effect are water, glass, arsenic selenide ( $\text{As}_2\text{Se}_3$ ), and crystalline tellurium dioxide ( $\text{TeO}_2$ ).

An acoustic wave of frequency  $F$  has a wavelength given by Eq. (5), where  $C$  is the acoustic velocity,

$$\Lambda = \frac{C}{F} \quad (5)$$

which is typically  $3 \times 10^3$  m/s. Such a wave produces a spatially periodic strain and corresponding optical phase grating with period  $\Lambda$  that diffracts an optical beam as shown in Fig. 2. A short grating ( $Q < 1$ ) produces many diffraction orders (Raman-Nath regime), whereas a long grating ( $Q > 10$ ) produces only one diffracted beam (Bragg regime). The quantity  $Q$  is defined by Eq. (6). If one detects the

$$Q = \frac{2\pi L\lambda}{n\Lambda^2} \quad (6)$$

Bragg-diffracted beam (diffraction order +1), its intensity can be switched on and off by turning the acoustic power on and off.

The Bragg angle  $\theta$  through which the incident beam is diffracted is given by Eq. (7). Thus the beam

$$\sin \frac{\theta}{2} = \frac{\lambda}{2\Lambda} \quad (7)$$

angle can be scanned by varying  $F$ .

Acoustooptic devices operate satisfactorily at speeds up to several hundred megahertz.

**Optical waveguide devices.** The efficiency of the modulators described above can be put in terms of

the electrical modulating power required per unit bandwidth to achieve, say, 70% intensity modulation or one radian of phase modulation. For very good bulk electrooptic or acoustooptic modulators, approximately 1 to 10 mW is required for a 1-MHz bandwidth; that is, a figure of merit of 1 to 10 mW/MHz. This figure of merit can be improved by employing a dielectric waveguide to confine the light to the same small volume occupied by the electric or acoustic modulating field, without the diffraction that is characteristic of a focused optical beam. This optical waveguide geometry is also compatible with optical fibers and lends itself to integrated optical circuits. See INTEGRATED OPTICS; OPTICAL COMMUNICATIONS; OPTICAL FIBERS.

An optical wave can be guided by a region of high refractive index surrounded by regions of lower index. A planar guide confines light to a plane, but permits diffraction within the plane; a strip or channel guide confines light in both transverse dimensions without any diffraction.

Planar guides have been formed in semiconductor materials such as  $\text{Al}_x\text{Ga}_{1-x}\text{As}$  and  $\text{In}_x\text{Ga}_{1-x}\text{As}_y\text{P}_{1-y}$  by growing epitaxial crystal layers with differing values of  $x$  and  $y$ . Since the refractive index decreases as  $x$  increases, a thin GaAs layer surrounded by layers of  $\text{Al}_{0.3}\text{Ga}_{0.7}\text{As}$ , for example, will guide light. If one outer layer is  $p$ -type and the other two  $n$ -type, a reverse bias may be applied to the  $pn$  junction to produce a large electric field in the GaAs layer. Very efficient electrooptic modulators requiring about 0.1 mW/MHz at  $\lambda = 1 \mu\text{m}$  have been realized in such junctions. Further improvement has been realized by loading the planar guide with a rib or ridge structure to provide lateral guiding within the plane. Ridge devices 1 mm long require about 5 V to switch light on or off, with a figure of merit of 10  $\mu\text{W}/\text{MHz}$ .

Both planar and strip guides have been formed in  $\text{LiNbO}_3$  by diffusing titanium metal into the surface in photolithographically defined patterns. Planar guides with surface acoustic wave transducers (Fig. 3) have been employed in waveguide

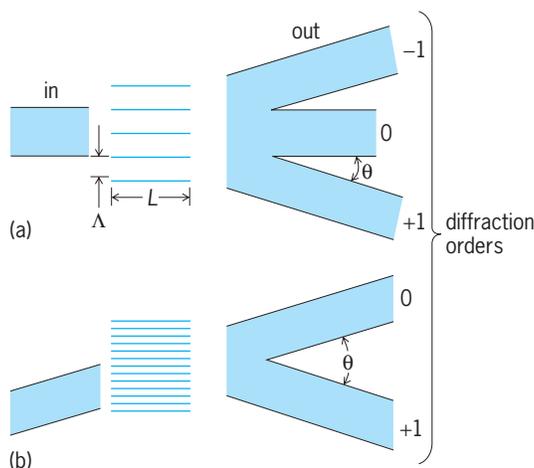


Fig. 2. Acoustooptic modulator-deflector. (a) Raman-Nath regime. (b) Bragg regime.

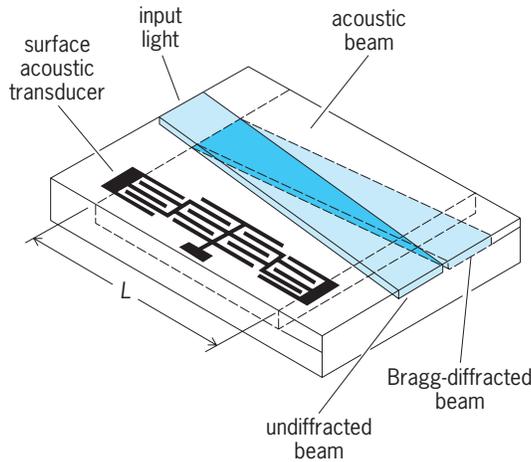


Fig. 3. Waveguide acoustooptic modulator-deflector.

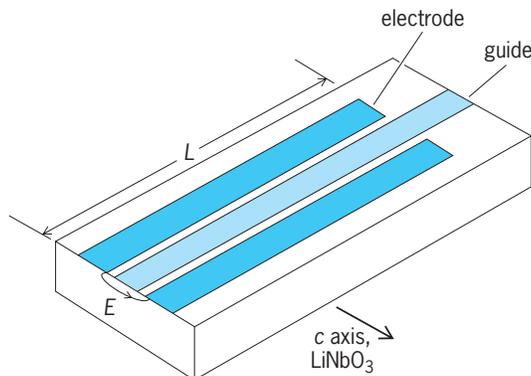


Fig. 4. Titanium-diffused  $\text{LiNbO}_3$  strip waveguide with coplanar electrodes for electrooptic phase modulation.

acoustooptic devices. Strip guides (Fig. 4) have been used to make electrooptic phase modulators 3 cm long requiring 1 V to produce  $\pi$  radians of phase shift at  $\lambda = 0.63 \mu\text{m}$ ; the figure of merit is  $2 \mu\text{W}/\text{MHz}$ .

**Lithium niobate waveguide modulation.** In order to make an intensity modulator or a  $2 \times 2$  switch, a directional coupler waveguide pattern is formed, and electrodes are applied as shown in Fig. 5. With no voltage applied and  $L$  the proper length for a given separation between waveguides, light entering guide 1 will be totally coupled to and exit from guide 2; with the proper applied voltage, the phase match between the two guides will be destroyed, and no light will be coupled to guide 2, but will all exit from guide 1. Switches and modulators operating at speeds beyond 10 GHz have been realized in this way.

The directional coupler switch is difficult to manufacture reproducibly because of the small separation between guides, a critical dimension. A more robust modulator structure is based on a waveguide version of the Mach-Zehnder interferometer (Fig. 6). In a Mach-Zehnder modulator, the input light is divided equally by a Y-branch splitter and travels along two identical paths before recombining in a second Y-branch. To intensity-modulate the output, equal and opposite phases are induced electrooptically in the two arms by the voltages  $+V$  and  $-V$  applied to the

electrodes. When the phase difference is  $\pi$ , light in the Y-branch combiner interferes destructively and radiates into the substrate with no signal appearing in the output guide. When the phase difference is 0 or  $2\pi$ , the signals in the two arms add constructively and all the light emerges in the output guide. See INTERFEROMETRY.

Lithium niobate Mach-Zehnder modulators are often used in long-distance digital lightwave systems, operating at speeds up to 10 Gb/s. They are well suited to this application because they produce intensity modulation without any associated phase modulation. Any extraneous phase modulation would cause a widening of the laser spectrum, which in turn would cause the optical pulses to spread in time as they traveled along the fiber due to chromatic dispersion in the fiber. Spurious phase modulation is known as chirp. Chirp limits the maximum bit rate and span length of a light-wave system to the point at which the pulse spreading remains less than about one-fourth the width of the data pulses. Chirp is measured as the ratio  $\alpha$  of the phase modulation to amplitude modulation. Commercial Mach-Zehnder modulators operate with  $\alpha = 0$  and a bit rates in excess of 10 Gb/s.

**Semiconductor laser modulation.** A semiconductor laser comprises a forward-biased  $pn$  junction, a waveguide region, and a resonator (Fig. 7a). The dc injection current provides gain and determines the laser output intensity. An additional modulating current may be inserted to produce a directly modulated laser, operating at speeds of a few gigahertz (Fig. 7b). This is a low-cost method for providing a modulated output, but its application is limited to lower bit rates and shorter spans by chirp associated with modulation of the waveguide refractive index by the modulating current. Typical values of  $\alpha$  in directly modulated lasers are about 4. See OPTICAL GUIDED WAVES.

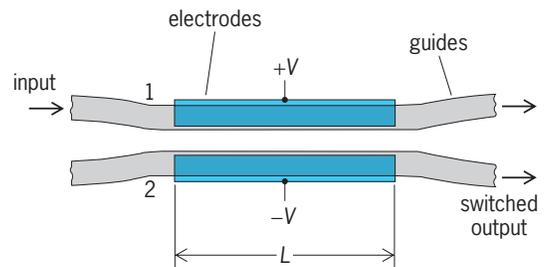


Fig. 5. Directional coupler waveguide modulator switch.

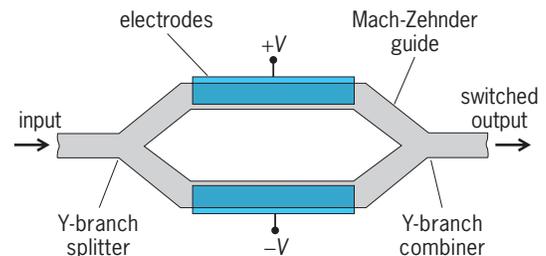


Fig. 6. Mach-Zehnder electrooptic waveguide modulator.

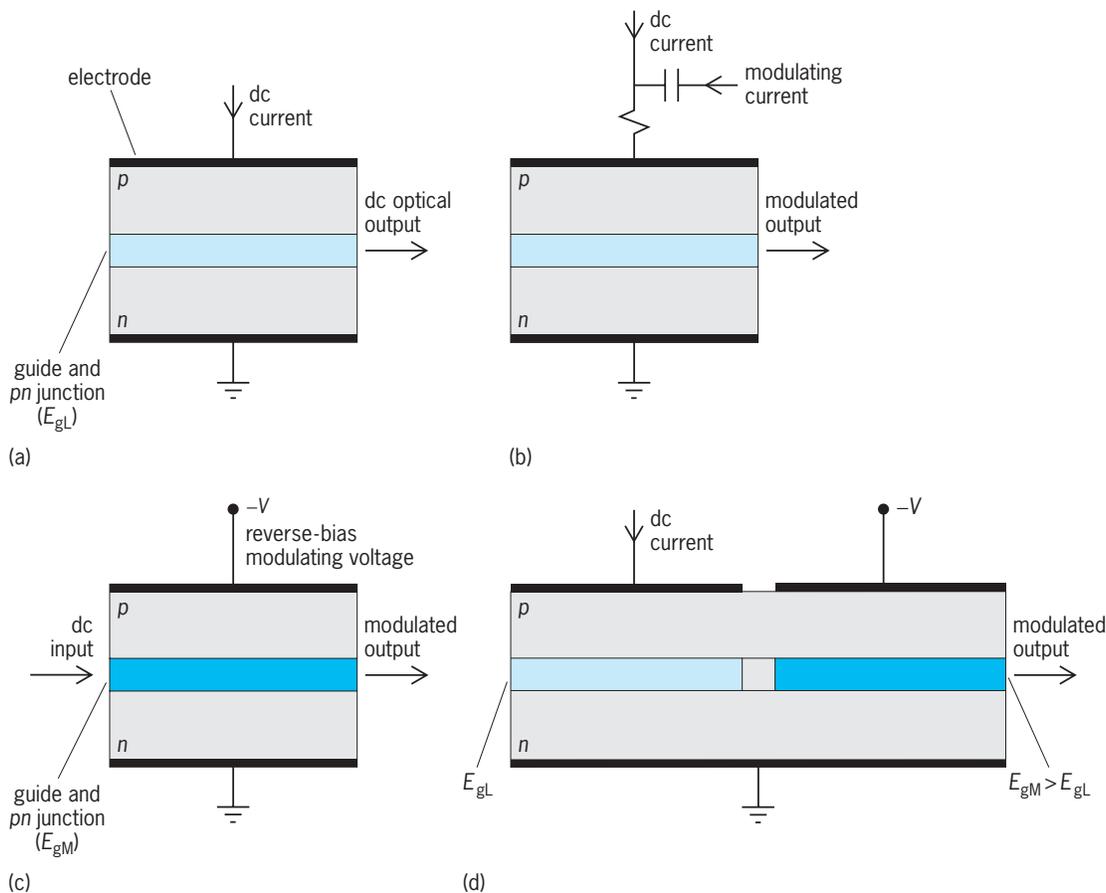


Fig. 7. Semiconductor laser modulation. (a) Unmodulated semiconductor laser. (b) Directly modulated laser. (c) Electroabsorption modulator. (d) Integrated externally modulated laser.

When chirp is critical, one must use an external modulator such as a lithium niobate Mach-Zehnder modulator. An alternative is the semiconductor electroabsorption modulator, consisting of a reverse-biased  $pn$  junction with bias voltage  $-V$  and an associated waveguide (Fig. 7c). The absorption edge energy  $E_{gM}$  of the electroabsorption modulator's waveguide material is designed to be just above the energy of the transmitted laser light  $E_{gL} [= hc/\lambda]$ . Under these conditions, light entering the waveguide is transmitted to the output when  $V = 0$ . However, when the reverse-bias  $V$  is increased, the absorption edge is shifted to lower energy and reduces the transmission, providing intensity modulation. Electroabsorption modulators can be designed with small but nonzero values of  $\alpha$ .

The coupling of an unmodulated semiconductor laser to an external electroabsorption (or Mach-Zehnder) modulator is complicated by requirements for precise alignment of waveguides and minimal reflections from interfaces, which are overcome by integrating on a single chip an unmodulated laser and an electroabsorption modulator (Fig. 7d). In a commercial electroabsorption modulator, the substrate is InGaAsP and the laser operates at 1550 nm, the wavelength at which optical fiber has minimum attenuation. The absorption edge in the laser section is near 1550 nm, but the absorption edge in the elec-

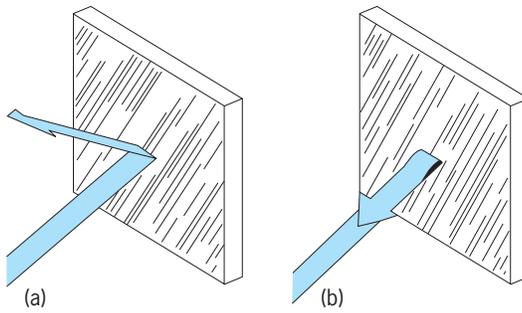
troabsorption section is designed to be at a slightly higher energy or, equivalently, shorter wavelength of about 1500 nm to make it transparent to the laser wavelength. Commercial versions of this integrated external modulator laser operate at 2.5 Gb/s and are capable of higher bit rates. See INTEGRATED OPTICS; SURFACE-ACOUSTIC-WAVE DEVICES.

Ivan P. Kaminow

Bibliography. I. P. Kaminow and T. L. Koch (eds.), *Optical Fiber Telecommunications IIIB*, Academic Press, 1997; S. E. Miller and I. P. Kaminow, *Optical Fiber Telecommunications II*, Academic Press, 1988; T. Tamir (ed.), *Guided-Wave Optoelectronics*, 2d ed., Springer-Verlag, 1990.

## Optical phase conjugation

A process that involves the use of nonlinear optical effects to precisely reverse the direction of propagation of each plane wave in an arbitrary beam of light, thereby causing the return beam to exactly retrace the path of the incident beam. The process is also known as wavefront reversal or time-reversal reflection. The unique features of this phenomenon suggest widespread application to the problems of optical beam transport through distorting or inhomogeneous media. Although closely related, the field



**Fig. 1.** Comparison of reflections (a) from a conventional mirror and (b) from an optical phase conjugator. (After V. J. Corcoran, ed., *Proceedings for the International Conference on Laser '78 for Optical and Quantum Electronics*, STS Press, McLean, Virginia, 1979)

of adaptive optics will not be discussed here. See ADAPTIVE OPTICS.

**Fundamental properties.** Optical phase conjugation is a process by which a light beam interacting in a nonlinear material is reflected in such a manner as to retrace its optical path. As Fig. 1 shows, the image-transformation properties of this reflection are radically different from those of a conventional mirror. The incoming rays and those reflected by a conventional mirror (Fig. 1a) are related by reversal of the component of the wave vector  $\vec{k}$  which is normal to the mirror surface. Thus a light beam can be arbitrarily redirected by adjusting the orientation of a conventional mirror. In contrast, a phase-conjugate reflector (Fig. 1b) inverts the vector quantity  $\vec{k}$  so that, regardless of the orientation of the device, the reflected conjugate light beam exactly retraces the path of the incident beam. This retracing occurs even though an aberrator (such as a piece of broken glass) may be in the path of the incident beam. Looking into a conventional mirror, one would see one's own face, whereas looking into a phase-conjugate mirror, one would see only the pupil of the eye. This is because any light emanating from, say, one's chin would be reversed by the phase conjugator and would return to the chin, thereby missing the viewer's eye. A simple extension of the arrangement in Fig. 1b indicates that the phase conjugator will reflect a diverging beam as a converging one, and vice versa. These new and remarkable image-transformation properties (even in the presence of a distorting optical element) open the door to many potential applications in areas such as laser fusion, atmospheric propagation, fiber-optic propagation, image restoration, real-time holography, optical data processing, nonlinear microscopy, laser resonator design, and high-resolution nonlinear spectroscopy. See MIRROR OPTICS.

**Optical phase conjugation techniques.** Optical phase conjugation can be obtained in many nonlinear materials (materials whose optical properties are affected by strong applied optical fields). The response of the material may permit beams to combine in such a way as to generate a new beam that is the phase-conjugate of one of the input beams. Processes associated with degenerate four-wave mixing,

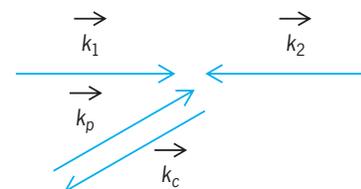
scattering from saturated resonances, stimulated Brillouin scattering, stimulated Raman scattering, photoreactive phenomena, surface phenomena, and thermal scattering have all been utilized to generate optical phase-conjugate reflections.

*Kerr-like degenerate four-wave mixing.* As shown in Fig. 2, two strong counterpropagating (pump) beams with  $k$ -vectors  $k_1$  and  $k_2$  (at frequency  $\omega$ ) are directed to set up a standing wave in a clear material whose index of refraction varies linearly with intensity. This arrangement provides the conditions in which a third (probe) beam with  $k$ -vector  $\vec{k}_p$ , also at frequency  $\omega$ , incident upon the material from any direction would result in a fourth beam with  $k$ -vector  $\vec{k}_c$  being emitted in the sample precisely retracing the third one. (The term degenerate indicates that all beams have exactly the same frequency.) In this case, phase matching (even in the birefringent materials) is obtained independent of the angle between  $\vec{k}_p$  and  $\vec{k}_1$ . The electric field of the conjugate wave  $E_c$  is given by the equation below, where  $\delta n$  is the change in the index of re-

$$E_c = E_p^* \tan\left(\frac{2\pi}{\lambda_0} \delta n l\right)$$

fraction induced by one strong counterpropagating wave,  $\lambda_0$  is the free-space optical wavelength, and  $l$  is the length over which the probe beam overlaps the conjugation region. The conjugate reflectivity is defined as the ratio of reflected and incident intensities, which is the square of the above tangent function. The essential feature of phase conjugation is that  $E_c$  is proportional to the complex conjugate of  $E_p$ . Although degenerate four-wave mixing is a nonlinear optical effect, it is a linear function of the field one wishes to conjugate. This means that a superposition of  $E_p$ 's will generate a corresponding superposition of  $E_c$ 's; therefore faithful image reconstruction is possible. See KERR EFFECT.

To visualize the degenerate four-wave mixing effect, consider first the interaction of the weak probe wave with pump wave number two. The amount by which the index of refraction changes is proportional to the quantity  $(E_p + E_2)^2$ , and the cross term corresponds to a phase grating (periodic phase disturbance) appropriately oriented to scatter the pump wave number one into the  $\vec{k}_c$  direction. Similarly, the very same scattering process occurs with the roles of two pump waves reversed. Creating these phase gratings can be thought of as "writing" of a hologram, and the subsequent scattering can be thought of as "reading" the hologram. Thus the four-wave mixing process is equivalent to volume holography in which



**Fig. 2.** Geometry of  $k$ -vectors for optical phase conjugation using degenerate four-wave mixing.

the writing and reading are done simultaneously. *See* DIFFRACTION GRATING; HOLOGRAPHY.

*Conjugation using saturated resonances.* Instead of using a clear material, as outlined above, the same four-wave mixing beam geometry can be set up in an absorbing or amplifying medium partially (or totally) saturated by the pump waves. When the frequency of the light is equal to the resonance frequency of the transition, the induced disturbance corresponds to amplitude gratings which couple the four waves. When the single frequency of the four waves differs slightly from the resonance frequency of the transition, both amplitude gratings and index gratings become involved. Because of the complex nature of the resonant saturation process, the simple  $\tan^2 [(2\pi/\lambda_0)(\delta n 2D)]$  expression for the conjugate reflectivity is no longer valid. Instead, this effect is maximized when the intensities of the pump waves are about equal to the intensity which saturates the transition.

*Stimulated Brillouin and Raman scattering.* Earliest demonstrations of optical phase conjugation were performed by focusing an intense optical beam into a waveguide containing materials that exhibit backward stimulated Brillouin scattering. More recently, this technique has been extended to include the backward stimulated Raman effect. In both cases, the conjugate is shifted by the frequencies characteristic of the effect. *See* RAMAN EFFECT.

Backward stimulated Brillouin scattering involves an inelastic process where an intense laser beam in a clear material is backscattered through the production of optical phonons. Here, the incoming and outgoing light waves interfere to beat at the frequency and propagation vector of the sound wave, and the incoming light wave and the sound wave interfere to produce a radiating nonlinear polarization at the frequency and propagation vector of the outgoing light wave. The production of the sound wave involves the electrostrictive effect in which the Brillouin scattering medium is forced into regions of high electric field. The scattering of the light by the sound wave can be viewed as Bragg scattering from the sound wave grating with a Doppler shift arising because the sound wave is moving. The coupled process of stimulated backward Brillouin scattering can then be understood by considering that the forward-going beam scatters off of a small amount of sound wave to make more backward-going light, and at the same time the forward-going beam and a small amount of backward-going beam interfere to promote the growth of the sound wave. This doubly coupled effect rapidly depletes the incoming laser beam while producing the backward beam and while leaving a strong sound wave propagating in the medium. *See* DOPPLER EFFECT; ELECTROSTRICTION; LATTICE VIBRATIONS; SCATTERING OF ELECTROMAGNETIC RADIATION.

*Photorefractive effects.* A degenerate four-wave mixing geometry can be set up in a nonlinear crystal which exhibits the photorefractive effect. In such a material, the presence of light releases charges which migrate until they are trapped. If the light is in the form of an interference pattern, those migrating charges

which get trapped in the peaks are more likely to be re-released than would those which become recombined in the valleys. This process leads to the charges being forced into the valleys of the interference pattern. This periodic charge separation sets up large periodic electric fields which, through the Pockels effect, produce a spatially periodic index of refraction. Hence a light-wave grating is converted into an index grating which can then scatter a pump wave into the conjugate direction. This process requires very little power, and therefore can be studied without the use of high-power lasers. *See* ELECTRON-HOLE RECOMBINATION; ELECTROOPTICS; INTERFERENCE OF WAVES; TRAPS IN SOLIDS.

*Surface phase conjugation.* Here a strong pump wave is directed normal to a surface which exhibits nonlinear behavior. Phenomena which can couple the light waves include electrostriction, heating, damage, phase changes, surface charge production, and liquid crystal effects. In general, the surface phase-conjugation process is equivalent to two-dimensional (thin) holography. *See* LIQUID CRYSTALS; SURFACE PHYSICS.

*Thermal scattering effects.* Here an optical interference pattern in a slightly absorbing material will be converted to an index grating if the material has a temperature-dependent index of refraction. In the four-wave mixing geometry, these thermal gratings can then scatter pump waves into the conjugate wave direction.

**Practical applications.** Many practical applications of optical conjugators utilize their unusual image-transformation properties. Because the conjugation effect is not impaired by interposition of an aberrating material in the beam, the effect can be used to repair the damage done to the beam by otherwise unavoidable aberrations. This technique can be applied to improving the output beam quality of laser systems which contain optical phase inhomogeneities or imperfect optical components. In a laser, one of the two mirrors could be replaced by a phase-conjugating mirror, or in laser amplifier systems, a phase-conjugate reflector could be used to reflect the beam back through the amplifier in a double-pass configuration. In both cases, the optical-beam quality would not be degraded by inhomogeneities in the amplifying medium, by deformations or imperfections in optical elements, windows, mirrors, and so forth, or by accidental misalignment of optical elements. *See* IMAGE PROCESSING.

Aiming a laser beam through an imperfect medium to strike a distant target may be another application. The imperfect medium may be turbulent air, an air-water interface, or the focusing mirror in a laser-fusion experiment. Instead of conventional aiming approaches, one could envision a phase-conjugation approach in which the target would be irradiated first with a weak diffuse probe beam. The glint returning from the target would pass through the imperfect medium, and through the laser amplifier system, and would then strike a conjugator. The conjugate beam would essentially contain all the information needed to strike the target after

passing through both the amplifier and the imperfect medium a second time. Just as imperfections between the laser and the target would not impair the results, neither would problems associated with imperfections in the elements within the laser amplifier. See NUCLEAR FUSION.

Other applications are based upon the fact that four-wave mixing conjugation is a narrow-band mirror that is tunable by varying the frequency of the pump waves. There are also applications to fiber-optic communications. For example, a spatial image could be reconstructed by the conjugation process after having been “scrambled” during passage through a multimode fiber. Also, a fiber-optic communication network is limited in bandwidth by the available pulse rate; this rate is determined to a large extent by the dispersive temporal spreading of each pulse as it propagates down the fiber. The time-reversal aspect of phase conjugation could undo the spreading associated with linear dispersion and could therefore increase the possible data rate. See LASER; NONLINEAR OPTICS; OPTICAL COMMUNICATIONS; OPTICAL FIBERS; OPTICAL INFORMATION SYSTEMS.

Robert A. Fisher; Barry J. Feldman

Bibliography. R. A. Fisher (ed.), *Optical Phase Conjugation*, 1983; M. Gower and D. Proch, *Optical Phase Conjugation*, 1994; J.-I. Sakai, *Phase Conjugate Optics*, 1992.

## Optical prism

A simple component, made of a light-refracting and transparent material such as glass and bounded by two or more plane surfaces at an angle (Fig. 1), that is used in optical devices, especially to change the direction of light travel, to accomplish image rotation or inversion, and to disperse light into its constituent colors. Once light enters a prism, it can be reflected one or more times before it exits the prism.

A variety of prisms can be classified according to their function. Some prisms, such as the Dove prism, can be used to rotate an image and to change its parity. Image inversion by prisms in traditional binoculars is a typical application (Fig. 2). Some prisms take advantage of the phenomenon of total internal reflection to deviate light, such as the right-angle prism and the pentaprism used in single lens

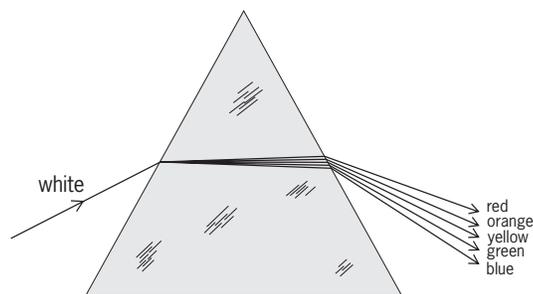


Fig. 1. Optical prism, consisting of a refracting medium bounded by two surfaces. Dispersion of light into its component colors is shown. (After W. J. Smith, *Modern Optical Engineering*, 3d ed., SPIE Press–McGraw-Hill, 2000)

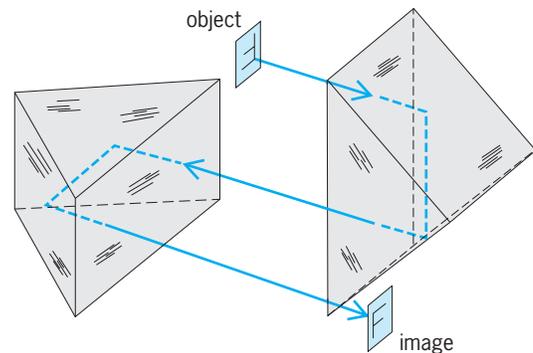


Fig. 2. System of two Porro prisms, used to invert the image in traditional binoculars. (After R. E. Hopkins and R. Hanau, *U.S. Military Handbook for Optical Design*, republished by Sinclair Optics, 1987)

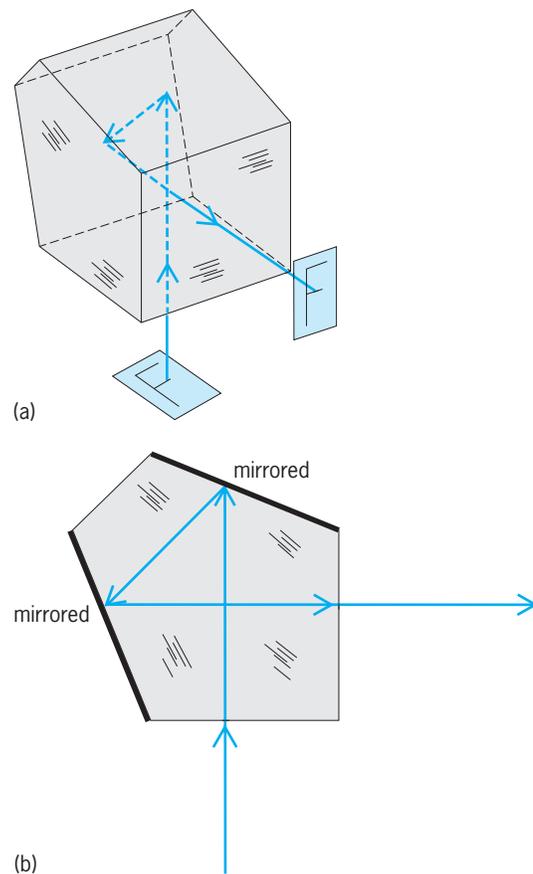


Fig. 3. Pentaprism, used in single lens reflex cameras. (a) Perspective view, showing how the prism deviates light through an angle of  $90^\circ$  without inverting or reverting the image. (b) Ray diagram. (After R. E. Hopkins and R. Hanau, *U.S. Military Handbook for Optical Design*, republished by Sinclair Optics, 1987)

reflex cameras (Fig. 3). A thin prism is known as an optical wedge; it can be used to change slightly the direction of light travel, and therefore it can be used in pairs as an alignment device (Fig. 4). Optical wedges are also used in stereoscopic instruments to allow the viewer to observe the three-dimensional effect without forcing the eyes to point in different directions. A variable wedge can be integrated into a commercial pair of binoculars to stabilize the line of sight in the presence of the user's slight hand

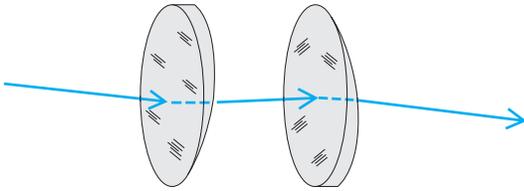


Fig. 4. Set of two thin prisms, known as optical wedges, which can be used as an alignment device. (After P. R. Yoder, Jr., *Non-image forming optical components, in Geometrical Optics, SPIE Proc., 531:206–220, 1985*)

movements. Other prisms such as corner-cubes can be used to reflect light backward, and are fabricated in arrays for car and bicycle retroreflectors. See BINOCULARS; MIRROR OPTICS; PERISCOPE; RANGEFINDER (OPTICS); REFLECTION OF ELECTROMAGNETIC RADIATION.

**Light dispersion.** When light enters at an angle to the face of a prism, according to Snell's law it is refracted. Since the index of refraction depends on the wavelength, the light is refracted at different angles and therefore it is dispersed into a spectrum of colors. The blue color is refracted more than the red. When light reaches the second face of the prism, it is refracted again and the initial dispersion can be added to or canceled, depending on the prism angle. An important application of a prism is to disperse light (Fig. 1). A combination of prisms in tandem can increase the amount of light dispersion. Dispersing prisms have been used in monochromators and spectroscopic instruments. With two prisms of different materials, it is possible to obtain light deviation without dispersion (an achromatic prism) or dispersion without deviation. See DISPERSION (RADIATION); REFRACTION OF WAVES; SPECTROSCOPY.

**Image manipulation.** An anamorphic magnification (magnification whose magnitude is different in each of two perpendicular directions) can be imparted to an image by prisms. For example, the aspect ratio of a beam becomes different as it enters and exits a system of Brewster prisms, and a pair of such prisms is often used to transform the cross section of a diode laser beam from elliptical to circular. (In a Brewster prism, light is incident at Brewster's angle, so that reflection losses for plane-polarized light, that is, light vibrating in the plane of incidence, are minimized.) The angular magnification will also be anamorphic. A prism can also introduce aberrations into a beam of light, depending on how the beam enters and exits the prism. Some prisms can be considered as parallel plates of glass, and therefore they can shift axially the position of an image. See ABERRATION (OPTICS); GEOMETRICAL OPTICS; POLARIZED LIGHT.

**Other applications.** Prisms find many other applications. For example, they have been used to measure the index of refraction, taking advantage of the phenomenon of minimum deviation; they find application in scanning optical systems; and they are used in fingerprint reading devices. See FINGERPRINT.

Jose M. Sasian

Bibliography. E. Hecht, A. Zajac, and K. Guardino, *Optics*, 3d ed., Addison-Wesley, 1997; W. J. Smith,

*Modern Optical Engineering*, 3d ed., SPIE Press-McGraw-Hill, 2000.

## Optical projection systems

Optical projection is the process whereby a real image of a suitably illuminated object is formed by an optical system in such a manner that it can be viewed, photographed, or otherwise observed. Essential equipment in an optical projection system consists of a light source, a condenser, an object holder, a projection lens, and (usually) a screen on which the image is formed (Fig. 1).

The luminance of the image in the direction of observation will depend upon (1) the average luminance of the image of the light source as seen through the projection lens from the image point under consideration, (2) the solid angle subtended by the exit pupil of the projection lens at this image point, and (3) the reflective or transmissive characteristics of the screen. Usually it is desirable to have this luminance as high as possible. Therefore, with a given screen, lens, and projection distance, the best arrangement is to have the light source imaged in the projection lens, with its image filling the exit pupil as completely and as uniformly as possible.

The object is placed between the condenser and the projection lens. If transparent, it can be inserted directly in the light beam; however, it should be positioned, and the optical system should be so designed that it does not vignette (cut off) any of the image of the light source in the projection lens. If the object is opaque, an arrangement known as an epidiascope (Fig. 2) is used. A difficulty in the design of this system is to illuminate the object so that all portions

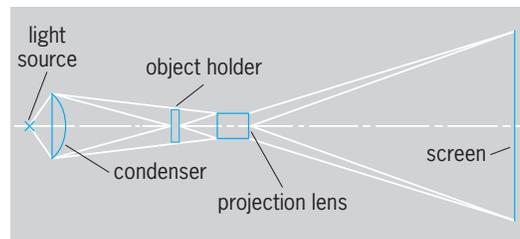


Fig. 1. Simple optical projection system.

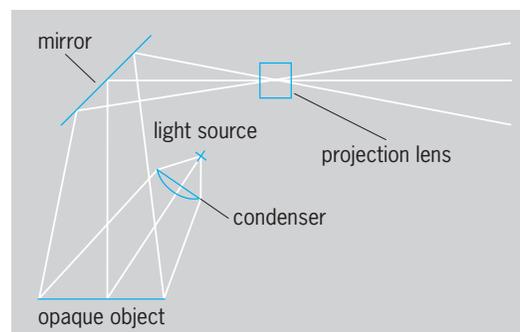


Fig. 2. An epidiascope, or system for projecting an image of an opaque object.

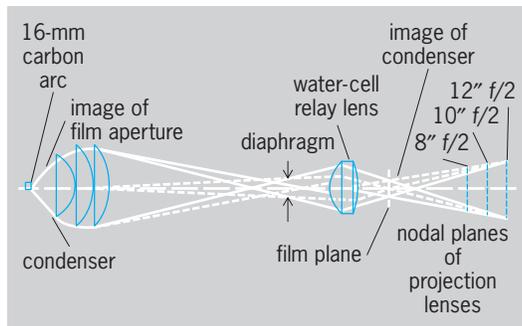


Fig. 3. Relay condenser system having water cell incorporated in second stage. 1" = 25 mm; 16 mm = 0.63 in.

will show well in the projected image, without excessive highlights or glare.

If a small uniform source which radiates in accordance with Lambert's law is projected through a well-corrected lens to a screen which is perpendicular to the optic axis of the lens, maximum illuminance of the image will occur on this axis, and illuminance away from the axis will decrease in proportion to the fourth power of the cosine of the angle subtended with the axis at the projection lens. In practice, it is possible to design distortion into the condenser so that the illuminance is somewhat lower on the axis, and considerably higher away from the axis than is given by this fourth-power law. Acceptable illumination for most visual purposes can allow a falloff from center to side in the image of as much as 50%, particularly if the illuminance within a circle occupying one-half of the image area does not drop below 80% of the maximum value. See PHOTOMETRY.

**Light source.** Usually, either an incandescent or an arc lamp is used as the light source. To keep luminance high, incandescent projection lamps operate at such high temperatures that their life is comparatively short. Also, they must be well cooled; all except the smallest sizes require cooling fans for this

purpose. Filaments are finely coiled and accurately supported, usually being carefully aligned in a pre-focus base so that they will be precisely positioned in the optical system. Spacing between coils is such that a small spherical mirror can be used in back of the lamp to image the coils in the spaces between the coils, thus increasing usable light output nearly twofold. See LAMP.

When a highly uniform field is required, a lamp consisting of a small disk of ceramic material, heated to incandescence by radio-frequency induction, is available. With this, it is possible to maintain illuminance of a projection field of several square inches with a variation of only 2-3%. See INCANDESCENT LAMP.

Arc lamps are used when incandescent lamps cannot provide sufficient flux to illuminate the screen area satisfactorily. Carbon electrodes with special feed control, magnetic arc stabilization, and other devices are usually used to keep the arc as accurately positioned and as flicker-free as possible. The high-pressure xenon arc lamp is also used as a projection light source. It has a color which more accurately duplicates sunlight than the carbon arc. Its intensity is considerably higher than that of any incandescent lamp, and it avoids most of the problems attendant on the burning of carbons. A special power supply, which ordinarily incorporates a striking circuit, is needed. See ARC LAMP.

A shield between the arc and the object, called a douser, is used to protect the object while the arc is ignited and to provide a quick shutoff for the light beam. Often water cells and other heat-filtering devices are inserted in the beam to keep the heat on the object as low as possible.

**Condenser.** The condenser system is used to gather as much of the light from the source as possible and to redirect it through the projection lens. Both reflective and refractive systems are used. Reflectors can be of aluminum, although the better ones are of glass with aluminized coatings. They are usually elliptical in shape, with the light source at one focus and the image position in the projection lens at the other.

Refractive systems may be of heat-resistant glass or fused quartz. With arc lamps particularly, the condenser lens is very close to the source in order to provide the high magnification required if the image is to fill the projection lens. Usually, therefore, the condenser requires special cooling. In larger projectors, several elements are used to give the required magnification; these are often aspherical in shape to give the required light distribution.

A well-designed condenser can pick up light in a cone having a half-angle in excess of  $50^\circ$ . This means that, with a well-designed arc lamp or with an incandescent lamp using an auxiliary spherical mirror, more than one-third of the total luminous flux radiated by the source can be directed through the projector.

To obtain the high magnification required with arc sources and large-aperture lenses, a relay type of condenser (Fig. 3) may be used. This images the source

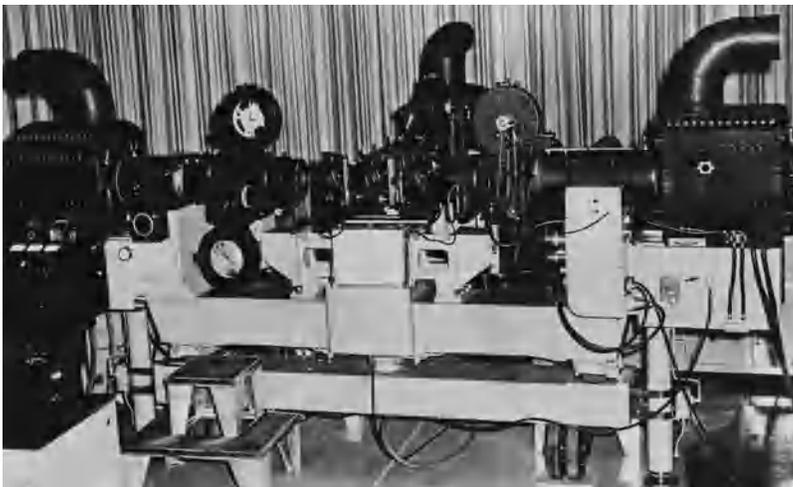


Fig. 4. Triple-head projector with three 250-A arc lamps and 12-in. (30-mm)  $f/2$  projection lenses, incorporating iris diaphragms for intensity control. (Paramount Pictures)

beyond the first condenser system, and then uses a second lens to relay this image to the projection lens. This arrangement allows for better light-distribution control in the screen image with less waste of light at the object position. Also, an Inconel "cat's-eye" diaphragm at the first image point gives a convenient intensity control for the light beam. For additional information on condensers *see* OPTICAL MICROSCOPE

**Object holder.** The function of the object holder is to position the object exactly at the focal point of the projection lens. Slight motion or vibration will reduce image sharpness. Therefore, proper mechanical design of this part of the system is most important. When a succession of objects is to be projected, the holder must be able to position these objects precisely and clamp them firmly in a minimum time. Some designs have been suggested and a few tried which allow the object to move while the system projects a fixed image. However, this requires a motion in some portion of the optical system, which invariably reduces the quality of the projected image. Because of this, such systems have not found wide favor.

The object holder must also provide for protection of the object from unwanted portions of the beam, for cooling the object, where this may be required, and for accurately adjusting its position laterally with respect to the beam. In most systems, focusing of the projection lens is provided by moving the lens itself rather than by disturbing the position of the object longitudinally in the system.

**Projection lens.** The function of the projection lens is to produce the image of the object on the screen. Its design depends upon the use to be made of the projected image. As examples, a profile or contour projector requires a lens which is well corrected for the aberrations known as distortion and field curvature; an optical printer (or enlarger) requires a lens carefully designed to work at comparatively short conjugate focal distances; and picture projectors must preserve the quality which has been captured by the camera lens. *See* ABERRATION (OPTICS).

Since projection lenses usually transmit relatively intense light beams, they must be designed to be heat-resistant. Their surfaces are usually not cemented. Optical surfaces are coated to reduce reflection, but these coatings must be able to withstand heat without deterioration. Because camera lenses are not designed for this type of operation, they should not be employed as projection lenses unless the projector is specifically designed to use them.

An ideal position at which to control intensity is at the projector lens. Large-aperture lenses containing iris diaphragms are used in the large process projectors of motion picture studios (**Fig. 4**).

**Projection screen.** Usually a projection screen is used to redirect the light in the image for convenient observation. Exceptions are systems which project the image directly into other optical systems for further processing, or which form an aerial image which is to be viewed only from a localized position.

Screens may be either reflective or translucent, the latter type being used when the image is to be observed or photographed from the side away from the projector. An example is the so-called self-contained projector, which has the screen on the housing holding the other elements. Reflective screens may be matte, having characteristics approaching Lambert reflection; directional, producing an image which will appear brighter in the direction of specular reflection; or reflexive, directing most of the light back toward the projector and giving an image of relatively high luminance with low projection intensity, but with a very confined viewing angle. *See* CINEMATOGRAPHY.

Armin J. Hill

## Optical pulses

Bursts of electromagnetic radiation of finite duration. Optical pulses are used to transmit information or to record the chronology of physical events. The simplest example is the photographic flash. This was probably first developed by early photographers who used flash powder that, when ignited, produced a short burst of intense light. This was followed by the flash lamp, in which a tube filled with an inert gas such as xenon is excited by a brief electrical pulse. A great advance in the creation of short optical pulses came with the invention of the laser. Lasers are now the most common and effective way of generating a variety of short optical pulses, of different durations, energies, and wavelengths. *See* LASER; STROBOSCOPIC PHOTOGRAPHY.

**Q-switching.** Pulses of millisecond ( $10^{-3}$  s) duration are very simply generated by mechanically modulating a constant light source such as a lamp or a continuous-wave laser. This can be done, for example, by placing a rotating disk with holes in it in front of the light source. Shorter laser pulses, of microsecond ( $10^{-6}$  s) or nanosecond ( $10^{-9}$  s) duration, are generated by using a technique known as Q-switching. A modulating device is incorporated inside the laser cavity that allows the buildup of the laser radiation inside the cavity and then switches it out in an instant. The modulating device is usually controlled by external electrical pulses. Semiconductor diode lasers, which are used to transmit information (voice or data) over a fiber-optic cable, are pumped by electricity and can be directly pulsed by applying to them a pulsed electrical signal. *See* OPTICAL COMMUNICATIONS; OPTICAL FIBERS.

**Mode locking.** Ultrashort laser pulses, with durations of the order of picoseconds ( $1 \text{ ps} = 10^{-12}$  s) or femtoseconds ( $1 \text{ fs} = 10^{-15}$  s), are generated by using a general principle known as mode locking. Lasers that directly produce these pulses employ either active or passive mode locking. The Fourier transform relationship implies that an optical pulse must contain a certain number of frequency modes, or equivalently spectral components, which is inversely proportional to its time duration. Mode locking is a technique whereby several frequency modes of the laser structure are made to resonate simultaneously

and with a well-orchestrated relationship so as to form a short-duration pulse at the laser output.

In the case of active mode locking, the laser is stimulated with external pulses, either electrical or optical. For example, the excitation that drives the laser can be another short laser pulse, which has a repetition rate precisely adjusted to the length of the mode-locked laser. Another active technique consists of incorporating into the laser itself an optical element that modifies the laser's properties when the laser is stimulated with electrical pulses, again with a well-adjusted repetition rate.

In the case of passive mode locking, a conceptually more difficult technique to understand, the laser is not actively stimulated with some form of external pulses. Rather, the mode locking originates from the complex interaction of the various elements of the laser cavity. The laser can be pumped by a continuous source of energy, yet still produce ultrashort pulses of light. One of the more standard ways of realizing passive mode locking is by incorporating into the laser resonator a lossy material known as a saturable absorber, in addition to the laser gain medium. The saturable absorber absorbs the front edge of the pulse and then saturates, becoming transparent to the remainder of the pulse. The combination of the gain and the saturable absorber as the pulse makes round trips in the laser cavity leads to a stable short pulse.

There are other, and more exotic, forms of passive mode locking. Typically, lasers mode-locked in this way contain a nonlinear element that acts much as a saturable absorber would. The titanium:sapphire laser (a sapphire crystal doped with the trivalent ion  $Ti^{3+}$ ) produces extremely short pulses over a wide range of wavelengths (700–1000 nm). The nonlinear interaction that leads to the passive mode locking is believed to originate in the titanium:sapphire crystal itself, which is in fact also the laser gain medium. Lasers fabricated from glass fibers similar to those used in fiber-optic networks can also be passively mode-locked and form very compact sources, as they can be pumped with miniature laser diodes. See FOURIER SERIES AND TRANSFORMS; NON-LINEAR OPTICS.

**Shortest durations.** Pulses as short as 11 fs have been produced directly by a passively mode-locked titanium:sapphire laser. The titanium:sapphire laser has also allowed the extension of ultrashort optical pulses to other wavelength ranges, such as the near-infrared (2–10  $\mu\text{m}$ ). Dye lasers, based on organic dyes in solution, have achieved durations as short as 27 fs. Ultrashort diode laser pulses have been obtained by active and passive mode locking and produce pulses as short as a few hundred femtoseconds. They are more commonly operated so as to give rise to pulses in the picosecond range, appropriate for optical communication systems.

Optical pulses from many of these lasers can be shortened even further by pulse compression. This is accomplished by passing the pulses through a series of optical elements that include a nonlinear medium such as an optical fiber. Laser pulses as short as 6 fs

have been generated in this fashion. The duration of these pulses is equivalent to three cycles of light at the center wavelength of the pulse, 620 nm. These pulses have a correspondingly large spectral bandwidth, about 60 nm, or one-tenth of the entire visible spectrum.

At present, no optical detectors can measure pulse durations of less than 1 ps. The only way, therefore, to accurately measure the duration of a subpicosecond optical pulse is to use the pulse itself. The technique used, autocorrelation, involves splitting the pulse into two equal parts and then recombining them with a varying time delay between the two pulses. The resulting sum signal, which is a function of the relative time delay between the two pulses, can then be analyzed to obtain the temporal duration of the pulse.

**Time-resolved spectroscopy.** The generation of ultrashort laser pulses has been motivated by the quest for ever better resolution in the study of the temporal evolution and dynamics of physical systems, events, and processes. Such laser pulses are capable of creating snapshots in time of many events that occur on the atomic or molecular scale, a technique known as time-resolved spectroscopy. This stroboscopic aspect of ultrashort laser pulses is their most important scientific application and is used in physics, engineering, chemistry, and biology. For example, ultrashort pulses can excite and take snapshots of molecular vibrations and deformations. They can track the passage of charge carriers through a microscopic semiconductor device. This ability to understand the dynamics of the more elemental building blocks of nature can in turn make it possible to build ever faster devices for use in information processing and information transmission, in addition to providing a better understanding of the physical world. See LASER PHOTOCHEMISTRY; LASER SPECTROSCOPY; OPTICAL INFORMATION SYSTEMS; ULTRAFAST MOLECULAR PROCESSES.

**High-intensity pulses.** Ultrashort pulses also produce extremely high instantaneous intensities of light, since the entire energy of the pulse is compressed into such a short time duration. High-intensity pulses may have important applications in x-ray laser production, high-resolution lithography, and fundamental physics at high intensities. The method used to amplify short pulses to very high intensities is known as chirped pulse amplification. It involves stretching out the pulse in time by spreading out its spectral components, followed by the amplification stage, and finally a compression of the pulse to close to its initial duration. This method avoids the deleterious effects that can accompany the amplification of an ultrashort pulse in a high-energy amplifier. Titanium:sapphire lasers amplifiers contribute significantly to the production of ultrahigh-intensity pulses. Pulses with a duration of 21 fs and an energy per pulse of 0.5 mJ have been demonstrated.

Philippe C. Becker

**High-power laser interaction studies.** The advent of short-pulse, high-power lasers has opened up new areas of research in atomic physics under high-field

or high-density conditions. In particular, the interaction of high-power lasers with appropriate targets can give rise to incoherent radiation sources for x-ray spectroscopy and x-ray diffraction studies, as well as to coherent soft x-ray sources for x-ray holography. Such studies can enhance understanding of the dynamical properties of material or biological systems.

High-power lasers (in the ultraviolet, visible, or infrared parts of the spectrum) have peak powers of the order of terawatts ( $1 \text{ TW} = 10^{12} \text{ W}$ ). Most of them fall into two groups. Lasers of pulse duration of about 1 ns have been developed primarily for nuclear fusion studies but are also used to generate coherent and incoherent x-ray sources. Lower-energy, smaller lasers of pulse duration of order 1 ps or shorter can give rise to very high fields (by focusing the beam to a very small spot), thereby accessing a new regime of atomic spectroscopy.

*High-density physics.* In nuclear fusion studies, a multi-beam high-power laser is used to spherically compress a target core to densities of the order of  $10^{24}$  electrons per cubic centimeter and temperatures above 1 keV ( $10^7 \text{ K}$ ). Under these conditions, similar to those prevailing in stellar interiors, x-ray spectral lines of core elements (heavier than hydrogen) are strongly broadened. This broadening is due to the electric fields produced by neighboring plasma particles (the Stark effect), and it can be used to deduce the plasma density. A stronger effect is associated with the compressed shell surrounding the core, which typically achieves high densities at lower temperatures. Here the plasma ions are in the strongly coupled regime. The electron energy, for the most part, conforms to the Thomas-Fermi distribution. Such experiments can thus be used to study the properties (such as transport and equation of state) of Fermi-degenerate plasmas; a similar state (at much higher densities) exists in white dwarf stars. *See* FERMI-DIRAC STATISTICS; NUCLEAR FUSION; STARK EFFECT; STELLAR EVOLUTION; WHITE DWARF STAR.

*Multiphoton ionization.* When a high-power, very short-pulse laser is focused onto a low-pressure gas, the high laser flux at focus (above  $10^{13} \text{ W/cm}^2$ ) corresponds to a high-intensity radiation field, and the low gas density isolates the radiation interaction with individual atoms from interparticle interactions. Experiments show a phenomenon termed above-threshold ionization (ATI): the escaped photoelectrons have discrete energies separated by the photon energy of the laser. Each successive peak corresponds to a multiphoton ionization event involving the absorption of one additional photon (generally, a nonresonant process, insensitive to atomic structure). Peaks corresponding to the simultaneous absorption of tens of photons have been observed. At much higher flux levels (above  $10^{15} \text{ W/cm}^2$ ), the charge state of escaping ions indicates the simultaneous absorption of hundreds of photons. When subpicosecond pulses are used, the above-threshold ionization peaks are split into components, related to the fact that high-lying, Rydberg states are shifted (upward) in the presence of a high field. As the intensity changes during the laser pulse, various Rydberg states move into res-

onance with the total energy of the absorbed photons (corresponding to a given above-threshold ionization peak), thus enhancing the photoionization and giving rise to the component peaks. The short pulse assures that the energy of the escaping photoelectron is not modified by the field. This is therefore a method of measuring atomic level positions in the presence of high radiation fields. *See* ATOMIC STRUCTURE AND SPECTRA; RYDBERG ATOM.

*X-ray holography and diffraction.* Soft x-ray lasers can be realized by focusing a high-power laser into an elongated focal line on an appropriate target. The high-power laser can form a plasma of the desired ionization state in which plasma electrons or plasma photons can excite population inversion between pairs of ionic energy levels. Spontaneous emission of transitions between such levels is then amplified along the direction of the focal line. Significant coherent output has been achieved at wavelengths from above 20 nm down to below 10 nm (with pulses of energy up to 1 millijoule at the longer end of this range). Studies have aimed at improving the spatial coherence and extending lasing to wavelengths below 5 nm, where the scattering contrast between water and proteins is maximal. These sources can be used for dynamical, in-line holography of hydrated (living) biological samples, with spatial resolution better than 0.1 micrometer. *See* HOLOGRAPHY.

In other examples of laser-based studies, the intense, incoherent x-ray emission from nanosecond-laser-irradiated targets was used for dynamical structure studies of living samples by diffracting off periodic protein molecules; the electron pulse from a picosecond-laser irradiated photocathode was used to dynamically study the annealing of metals by de Broglie diffraction; and a repetitive, picosecond-pulse laser was used to measure the fluorescence decay time of chlorophyll macromolecules. *See* ELECTRON DIFFRACTION; FLUORESCENCE; X-RAY DIFFRACTION.

*Plasma-wave spectroscopy.* When a high-power laser impinges on a solid target, forming a plasma, a rich spectrum of inelastically scattered radiation is observed. The frequencies in this spectrum are various combinations of the laser frequency and frequencies of natural modes or waves excited in the plasma by the laser itself. These include electron plasma waves (giving rise to Raman scattering, so called in analogy to the Raman scattering from molecular energy levels) and ionic waves (giving rise to Brillouin scattering). The density gradient, characteristic of plasma expanding from an irradiated target, permits the excitation of a wide range of frequencies, enriching the spectrum. The characteristics of such spectra are a source of information on the properties of the laser-heated plasma, in particular the temperature and the density gradient. *See* PLASMA (PHYSICS); RAMAN EFFECT; SCATTERING OF ELECTROMAGNETIC RADIATION.

*Relativistic and quantum effects.* At moderate flux levels (less than  $10^{15} \text{ W/cm}^2$ ), a free electron oscillates harmonically in a laser field as a mass on a spring. At higher flux levels, the electron's oscillatory

velocity in the field approaches the speed of light and its motion becomes anharmonic. Its accompanying radiation includes the second and higher harmonics of the laser frequency. For counterpropagating electron-laser collisions, this process is multiphoton Compton scattering. Under these conditions, the electric field of the laser is Lorentz-boosted in the rest frame of the electrons. With high-intensity lasers (flux greater than  $10^{18}$  W/cm<sup>2</sup>) and relativistic electron beams (energy greater than 50 GeV), the electric field can be greater than the Schwinger critical field, which is the electric field ( $1.3 \times 10^{16}$  V/cm) at which an electron is accelerated to an energy equivalent to its rest mass in one Compton wavelength. See ANHARMONIC OSCILLATOR; COMPTON EFFECT; LORENTZ TRANSFORMATIONS; QUANTUM ELECTRODYNAMICS; RELATIVITY.

Robert L. McCrory

Bibliography. V. R. Berestetskii, E. M. Lifshitz, and L. P. Pitaevskii, *Quantum Electrodynamics*, 2d ed., 1982; R. W. Falcone and J. Kirz (eds.), *Short Wavelength Coherent Radiation: Generation and Applications*, Optical Society of America, vol. 2, 1988; J. Hermann and B. Wilhelmi, *Lasers for Ultrashort Pulses*, 1987; H. Hora and G. H. Miley (eds.), *Laser Interactions and Related Plasma Phenomena*, vols. 6–10, 1984–1993; W. Kaiser (ed.), *Ultrashort Laser Pulses: Generation and Applications*, 2d ed., 1993; W. L. Kruer, *The Physics of Laser Plasma Interactions*, 1988.

## Optical pumping

The process of causing strong deviations from thermal equilibrium populations of selected quantized states of different energy in atomic or molecular systems by the use of optical radiation (that is, light of wavelengths in or near the visible spectrum), called the pumping radiation.

At thermal equilibrium at temperature  $T$  K, the relative numbers of atoms,  $N_2/N_1$ , in quantized levels  $E_2$  and  $E_1$ , respectively, where  $E_2$  is the higher, are given by  $N_2/N_1 = e^{-(E_2 - E_1)/kT}$ , where  $k$  is Boltzmann's constant. The number of atoms in the higher level is, at equilibrium, always less than that in the lower, and as the energy difference between the two levels increases the number in the higher level becomes very small indeed. By exposing a suitable system to optical radiation, one can, so to speak, pump atoms from a lower state to an upper state so as greatly to increase the number of atoms in the upper state above the equilibrium value. See ENERGY LEVEL (QUANTUM MECHANICS).

In an early application of the principle, the levels  $E_2$  and  $E_1$  were not far apart, so that the equilibrium populations of the atoms in the two levels were not greatly different. The system was chosen to possess a third level  $E_3$ , accessible from  $E_1$  but not from  $E_2$  by the absorption of monochromatic visible light. (The states involved were the paramagnetic Zeeman components of atomic states, and the necessary selectivity of the transitions excitable by the visible light was secured by appropriate choice of the state of polar-

ization of this light.) The visible light excites atoms from  $E_1$  to  $E_3$ , from which they return, with spontaneous emission, with about equal probability, to the lower states  $E_2$  and  $E_1$ . After a period of time, provided there has been sufficiently intense excitation by the visible light, most of the atoms are in state  $E_2$  and few are in the lower state  $E_1$ —atoms have been pumped from  $E_1$  to  $E_2$  by way of the highly excited state  $E_3$ . See ATOMIC STRUCTURE AND SPECTRA; ZEEMAN EFFECT.

Optical pumping is vital for light amplification by stimulated emission in an important class of lasers. For example, the action of the ruby laser involves the fluorescent emission of red light by a transition from an excited level  $E_2$  to the ground level  $E_1$ . In this case  $E_2$  is relatively high above  $E_1$  and the equilibrium population of  $E_2$  is practically zero. Amplification of the red light by laser action requires that  $N_2$  exceed  $N_1$  (population inversion). The inversion is accomplished by intense green and violet light from an external source which excites the chromium ion in the ruby to a band of levels,  $E_3$ , above  $E_2$ . From  $E_3$  the ion rapidly drops without radiation to  $E_2$ , in which its lifetime is relatively long for an excited state. Sufficiently intense pumping forces more luminescent ions into  $E_2$  by way of the  $E_3$  levels than remain in the ground state  $E_1$ , and amplification of the red emission of the ruby by stimulated emission can then occur. See LASER.

William West

Bibliography. O. Svelto, *Principles of Lasers*, 4th ed., 1998; J. T. Verdeyen, *Laser Electronics*, 3d ed., 1993; A. Yariv, *Optical Electronics*, 4th., 1991.

## Optical recording

The process of recording signals on a medium through the use of light, so that the signals may be reproduced at a subsequent time. Photographic film has been widely used as the medium, but in the late 1970s development of another medium, the optical disk, was undertaken. The introduction of the laser as a light source greatly improves the quality of reproduced signals. Digital storage techniques make it possible to obtain extremely high-fidelity reproduction of sound signals in optical disk recording systems. This article first describes optical film recording and then some of the more modern optical recording methods.

### Optical Film Recording of Sound

Optical film recording is also termed motion picture recording or photographic recording. A sound motion picture recording system consists basically of a modulator for producing a modulated light beam and a mechanism for moving a light-sensitive photographic film relative to the light beam and thereby recording signals on the film corresponding to the electrical signals. A sound motion picture reproducing system is basically a combination of a light source, an optical system, a photoelectric cell, and a mechanism for moving a film carrying an optical record by means of which the recorded photographic

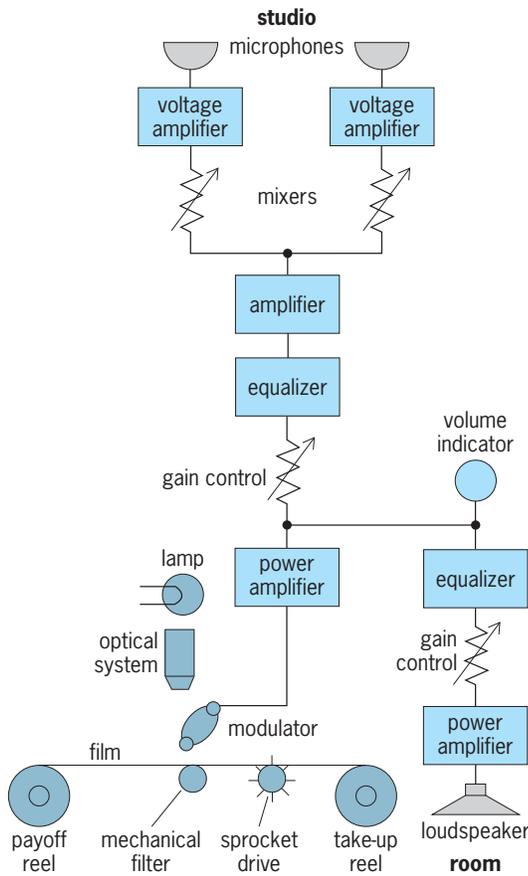


Fig. 1. Schematic arrangement of apparatus in a complete optical sound motion picture recording system.

variations are converted into electrical signals of approximately similar form.

**Recording system.** In a monophonic sound motion picture recording system (Fig. 1), the output of each microphone is amplified and fed to a mixer, a device having two or more inputs and a common output. If more than one microphone is used, such as a microphone for each of two actors, the outputs of the microphones may be adjusted for the proper balance by means of the mixers. An electronic compressor is used to reduce the amplitude range to that suitable for reproduction in the home. An equalizer provides the standard motion picture recording characteristic. A gain control provides means for controlling the overall signal level fed to the power amplifier. A light modulator, actuated by the amplifier, records a photographic image upon the film corresponding to the electrical input. The narrow bands used for the sound record on motion picture film are called sound tracks. A monitoring system consisting of a volume indicator, complementary equalizer, gain control, power amplifier, and loudspeaker or headphone is used to control the recording operation. See SOUND RECORDING.

**Modulator.** In the variable-area recording system (Fig. 2), the transmitted light amplitude is a function of the amount of unexposed area in the positive print. This type of sound track is produced by means

of a mirror galvanometer which varies the width of the light slit under which the film passes. A triangular aperture is uniformly illuminated by means of a lamp and lens system. The image of this triangular aperture is reflected by the galvanometer mirror focused on a mechanical slit, which in turn is focused on the film. The triangular light image on the mechanical slit moves up and down on the mechanical slit. The result is that the width of the exposed portion of the negative sound track corresponds to the rotational vibrations of the galvanometer. In the positive record, the width of the unexposed portion corresponds to the signal. See GALVANOMETER.

A variable-density system has also been used, but it has been largely replaced by the variable-area system. The reproducer can play either one.

**Recording film transport.** The film-transport mechanism used in recording sound on film consists of a positive drive of the perforated film and a constant-speed drive of the film where the modulated light beam strikes the film. Positive drive of the film is obtained by the sprocket drive, which is interlocked with the camera drive so that synchronism of picture and sound will be obtained. When the film passes over the sprocket drive, variations in the motion of the film at the sprocket-hole frequency are produced. These variations in the film speed must be removed at the recording point to eliminate spurious frequency modulation of the image on the film. Uniform speed

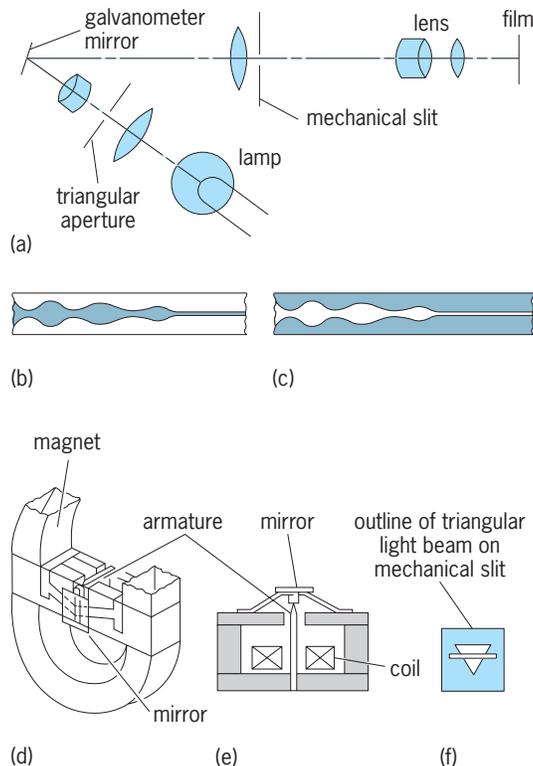


Fig. 2. Elements of a variable-area sound motion picture film recording system. Transmitted light amplitude is a function of the amount of unexposed area in the positive print. (a) Schematic of a recording system. (b, c) Negative and positive sound tracks. (d) Perspective view. (e) Sectional view. (f) Mechanical slit.

at the recording point is provided by a mechanical contrivance called a filter. It is located between the sprocket drive and recording point and consists of the inertia of the recording drum and the compliance of the film between the recording drum and the sprocket drive. The recording drum is driven by a magnetic system from the motor which drives the sprocket and thereby imparts a slight amount of drive to the film. The magnetic drive isolates the variations in the rotational speed of the motor drive from the rotating drum. The combination of the isolating filter and magnetic drive provides a system with very uniform motion of the surface of the drum. The image of the modulator is focused on the film while the film is in contact with the drum.

*Film and sound track.* In the recording of sound motion pictures, the picture and sound are recorded on separate photographic films. The camera and sound recorder must be synchronized. This is accomplished by the use of an interlock system between the camera and sound recorder and the use of perforated film in the form of sprocket holes along the two edges of the film for both the camera and sound recorder. The sound track on 35-mm film occupies a space about 0.1 in. (2.5 mm) wide just inside the sprocket holes.

*Film developing and printing.* The mass production of motion picture positive prints proceeds in a series of steps. The negative record is developed, and then the required number of positive prints of both picture and sound is printed from the negative record. These positive records are developed and are used for sound reproduction and picture projection in the theater.

**Reproducing system.** In a sound motion picture reproducing system (Fig. 3), the first element is the optical system, consisting of a lamp and a lens arrangement which produces an illuminated slit upon the film. The light beam passes through the film and falls upon a photoelectric cell. When the film is pulled past the slit, the variations in light, which are due to the variable-density or variable-area recording on the film, fall upon the photoelectric cell and are converted into the corresponding electrical variations. The output of the photoelectric cell is fed to an amplifier followed by a filter, which is used to cut the ground noise (residual noise in the absence of the signal) due to the film above the upper limit of reproduction, and by equalizers, which are used to adjust the frequency characteristic to that suitable for the best sound reproduction in the theater. A volume control (gain control) is used for adjusting the level of sound output. The output of a power amplifier feeds stage loudspeakers, located behind the screen, and a monitoring loudspeaker. Except for the stage loudspeakers, the entire equipment including the monitoring loudspeaker is located in the projection booth. See CINEMATOGRAPHY; SOUND-REPRODUCING SYSTEMS.

*Optical electronic reproducer.* In a motion picture film sound-reproducing system, the light source, in the form of an incandescent lamp, is focused upon a mechanical slit by means of a condensing lens. The me-

chanical slit in turn is focused on the negative film. The height of the image on the film is usually about 0.00075 in. (19 micrometers). Under these conditions the amount of light which impinges upon a photocell behind the film is proportional to the unexposed portion of the sound track in variable-area recording. When the film is in motion, the resultant light undulations which fall upon the photocell correspond to the voltage variations applied to the recording galvanometer. The voltage output of the photocell is proportional to the amount of light which falls upon the cathode.

*Reproducing film transport.* The film transport used in reproducing sound on photographic film consists of a positive drive of the perforated film and a constant-speed drive where the light passes through the film to the photoelectric cell. Positive drive of the film is obtained by means of two sprocket drives. The sprocket drives are geared with the positive picture drive so that a constant loop of film is maintained between the sound head and the picture head. The positive drive also ensures that the film speed in reproduction will be the same as that in recording. There is a loose loop of film between the picture head and the sound head, so that variations in the picture drive will not be imparted to the sound head.

After the film enters the sound head, it passes over a drum. The light beam of the reproducing

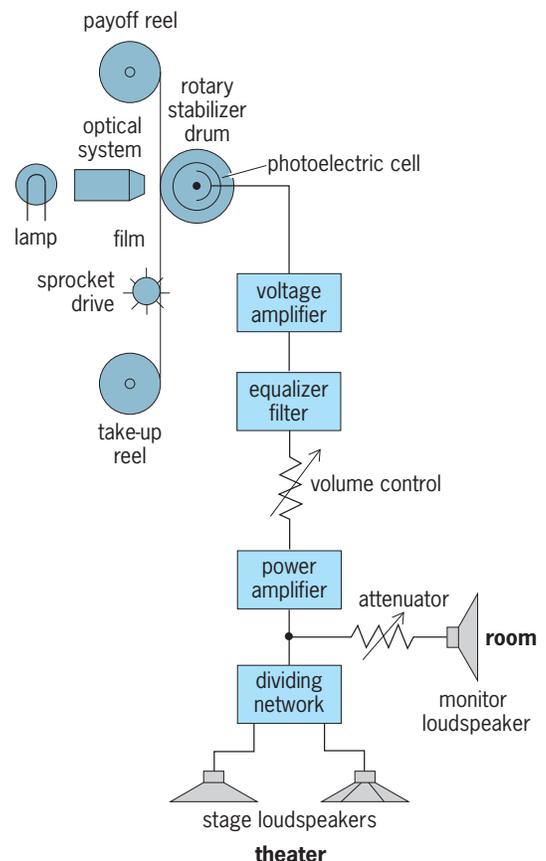


Fig. 3. Schematic arrangement of apparatus in a complete optical sound motion picture reproducing system.

system passes through the film to the photocell located inside the drum. The drum is driven by the first sprocket drive. The compliance of the film between the film and the sprocket provides a mechanical filter system and thereby reduces the sprocket-hole ripple at the drum. Under these conditions, the drum is rotated at a constant speed, and as a consequence the film will move past the light beam at a constant speed. The second sprocket isolates the take-up reel from the reproducing system.

**Distortion and noise in reproduction.** Most commonly, distortion in an optical reproducing system is due to the inherent nonlinear characteristics of the photographic process. This type of distortion can be reduced to a low value by the use of proper illumination in the recording or duplicating process. The developing processes must also be accurately controlled in order to achieve a low value of nonlinear distortion.

Noise in clean film is due to the inherent grain structure of the photographic medium. Scratches and foreign particles on the film add additional noise.

Another source of distortion is a nonuniform motion of the film in the recording and reproducing process. This is manifested as a frequency modulation of the reproduced signal and is termed flutter and wow.

**Laser-beam film recording.** In laser-beam film recording, an optical film system utilizes a laser as a light source, a combination of an acoustooptical modulator (AOM) and an acoustooptical deflector (AOD) instead of a galvanometer. A 100-kHz pulse-width modulation (PWM) circuit converts the audio input signal into a PWM signal. The laser beam is made to continuously scan the sound track area at right angles to the direction of the film transport. This is done by means of the acoustooptical deflector, which in turn is driven by a 100-kHz sawtooth signal. Simultaneously, the laser beam is pulse-width-modulated by means of the acoustooptical modulator, which is driven by a 100-kHz PWM signal. The scanning signal and the pulse-width-modulated signal combine and generate the variable-area sound track exposure on the film (Fig. 4). The traces of

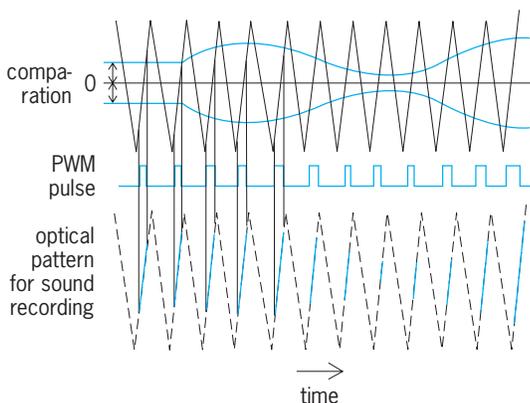


Fig. 4. Schematics of the sound track waveform generation of the laser sound recorder. Actual distance between successive scans is much smaller than shown. (After T. Taneda et al., *A high quality optical sound recording system using a scanned laser beam*, SMPTE J., 89:95-97, 1980)

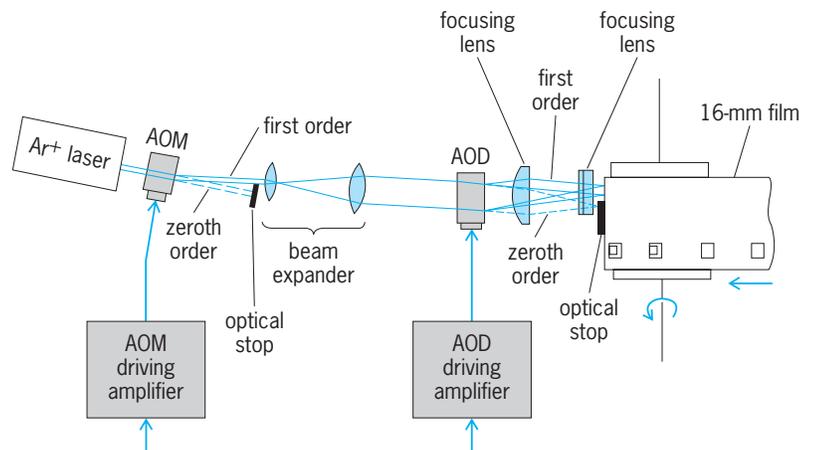


Fig. 5. Basic layout of the main optical components of the laser-beam sound recorder. (After T. Taneda et al., *A high quality optical sound recording system using a scanned laser beam*, SMPTE J., 89:95-97, 1980)

successive scans are fused into a pattern of variable-area recording.

The light from the laser (Fig. 5) is at first pulse-width-modulated in the acoustooptical modulator at 10-microsecond (100-kHz) intervals in response to the amplitude of the audio signal. As is normal for acoustooptical modulation, the modulated beam is separated into a zeroth-order (nondeflected) and a first-order (deflected) beam. Only the first-order beam is used, and the zeroth-order beam is blocked by an optical stop. Next, the diameter of the first-order beam is expanded to the size of the aperture. The beam is then deflected by the acoustooptical deflector and is again separated into a first-order and zeroth-order beam. Both beams are now converged to form spot images by means of a pair of cylindrical lenses. The first-order beam is focused on the film, while the zeroth-order beam is blocked by a fixed optical spot just in front of the film. Owing to the nature of the recording system, a wide-band frequency response can be obtained. See ACOUSTOOPTICS; LASER; PULSE MODULATION. H. Date

### Optical Data Storage

Optical data storage involves placing information in a medium so that, when a light beam scans the medium, the reflected light can be used to recover the information. There are many forms of storage media, and many types of systems are used to scan data.

**Storage principles.** The storage of data on optical media involves two steps. First, data are recorded on the medium. Recorded data have the property that the reflected light is modulated compared to regions without data. Second, a light beam scans the medium, and modulation in the reflected light is used to detect the data pattern under the scanning spot. See MODULATION.

In the recording process (Fig. 6), an input stream of digital information is converted with an encoder and modulator into a drive signal for a laser source. The laser source emits an intense light beam that is

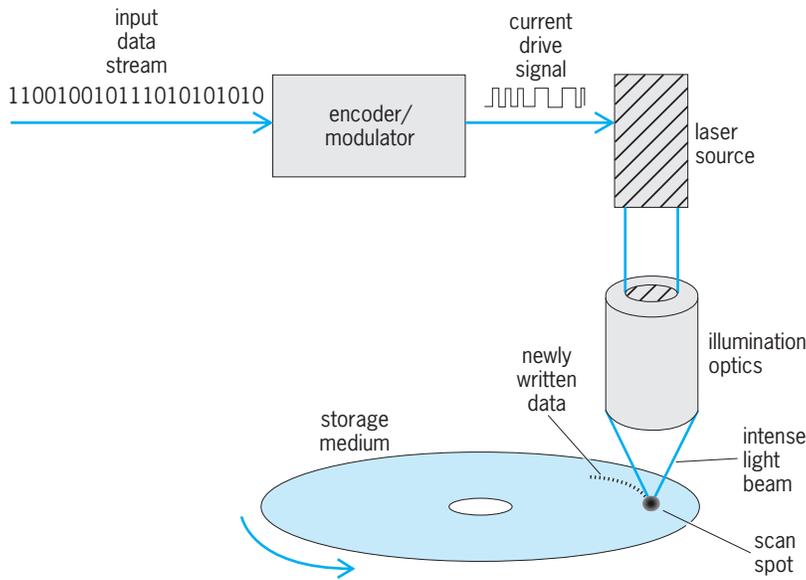


Fig. 6. Recording process for a simple optical medium. Writing data into the recording layers involves modulating an intense laser beam as the layers move under the scan spot.

directed and focused into the storage medium with illumination optics. As the medium moves under the scanning spot, energy from the intense scan spot is absorbed, and a small localized region heats up. The storage medium, under the influence of the heat, changes its reflective properties. Since the light beam is modulated in correspondence to the input data stream, a circular track of data marks is formed as the medium rotates. After every revolution, the path

of the scan spot is changed slightly in radius to allow another track to be written.

In readout of the medium (Fig. 7), the laser is used at a constant output power level that will not heat the medium beyond its thermal writing threshold. The laser beam is directed through a beam splitter into the illumination optics, where the beam is focused into the medium. As the data to be read pass under the scan spot, the reflected light is modulated. The modulated light is collected by the illumination optics and directed by the beam splitter to the servo and data optics, which converge the light onto detectors. The detectors change the light modulation into current modulation that is amplified and decoded to produce the output data stream.

**Configurations for optical media.** Optical media can be produced in several different configurations. The most common configuration is the single-layer disk (Fig. 8a), such as the compact disk (CD), where data are recorded in a single storage layer. A substrate provides mechanical support for the storage layer. The substrate is transparent and also provides a measure of contamination protection, because light is focused through the substrate and into the recording layer. Dust particles on the surface of the substrate only partially obscure the focused beam, so enough light can penetrate for adequate signal recovery. See COMPACT DISK.

In order to increase data capacity of the disk, several layers can be used (Fig. 8b). Each layer is partially transmitting, which allows a portion of the light to penetrate throughout the thickness of the layers. The scan spot is adjusted by refocusing the

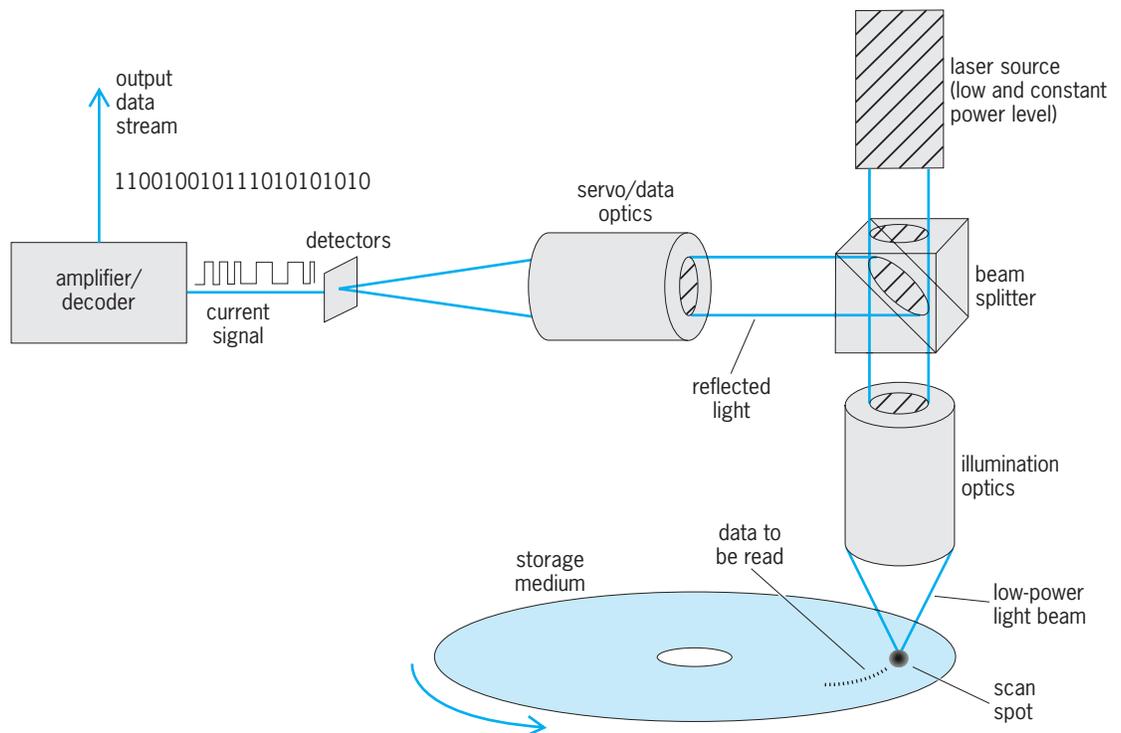


Fig. 7. Readout of an optical medium. Low-power laser beam illuminates the recording layers, and modulation of the reflected light is observed with the detectors. Beam splitter serves to direct a portion of the reflected light to detectors.

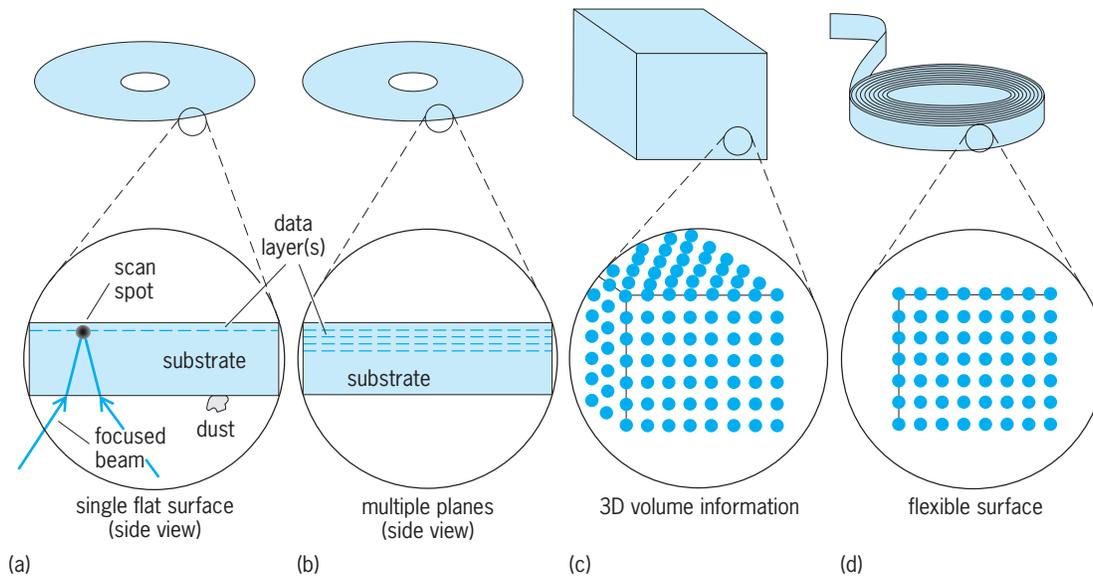


Fig. 8. Configurations for optical media. (a) Single-layer disk. (b) Multiple-layer disk. (c) Volumetric. (d) Ribbon.

illumination optics so that only one layer is read out at a time.

Data can also be recorded in volumetric configurations (Fig. 8c). As with the multiple-layer disk, the scan spot can be refocused throughout the volume of material to access information. Volumetric configurations offer the highest efficiency for data capacity, but they are not easily paired with simple illumination optics.

The final configuration is to place the information on a flexible surface, such as ribbon or tape (Fig. 8d). As with magnetic tape, the ribbon is pulled under the scan spot and data are recorded or retrieved. Flexible media have about the same capacity efficiency as volumetric storage. The advantage of a flexible medium over a volumetric medium is that no refocusing is necessary. The disadvantage is that a moderately complicated mechanical system must be used to move the ribbon.

**Commercial media and optical systems.** There are several types of optical storage media. The most popular media are based on pit-type, magneto-optic, phase-change, and dye-polymer technologies. CD and digital versatile disc (DVD) products use pit-type technology. Erasable disks using magneto-optic (MO) technology are popular for workstation environments. Compact-disk-rewritable (CD-RW) products [also known as compact-disk-erasable (CD-E)] use phase-change technology, and compact-disk-recordable (CD-R) products use dye-polymer technology. CD and DVD products are read-only memories (ROMs); that is, they are used for software distribution and cannot be used for recording information. CD-R products can be used for recording information, but once the information is recorded, they cannot be erased and reused. Both CD-RW and MO products can be erased and reused.

Pit-type technology for CD and DVD products is based on a simple scattering phenomenon. Data are recorded as small pits in the surface of a disk. The pits

are arranged in spiral tracks around the center of the disk. The pit length is of the order of the wavelength of light, or about 0.5–1 micrometer. The widths of the pits along a track are nearly uniform and measure about 0.5–0.8  $\mu\text{m}$ . Information is retrieved by scanning a focused optical spot, about 2  $\mu\text{m}$  wide and 2  $\mu\text{m}$  long, over the pits in a track as the disk spins. As the light spot passes over a pit, the reflected light scatters away from the illumination optics. The remaining light collected is small compared to the amount of light that gets collected when the spot is over a smooth portion of the track, where the disk acts as a mirror to the focused light. The data signal is derived from the detector that senses the amount of collected light.

MO products store information in small magnetic marks, which are about the same size as pits on a CD. Each mark contains magnetic domains oriented in the opposite direction compared to the magnetic domains of the background. The marks have the property that, as a focused light spot passes over it, the polarization of the reflected light is rotated. In order to detect the data signal, a detector is used to sense the change in polarization of the reflected light. To record or erase marks, a higher-power focused spot is used to locally heat the medium, and with the application of an external magnetic field, the magnetic domains of the material can be switched in micrometer-sized regions. A major difference between CD and MO products is that the MO marks are produced in a track with an almost undetectable change in the topology of the track. That is, there is almost no mechanical deformation of the track as the marks are recorded or erased. This property enables MO products to exhibit over  $10^6$  erase cycles with little if any degradation in performance. See MAGNETOOPTICS; POLARIZED LIGHT.

CD-RW products are similar to MO products in that they can be erased with multiple cycles before degradation. However, phase-change technology is based

on the differences of the crystalline and amorphous states of the medium. Marks representing data are stored along tracks. Usually, marks are in the amorphous state and the background is in the crystalline state. Amorphous and crystalline states reflect different amounts of light back into the optical system. As in a CD player, the detector in a CD-RW player senses the amount of collected light. The data signal is derived from the detector current. To record or erase phase-change marks, a higher-power focused spot is used to locally heat the medium in micrometer-sized regions. The thermal cycle of the local regions determines if the region will stabilize in a crystalline or amorphous state. By controlling the energy in the focused spot, the thermal cycle and the state of the material can be controlled. The phase-change process inevitably involves a mechanical deformation of the material. Therefore, the number of erase cycles is limited to several thousand. See AMORPHOUS SOLID.

The dye polymers or dye monomers used in CD-R products are organic films that are ablated to form pits along tracks. To form a pit, a high-power focused spot locally heats a micrometer-sized area. The dye polymer absorbs a large percentage of the

laser energy. Due to the low thermal conductivity of dye polymers, extremely high temperatures can be reached. In the heated area, the dye material is vaporized or heated to the point that material flows to form a pit. To read data, a low-power laser beam scans the track, and the collected light is sensed with a simple detector. The optical system of a CD-R player is similar to that of a CD-RW player.

**Performance.** Three important performance characteristics of optical data storage devices are the capacity, data rate, and access time.

Capacity is the maximum amount of data that can be stored on a single disk. Capacity is usually specified in terms of gigabytes. (1 GB =  $10^9$  bytes; 1 byte = 8 bits.) See BIT.

Data rate is the number of digital bits per second that are recorded or retrieved from a device during transfer of a large data block. Data rate is usually specified in terms of megabits per second. (1 Mbps =  $10^6$  bits per second.)

Access time is the latency (time lapse) experienced between when a request is made to access data and when the data start flowing through the communication channel. Access time is usually specified in terms of milliseconds. (1 ms =  $10^{-3}$  second.)

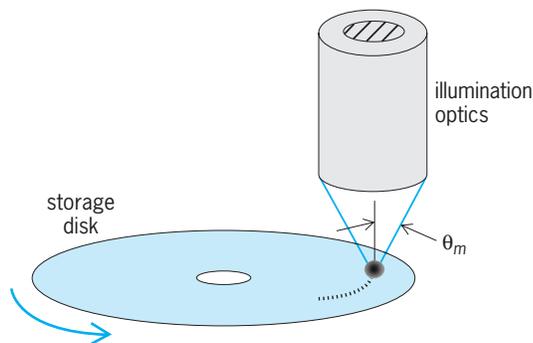
Together, data rate and access time determine the throughput of the device. That is, throughput determines the time required to locate and transmit data to and from the storage device.

The capacity is primarily determined by the size  $s$  of the scan spot (the diameter of a circle at which the light intensity drops to  $1/e^2$  of its maximum value). As  $s$  decreases, smaller transitions between marks and the regions around marks can be sensed. Therefore, marks can be made smaller and the density increases. A useful approximation is  $s \sim \lambda/NA$ , where  $\lambda$  is the wavelength of the laser in air and NA is the numerical aperture of the focused beam. Numerical aperture is defined as the sine of the focusing-cone half angle (Fig. 9). CD systems exhibit  $\lambda = 0.78 \mu\text{m}$  and  $NA = 0.45$ , which produces  $s = 1.7 \mu\text{m}$ . DVD systems exhibit  $\lambda = 0.65 \mu\text{m}$  and  $NA = 0.60$ , with  $s = 1.1 \mu\text{m}$  (Fig. 10).

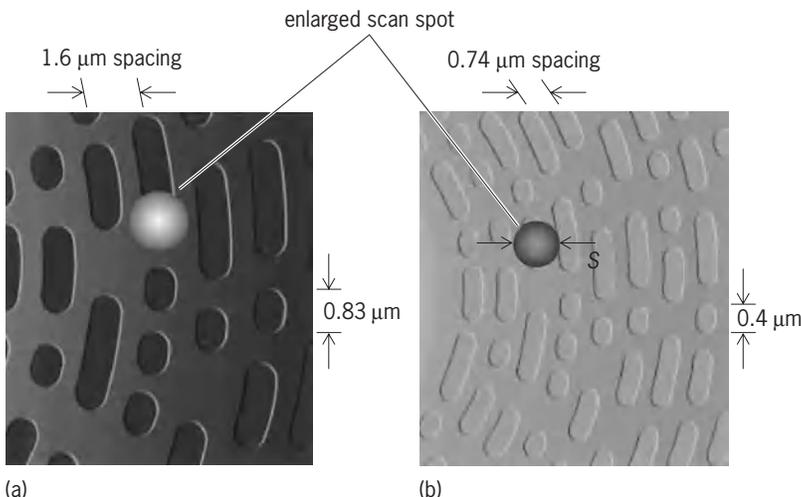
The data rate can be different for writing and reading data on a disk. During writing, the data rate is determined by the highest medium velocity that produces clearly defined marks. During reading, the data rate is determined by the highest medium velocity that produces sufficient signal-to-noise ratio. One straightforward way to increase data rate is to use more than one laser beam at a time. The increase in data rate is nearly proportional to the number of beams.

The access time is determined by the latency due to the time required for disk rotation. The highest latency is the time it takes the disk to make one revolution. Reduction of latency requires spinning the disk faster.

Important considerations for storage are the performance requirements of new and existing applications. For example, as introduced in 1991, the CD-ROM exhibited a capacity of 0.64 GB and a data rate of 1.2 Mbps. By 2000, although the CD-ROM still



**Fig. 9.** Geometry that determines the size  $s$  of the scan spot. The spot size is a function of the laser wavelength  $\lambda$  and the maximum angle,  $\theta_m$ , of the focused cone of light illuminating the disk:  $s = \lambda/NA = \lambda/\sin \theta_m$ , where NA is the numerical aperture.



**Fig. 10.** Comparison of scan spot sizes and sizes of data marks for (a) CD and (b) DVD media. Spot size  $s$  is about  $1.1 \mu\text{m}$  for CD and  $1.7 \mu\text{m}$  for DVD.

Performance data of disk-based products					
	CD-1X	CD-40X	DVD-1X	DVD-40X	BD-1X
Capacity, GB	0.64	0.64	4.7	4.7	25
Data rate, Mbps	1.2	48	10	400	36
Capacity-rate product (CRP)	0.77	30.7	47	4700	900
Retrieval time, min	70	1.7	62.7	1.6	92.6

had the same capacity, its data rate had been raised to over 50 Mbps. Its higher speed enabled its acceptance for computer applications over introductory DVD products, which had higher capacity but exhibited a data rate of only 10 Mbps.

A serious limitation exists with disk-based optical data storage. As the data rate increases, the playing time for a fixed capacity decreases. For applications that require long playing times (and correspondingly high capacities), this limitation necessitates compression, which is lossless only for compression factors of 2 to 4. For example, a CD-ROM drive operating at 50 Mbps takes only 102 seconds to read the entire disk. Correspondingly, a hypothetical DVD-ROM drive operating at 400 Mbps (a similar speed multiplier compared to the fast CD drive) takes less than 100 seconds to read a 4.7 GB disk.

A useful figure of merit is the capacity-rate product (CRP), which is the product of the capacity in GB and the data rate in Mbps. The CRP and other performance characteristics of disk-based products are given in the table. The data-rate speedup factor is shown as 1X or 40X, where 1X refers to the data rate of products first introduced, such as the CD-ROM in 1991, and 40X refers to a data rate that is 40 times faster than the 1X rate. Also included in the table are preliminary data concerning the Blue-Ray Disc (BD), which is now commercially available. See COMPUTER STORAGE TECHNOLOGY.

Tom D. Milster; Glenn T. Sincerbox

**Bibliography.** T. W. McDaniel and R. Victoria, *Handbook of Magneto-Optical Data Recording: Materials, Subsystems, Techniques*, Noyes Publications, Park Ridge, NJ, 1997; A. B. Marchant, *Optical Recording: A Technical Overview*, Addison-Wesley, Reading, MA, 1990; T. D. Milster, Near-field optics: A new tool for data storage, *Proc. IEEE*, 88(9):1480-1490, 2000; D. Psaltis, D. G. Stinson, and G. S. Kino, Introduction by T. D. Milster, Optical data storage: Three perspectives, *Optics & Photonics News*, pp. 35-39, November 1997.

## Optical rotatory dispersion

The change in rotation as a function of wavelength experienced by linearly polarized light as it passes through an optically active substance. See OPTICAL ACTIVITY; POLARIZED LIGHT.

**Optically active materials.** Substances that are optically active can be grouped into two classes. In the first the substances are crystalline and the optical activity depends on the arrangement of nonop-

tically active molecular units. When these crystals are dissolved or melted, the resulting liquid is not optically active. In the second class the optical activity is a characteristic of the molecular units themselves. Such materials are optically active as liquids or solids. A typical substance in the first category is quartz. This crystal is optically active and rotates the plane of polarization by an amount which depends on the direction in which the light is propagated with respect to the optic axis. Along the axis the rotation is  $29.73^\circ/\text{mm}$  ( $755^\circ/\text{in.}$ ) for light of wavelength 508.6 nanometers. At other angles the rotation is less and is obscured by the crystal's linear birefringence. Molten quartz, or fused quartz, is isotropic. Turpentine is a typical material of the second class. It gives rotation of  $-37^\circ$  in a 10-cm (3.94-in.) length for the sodium D lines. See CRYSTAL OPTICS.

**Reasons for variation.** In all materials the rotation varies with wavelength. Optical activity is actually circular birefringence. In other words, a substance which is optically active transmits right circularly polarized light with a different velocity from left circularly polarized light.

Any type of polarized light can be broken down into right and left components. Let these components be  $R$  and  $L$ . The lengths of the rotating light vectors will then be  $R/\sqrt{2}$  and  $L/\sqrt{2}$ . At  $t = 0$ , the  $R$  vector may be at an angle  $\psi_r$  with the  $x$  axis and the  $L$  vector at an angle  $\psi_l$ . Since the vectors are rotating at the same velocity, they will coincide at an angle  $\beta$  which bisects the difference as in Eq. (1). If  $R =$

$$\beta = \frac{\psi_r + \psi_l}{2} \quad (1)$$

$L$ , the sum of these two waves will be linearly polarized light vibrating at an angle  $\gamma$  to the axes given by Eq. (2).

$$\gamma = \frac{\psi_r - \psi_l}{2} \quad (2)$$

If, in passing through a material, one of the circularly polarized beams is propagated at a different velocity, the relative phase between the beams will change in accordance with Eq. (3), where  $d$  is

$$\psi'_r - \psi'_l = \frac{2\pi d}{\lambda}(n_r - n_l) + \psi_r - \psi_l \quad (3)$$

the thickness of the material,  $\lambda$  is the wavelength, and  $n_r$  and  $n_l$  are the indices of refraction for right and left circularly polarized light. The polarized light

incident at an angle  $\gamma$  has, according to this equation, been rotated an angle  $\alpha$  given by Eq. (4).

$$\alpha = \frac{\pi d}{\lambda}(n_r - n_l) \quad (4)$$

This shows that the rotation would depend on wavelength, even in a material in which  $n_r$  and  $n_l$  were constant and which thus had no dispersion of circular birefringence. In addition to this pseudodispersion, there is a true rotatory dispersion which depends on the variation with wavelength of  $n_r$  and  $n_l$ .

From Eq. (4) it is possible to compute the circular birefringence for various materials. This quantity is of the order of magnitude of  $10^{-8}$  for many solutions and  $10^{-5}$  for crystals. It is  $10^{-1}$  for linear birefringent crystals.

Bruce H. Billings

**Specific and molar rotations.** For purposes of comparison and compilation, the specific rotation  $[\alpha]$ , which is the optical rotation that would be given by a sample 1 decimeter (0.1 m) thick and at a concentration of 1 kg/dm<sup>3</sup>, is calculated from the observed rotation, as in Eq. (5), where  $d$  is the thickness in

$$[\alpha] = \frac{\alpha}{10dw} \quad (5)$$

meters and  $w$  is the concentration of optically active solute in kg/dm<sup>3</sup>. Molar rotation ( $[M]$ ), the optical rotation of a sample 1 m thick containing 1 mole/dm<sup>3</sup> of the optically active solute, is defined by Eq. (6),

$$[M] = \frac{[\alpha]M}{100} = \frac{\alpha}{dc} \quad (6)$$

where  $M$  is the molecular weight of the molecule or, in the case of a polymer, the molecular weight of the average monomer unit, and  $c$  is the solute concentration in moles per liter.

**Circular dichroism.** At wavelengths that are far from any absorption bands, the optical rotation generally shows a monotonic increase in magnitude as the wavelength decreases. The Drude equation (7) describes the optical rotatory disper-

$$[M] = \frac{A}{\lambda^2 - \lambda_0^2} \quad (7)$$

sion (ORD) in such cases, where  $A$  and  $\lambda_0$  are empirical constants, with  $\lambda_0$  generally being a wavelength in the far-ultraviolet region. Most colorless optically active substances obey the Drude equation in the visible and near-ultraviolet regions.

Wavelengths that are absorbed by the optically active sample, the two circularly polarized components will be absorbed to differing extents. This unequal absorption is known as circular dichroism (CD) and is given by Eq. (8), where  $\epsilon_l$  and  $\epsilon_r$  are the molar ex-

$$\Delta\epsilon = \epsilon_l - \epsilon_r \quad (8)$$

inction coefficients for left and right circularly polarized light, respectively. Circular dichroism causes incident linearly polarized light to become elliptically polarized. Optical rotation can still be measured as the angle between the plane of polarization of the

incident light and the major axis of the elliptically polarized light. The elliptically polarized light is also characterized by a second angle, the ellipticity  $\theta$ , which is an alternative measure of circular dichroism. The molar ellipticity, defined by analogy to the molar rotation [Eqs. (6) and (7)], is proportional to  $\Delta\epsilon$ , as in Eq. (9), where  $\theta$  is the observed ellipticity,

$$[\theta] = \frac{\theta}{dc} = 3300\Delta\epsilon \quad (9)$$

$d$  is the sample thickness in meters, and  $c$  is the molar concentration of the solute. See ABSORPTION OF ELECTROMAGNETIC RADIATION.

If a single absorption band is considered, the circular dichroism spectrum has the same shape as the absorption spectrum, whereas the optical rotatory dispersion spectrum shows more complex behavior. For a positive circular dichroism band, the optical rotatory dispersion passes through a maximum on the long-wavelength side of the circular dichroism and absorption maximum ( $\lambda_{\max}$ ), vanishes at the absorption maximum, then passes through a minimum on the short-wavelength side of the absorption maximum. For a negative circular dichroism band, characteristic of the mirror-image molecule (enantiomer), opposite behavior is observed. The characteristic shapes of optical rotatory dispersion and circular dichroism spectra in the vicinity of absorption bands are referred to as Cotton effects, with a positive optical rotatory dispersion Cotton effect corresponding to a positive circular dichroism band. See COTTON EFFECT.

Optical rotatory dispersion and circular dichroism are closely related, just as are ordinary absorption and dispersion. If the entire optical rotatory dispersion spectrum is known, the circular dichroism spectrum can be calculated, and vice versa. Circular dichroism bands are narrow, confined to regions of absorption, in contrast to optical rotatory dispersion bands, which decrease slowly away from the absorption maximum. Circular dichroism is generally preferred because the contributions of individual chromophores are more easily distinguished. Weak circular dichroism bands can be detected and measured without interference from strong bands at shorter or longer wavelengths. With optical rotatory dispersion, weak Cotton effects are difficult to study, because they are often small inflections on a large and strongly curving background. Optical rotatory dispersion still proves useful for molecules such as hydrocarbons and carbohydrates, for which circular dichroism bands are confined to the deep-ultraviolet.

**Relationships to molecular structure.** In order for a molecule (or crystal) to exhibit circular birefringence and circular dichroism, it must be distinguishable from its mirror image. An object that cannot be superimposed on its mirror image is said to be chiral, and optical rotatory dispersion and circular dichroism are known as chiroptical properties. Helical molecules represent especially clear examples of chiral objects and include the  $\alpha$ -helix in polypeptides and proteins, the double helix of deoxyribonucleic acid (DNA), the helix of glucose units in amylose

(starch), and the helices of oxygen-silicon-oxygen (O—Si—O) units in quartz crystals. Molecules that contain a carbon (or other atom) with four different substituents in a tetrahedral arrangement exist in two mirror-image forms and are therefore also chiral, and the two enantiomers have mirror optical rotatory dispersion and circular dichroism spectra. A carbon or other atom that has a chiral arrangement of substituents is said to be an asymmetric or chiral center. In many complexes of transition-metal ions, the metal ion is a chiral center. Most biological molecules have one or more chiral centers and undergo enzyme-catalyzed transformations that either maintain or reverse the chirality at one or more of these centers. Still other enzymes produce new chiral centers, always with a high specificity. These properties account for the fact that optical rotatory dispersion and circular dichroism are widely used in organic and inorganic chemistry and in biochemistry. See ENZYME; STEREOCHEMISTRY.

At one time, measurements of optical rotatory dispersion and circular dichroism were confined to visible and ultraviolet wavelengths, from about 185 to 700 nm. Developments in instrumentation for measuring circular dichroism have opened the very far-ultraviolet (down to about 140 nm) and the infrared (wavelengths up to 15 micrometers). The deep-ultraviolet has proven useful for studies of proteins, nucleic acids, and carbohydrates. Circular dichroism in the infrared derives from transitions between vibrational energy levels, rather than electronic energy levels as in visible-ultraviolet circular dichroism, and it is known as vibrational circular dichroism (VCD). Measurements of vibrational circular dichroism, though technically demanding, promise to provide much information about the stereochemistry and conformation of both large and small molecules. See INFRARED RADIATION.

**Magnetic effects.** In the absence of magnetic fields, only chiral substances exhibit optical rotatory dispersion and circular dichroism. In a magnetic field, even substances that lack chirality rotate the plane of polarized light, as shown by M. Faraday. Magnetic optical rotation is known as the Faraday effect, and its wavelength dependence is known as magnetic optical rotatory dispersion (MORD). In regions of absorption, magnetic circular dichroism (MCD) is observable. In addition to the other advantages of circular dichroism over optical rotatory dispersion, solvent and sample cells do not contribute to the magnetic circular dichroism signal but do contribute to magnetic optical rotatory dispersion, making measurements of the latter on dilute solutions difficult. Magnetic circular dichroism has been used to assign electronic transitions and thus determine the nature of excited states in both transition-metal complexes and aromatic hydrocarbons. In biochemistry, magnetic circular dichroism has provided important information about the geometry of the coordination sites of transition-metal ions such as iron and copper in metalloproteins. See COORDINATION CHEMISTRY; COORDINATION COMPLEXES; FARADAY EFFECT; MAGNETOOPTICS.

R. W. Woody

**Bibliography.** L. D. Barron, *Molecular Light Scattering and Optical Activity*, 1983; S. F. Mason, *Molecular Optical Activity and the Chiral Discriminations*, 1982; K. Nakanishi, N. Berova, and R. Woody, *Circular Dichroism: Principles and Applications*, 1994; N. Purdie (ed.), *Analytical Applications of Circular Dichroism*, 1993.

## Optical surfaces

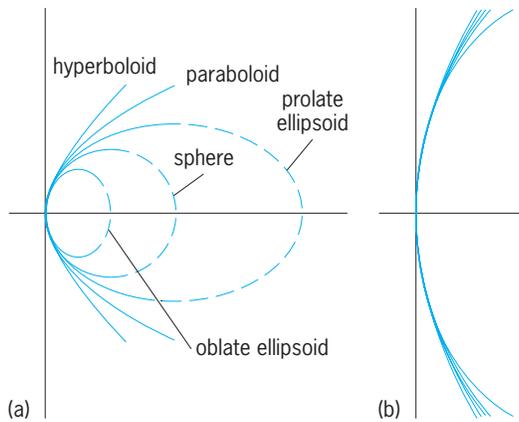
Interfaces between different optical media at which light is refracted or reflected. From a physical point of view, the basic elements of an optical system are such things as lenses and mirrors. However, from a conceptual point of view, the basic elements of an optical system are the refracting or reflecting surfaces of such components, even though they cannot be separated from the components. Surfaces are the basic elements of an optical system because they are the elements that affect the light passing through the system. Every wavefront has its curvature changed on passing through each surface so that the final set of wavefronts in the image space may converge on the appropriate image points. Also, the aberrations of the system depend on each surface, the total aberrations of the system being the sum of the aberrations generated at the individual surfaces. See ABERRATION (OPTICS); REFLECTION OF ELECTROMAGNETIC RADIATION; REFRACTION OF WAVES.

Optical systems are designed by ray tracing, and refraction at an optical surface separating two media of different refractive index is the fundamental operation in the process. The transfer between two surfaces is along a straight line if, as is usually the case, the optical media are homogeneous. The refraction of the ray at a surface results in a change in the direction of the ray. This change is governed by Snell's law.

**Spherical and aspheric surfaces.** The vast majority of optical surfaces are spherical in form. This is so primarily because spherical surfaces are much easier to generate than nonspherical, or aspheric, surfaces. Not only is the sphere the only self-generating surface, but a number of lenses can be ground and polished on the same machine at the same time if they are mounted together on a common block so that the same spherical surface is produced simultaneously on all of them.

Although aspheric surfaces can potentially improve the performance of a lens system, they are very rarely used. High-quality aspheric surfaces are expensive to produce, requiring the services of a highly skilled master optician. Moreover, lens systems seldom need aspherics because the aberrations can be controlled by changing the shape of the component lenses without changing their function in the system, apart from modifying the aberrations. Also, many lens components can be included in a lens system in order to control the aberrations. See LENS (OPTICS).

On the other hand, mirror systems usually require aspheric surfaces. Unlike lenses, where the shape



**Conics of revolution.** (a) Cross sections of entire surfaces. (b) Cross sections of portions near the optical axis.

can be changed to modify the aberrations, mirrors cannot be changed except by introducing aspheric surfaces. Mirror systems are further constrained by the fact that only a few mirrors, usually two, are used in a system because each successive mirror occludes part of the beam going to the mirror preceding it. See MIRROR OPTICS.

**Conics of revolution.** The most common form of rotationally symmetric surface is the conic of revolution. This is obtained conceptually by rotating a conic curve (ellipse, parabola, or hyperbola) about its axis of symmetry. There are two forms of ellipsoid, depending on whether the generating ellipse is rotated about its major or its minor axis. In the first case it is a prolate ellipsoid, and in the second it is an oblate ellipsoid. There is only one form of paraboloid, and only the major axis is used in generating the hyperboloid. The departure of conic surfaces from spherical form is shown in the illustration. See CONIC SECTION.

The classical virtue of the conics of revolution for mirrors is the fact that light from a point located at one focus of the conic is perfectly imaged at the other focus. If these conic foci are located on the axis of revolution, the mirror is free of spherical aberration for such conjugate points. For example, the classical Cassegrain design consists of two mirrors, a paraboloidal primary mirror and a hyperboloidal secondary mirror. The paraboloid forms a perfect image of a point at infinity on its axis. When the light converging to this image is intercepted by the convex hyperboloid with its virtual conic focus at the image point, the final image will be formed at the real conic focus of the hyperboloid. Thus, in the classical Cassegrain design the two mirrors are separately corrected for spherical aberration, and so is the system. However, all other monochromatic aberrations remain uncorrected.

Instead of using the two aspheric terms to correct the same aberration for the two mirrors separately, it is more effective to use the two aspherics to correct two aberrations. A Cassegrain system in which both spherical aberration and coma are corrected by the two aspherics is aplanatic, and is identified

as a Ritchey-Chrétien system. The aspheric on the primary is a hyperboloid slightly stronger than the paraboloid of the classical primary, and the aspheric on the secondary is also a hyperboloid, slightly different from that for the classical secondary.

**General aspherics of revolution.** A more general aspheric of revolution is frequently used where terms in even powers of the distance from the axis, usually from the fourth to the twelfth power, are added to the spherical or conic term. This gives greater control over the aberrations, especially if higher-order aberrations are significant. The added terms are necessary also for describing an aspheric departure from a plane surface where a conic cannot be used, as with a Schmidt corrector plate. See SCHMIDT CAMERA.

**Nonrotationally symmetric surfaces.** Occasionally nonrotationally symmetric surfaces are used in optical systems. The most common varieties are cylindrical and toric surfaces, where the cross sections are circular. They are employed, for example, in the optical systems for taking and projecting wide-screen motion pictures, where the image recorded on the film is compressed in the horizontal direction but not in the vertical. The camera lens compresses the image and the projector lens expands it to fill the wide screen. Noncircular cross sections can also be specified by adding terms in even powers of  $x$  and  $y$ , the coordinates perpendicular to the optical axis, to the description of the surface, but these surfaces are very difficult to make. See CINEMATOGRAPHY.

**Eccentric mirrors.** Another significant class of aspheric surfaces consists of those which are eccentric portions of rotationally symmetric aspheric surfaces. These are used in mirror systems to eliminate the obscuration of the incoming beam by succeeding mirrors. For small mirrors they are cut out of a large rotationally symmetric mirror, but for larger mirrors they can be made by computer-controlled generating and polishing machines. See GEOMETRICAL OPTICS.

Roland V. Shack

**Bibliography.** D. Malacara, *Optical Shop Testing*, 2d ed., 1992; D. Malacara and A. Malacara, *Handbook of Lens Design*, 1994; Optical Society of America, *Handbook of Optics*, 2 vols., 2d ed., 1995; D. C. O'Shea, *Elements of Modern Optical Design*, 1985.

## Optical tracking systems

Multipurpose instruments used to make measurements on a remote object, often an airborne vehicle. These systems are used to provide two basic types of data: accurate measurement of the position, velocity, and other motion parameters of the target; and information about the target, such as images or optical spectra.

Many optical tracking systems are used for weapon system qualification. A system tracking a towed aerial target can accurately determine the miss distance of an interceptor missile. One tracking a military aircraft can measure its rate of climb or the release characteristics of its ordnance.



Fig. 1. Little Bright Eyes. (White Sands Missile Range)

Another common use of optical tracking systems is for range safety at missile and rocket launch sites. A remotely operated optical tracking system can provide real-time position data and images of the launch vehicle, assuring that all technical parameters are within specification without risk to personnel.

**Development.** The first optical tracking mount was developed in Germany in 1937 at the Peenemünde Rocket Launch Site by associates of Wernher von Braun. Development continued to support the German rocket development program until the production of the first commercial phototheodolite in 1940. The United States government issued a specification the next year for a similar instrument, but its development was not completed before the end of the war.

With the 1945 arrival of von Braun, his scientific crew, and their V2 rockets in the United States, a longer-range tracking mount was required. James Edson, an anti-aircraft artillery officer, married a 35-mm film camera to a gun mount and created the forerunner of modern optical tracking mounts, called Little Bright Eyes (Fig. 1). Test ranges incorporated a succession of improvements in each new instrument, but all were based on gun mounts. This changed with the third generation of the Intercept Ground Optical Recorder (IGOR-III), developed and produced in 1960. IGOR-III was designed from the beginning to be a precision optical instrument and incorporated many design features still in use.

Two operators rode on the IGOR system, one controlling the elevation angle and the other the azimuth or bearing direction. This was common practice on optical tracking systems until the development of the Recording Optical Tracking Instrument (ROTI). This instrument was similar in size and capability to the IGOR, but could be directed by a single operator using a joystick. The development of these large systems concluded with production of the Distant Object Attitude Measurement System (DOAMS; Fig. 2) in 1976.

IGOR, ROTI, and DOAMS are the high end of optical trackers, with apertures up to 20 in. (0.5 m) and 400-in. (10-m) focal length. Coupled with large-format 70-mm high-speed film cameras, these instru-

ments were perfectly suited for documenting high-altitude and long-range intercepts. These missions constitute only a small portion of range test requirements; most of the rest could be easily satisfied by a smaller instrument.

The cinetheodolite (Fig. 3), familiarly referred to as a "cine", was developed in 1952. These instruments quickly became standard equipment at test ranges. Various models were in production for 30 years, with over 400 systems built. The most common of these, the EOTS-F, had a 7.5-in. (190-mm)



Fig. 2. Distant Object Attitude Measurement System (DOAMS). (L-3 Communications Corp.)



Fig. 3. Cinetheodolite. (L-3 Communications Corp.)

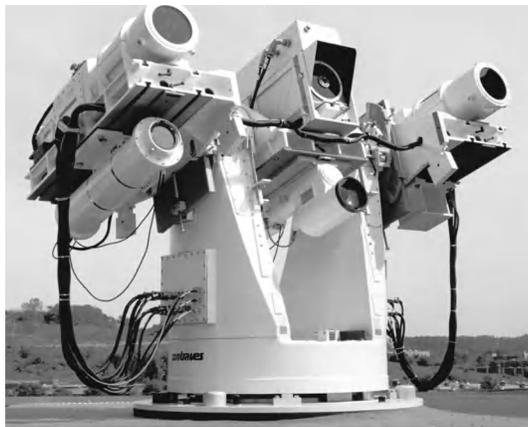


Fig. 4. KINETO tracking mount. (L-3 Communications Corp.)

aperture and a 35-mm high-speed film camera. Nearly 100 of these systems are still in use.

The cinetheodolites were extremely accurate, but required specially prepared sites and had a single optical configuration. There was a need for a more versatile instrument, and in 1963 the Cine Sextant, a precision two-axis tracking system mounted on a trailer, was introduced. These systems could carry multiple instruments and be reconfigured for each test. They have been largely supplanted by the Compact Tracking Mount (CTM) and the KINETO tracking mount (Fig. 4). These somewhat smaller mounts have all-electric drives instead of the hydraulic/electric combination on the Cine Sextant and as a result are considerably less expensive. Over 150 KINETOs have been produced, making it the most common system in use at test ranges.

**Characteristics.** To achieve the basic objective, directing an optical instrument at an agile target, an optical tracking system must meet several key performance parameters, as summarized in the table.

The aperture of the optical system sets the overall sensitivity; a larger lens collects more light. The focal length determines the size of the image, as in a normal film camera. Optical tracking systems use long-focal-length, telephoto-type lenses. The system resolution is the angular extent of the smallest discernible features. See FOCAL LENGTH; GEOMETRICAL OPTICS; LENS (OPTICS); RESOLVING POWER (OPTICS); TELESCOPE.

The angular velocity is a measure of how fast the system rotates. A jet flying at the speed of sound 1 mi (1.6 km) from the tracking system requires a line-of-

sight rate of  $12^\circ/\text{second}$ . A KINETO tracking mount can move at eight times that rate.

**Tracking.** Optical tracking systems have two basic modes of operation, feedback tracking and ephemeris tracking. With feedback tracking, the position of the target in the field of view of the tracking sensor provides information that corrects the trajectory of the system. Feedback tracking can be as simple as an operator moving a joystick based on what he or she sees on a screen, or it can employ an automatic video tracker (AVT), a specialized image-processing computer that measures the target position in the sensor field of view. The most common type of AVT is called a centroid tracker because it determines the outline of the target and then calculates its center. Some types of objects that have variable size, such as a missile with a plume, are better tracked with an edge tracker. A third type, called a correlation tracker, overlays the current image with the previous one and adjusts the offsets to minimize the differences between them. This is useful in high-clutter environments.

An ephemeris is a mathematical prediction of the path of a celestial body, the usual means of tracking a satellite. Most satellites are visible only when the Sun is shining on them and the angle between the Sun and the observer is favorable. By accurately calibrating the pointing of the tracking system and commanding it from the same equations that determine the motion of the satellite, the line of sight can be directed correctly without feedback. This could be important, for instance, in maintaining a laser communications link to a faint satellite. See EPHEMERIS.

**Instrumentation.** Historically, the instrument of choice for optical tracking systems has been the high-speed film camera. Under favorable conditions, a film could deliver 0.00025-in. ( $10\text{-}\mu\text{m}$ ) resolution in a 70-mm format at 1000 frames per second. To do the same, a digital camera would need 50 megapixels and an 800-gigabit-per-second data rate. No commercial camera has reached this level of performance, but there have been great advances. Delivery of 1000 pictures per second at 2 megapixels has been achieved. See CAMERA; PHOTOGRAPHY.

Digital cameras have largely replaced film cameras on the ranges, partly due to the desire for real-time data and partly due to the expense and hazard of film processing. This replacement has been facilitated by the development of standard camera interfaces.

An increasing number of optical tracking systems employ infrared cameras, sometimes called FLIRs after their original application as the forward-looking infrared sensors in aircraft. These sensors do not have the same resolution as visible-band cameras, but have the advantage of being able to track objects that are not sunlit or that have poor contrast. See INFRARED IMAGING DEVICES.

It is often necessary to measure the range to a target. This is accomplished by adding a radar ranging system to the optical tracking system or by incorporating a laser rangefinder. Both determine the distance by measuring the time necessary for a pulse to travel to the target and be reflected back to its source.

Typical tracking system performance		
Parameter	Typical value	Units
Aperture	12 (300)	inches (mm)
Focal length	50 (1250)	inches (mm)
Resolution	1–5	arc-seconds*
Angular velocity	30	degrees/second
Pointing accuracy	5–10	arc-seconds*

\*1 arc-second =  $(1/3600)^\circ = 0.00028^\circ$ .

Most laser rangefinders now meet standards for eye safety and can easily be added to a KINETO or other configurable tracking system. See LASER; RADAR.

**Special applications.** Four applications are described below.

*Time space-position information (TSP).* One or more optical tracking systems can be used to accurately fix the position of a target in space. To effectively use this information, the system locations must be accurately leveled and precisely located by survey or the Global Positioning System. One approach is to use a single optical tracking system that is fitted with a laser rangefinder and to convert measurements in this azimuth-elevation-range (AER) coordinate system into Cartesian (XYZ) coordinates so that a target trajectory is established.

By time-tagging the data (accurately determining the time at which the position is recorded), several instruments can make coordinated measurements. By using techniques similar to a surveyor, angle-only measurements from two, three, or four optical tracking systems can determine the Cartesian position of the target. The use of a greater number of tracking systems can reduce target location errors through advanced processing. See SURVEYING.

*Fire-control systems.* The ability of an optical tracking system to accurately determine the position and trajectory of a target leads naturally to its incorporation into air defense systems. The optical line of sight is corrected for the ballistic characteristics of the weapon being used and the calculated motion of the target until intercept.

*Space surveillance.* The United States government's Groundbased Electro-Optical Deep Space Surveillance (GEODSS) telescopes perform a critical function by mapping the location of artificial and natural objects in high orbits around the Earth. Objects in low-earth orbits are tracked by radar, but this optical tracking system was uniquely configured to detect these small objects at ranges of thousands of kilometers. The tracking system slowly and precisely changes its direction to compensate for the rotation of the Earth, causing the stars to appear stationary in its camera. Satellites, space debris, and minor asteroids appear as short streaks since they move relative to the stars. Once a small part of the sky has been investigated, the system quickly repositions to another location and repeats the process. One GEODSS site has detected over 3 million minor asteroids, of which 200,000 were previously unknown. See ASTEROID.

*Satellite laser ranging.* There is a network of ground-based satellite laser ranging systems run by the National Aeronautics and Space Administration (NASA), foreign governments, and foreign and domestic universities. Each of these systems continuously measures the distance from its location to specially instrumented reflective satellites orbiting the Earth, often day and night. A 4000-mi (6400-km) measurement can be made with an accuracy of better than 0.04 in. (1 mm). The accuracy of these measurements, coupled with the worldwide distribution of the stations, the extreme stability of the satellite orbits, and the long history of observations, allows the motion of

the Earth's individual crustal plates, amounting to millimeters per year, to be precisely measured. See GEODESY.

James E. Kimbrell

Bibliography. R. Delgado, Photo-optical range instrumentation: An overview, *Opt. Eng.*, 20(5):701-711, September-October 1981.

## Optics

Narrowly, the science of light and vision; broadly, the study of the phenomena associated with the generation, transmission, and detection of electromagnetic radiation in the spectral range extending from the long-wave edge of the x-ray region to the short-wave edge of the radio region. This range, often called the optical region or the optical spectrum, extends in wavelength from about 1 nanometer ( $4 \times 10^{-8}$  in.) to about 1 mm (0.04 in.). For information on the various branches of optics see GEOMETRICAL OPTICS; METEOROLOGICAL OPTICS; PHYSICAL OPTICS; VISION.

In ancient times there was some isolated elementary knowledge of optics, but it was the discoveries of the experimentalists of the early seventeenth century which formed the basis of the science of optics. The statement of the law of refraction by W. Snell, Galileo Galilei's development of the astronomical telescope and his discoveries with it, F. M. Grimaldi's observations of diffraction, and the principles of the propagation of light enunciated by C. Huygens and P. de Fermat all came in this relatively short period. The publication of Isaac Newton's *Opticks* in 1704, with its comprehensive and original studies of refraction, dispersion, interference, diffraction, and polarization, established the science.

So great were the contributions of Newton to optics that a hundred years went by before further outstanding discoveries were made. In the early nineteenth century many productive investigators, foremost among them Thomas Young and A. J. Fresnel, established the transverse-wave nature of light. The relationship between optical and magnetic phenomena, discovered by M. Faraday in the 1840s, led to the crowning achievement of classical optics—the electromagnetic theory of J. C. Maxwell. Maxwell's theory, which holds that light consists of electric and magnetic fields propagated together through space as transverse waves, provided a general basis for the treatment of optical phenomena. In particular, it served as the basis for understanding the interaction of light with matter and, hence, as the basis for treatment of the phenomena of physical optics. In the hands of H. A. Lorentz, this treatment led at the end of the nineteenth century and the beginning of the twentieth to an explanation of many optical phenomena, such as the Zeeman effect, in terms of atomic and molecular structure. The theories of Maxwell and Lorentz are regarded as the culmination of classical optics. See ELECTROMAGNETIC RADIATION; LIGHT; MAXWELL'S EQUATIONS.

In the twentieth century, optics has been in the forefront of the revolution in physical thinking caused by the theory of relativity and especially by

the quantum theory. To explain the wavelength dependence of heat radiation, the photoelectric effect, the spectra of monatomic gases, and many other phenomena of physical optics, radical departure from the ideas of Lorentz and Maxwell about the mechanism of the interaction of radiation and matter and about the nature of radiation itself has been found necessary. The chief early quantum theorists were M. Planck, A. Einstein, and N. Bohr; later came L. de Broglie, W. Heisenberg, P. A. M. Dirac, E. Schrödinger, and others.

The science of optics finds itself in a position that is satisfactory for practical purposes but less so from a theoretical standpoint. The theory of Maxwell is sufficiently valid for treating the interaction of high-intensity radiation with systems considerably larger than those of atomic dimensions. The modern quantum theory is adequate for an understanding of the spectra of atoms and molecules and for the interpretation of phenomena involving low-intensity radiation, provided one does not insist on a very detailed description of the process of emission or absorption of radiation. However, a general theory of relativistic quantum electrodynamics valid for all conditions and systems has not been worked out. See QUANTUM ELECTRODYNAMICS.

The development of the laser has been an outstanding event in the history of optics. The theory of electromagnetic radiation from its beginnings was able to comprehend and treat the properties of coherent radiation, but the controlled generation of coherent monochromatic radiation of high power was not achieved in the optical region until the work of C. H. Townes and A. L. Schawlow in 1958 pointed the way. Many achievements in optics, such as holography and interferometry over long paths, have resulted from the laser. See HOLOGRAPHY; INTERFEROMETRY; LASER.

Richard C. Lord

Bibliography. M. Born and E. Wolf, *Principles of Optics*, 7th ed., 1999; B. D. Guenther, *Modern Optics*, 1990; F. A. Jenkins and H. E. White, *Fundamentals of Optics*, 4th ed., 1976; J. Meyer-Arendt, *Introduction to Classical and Modern Optics*, 4th ed., 1995; Optical Society of America, *Handbook of Optics*, 2 vols., 2d ed., 1995; O. Svelto, *Principles of Lasers*, 4th ed., 1998.

### Optimal control (linear systems)

A branch of modern control theory that deals with designing controls for linear systems by minimizing a performance index that depends on the system variables. Under some mild assumptions, making the performance index small also guarantees that the system variables will be small, thus ensuring closed-loop stability. See CONTROL SYSTEM STABILITY.

Classical control theory applies directly in the design of controls for systems that have one input and one output. Complex modern systems, however, have multiple inputs and outputs. Examples include aircraft, satellites, and robot manipulators, which, though nonlinear, can be linearized about a desired

operating point. Modern control theory was developed, beginning about 1960, for these multivariable systems. It is characterized by a state-space description of systems, which involves the use of matrices and linear algebra, and by the use of optimization techniques to determine the control policies. These techniques facilitate the use of modern digital computers, which lead to an interactive approach to design of controls for complex systems. See DIGITAL COMPUTER; LINEAR ALGEBRA; MATRIX THEORY; MULTIVARIABLE CONTROL; OPTIMIZATION.

The multivariable state-variable description is of the form of Eqs. (1), where  $u(t)$  is an  $m$ -dimensional

$$\dot{x} = Ax + Bu \quad z = Hx \quad (1)$$

control input vector and  $z(t)$  is a performance output vector for which there are specified performance requirements. The state  $x(t)$  is an internal variable of dimension  $n$  that describes the energy storage properties of the system. Matrices  $A$  and  $B$  describe the system dynamics and are determined by a physical analysis using, for example, Newton's laws of motion. They may in general be time-varying. Matrix  $H$  is chosen to select the variables of importance as the performance outputs. In the case of nonlinear systems, the description of Eq. (1) results when the system is linearized about a desired operating point. See LINEAR SYSTEM ANALYSIS; NONLINEAR CONTROL THEORY.

**State-variable feedback design.** Traditionally, modern control-system design assumes that all the states  $x(t)$  are available for feedback so that the control input is of the form of Eq. (2), where  $K(t)$  is an  $m \times n$

$$u = -Kx \quad (2)$$

feedback gain matrix, generally time-varying. Substituting the control of Eq. (2) into the system of Eq. (1) yields the closed-loop system given in Eqs. (3). The

$$\dot{x} = (A - BK)x \quad z = Hx \quad (3)$$

control design problem is to choose the  $mn$  entries of the feedback matrix  $K$  to yield a desired closed-loop behavior of the performance output  $z(t)$ .

**Linear quadratic regulator.** To obtain satisfactory performance of the system of Eq. (1), a quadratic performance index of the form of Eq. (4) may be chosen,

$$J = \frac{1}{2}x^T(T)S(T)x(T) + \frac{1}{2} \int_0^T (x^T Qx + u^T R u) dt \quad (4)$$

where  $[0, T]$  is the time interval of interest and the symmetric weighting matrices  $S(T)$ ,  $Q$ , and  $R$  are design parameters that are selected to obtain the required performance. They must be chosen so that  $x^T(T)S(T)x(T) \geq 0$  and  $x^T Qx \geq 0$  for all  $x(t)$ , and  $u^T R u > 0$  for all  $u(t)$ . That is,  $Q$  and  $S(T)$  should be positive semidefinite while  $R$  should be positive definite.

The positive semidefiniteness of the weighting matrices ensures that the performance index  $J$  is

nonnegative. Then, the performance index can be considered as a generalized energy function, so that, if it is kept small, the states  $x(t)$  and controls  $u(t)$  that are weighted in Eq. (4) are also small over the time interval  $[0, T]$ .

The selection of the design parameters  $S(T)$ ,  $Q$ , and  $R$  is generally not straightforward, and presents one of the challenges in the use of modern design techniques. There has been a large amount of work on the topic, which involves consideration of time-domain responses, frequency-domain responses, and the closed-loop poles.

The optimal control design problem is now to select the feedback gain matrix  $K$  in such a way that the performance index is minimized. This is called the linear-quadratic regulator (LQR) problem because the system of Eq. (1) is linear, the performance index of Eq. (4) is a quadratic function, and minimizing  $J$  corresponds to keeping the state small or regulating it to zero.

The optimal linear-quadratic feedback gain  $K$  may be determined by introducing a Lagrange multiplier  $\lambda(t)$  and applying the calculus of variations. By using well-known techniques, it can be shown that the optimal gain  $K$  may be computed by solving the system equation (1), in conjunction with the Euler-Lagrange equations (5) and (6). To solve these equations, it is

$$\lambda = -A^T \lambda - Qx \tag{5}$$

$$0 = B^T \lambda + Ru \tag{6}$$

necessary to take into account the boundary equations (7). While the system of Eq. (1) is solved forward in time, given an initial condition, the Euler-Lagrange equations must be solved backward in time, given a final condition. See CALCULUS OF VARIATIONS.

$$x(0) \text{ given} \tag{7a}$$

$$\lambda(T) = S(T)x(T) \tag{7b}$$

Solving Eq. (6) for  $u(t)$  and substituting into Eq. (1) yields the hamiltonian system in Eqs. (8). Solving this

$$x = Ax - BR^{-1}B^T \lambda \tag{8a}$$

$$\lambda = -A^T \lambda - Qx \tag{8b}$$

system with the boundary conditions of Eqs. (7) is a two-point boundary value problem, since  $x(0)$  is specified at the initial time, but Eq. (7b) must hold at the final time. By introducing an auxiliary time-varying  $n \times n$  matrix variable  $S(t)$  and using the so-called sweep method, this can be converted to an easier problem. Thus, the optimal linear-quadratic feedback gain  $K$  in the feedback control law given in Eq. (2) may be determined by using Eq. (9), with

$$K = -R^{-1}B^T S \tag{9}$$

$S(t)$  determined by solving the matrix Riccati equation (10), where the final condition  $S(T)$  is the design

$$-S = A^T S + SA + Q - SBR^{-1}B^T S \tag{10}$$

matrix in the performance index  $J$ .

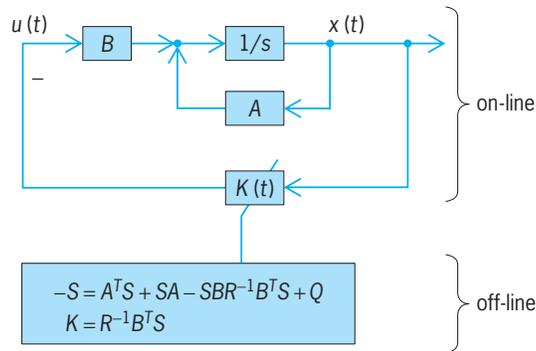


Fig. 1. Feedback formulation of the linear quadratic regulator.

The structure of the linear-quadratic regulator is shown in Fig. 1, where  $s$  is the Laplace transform variable. It is apparent that it consists of an inner control loop of feedback gains  $K(t)$ , along with an outer loop required to determine  $K(t)$  by solving the quadratic matrix Riccati equation. This hierarchical structure, with a nonlinear outer loop, is typical of modern control systems.

Although the structure of the control system appears complicated, it is not difficult to implement by using modern digital computers. In fact, the Riccati equation may be solved off-line before the control run and the feedback gain  $K(t)$  stored in computer memory. Then, during the control run these gains are switched into the system control loop.

It can be shown that, by using the optimal gain  $K(t)$  just determined, the performance index takes on the minimum value given by Eq. (11). Thus,  $J_{\min}$  may be

$$J_{\min} = x^T(0)S(0)x(0) \tag{11}$$

computed by knowing only the initial state of the system before the control is actually applied to the system. If this value for the performance index is not suitable, new values for the design parameters  $S(T)$ ,  $Q$ , and  $R$  may be selected and the design repeated. This is an important feature of the linear-quadratic regulator. See OPTIMAL CONTROL THEORY.

*Constant feedback gains.* The optimal linear-quadratic feedback gain  $K(t)$  is time-varying. This is in sharp contrast to the situation in classical control theory, where feedback gains are designed by using Laplace transform theory and so are constant. Since constant-feedback gains are easier to implement than time-varying gains, it may be of interest to obtain constant-feedback gains for multivariable systems.

If the final time  $T$  approaches infinity, then the time interval of interest becomes  $[0, \infty)$ , so that the feedback law tries to minimize the performance index and hence regulate the system to zero over all time. This is called the infinite-horizon linear-quadratic control problem. In this case, if the system and performance-index matrices  $A$ ,  $B$ ,  $Q$ , and  $R$  are time-invariant, then under certain conditions the Riccati equation (10) has a limiting solution  $S_{\infty}$  that is constant. If this occurs, then  $S; ' = 0$ , so that

Eq. (10) becomes the algebraic Riccati equation (12).

$$0 = A^T S + SA + Q - SBR^{-1}B^T S \quad (12)$$

Then,  $S_\infty$  may be determined directly by solving the algebraic Riccati equation, for which numerical techniques are available using digital computers.

The system of Eq. (1) is called controllable if all of its modes may be independently controlled by using the input  $u(t)$ . If  $L$  is any square root of  $Q$ , so that  $Q = L^T L$ , then  $(A, L)$  is called observable if all of the system modes may be independently reconstructed by measuring the linear combination  $Lx$  of the state's components. This means basically that all system modes appear in the performance index.

A fundamental result of modern control theory pertains to the case in which the system is time-invariant,  $Q$  and  $R$  are constant, Eq. (1) is controllable, and  $(A, L)$  is observable, where  $L$  is any square root of  $Q$ . Then, there exists a positive definite limiting solution  $S_\infty$  to the Riccati equation that is also the unique positive definite solution to the algebraic Riccati equation. Moreover, the optimal linear-quadratic feedback gain  $K$  given by Eq. (9), with  $S_\infty$  used there, is constant and guarantees that the closed-loop system of Eq. (3) is stable.

This is an extremely important result, for it shows how to solve the difficult problem of stabilizing a system with multiple inputs and outputs. In fact, under the conditions of controllability and observability, the closed-loop system is stable for any choice of the design parameters  $Q$  and  $R$ .

*Minimum-time control.* To obtain different performance objectives, performance indices other than the quadratic performance index in Eq. (4) may be used. If it is desired to drive the system state from a given initial value  $x(0)$  to a specified value  $x(T)$  in minimum time, the performance index in Eq. (13)

$$J = \int_0^T dt = T \quad (13)$$

is suitable. Choosing the feedback gain  $K$  so as to minimize this performance index will result in the minimum time  $T$ .

A difficulty with minimum-time control of linear systems is that the solution to the problem of mov-

ing from one state to another as quickly as possible is to use infinite control energy. Thus, in order for the problem to have a solution, it is necessary to assume that the control inputs are bounded so that, for example,  $|u(t)| \leq 1$ . The solution to the minimum-time control problem is then found to be the bang-bang control strategy, so called because it involves using a control input that is always at either its maximum or minimum value. The maximum input is used first to accelerate the system quickly, with the minimum control value used to decelerate it so that it comes to rest at the desired final state.

*Observer and regulator design.* To implement the state feedback control law of Eq. (2), all  $n$  components of the state  $x(t)$  must be available. However, in practice it is too difficult or expensive to measure all the states. Instead, there is generally available a measured output of the form in Eq. (14) which has only  $p < n$

$$y = Cx \quad (14)$$

components. Therefore, if the measured  $y(t)$  is used, it is necessary to reconstruct the full state  $x(t)$  for use in Eq. (2). This is called the state observer or estimator problem.

One sort of observer is given in Eq. (15). It uses

$$\dot{\hat{x}} = A\hat{x} + Bu + L(y - C\hat{x}) \quad (15)$$

$u(t)$  and  $y(t)$  as inputs, from which the estimate  $\hat{x}(t)$  of the state  $x(t)$  is reconstructed. The structure of the system and observer is shown in Fig. 2. Since  $\hat{y} = C\hat{x}$  is the estimate of the output, the observer gain  $L$  multiplies the output error  $e_y = y - \hat{y}$ .

By defining the estimation error  $e(t)$  as in Eq. (16), Eq. (15) may be subtracted from Eq. (1) to obtain the error dynamics in Eq. (17). Therefore, to make

$$e = x - \hat{x} \quad (16)$$

$$\dot{e} = (A - LC)e \quad (17)$$

the error go to zero, it is necessary only to select the observer gain  $L$  so that  $(A - LC)$  is stable, which may be accomplished if  $(A, C)$  is observable. This is called the output injection problem and is dual to the state feedback problem, as may be seen by comparing Eq. (17) and Eq. (3).

The state-variable feedback law, Eq. (2), where  $K$  has been found by using optimal control theory, may be replaced by the control in Eq. (18), which uses

$$u = -K\hat{x} + v \quad (18)$$

the estimate of  $x(t)$  reconstructed from the available measurements  $y(t)$ . An external input  $v(t)$  is included so that the closed-loop system may be manipulated. The structure of the system is shown in Fig. 2, from which it is evident that the observer-plus-feedback arrangement is nothing but a dynamic compensator for the system. It is called a regulator.

It can be shown that the poles of the closed-loop system are just the poles of  $(A - BK)$ , which depend on the optimal control design, along with the poles of the observer  $(A - LC)$ . Thus, the feedback

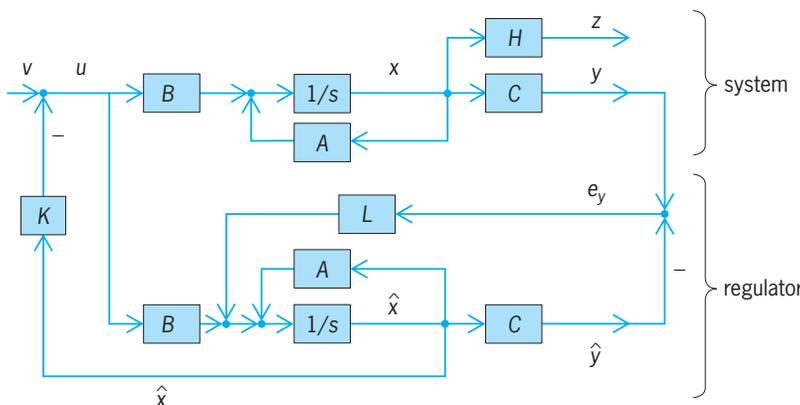


Fig. 2. Structure of state observer and feedback in the regulator configuration.

and the observer may be designed separately. This is the important separation principle of modern control theory.

Other design techniques result in reduced-order observers that have fewer than  $n$  states and so are easier to implement than the one described here. If there is noise in the system or measurements, then stochastic notions must be used in the observer design. The resulting estimator is called the Kalman filter. See ESTIMATION THEORY; STOCHASTIC CONTROL THEORY.

**Output feedback design.** An alternative to state-variable feedback design, which generally requires the use of an observer, is to perform an optimal control design directly in terms of the available outputs of Eq. (14). This can result in significant simplifications in the control law, since the additional dynamics of an observer do not need to be built.

If, instead of the state feedback law of Eq. (2), the output feedback control law in Eq. (19) is used

$$u = -Ky + v \quad (19)$$

[where  $v(t)$  is an external input], then the closed-loop system is given by Eq. (20). The output feed-

$$\dot{x} = (A - BKC)x + Bv \equiv A_c x + Bv \quad (20)$$

$$z = Hx$$

back linear quadratic regulator problem is to select the output feedback gain  $K$  so that the performance index of Eq. (4) is minimized.

In the infinite-horizon case, the gain  $K$  is constant. Here it can be shown that the optimal output feedback gain  $K$  may be found by solving Eqs. (21), where  $X$ , given by Eq. (22), is the expected initial

$$0 = A_c^T P + PA_c + C^T K^T R K C + Q \quad (21a)$$

$$0 = A_c S + SA_c^T + X \quad (21b)$$

$$K = R^{-1} B^T P S C^T (C S C^T)^{-1} \quad (21c)$$

$$X = E\{x(0)x^T(0)\} \quad (22)$$

mean-square state and  $S$  and  $P$  are matrix variables. Eqs. (21) are three coupled nonlinear matrix equations. These may be readily solved on a digital computer if the output of the system can be stabilized, that is, if there exists a gain  $K$  that makes  $A_c$  stable.

In output feedback design it is necessary to have some information about the initial state, namely  $X$ . This is not required in state feedback design. This means that the expected value of  $J$ , not  $J$  itself, must be minimized.

The optimal value of the expected value of the performance index using the optimal linear quadratic gain  $K$  is given in Eq. (23) [the trace of a matrix

$$J = \frac{1}{2} \text{trace}(PX) \quad (23)$$

is the sum of its diagonal elements]. As in the state feedback case, it may be computed off-line before the control input is ever applied to the system. Then, if

it is too high, new values for  $Q$ ,  $R$ , and  $S(T)$  may be selected and the design repeated.

**Tracking a reference input.** The function of the linear quadratic regulator is to hold the performance output near zero. If it is desired for  $z(t)$  to follow a nonzero reference input  $r(t)$ , then the control law must be modified. This is called the tracking or servo design problem.

It is possible to perform an optimal tracker design, where the tracking error  $e(t)$  defined in Eq. (24) is

$$e = r - z \quad (24)$$

weighted in the performance index instead of  $x(t)$ . That is,  $x^T Q x$  is replaced by  $e^T Q e$  and  $x^T(T)S(T)x(T)$  is replaced by  $e^T(T)S(T)e(T)$ . However, the result is a noncausal system that is difficult to implement.

Alternatively, the tracker problem may be solved by converting it first into a regulator problem. Suppose the system of Eq. (1) with the measured output of Eq. (14) has the applied control of Eq. (19). For perfect tracking, there must exist an ideal plant state  $x^*$  and an ideal plant input  $u^*$  such that the system of Eq. (1) has a performance output  $z(t)$  equal to  $r(t)$ . If this is not so, then tracking with zero error is impossible. If the state, control, and output deviations are defined as in Eqs. (25), then it may be shown that the deviation has dynamics given by Eqs. (26).

$$\begin{aligned} \tilde{x} &= x - x^* & \tilde{u} &= u - u^* \\ \tilde{y} &= y - y^* & \tilde{z} &= z - z^* \end{aligned} \quad (25)$$

$$\dot{\tilde{x}} = A\tilde{x} + B\tilde{u} \quad (26a)$$

$$\tilde{y} = C\tilde{x} \quad (26b)$$

$$\tilde{z} = H\tilde{x} = -e \quad (26c)$$

Therefore, to keep the tracking error  $e(t)$  small it is only necessary to design a regulator to regulate the state of the deviation system of Eqs. (26) to zero. This may be accomplished by finding the output feedback gain  $K$  that minimizes a quadratic performance index in terms of  $\tilde{x}(t)$  and  $\tilde{U}(t)$ .

If constant output feedback gains  $K$  have been obtained that are optimal with respect to Eqs. (26), then Eq. (27) holds, so that the required control for the plant is given by Eq. (28). That is, the resulting con-

$$\tilde{u} = -K\tilde{y} \quad (27)$$

$$u = u^* + \tilde{u} = -Ky + u^* + Ky^* \quad (28)$$

trol for the servo or tracker problem is the optimal regulator control  $-Ky$  plus some feedforward terms that are required to guarantee perfect tracking. These terms may be found by a transfer function analysis of the closed-loop system.

If  $r(t)$  is a unit step, then the ideal responses are the steady-state responses, since  $x^* = 0$ . Then, the feedforward terms  $u^* + Ky^*$  involve the inverse of the zero-frequency (direct-current) gain. That is, defining the closed-loop transfer function as in Eq. (29),

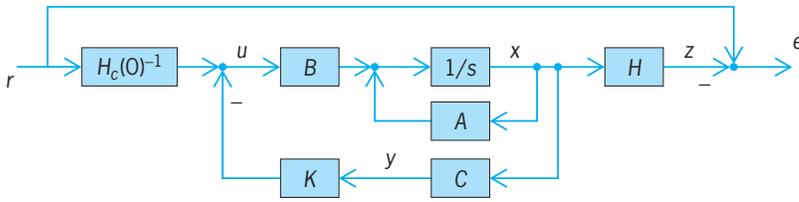


Fig. 3. Structure of tracker (servo) control system.

the servo control of Eq. (28) is given by Eq. (30).

$$H_c(s) = H[sI - (A - BKC)]^{-1}B \quad (29)$$

$$u = -Ky + H_c^{-1}(0)r \quad (30)$$

The structure of the servo control system is shown in Fig. 3. See SERVOMECHANISM.

**Robustness and frequency-domain techniques.** In addition to stability of the closed-loop system, an important property is guaranteed stability and performance in the presence of disturbance inputs, sensor noise, and variations in the plant parameters. This property is called robustness and is conveniently studied in the frequency domain.

Many of the classical frequency-domain design techniques may be extended to multivariable systems by using the notions of the singular values and principal phrases of a complex matrix. These quantities extend to nonscalar transfer functions the ideas of the transfer function magnitude and phase, respectively, and may be computed in a straightforward fashion by using modern computer software.

The loop gain of a multivariable closed-loop control system is the matrix  $K(s)G(s)$ , with  $G(s)$  the plant and  $K(s)$  the compensator, which may be designed by modern linear quadratic techniques. For good closed-loop disturbance rejection properties, the singular values of the loop gain should be large at low frequencies, for then the sensitivity is small. On the other hand, for good noise rejection, the singular values of the loop gain should be small at high frequencies. A representative plot of the maximum and minimum singular values of the loop gain (and  $\bar{\sigma}$  and  $\underline{\sigma}$ , respectively) is shown in Fig. 4.

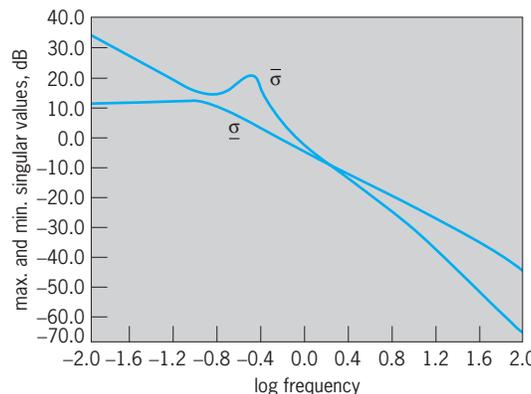


Fig. 4. Plot of maximum and minimum singular values of the loop gain,  $\bar{\sigma}$  and  $\underline{\sigma}$ , as functions of the logarithm of frequency (Bode magnitude plot), for a representative multivariable closed-loop control system.

If the regulator  $K(s)$  is a constant full state-variable feedback matrix  $K$  designed by using linear quadratic techniques, then the closed-loop system has important guaranteed robustness properties. In fact, it can then be shown that the return difference [that is,  $I + K(s)G(s)$ ] satisfies the constraint in inequality (31).

$$\underline{\sigma} [I + KG(j\omega)] \geq 1 \quad (31)$$

Thus, the linear quadratic regulator always results in decreased sensitivity.

The linear quadratic constraint of inequality (31) may be used to show that the linear quadratic regulator has an infinite gain margin and a phase margin of at least  $60^\circ$ . The conclusion is that optimality is a stronger condition than stability.

**Discrete-time systems.** In order to control the system of Eq. (1) by using modern microprocessors, it must first be converted to a discrete-time system by sampling. Optimal control theory for discrete-time systems is well developed, and is in many respects similar to the theory outlined here for continuous-time systems. See CONTROL SYSTEMS; MICROPROCESSOR; SAMPLED-DATA CONTROL SYSTEM.

Frank L. Lewis

Bibliography. B. Anderson and J. B. Moore, *Optimal Control: Linear Quadratic Methods*, 1989; M. J. Grimble and M. A. Johnson, *Optimal Control and Stochastic Estimation: Theory and Applications*, 2 vols., 1988; S. Holly and A. H. Hallett, *Optimal Control, Expectations and Uncertainty*, 1989; F. L. Lewis, *Applied Optimal Control and Estimation*, 1992; F. L. Lewis, *Optimal Control*, 2d ed., 1995; E. Mosca, *Optimal, Predictive, and Adaptive Control*, 1995; E. O. Roxin, *Modern Optimal Control*, 1989.

## Optimal control theory

An extension of the calculus of variations for dynamic systems with one independent variable, usually time, in which control (input) variables are determined to maximize (or minimize) some measure of the performance (output) of a system while satisfying specified constraints. The theory may be divided into two parts: optimal programming, where the control variables are determined as functions of time for a specified initial state of the system, and optimal feedback control, where the control variables are determined as functions of the current state of the system. See CALCULUS OF VARIATIONS.

Examples of optimal control problems are (1) determining paths of vehicles or robots between two points to minimize fuel or time, (2) determining time-varying feedforward/feedback logic to bring a vehicle, robot, or industrial process to a desired terminal state in a finite time using acceptable control magnitudes, and (3) determining constant-gain feedback control logic for vehicles or industrial processes to keep them near a desired operating point in the presence of disturbances with acceptable control magnitudes.

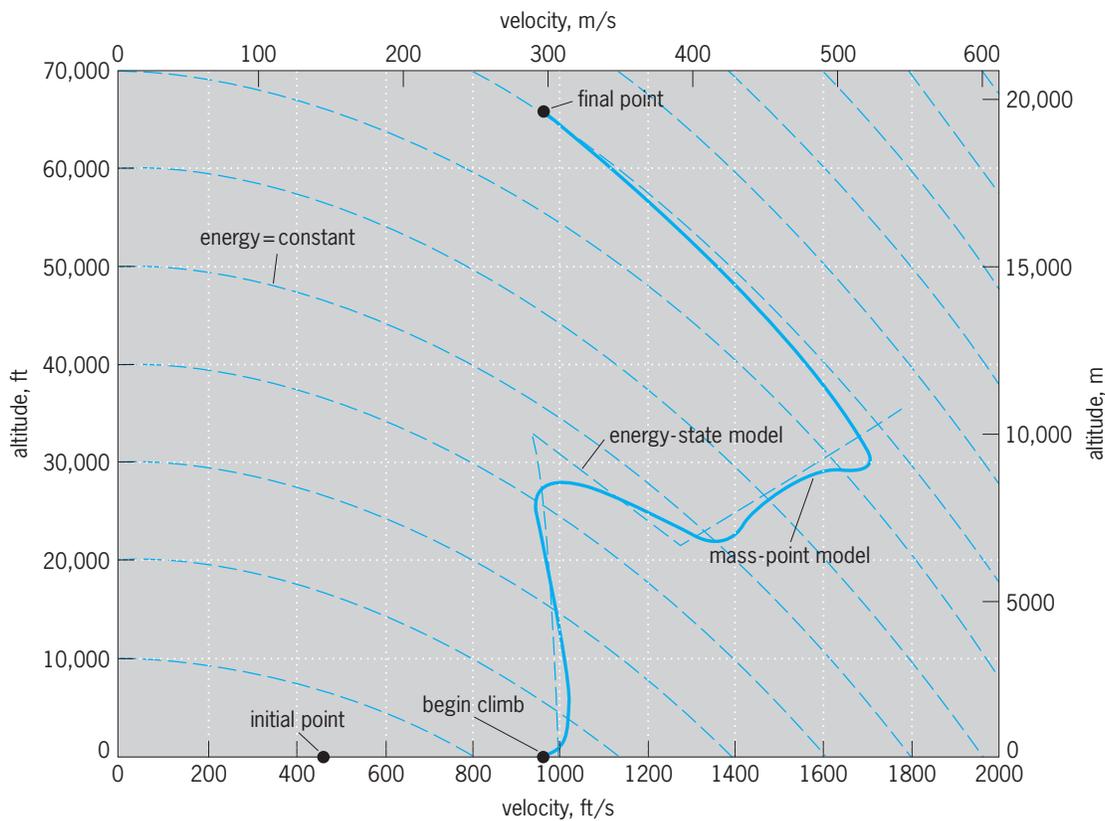


Fig. 1. Minimum-time flight path of a supersonic aircraft from takeoff to an altitude of 20,000 m (65,600 ft), velocity of Mach 1.0, and horizontal flight.

Dynamic systems may be divided into two categories: continuous dynamic systems, where the control and state variables are functions of a continuous independent variable, such as time or distance; and discrete dynamic systems, where the independent variable changes in discrete increments. Many discrete systems are discretized versions of continuous systems; the discretization is often made so that (1) the system can be analyzed or controlled by digital computers (or both), or (2) measurements of continuous outputs are made at discrete intervals of time (sampled-data systems) in order to share data transmission channels.

A large class of interesting optimal control problems can be described as follows: the dynamic system is described by a set of coupled first-order ordinary differential equations of the form of Eq. (1), where  $x$

$$\dot{x} = f(x, u, t) \quad (1)$$

(the state vector) represents the state variables of the system (such as position, velocity, temperature, voltage, and so on);  $u$  (the control vector) represents the control variables of the system (such as motor torque, control-surface deflection angle, valve opening, and so on);  $t$  is time; and  $f$  represents a set of functions of  $x$ ,  $u$ , and  $t$ . The performance index  $J$ , which one desires to minimize, is a scalar function of the final time  $t_f$  and the final state  $x(t_f)$  plus an integral from the initial time  $t_0$  to the final time  $t_f$  of a scalar function of the state and control vectors and

time, as given in Eq. (2). Possible constraints include

$$J = \varphi[x(t_f), t_f] + \int_{t_0}^{t_f} L[x(t), u(t), t] dt \quad (2)$$

(a) specified vector functions constraining the initial time  $t_0$  and the initial state  $x(t_0)$ , as in Eq. 3; (b) spec-

$$\alpha[x(t_0), t_0] = 0 \quad (3)$$

ified vector functions constraining the final time and final state, as in Eq. (4); (c) specified vector functions

$$\psi[x(t_f), t_f] = 0 \quad (4)$$

constraining the control variables [inequality (5)];

$$C[u(t), t] \leq 0 \quad (5)$$

(d) specified vector functions constraining both control and state variables [inequality (6)]; (e) specified

$$CS[x(t), u(t), t] \leq 0 \quad (6)$$

vector functions constraining only the state variables [inequality (7)]. Constraints  $a$  and  $b$  are equality con-

$$S[x(t), t] \leq 0 \quad (7)$$

straints, whereas constraints  $c$ ,  $d$ , and  $e$  are inequality constraints. See OPTIMIZATION.

**Necessary conditions.** The calculus of variations is the basic technique of optimal control theory. To

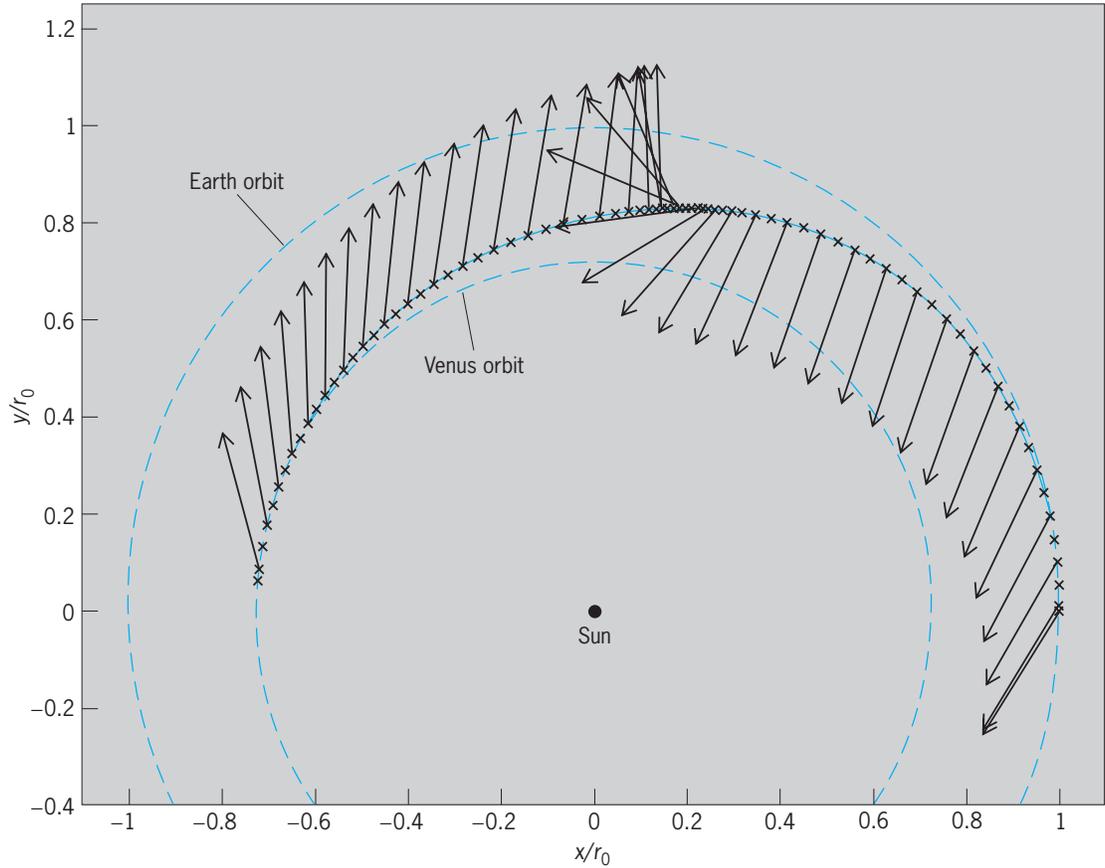


Fig. 2. Minimum-time transfer orbit of a spacecraft from Earth orbit to Venus orbit with low constant thrust. Arrows show direction of thrust.

illustrate the use of necessary and sufficient conditions for a minimum of  $J$ , consider the case where (1) the functions of  $f(x, u, t)$ ,  $\varphi(x, t)$ , and  $\psi(x, t)$  are continuous and have continuous first and second partial derivatives; (2)  $t_0$ ,  $t_f$ , and  $x(t)$  must satisfy Eqs. (8)–(11), where notation (12) is used ( $a \triangleq b$  indicates

$$\dot{x} = f(x, u, t) \quad x(t_0) \text{ specified} \quad (8)$$

$$0 = \frac{\partial H}{\partial u} \rightarrow u = u(x, \lambda, t) \quad (9)$$

$$\dot{\lambda}^T = -\frac{\partial H}{\partial x} \quad \lambda^T(t_f) = \frac{\partial \Phi}{\partial x(t_f)} \quad (10)$$

$$\psi[x(t_f)] = 0 \quad (11)$$

$$\begin{aligned} H &\triangleq L + \lambda^T f \\ \Phi &\triangleq \phi + v^T \psi \end{aligned} \quad (12)$$

that  $a$  is defined as equal to  $b$ );  $H$  is called the variational hamiltonian,  $\lambda^T(t)$  is a row vector of influence functions (also called Lagrange multiplier functions, adjoint functions, and co-state functions);  $v^T$  is a row vector of influence parameters; and the partial derivatives are evaluated on the optimal path. Equations (8)–(11) constitute a two-point boundary-value problem. Differential equations (8) and (10) are coupled, since  $x$  depends on  $u$  and, from Eq. (9),  $u$  depends on  $\lambda$ , whereas  $\lambda$  depends on  $x$  because

$\partial f/\partial x$  and  $\partial L/\partial x$  are functions of  $x$ . The parameters  $v$  must be chosen so that Eq. (11) is satisfied. Equations (9) and (10) are called the Euler-Lagrange equations [although some authors refer to the whole set, Eqs. (8)–(11), by this name]. A stationary solution is one that satisfies Eqs. (8)–(11). Such solutions may not exist for a given problem, and even if a solution is found, it may not be minimizing. A minimizing solution must also satisfy the condition of inequality (13)

$$\frac{\partial^2 H}{\partial u^2} \geq 0 \quad t_0 \leq t \leq t_f \quad (13)$$

[due to A. M. Legendre and A. Clebsch]. If a solution to Eqs. (8)–(11) satisfies the strong version of inequality (13), that is, inequality (14), then it is a locally minimizing solution.

$$\frac{\partial^2 H}{\partial u^2} \geq 0 \quad t_0 \leq t \leq t_f \quad (14)$$

Occasionally, more than one locally minimizing solution exists for a given problem. The locally minimizing solution that gives the smallest  $J$  is called the globally minimizing solution. Unfortunately, one cannot always be sure that all of the locally minimizing solutions have been found.

**Gradient and shooting algorithms.** Realistic optimal control problems must be solved numerically with a digital computer. Gradient algorithms start with a guess of the optimal control history  $u(t)$ , which determines a nominal state history  $x(t)$ . Backward

integration of  $\lambda(t)$  over this nominal path yields the gradient  $\partial H/\partial u$ , which determines small changes in  $u(t)$  that bring the path closer to feasibility and optimality. After 10–50 iterations, the gradient is close to zero over the whole path and the terminal conditions are met. One common shooting algorithm integrates the coupled Euler-Lagrange equations forward with several initial guesses of  $\lambda(t_0)$ , determining  $u(t)$  from Eq. (9) at each integration step. The correct values of  $\lambda(t_0)$  are then interpolated by zeroing the errors in the terminal conditions. Good initial guesses (for examples, from a gradient solution) are required for shooting algorithms to converge since the Euler-Lagrange equations are inherently unstable.

Examples of gradient solutions are shown in Figs. 1 and 2. Figure 1 shows the minimum-time flight path of a supersonic aircraft from takeoff to an altitude of 20 km (12 mi), arriving at Mach number 1.0 and horizontal flight. Full throttle is used all the way; the control is angle-of-attack, which can be changed very rapidly using the elevators. Two solutions are shown, one using a mass-point model with five states and another using a less precise energy-state model with only two states. Figure 2 shows the minimum-time flight path of a spacecraft from Earth orbit to the orbit of Venus. The control is the direction of the low constant thrust force which is shown by arrows in the figure. The spacecraft arrives with the Venus orbital velocity and tangent to the Venus orbit.

**Minimum principle.** The minimum principle is a generalization of Eq. (14), due to L. S. Pontryagin. It is a statement of sufficient conditions for a local minimum of  $J$  for problems with (or without) control variable constraints. In the previous development, Eq. (9) and inequality (13) are replaced by Eq. (15),

$$u = \arg \min_u H[x(t), \lambda(t), u(t), t] \quad (15)$$

where  $H$  is, again, the variational hamiltonian defined in Eq. (12), and the minimization is carried out over the entire admissible range of  $u$ , holding  $x$ ,  $\lambda$ , and  $t$  fixed.

Nonlinear programming algorithms may be used to solve optimal control problems using the minimum principle by parametrizing the control or output variables. They are especially useful for problems with control or state variable inequality constraints since several professional nonlinear programming codes are available. Figure 3 shows an example, the minimum-time erection of a pendulum from hanging straight down with no angular velocity to straight up with no angular velocity using bounded horizontal force (the control) on the supporting cart. The cart starts and ends at the same position with zero velocity. This is an example of “bang-bang control” since the specific force changes from 1  $g$  (the acceleration of gravity) to the left to 1  $g$  to the right five times during the erection. The figure is a stroboscopic movie of the motion; the square symbols indicate the changing position of the cart while the circle symbols indicate the position of the pendulum bob. See NONLINEAR CONTROL THEORY; NONLINEAR PROGRAMMING.

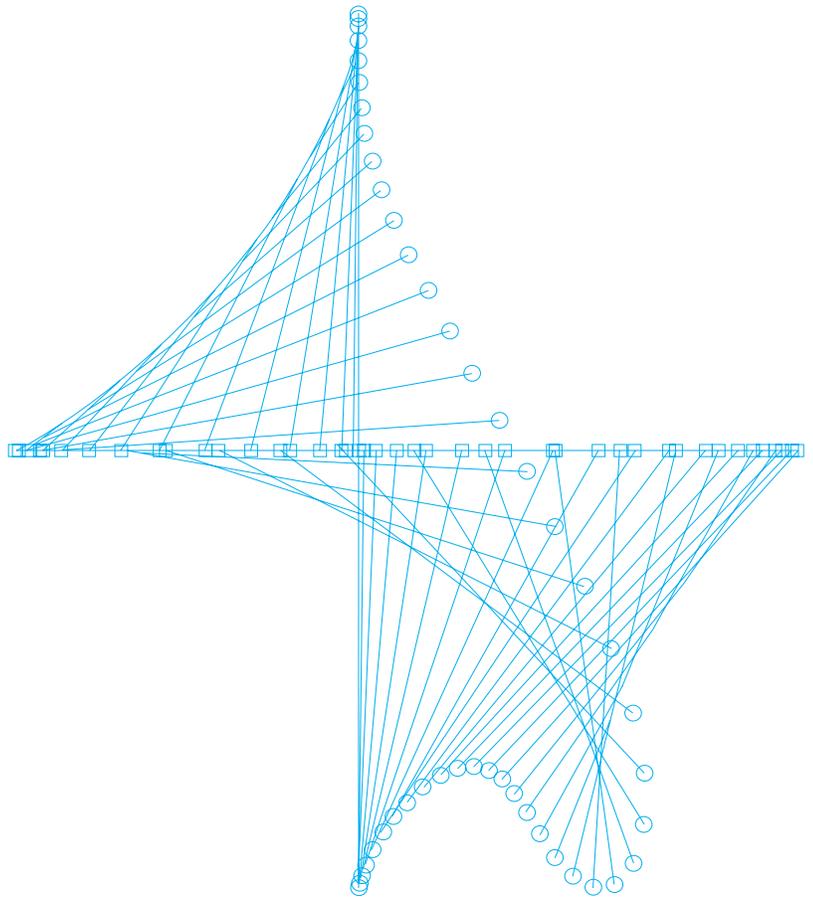


Fig. 3. Minimum-time erection of an inverted pendulum using bounded horizontal force on the supporting cart.

**Dynamic programming.** This is a technique for determining families of optimum solutions in feedback form, that is,  $u = u(x)$ . It is based on the principle of optimality, which may be stated as follows: if an optimal path is known from some point in the state-time space  $[x(t_0), t_0]$  to a terminal manifold  $\{x(t_f, t_f) = 0\}$ , then this path is also the optimal path to this terminal manifold from any intermediate point on the path. This simple and rather obvious principle can be used to derive a first-order partial differential equation for the optimal return function,  $J^{(0)}[x(t), t]$ , which is defined as the value of the performance index  $J$  in Eq. (2), starting from  $x(t)$  at time  $t$  and proceeding on an optimum path to the terminal manifold. The partial differential equation is Eq. (16), where Eqs. (17) are

$$\frac{\partial J^{(0)}}{\partial t} = -H^{(0)}\left(x, \frac{\partial J^{(0)}}{\partial x}, t\right) \quad (16)$$

$$H^{(0)}\left(x, \frac{\partial J^{(0)}}{\partial x}, t\right) \triangleq \min_u H\left[x, \frac{\partial J^{(0)}}{\partial x}, u, t\right] \quad (17)$$

$$H \triangleq L(x, u, t) + \frac{\partial J^{(0)}}{\partial x} f(x, u, t)$$

used and the solution must satisfy the terminal condition of Eq. (18). An important by-product of solving

$$J^{(0)}(x, t) = \varphi(x, t) \quad \text{or} \quad \psi(x, t) = 0 \quad (18)$$

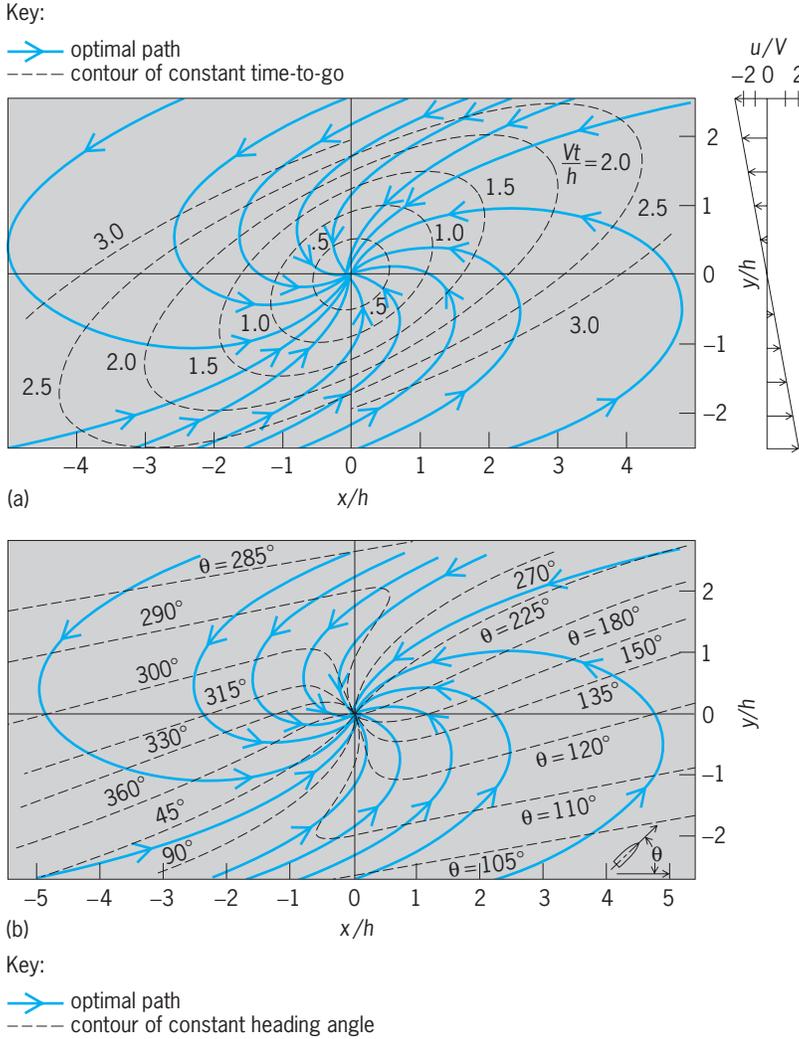


Fig. 4. Minimum-time ship routing with linear variation in current. (a) Contours of constant time-to-go and optimal paths. (b) Contours of constant heading angle  $\theta$  and optimal paths.

Eqs. (16) and (17) is the optimal control law (19),

$$u^{(0)}(x) \text{ or } u^{(0)}(x, t) \tag{19}$$

which gives the control variables as functions of the state variables (or state variables and time if  $f$  or  $L$  depend explicitly on time). Comparison on Eqs. (12) and (17) yields Eq. (20), that is, the Lagrange multiplier functions introduced in Eqs. (10) are influence functions on the value of the optimal return function for infinitesimal changes in  $x$ . Solutions to Eqs. (17) and (19) are constructed backward from the terminal manifold  $\psi(x, t) = 0$  and have a wavelike character; that is, contours of constant  $J^{(0)}$  are like wavefronts and optimal paths are like rays.

$$\frac{\partial J^{(0)}}{\partial x} \equiv \lambda^T \tag{20}$$

As an example, consider the problem of finding minimum-time paths to  $x = y = 0$  for a ship that travels with constant velocity  $V$  with respect to the water, but there is a current parallel to the  $x$  axis whose magnitude is  $-V(y/b)$ ; hence, at  $y = b$ , the velocity of the current is in the negative  $x$  direction

and equals the ship's velocity with respect to the water. The problem may be stated as follows: find  $\theta(x, y)$ , the heading of the ship as function of position, to go to  $x = y = 0$  in minimum time where Eqs. (21) and (22) are given. Optimal paths and

$$\dot{x} = V \cos \theta - \frac{V}{b}y \tag{21}$$

$$\dot{y} = V \sin \theta \tag{22}$$

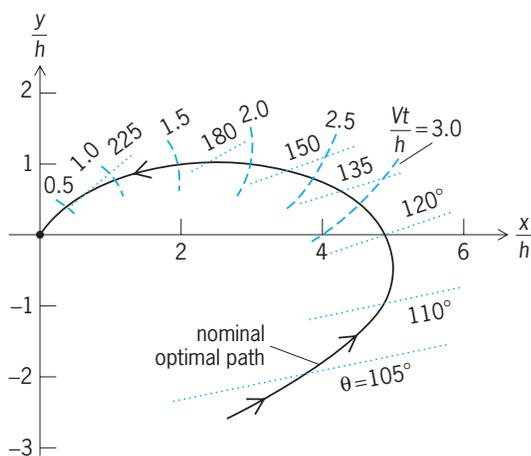
contours of constant time-to-go (measured in units of  $b/V$ ) are shown in Fig. 4a. Optimal paths and contours of constant heading angle are shown in Fig. 4b; this chart is the feedback control solution since it gives  $\theta(x, y)$ , that is, optimal heading as a function of position. For continuous dynamic systems with three or more state variables, solutions to Eqs. (17) and (19) require too much computation to be considered practical at present. If feedback solutions are desired for such systems, neighboring optimum feedback control (discussed below) is a more practical approach.

**Singular optimal control.** In some optimal control problems, extremal arcs ( $\partial H/\partial u = 0$ ) occur on which the matrix  $\partial^2 H/\partial u^2$  is only semidefinite; that is, it has one or more zero eigenvalues. Such arcs are called singular arcs. In the classical calculus of variations, an optimal path that contains one or more singular arcs is called a broken extremal since the control variables have discontinuities ("corners") at one or both ends of the singular arcs. Singular arcs may appear in optimal control problems involving cascaded dynamic systems if the direct effect of the control and the effect of a strongly controlled state have opposite signs. Cascaded linear systems of this type are called nonminimum phase (NMP) systems and are characterized by having right-half-plane transmission zeros in the transfer function matrix.

**NOFC.** Neighboring optimum feedback control (NOFC) about a nominal optimal path is a viable alternative to a "complete" optimal feedback control solution. NOFC constitutes an approximation to the optimal feedback control solution in a region neighboring the nominal optimal path; in the case of the ship routing problem illustrated in Fig. 4, an example of such region is shown in Fig. 5. NOFC is a combination of the nonlinear open-loop optimal path plus a linear-quadratic closed-loop solution in the neighborhood of this path. The control logic is of the form in Eq. (23), where  $u^{(N)}(t)$  and  $x^{(N)}(t)$  are

$$u = u^{(N)}(t) + C(t)[x(t) - x^{(N)}(t)] \tag{23}$$

the optimal control and state variable programs for the nominal initial and final conditions;  $x(t)$  is the current (perturbed) state and  $u(t)$  is the current control that should be used in order to proceed along a neighboring optimal path to the specified terminal manifold; and  $C(t)$  is the feedback gain matrix obtained by solving a slight generalization of the matrix Riccati equation, using weighting matrices that are second derivatives of the variational hamiltonian



Key:  
 - - - contours of constant time-to-go  
 ····· contours of constant heading angle

Fig. 5. Neighboring optimum feedback control about a nominal optimal path for the minimum-time ship-routing problem with linear variation in current.

[defined in Eq. (12)], evaluated on the nominally optimal path. See OPTIMAL CONTROL (LINEAR SYSTEMS).

If the final time  $t_f$  is determined implicitly by the point where the path intersects the terminal manifold [Eq. (4)], the NOFC law, Eq. (23), may be modified to use estimated time-to-go,  $t_f^{(e)} - t = \tau$ , as the index variable for  $u^{(N)}$ ,  $x^{(N)}$ , and  $C$ ; an additional set of feedback gains  $n$  must be determined such that Eq. (24) is satisfied. Equation (24) may have to be

$$t_f^{(e)} = t_f + n(\tau) [x(t) - x^{(N)}(\tau)] \quad (24)$$

used iteratively to determine a consistent  $\tau$ , that is, start with  $\tau_0 = t_f - t$  to find  $\tau_{f_0}^{(e)}$ , then use  $\tau_1 = t_{f_0}^{(e)} - t$ , then  $\tau = t_{f_1}^{(e)}$ , and so on, until  $\tau_{n+1} \cong \tau_n$ . See CONTROL SYSTEMS.

Arthur E. Bryson, Jr.

Bibliography. B. Anderson and J. B. Moore, *Optimal Control: Linear Quadratic Methods*, 1989; A. E. Bryson, *Dynamic Optimization*, 1998; D. N. Kirk, *Optimal Control Theory: An Introduction*, 1992; F. L. Lewis, *Optimal Control*, 2d ed., 1995; E. O. Roxin, *Modern Optimal Control*, 1989; R. F. Stengel, *Stochastic Optimal Control*, 1986.

## Optimization

The design and operation of systems or processes to make them as good as possible in some defined sense. The approaches to optimizing systems are varied and depend on the type of system involved, but the goal of all optimization procedures is to obtain the best results possible (again, in some defined sense) subject to restrictions or constraints that are imposed. While a system may be optimized by treating the system itself, by adjusting various parameters of the process in an effort to obtain better results, it generally is more economical to develop a model of the process and to analyze performance changes that result from adjustments in the model. In many

applications, the process to be optimized can be formulated as a mathematical model; with the advent of high-speed computers, very large and complex systems can be modeled, and optimization can yield substantially improved benefits. See DIGITAL COMPUTER.

Optimization is applied in virtually all areas of human endeavor, including engineering system design, optical system design, economics, power systems, water and land use, transportation systems, scheduling systems, resource allocation, personnel planning, portfolio selection, mining operations, blending of raw materials, structural design, and control systems. Optimizers or decision makers use optimization in the design of systems and processes, in the production of products, and in the operation of systems.

### Framework for Optimization

The framework for optimization includes forming a model, establishing and treating constraints, determining feasible solutions, and assigning performance measures.

**System models.** The first step in modern optimization is to obtain a mathematical description of the process or the system to be optimized. A mathematical model of the process or system is then formed on the basis of this description. Depending on the application, the model complexity can range from very simple to extremely complex. An example of a simple model is one that depends on only a single nonlinear algebraic function of one variable to be selected by the optimizer (the decision maker). Complex models may contain thousands of linear and nonlinear functions of many variables. As part of the procedure, the optimizer may select specific values for some of the variables, assign variables that are functions of time or other independent variables, satisfy constraints that are imposed on the variables, satisfy certain goals, and account for uncertainties or random aspects of the system.

System models used in optimization are classified in various ways, such as linear versus nonlinear, static versus dynamic, deterministic versus stochastic, or time-invariant versus time-varying. In forming a model for use with optimization, all of the important aspects of the problem should be included, so that they will be taken into account in the solution. The model can improve visualization of many interconnected aspects of the problem that cannot be grasped on the basis of the individual parts alone. A given system can have many different models that differ in detail and complexity. Certain models (for example, linear programming models) lend themselves to rapid and well-developed solution algorithms, whereas other models may not. When choosing between equally valid models, therefore, those that are cast in standard optimization forms are to be preferred. See MODEL THEORY.

**Constraints.** The model of a system must account for constraints that are imposed on the system. Constraints restrict the values that can be assumed by variables of a system. Constraints often are classified as being either equality or inequality constraints.

For example, Eq. (1) is a linear equality constraint,

$$x_1 + x_2 + x_3 = 50 \quad (1)$$

whereas inequality (2a) is a nonlinear inequality constraint and is equivalent to inequality (2b).

$$x_1^2 + x_2^2 + x_3^2 \leq 100 \quad (2a)$$

$$-x_1^2 - x_2^2 - x_3^2 \geq -100 \quad (2b)$$

Constraints (1) and (2) are algebraic constraints and depend on the three real variables  $x_1$ ,  $x_2$ , and  $x_3$ , which can be viewed as characterizing three-dimensional euclidean space. In the case of  $n$  real variables, the  $n$ -dimensional euclidean space  $R^n$  is of interest. Other optimization problems involve differential equation constraints, an example of which is Eq. (3).

$$\frac{dx_1}{dt} = -0.5x_1 + 3x_2 \quad (3)$$

See DIFFERENTIAL EQUATION.

The types of constraints involved in any given problem are determined by the physical nature of the problem and by the level of complexity used in forming the mathematical model.

Constraints that must be satisfied are called rigid constraints. Physical variables often are restricted to be nonnegative; for example, the amount of a given material used in a system is required to be greater than or equal to zero. Rigid constraints also may be imposed by government regulations or by customer-mandated requirements. Such constraints may be viewed as absolute goals.

In contrast to rigid constraints, soft constraints are those constraints that are negotiable to some degree. These constraints can be viewed as goals that are associated with target values. In place of Eq. (1), for example, a goal might be to make the left-hand side (a performance index perhaps) as close as possible to the right-hand side, the target value. The amount that the goal deviates from its target value could be considered in evaluating trade-offs between alternative solutions to the given problem.

**Feasible solutions.** When constraints have been established, it is important to determine if there are any solutions to the problem that simultaneously satisfy all of the constraints. Any such solution is called a feasible solution, or a feasible point in the case of algebraic problems. The set of all feasible points constitutes the feasible region. For example, if Eq. (1) and inequality (2) are the constraints that apply to a given problem, the feasible region is the intersection of two sets of points: the set of points on the plane where the  $x_i$ 's sum to 50 and the set of points within the sphere of radius 10 that is centered at the origin.

If no feasible solution exists for a given optimization problem, the decision maker may relax some of the soft constraints in an attempt to create one or more feasible solutions; a class of approaches to optimization under the general heading of goal programming may be employed to relax soft constraints

in a systematic way to minimize some measure of maximum deviations from goals.

**Performance measures.** A key step in the formulation of any optimization problem is the assignment of performance measures (also called performance indices, cost functions, return functions, criterion functions, and performance objectives) that are to be optimized. The success of any optimization result is critically dependent on the selection of meaningful performance measures. In many cases, the actual computational solution approach is secondary. Ways in which multiple performance measures can be incorporated in the optimization process are varied. Both single-criterion and multicriteria optimization methods are discussed below.

In some cases, all but one of the performance measures are assigned target values and are converted into soft constraints; the remaining performance measure, say  $f(x)$ , is then optimized (maximized or minimized, whichever is appropriate) subject to satisfying the constraints. The validity of Eq. (4) allows

$$\min[f(x)] = -\max[-f(x)] \quad (4)$$

an ordinary optimization problem to be cast as either a maximization or minimization problem. The resulting optimization problem is called a single-objective or single-criterion optimization problem, in contrast to multicriteria optimization problems.

The way in which a system is modeled often dictates approaches to obtaining optimal solutions. Unconstrained optimization, constrained optimization with a single performance measure, constrained optimization with multiple performance measures, and optimization of systems that evolve over time (dynamic optimization) each require fundamentally different approaches that will be discussed in turn.

### Unconstrained Optimization

A special class of system models are those in which there are no constraints. The independent variable is the column vector  $x = [x_1, x_2, \dots, x_n]^T$  of  $n$  variables in  $n$ -dimensional euclidean space  $R^n$  of real values; the superscript  $T$  signifies the transpose of a row vector to form the column vector  $x$ . Given a real scalar function  $f(x)$  of the  $n$  variables, points at which  $f(x)$  assumes local maximum and local minimum values are of interest.

**Points of zero slope.** If  $f(x)$  and its first derivatives are continuous, then the maximum and minimum of  $f$  (if they exist) must occur at points where the slope of the function is zero, that is, at points where the gradient of  $f$ ,  $\nabla f$ , satisfies Eq. (5). **Figure 1** illus-

$$\nabla f(x) \triangleq \left[ \frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \dots, \frac{\partial f}{\partial x_n} \right]^T = 0 \quad (5)$$

trates local maxima and local minima of a function of two variables. It contains a local maximum of 7 in addition to the global maximum of 9, and both points in the  $x_1, x_2$  plane satisfy the necessary condition of Eq. (5), as does the minimum point (where the

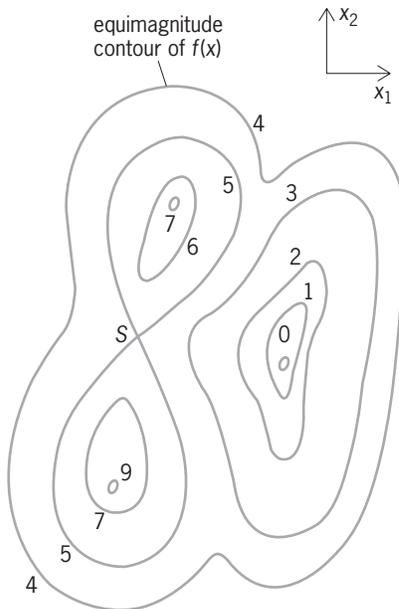


Fig. 1. Equipmagnitude contours of a function of two variables.

function equals 0). In addition, the point marked  $S$ , which is known as a saddle point, also satisfies Eq. (5). See CALCULUS OF VECTORS.

**Second-order conditions.** To determine if a point  $x^*$  that satisfies Eq. (5) is a local maximum point, the determinant of the  $n \times n$  matrix of second derivatives of  $f(x)$  [the hessian matrix of  $f(x)$ ] may be tested at  $x^*$ : if this hessian is negative definite,  $x^*$  is a local maximum point; if the hessian is negative semi-definite,  $x^*$  may be a local maximum point; otherwise  $x^*$  is not a local maximum point. The test for a local minimum point is similar, with “negative” in the preceding statement being replaced by “positive.”

**Global versus local.** To determine if a local maximum (minimum) is the global maximum (minimum), an exhaustive search over all local maxima (minima) may be conducted. If the function  $f(x)$  is known to be concave (convex), however, there is only one local maximum (minimum), and therefore also global maximum (minimum), assuming the function is bounded.

**Numerical solution methods.** A multitude of numerical methods called search techniques have been developed to determine maxima and minima of functions. The most appropriate method to use depends on the nature of  $f(x)$ . If  $f(x)$  is evaluated experimentally, a direct sequential search approach that does not require derivative information can be used. If  $f(x)$  is discontinuous, a factorial search or a random search (Monte Carlo method) may be useful. If  $f(x)$  and its derivatives are continuous, however, a method from the class of methods that exhibit quadratic convergence may be best. A maximization technique exhibits quadratic convergence if it can locate the exact maximum of a quadratic function (assuming it has a well-defined maximum) in a finite number of calculations. See MONTE CARLO METHOD; NUMERICAL ANALYSIS.

**Gradient-based methods.** The gradient vector  $\nabla f(x)$  is central to those search techniques that use derivative information. In Fig. 1, for example, at each point in the figure, the gradient vector is a vector that is perpendicular to the equipmagnitude contour and points in the direction of greatest incremental increase in  $f(x)$ . One simple search strategy is as follows: (1) evaluate a search direction  $r = \nabla f(x^0)$  at the current best  $x = x^0$ ; (2) let  $x = x^0 + vr$ , where  $v$  is a real scalar; (3) search for a maximum of  $f$  with respect to values of  $v \geq 0$ ; and (4) reassign the current best  $x$  and return to step 1. It can be shown, however, that this simple gradient search strategy is inefficient when the equipmagnitude contours are elongated or irregular in any way, and better gradient-based techniques are available.

### Constrained Optimization with Single Performance Measure

Optimization problems may be solved by a variety of techniques that come under the general heading of mathematical programming. The term programming in this regard originated in the sense of scheduling (programming) a solution to the problem and was introduced prior to the widespread availability of computers. A variety of problem types and solution methods are associated with nonlinear programming, linear programming, integer programming, quadratic programming, dynamic programming, goal programming, parametric programming, and so forth.

**Nonlinear programming.** A common optimization problem is the nonlinear programming problem given by expression (6), where the function  $f(x)$  is

$$\text{Maximize } f(x) \quad x \in S \quad (6)$$

a real scalar-valued function of  $n$  real variables contained in the vector  $x$ , and where  $x$  is constrained to be an element of the set  $S$  defined both by equality constraints given by Eq. (7) and by inequality constraints given by inequality (8). The constraint func-

$$p_i(x) = a_i \quad i = 1, 2, \dots, n_a < n \quad (7)$$

$$q_j(x) \leq b_j \quad j = 1, 2, \dots, n_b \quad (8)$$

tions  $p_i$ 's and  $q_j$ 's are real scalar-valued functions of  $x$ ; and the  $a_i$ 's and  $b_j$ 's are real constants.

Related minimization problems can be cast in the above form by use of Eq. (4). If the constraints of the problem define a convex set of points, and the performance measure is concave (convex), it can be shown that the only local maximum (minimum) is the global maximum (minimum) if it exists.

**Karush-Kuhn-Tucker conditions.** If the functions in expressions (6), (7), and (8) and the derivatives through second order of the functions are continuous, necessary conditions and sufficient conditions for constrained local maxima are of interest. Such conditions were used by J. L. Lagrange in his formulation of Lagrange multipliers for cases where all of the constraints are equality constraints. W. Karush first presented conditions for testing inequality constraints

with Lagrange multipliers in 1939, an approach that was extended and popularized by H. W. Kuhn and A. W. Tucker in 1951.

The Karush-Kuhn-Tucker conditions depend on the lagrangian function  $L$  given by Eq. (9), where

$$L(x, \alpha, \beta) \triangleq f + \sum_{i=1}^{n_a} (a_i - p_i)\alpha_i + \sum_{j=1}^{n_b} (b_j - q_j)\beta_j \quad (9)$$

$\alpha$  and  $\beta$  are column vectors of Lagrange multipliers, the  $\alpha_i$ 's and  $\beta_j$ 's, and where the other terms in Eq. (9) are defined by expressions (6), (7), and (8).

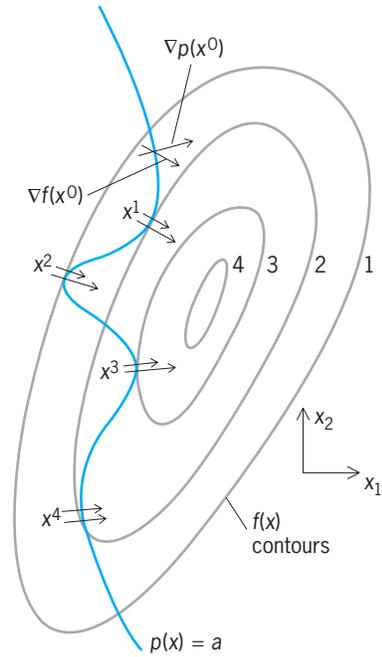
Technical details concerning Karush-Kuhn-Tucker conditions are simplified if the gradients of the active constraints are linearly independent at points of constrained local maxima. A given inequality constraint is said to be active at a point  $x$  if it is satisfied in an equality sense at  $x$ . Equality constraints necessarily are active constraints at points of constrained local maxima. A point at which the gradient  $\nabla f$  of  $f$  is nonzero and at which the gradients of the active constraints are linearly independent is called a regular point.

First-order necessary conditions for a regular point  $x^*$  to be a point of constrained local maximum are as follows: (1) all constraints of Eq. (7) and inequality (8) are satisfied at  $x^*$ ; (2)  $\nabla L(x^*, \alpha^*, \beta^*) = 0$  for some real vectors  $\alpha^*$  and  $\beta^*$ ; (3) all  $\beta_j^*$ 's are greater than or equal to zero; and (4) if the  $j$ th inequality constraint is not active at  $x^*$ , then  $\beta_j^* = 0$ . The  $\nabla L$  in condition (2) is the gradient of  $L$  with respect to  $x$ . In effect, these first-order conditions state that the gradient of the performance measure  $f$  is a linear combination of the gradients of the active constraints at  $x^*$ , with the inequality constraint proportionality factors, the  $\beta_j$ 's, necessarily being nonnegative at a constrained local maximum.

**Figure 2** depicts the case of one equality constraint and two variables  $x_1$  and  $x_2$ . At a typical nonoptimal point  $x^0$  that satisfies the constraint, no multiplier  $\alpha$  exists for which  $\nabla f = \alpha \nabla p$ . At points  $x^1$ ,  $x^2$ ,  $x^3$ , and  $x^4$ , however,  $\nabla f$  and  $\nabla p$  are collinear, and real  $\alpha$ 's exist for which  $\nabla f = \alpha \nabla p$ . The point  $x^1$  is a constrained local maximum;  $x^2$  is a constrained local minimum;  $x^3$  is the constrained global maximum; and  $x^4$  is a saddle point, which satisfies the first-order necessary conditions but which is neither a constrained maximum nor a constrained minimum.

**Figure 3** depicts the case of two inequality constraints and two variables  $x_1$  and  $x_2$ . In the interior of the shaded region, both inequality constraints are strictly satisfied (in which case the constraints are not active). On the common boundary point shown, however, both constraints are active,  $\nabla f$  is a positive linear combination of  $\nabla q_1$  and  $\nabla q_2$ , and the point is a constrained maximum point.

*Sensitivity results and higher-order conditions.* The Lagrange multipliers associated with a given local maximum can be shown to provide the following sensitivity information. If the right-hand sides of Eq. (7) and inequality (8) can be changed by small amounts  $\Delta a_i$

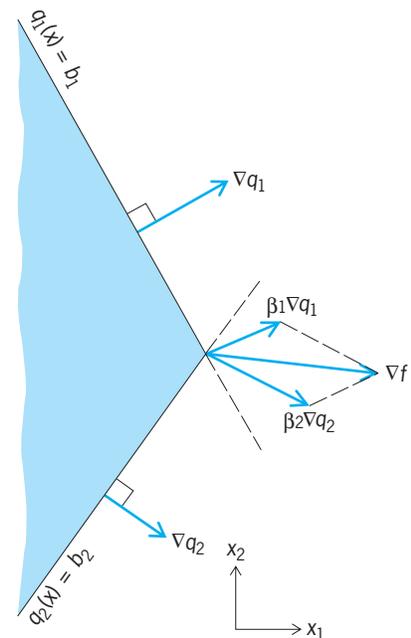


**Fig. 2. Conditions for local optima with an equality constraint.**

and  $\Delta b_j$ , respectively, then the constrained local maximum will change by an amount  $\Delta f$ , given by expression (10). Depending on the application, the

$$\Delta f \approx \sum_{i=1}^{n_a} \alpha_i^* \Delta a_i + \sum_{j=1}^{n_b} \beta_j^* \Delta b_j \quad (10)$$

Lagrange multipliers (the dual variables) in expression (10) might represent shadow prices or attribute costs. The costs associated with changes in  $a$ 's and  $b$ 's can be weighed against the improvements obtainable in  $f$ . The approximation (10) is based on



**Fig. 3. Conditions for local maximum with inequality constraints.**

first-order terms of Taylor series expansions and should not be used if the  $\Delta a_i$ 's and  $\Delta b_j$ 's are large enough to cause a change in constraint basis, that is, a case in which one or more active (inactive) inequality constraints become inactive (active).

Reasoning based on Taylor series expansions of problem functions can be used to obtain many other sensitivity results and to obtain second-order necessary conditions for local maxima, as well as second-order sufficient conditions.

**Solution methods.** Many numerical solution methods have been developed for nonlinear programming problems. Approaches to solution include the following: (1) gradient projection methods, in which the gradient of the performance measure is projected onto the constraint boundary as the search progresses; (2) penalty function methods, wherein a penalized performance measure is formed to equal the original performance measure when the constraints are satisfied, but to grow rapidly in magnitude as constraints are violated; and (3) augmented lagrangian methods, which modify the lagrangian of the problem in ways that ensure that the augmented lagrangian has unconstrained local maxima with respect to  $x$  at points of constrained local maxima.

When derivatives of problem functions are discontinuous, Monte Carlo or random search methods are popular approaches to solution. When integer or discrete value restrictions are imposed, various combinatorial solution methods may be appropriate. See COMBINATORIAL THEORY; NONLINEAR PROGRAMMING.

**Linear programming.** When all of the functions in expressions (6), (7), and (8) are linear, the problem reduces to a linear programming problem. Linear programming problems have a well-defined mathematical structure: the constraint set  $S$  is known to be convex; boundary surfaces of  $S$  are known to be planar and to intersect so as to define extreme points (vertices of simplexes); the global maximum of  $f$  (if it exists) is known to occur at a vertex of  $S$  (but the global maximum may also occur at neighboring boundary points); and the global maximum is the only maximum since no local maxima occur at other vertices. Every linear programming problem has a well-defined dual problem, the solution of which is related to the solution of the original problem and which provides sensitivity information regarding the original problem. See LINEAR PROGRAMMING.

### Multiobjective Optimization

When more than one performance measure is of interest, approaches to solution inevitably involve trade-offs. In this problem, there is a set  $\{f_1(x), f_2(x), \dots, f_m(x)\}$  of performance measures, each of which is in a form where maximization is desired with respect to the  $n$ -vector  $x$  of reals. Constraints of the form of Eq. (7) and inequality (8) may apply to the problem and define the feasible set  $S$  of allowed values of  $x$ .

**Efficient solutions.** A particular solution is said to dominate another solution if it achieves a higher value for one or more performance measures and a

lesser value for none. If a given feasible point  $x^*$  is not dominated by any other feasible point,  $x^*$  is said to be an efficient solution or a nondominated solution (also called a Pareto optimal solution, after V. Pareto, who developed the concept in 1906). An efficient solution, therefore, is a feasible solution from which an increase in the value of any single performance measure is obtainable only if there is a decrease in value of another performance measure. The set of all nondominated feasible solutions forms the efficient solution set. It is desirable to characterize the efficient solution set in order to apply trade-offs to select one or more members of the set as being most desirable. For linear problems, techniques are available for finding the efficient solution set. For large problems, however, the efficient solutions often consist of a continuum of values that are too numerous to be displayed effectively. Members of the efficient solution set can be obtained by multiparametric programming (discussed below).

**Approaches to solution.** The approach to solution for a given multiobjective optimization problem depends on the underlying model structure and on the preferences of the problem solver. A given approach may combine elements of the following: (1) treating some of the performance measures as soft equality constraints and using the sensitivity results of expression (10) to determine trade-offs in performance; (2) finding efficient solutions (Pareto optimal solutions); (3) using weighting factors and parametric programming; (4) assigning a goal for each performance measure and striving to minimize deviations from the goals; and (5) assigning rankings or priority levels, a prioritized vector of objective functions, and looking for a lexicographical optimum.

**Performance measure weighting.** One approach to multiobjective maximization is to assign a positive weighting factor  $w_i$  to each performance measure and to maximize the sum of the  $w_i f_i$ 's. Thus,  $f(x, w)$ , the single weighted measure to be maximized, is given by Eq. (11). If desired, the  $w_i$ 's can be normalized

$$f(x, w) = \sum_{i=1}^m w_i f_i(x) \quad w_i \geq 0 \quad (11)$$

by requiring that they sum to unity. If just one set of  $w_i$ 's is selected, it is difficult to tell in advance if they will result in a solution deemed as "good." An interactive approach to the problem often is used whereby the solution corresponding to one set of weights guides the selection of alternative weights for alternative solutions.

**Multiparametric programming.** Solutions corresponding to many different choices of the weighting factors may be of interest. The set of optimal solutions that correspond to all allowed combinations of  $w_i$ 's may be denoted  $V$ . Thus, feasible point  $x^*$  is a member of  $V$  if and only if there exists some  $w^*$  vector of  $w_i \geq 0$  for which  $f(x^*, w^*) \geq f(x, w^*)$  for all  $x$  contained in the feasible set  $S$ . The process of obtaining this set  $V$  is called multiparametric programming. If the  $f_i$ 's are concave and the feasible region  $S$  is convex, then  $V$  and the efficient solution set can be shown to be

identical. Also, the set of  $w$ 's can be divided into subsets, each of which is associated with a nondominated solution (or with a subset of nondominated solutions).

*Sensitivity and weighting factors.* If  $f$  is the weighted performance measure of Eq. (11), and if  $f^* = f(x^*, w^*)$  is a constrained local maximum of  $f$  with respect to  $x$ , for some fixed  $w^*$ , then incremental changes in the weighting factors can be made, giving  $w_i = w_i^* + \Delta w_i$ ,  $i = 1, 2, \dots, m$ . An additional term that should then be added to the right-hand side of expression (10) is given by expression (12), where

$$\sum_{i=1}^m [f_i^* + (\nabla f_i^*)^T \Delta x] \Delta w_i \quad (12)$$

$(\nabla f_i^*)^T$  is the transpose of the gradient of  $f_i$  with respect to  $x$  evaluated at  $x^*$ , and  $\Delta x$  is the change in the optimal solution point. As with the previous result (10), expression (12) should not be used if the incremental changes are large enough to cause a change in constraint basis. If the  $\Delta x_j \Delta w_i$  terms in expression (12) are sufficiently small in magnitude, the change in the weighted performance measure can be estimated prior to actually evaluating the new optimum point  $x^* + \Delta x$ .

*Goal programming.* The point of view taken in goal programming, introduced in 1961 by A. Charnes and W. W. Cooper, is that a decision maker, rather than trying to maximize the  $f_i(x)$ 's, should assign aspiration levels for the performance measures, thereby creating goals. Similarly, aspiration levels may be assigned for soft constraints. Interest is centered on cases where no solution exists that satisfies all constraints and achieves all goals; the objective then is to find solutions that come as close as possible to satisfying the soft constraints and to achieving the goals, while satisfying all rigid constraints. Different approaches to goal programming use different measures of how close the goals and the soft constraints are satisfied. Different goal programming methods may also include ways to account for prioritized goals.

In forming models for use with goal programming, information regarding the goals must be incorporated. Examples of such information include the priority structure of the goals, establishing which goals are on the same priority level and leading to ranked sets of goals; the minimum performance levels that would be acceptable for the performance measures; the penalties for not achieving goals (as a function of deviations from the goals); and the rates or the values to be assigned to overachievement and to underachievement. Some goals may have desired levels, above which no additional achievement is desired. The minimum acceptable levels for performance measures can be incorporated with rigid inequality constraints.

*Lexicographic screening.* In some applications, it is appropriate to assign goals and constraints to preemptive priority levels: goals and constraints at higher preemptive priority levels will always take precedence over goals and constraints at lower priority levels, regardless of any scalar weights assigned to the

lower-level goals. Rigid constraints can be viewed as being of top priority. Solution candidates for lower-priority levels are restricted to solutions that satisfy rigid constraints and that achieve higher-priority goals within specified tolerances. This approach to optimization is called lexicographical screening or preemptive goal programming.

### Dynamic Optimization

Many dynamic systems are characterized by differential equations or by difference equations. A typical example is a dynamic system that is modeled by a set of coupled first-order differential equations: the independent variable in the equations is time; the variables associated with the derivative terms are called state variables; and other variables in the equations may include control variables to be selected in an optimal manner over time. A typical performance measure to be minimized for such a system is the integral with respect to time of weighted values of errors squared and of control actions squared. One approach to the optimization of these systems is to use discrete approximations to produce algebraic models, in which case the methods of optimization described in previous sections apply. However, many useful concepts and solution approaches that are specific to the optimization of dynamic systems have been developed.

**Calculus of variations.** The classical calculus of variations is concerned with the optimization of performance measures that are functions of functions. For example, a performance measure that is the integral over time of the square of an error maps a function of time (the error squared) into a scalar value. Such performance measures are called functionals. The conditions for optimality associated with functionals involve differential equations. Conditions for optimality and solution methods for these problems have been developed. One of the modern applications of the calculus of variations deals with the optimal control of systems. See CALCULUS OF VARIATIONS; OPTIMAL CONTROL (LINEAR SYSTEMS).

In the mid 1950s, work by L. S. Pontryagin and others extended the classical calculus of variations to apply to cases where some of the variables to be selected could be restricted to closed and bounded sets and could exhibit step discontinuities as a function of time. This work has had major applications in the design of optimal nonlinear feedback controllers. See NONLINEAR CONTROL THEORY.

**Dynamic programming.** R. E. Bellman developed the framework for dynamic programming in the 1950s. The cornerstone of dynamic programming is the following principle of optimality: an optimal policy is one having the property that, regardless of the initial state of a system and the initial decision (control) applied, the remaining decisions must constitute an optimal policy with respect to the state resulting from the initial decision. This principle is especially helpful in solving problems that can be separated (separable programming) or decoupled in some fashion. Certain complicated problems can be embedded in a sequence of less complicated but

coupled problems. The solution of the original problem is obtained by solving a sequence of problems. See OPTIMAL CONTROL THEORY.

### Other Optimization Topics

Optimization includes a variety of other topics, such as stochastic optimization, multiple decision makers, combinatorial optimization, computer codes, and suboptimal solutions.

**Stochastic optimization.** Many systems contain random events and are modeled with random or stochastic variables. When such systems are optimized, performance measures and constraints may have to be defined in terms of expected values, variances, or other random variable measures. When the systems are modeled by algebraic stochastic relationships, stochastic programming may be useful for obtaining optimal solutions. For dynamic systems, stochastic optimal control and estimation algorithms may be appropriate. See ESTIMATION THEORY; STOCHASTIC CONTROL THEORY.

**Multiple decision makers.** In many applications, system performance is affected by more than one decision maker. The decision makers may have different performance objectives for the system. If goals conflict, or if the direct actions of any one decision maker are not known ahead of time by other decision makers, the resulting optimization problem is classified under the general heading of game theory. There are various classes of game theory problems, ranging from completely antagonistic games to cooperative games. In the latter case, the decision makers have the same goals for the system, but they may not have complete knowledge of the decisions being made by others.

A special case of game theory is the case where there are two decision makers, with one of them being "nature," and it is assumed that nature does everything in its power to make things worse for the ordinary decision maker; in such a case, a maximum strategy of "maximizing the worst case" may be appropriate. This also is called worst-case design. See DECISION THEORY; GAME THEORY.

**Combinatorial optimization.** Many optimization problems are associated with variables that are discrete in nature. For example, in some systems the absence of a certain entity would be represented by 0 and the presence of the entity by 1. Optimization involving such "no/yes" variables is called zero-one programming. Other system models may require certain variables to be integer values (integer programming) or to be selected from a specified set of discrete values. Discrete optimization problems include shortest-path problems that have applications in designing large-scale integrated electronic circuits and in scheduling routes for traveling sales persons. All such problems come under the general classification of combinatorial optimization. In general, solution methods for combinatorial optimization algorithms differ significantly from those discussed above.

**Computer codes for solving optimization problems.** Computer codes for solving optimization problems

can be obtained from many sources, including computer and software vendors, relevant journals and books, and government agencies. Independent evaluations of nonlinear programming codes are available. Relevant available material includes summaries of multiobjective linear programming packages, summaries of goal-programming packages, and codes for unconstrained search methods, the linear programming simplex method, and a combinatorial minimization method.

The continuing development of improved computer systems has a direct bearing on the development of numerical algorithms for optimization. Algorithms that incorporate parallel structures can take advantage of computer systems that have multiple processors executing in parallel. From the user's point of view, optimization codes should be user-friendly, should allow interactive participation, and should employ expert system structures to help the user both in obtaining and in interpreting results. Interactive solution approaches enable decision makers to play "what if" games and to determine sensitivity attributes of candidate solutions. See ALGORITHM; CONCURRENT PROCESSING; EXPERT SYSTEMS.

**Suboptimal solutions.** When dealing with an extremely large or complex system, global optimization may not be practical. In such a case, suboptimal solution may be achieved by reducing the system into subsystems, and by obtaining optimal results for each subsystem with some of the coupling terms ignored. Suboptimal solutions also may be acceptable, or even desirable, in cases where they provide answers with little complexity or where the sensitivity of the solution to uncontrolled parameter changes is reduced.

Donald A. Pierre

**Bibliography.** E. M. Beale, *Introduction to Optimization*, 1988; J. D. Dennis, Jr., and R. B. Schnabel, *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, 1983; A. V. Fiacco, *Introduction to Sensitivity and Stability Analysis in Nonlinear Programming*, 1983; S. French et al. (eds.), *Multi-Objective Decision Making*, 1983; P. E. Gill, W. Murry, and M. H. Wright, *Practical Optimization*, 1981; J. P. Ignizio, *Linear Programming in Single- and Multiple-Objective Systems*, 1982; M. O'Heigeartaigh, J. D. Lenstra, and A. H. G. Rinnooy Kan (eds.), *Combinatorial Optimization: Annotated Bibliographies*, 1985; D. A. Pierre, *Optimization Theory with Applications*, 1986; W. H. Press et al., *Numerical Recipes: The Art of Scientific Computing*, 2d ed., 1993.

## Oral glands

Glands located in the mouth that secrete fluids that moisten and lubricate the mouth and food and may initiate digestive activity; some perform other specialized functions. Fishes and aquatic amphibians have only solitary mucus (slime) secreting cells, distributed in the epithelium of the mouth cavity. Multicellular glands first appeared in land animals to keep the mouth moist and make food easier to swallow.

These glands occur in definite regions and bear distinctive names. Some glands of terrestrial amphibians have a lubricative secretion; others serve to make the tongue sticky for use in catching insects. Some frogs secrete a watery serous fluid that contains ptyalin, a digestive enzyme. The oral glands of reptiles are much the same, but are more distinctly grouped. In poisonous snakes and the single poisonous lizard, the Gila monster, certain oral glands of the serous type are modified to produce venom. Also many of the lizards have glands that are mixed in character, containing both mucous and serous cells. Oral glands are poorly developed in crocodilians and sea turtles. Birds bolt their food, yet grain-eaters have numerous glands, some of which secrete ptyalin.

**Mammals.** All mammals, except aquatic forms, are well supplied with oral glands. There are numerous small glands, such as the labial glands of the lips, buccal glands of the cheeks, lingual glands of the tongue, and palatine glands of the palate. Besides these, there are larger paired sets in mammals that are quite constant from species to species and are commonly designated as salivary glands (see **illus.**). The parotid gland, near each ear, discharges in front of the upper jaw. The submandibular (formerly submaxillary) gland lies along the posterior part of the lower jaw; its duct opens well forward under the tongue. The sublingual gland lies in the floor of the mouth, near the midplane. It is really a group of glands, each with its duct. Although not present in humans, the retrolingual gland, situated near the submandibular, is found in many mammals; its duct takes a course similar to that of the submandibular. Other occasional types are the molar gland of the hoofed mammals and the orbital gland of the dog family.

**Development.** All oral glands develop from the epithelial lining as branching buds. Each gland is organized somewhat after the pattern of a bush that bears berries on the ends of its twigs. The main stem and all branches of the bush correspond to a system of branching glandular ducts of various sizes; the termi-

nal berries correspond to the secretory end pieces. Actually these end pieces are more or less elongate, like the catkins of the willow or birch. The ducts are simple epithelial tubes. The end pieces specialize in different ways. Some elaborate a serous secretion that typically contains an enzyme; others secrete a mucous fluid; still others contain both types of secretory cells. In humans and most other mammals the parotid gland produces a purely serous secretion. The submandibular and sublingual glands of most mammals are mixed (seromucous). The secretion of the sublingual gland tends to be more highly mucous in composition than that of the submandibular.

**Secretions.** Saliva is a viscid fluid containing a mixture of all of the oral secretions. It contains mucus, proteins, salts, and the enzymes ptyalin and maltase. Most of the ptyalin in human saliva is furnished by the parotid gland. The digestive action of saliva is limited to starchy food. Other uses of saliva include the moistening of the mouth and food for easier manipulation by the tongue, the consequent facilitation of swallowing, and a lubrication by mucus that ensures a smoother passage of food down the esophagus to the stomach. The daily amount of saliva produced by an adult is about 1.5 quarts (1.4 liters), by the cow, 65 quarts (61.5 liters). See GLAND. Leslie B. Arey

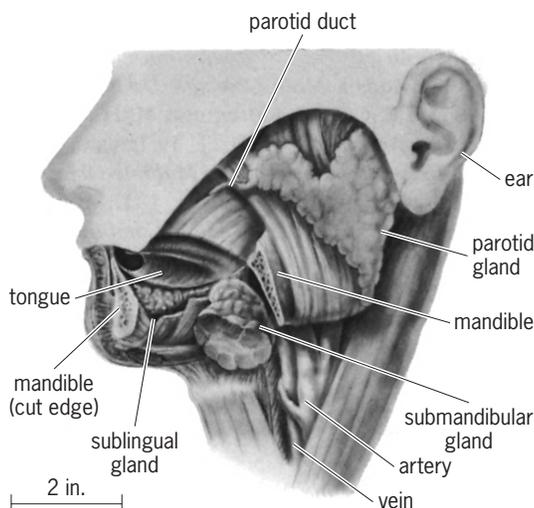
## Orange

The sweet orange (*Citrus sinensis*) is the most widely used species of citrus fruit and commercially is the most important. The sour or bitter oranges, of lesser importance, are distinct from sweet oranges and are classified as a separate species, *C. aurantium*. Citrus belonging to other species are also sometimes called oranges. This article deals only with the sweet orange. See SAPINDALES.

The sweet orange probably is native to northeastern India and southern China, but has spread to other tropical and subtropical regions of the world. The introduction of this delicious fruit into the Mediterranean region in the fifteenth century aroused much interest in citrus. The orange was first introduced into the Western Hemisphere by Columbus, and has become a major crop in Florida and California and in Brazil. The United States is the largest producer of oranges, followed by Spain, Italy, and Brazil. It is also a major crop in several other countries.

**Uses.** Sweet orange fruit is consumed fresh or as frozen or canned juice. A large portion of the crop, particularly in the United States, is used as frozen concentrate. After the juice is extracted, the peel and pulp are used for cattle feed. Peel oil is used in perfumes and flavoring, and citrus molasses is used as a livestock feed.

**Varieties.** The sweet orange tree is a moderately vigorous evergreen with a rounded, densely foliated top. The fruits are round (**Fig. 1**) or somewhat elongate and usually orange-colored when ripe. They can be placed in four groups: the common oranges, acidless oranges, pigmented oranges, and navel oranges. They may also be distinguished on the basis



Salivary glands, shown by a partial dissection of the human head. (After J. C. Brash, ed., *Cunningham's Textbook of Anatomy*, 9th ed., Oxford, 1951)



Fig. 1. Leaves, flowers, whole fruit, and cut fruit of *Citrus sinensis*.

of early, midseason, and late maturity. The common oranges are orange-colored and have varying numbers of seeds. Several, such as the Valencia, are important commercial cultivars throughout the world. The acidless oranges are prized in some countries because of their very low acidity, but generally are not major cultivars. The pigmented or "blood" oranges develop varying amounts of pink or red in the flesh and peel and have a distinctive aroma and flavor. Most originated in the Mediterranean area and are popular there, but are little grown in the United States. Navel oranges are so named because of the structure at the blossom end. Because they are usually seedless, some have become important cultivars for the fresh-fruit market. Many, if not all, of the cultivars appear to have been selected as bud sports, with the original type being small, seedy, and nonpigmented. See EVERGREEN PLANTS.

**Propagation.** Sweet oranges are propagated by budding or grafting on rootstocks in order to obtain uniform, high-yielding plantings. Many citrus species and relatives have been used as rootstocks. Diseases have severely restricted the use of some major rootstocks. Tristeza, a virus disease, has caused extensive worldwide losses of trees on sour orange rootstock. Root rot caused by soil-inhabiting fungi is also widespread, attacking many types of rootstocks. Large-scale selection and breeding programs have developed rootstocks that are tolerant to tristeza and root rot. See FRUIT; FRUIT, TREE.

R. K. Soost

**Diseases.** Sweet orange trees are affected by several bud-transmitted virus diseases. Psorosis, which causes scaling of the bark, is injurious to the tree irrespective of the rootstock used. Some other virus diseases are injurious only when infected trees are grown on certain rootstocks. For example, exocortis virus stunts trees growing on trifoliolate orange rootstock, whereas xyloporosis stunts trees on sweet lime rootstock. The aphid- and bud-transmitted tristeza virus causes a serious decline of trees grown on sour orange rootstock, and alternative rootstocks are essential for sweet orange production in countries where the efficient aphid vector of tristeza, *Toxoptera citricidus*, occurs.

Greening disease causes a devastating tree decline in parts of Asia and South Africa, regardless of the

rootstock used. It behaves like a virus disease, but is actually caused by an unnamed bacteriumlike organism. Greening disease is a problem only in those areas where the psyllid insect vectors are abundant. In South Africa, greening disease is controlled by injecting the tree trunks with antibiotics to kill the causal agent and by spraying the orchard with insecticides to control psyllids. Stubborn disease is another viruslike disease that seriously reduces fruit yields. It occurs in California and some Mediterranean countries. The mycoplasma-like organism, *Spiroplasma citri*, which causes stubborn disease is spread by infected budwood and certain leafhopper insects.

Blight is a serious wilt disease that rapidly renders orange trees unproductive. It behaves like an infectious disease; yet all attempts to transmit the disease by budding and grafting have failed. The cause of this disease is presently unknown. In Florida, blight is common in orchards planted on rough lemon rootstock, but is rarely seen on trees growing on sour orange.

Citrus canker, caused by *Xanthomonas citri*, is a bacterial disease that originated in Asia. It reached many other parts of the world through shipment of infected nursery trees, but it was later successfully eradicated in some areas, notably in the United States (Florida), Australia, and South Africa, through the costly destruction of all diseased trees.

Foot rot (trunk gummosis), caused by certain soil-borne *Phytophthora* fungi, occurs almost everywhere that orange trees are grown. Attacks commonly occur at the base of the trunk (Fig. 2). The bark and cambium are killed, thereby causing varying amounts of trunk girdling. Prevention depends on the use of resistant rootstocks, high budding to keep the highly susceptible sweet orange bark well above the ground, and good land drainage. These



Fig. 2. Foot rot on sweet orange portion of trunk, with root sprouts emerging from the unaffected disease-resistant rootstock.



Fig. 3. Greasy spot on sweet orange leaf.

same fungi cause brown rot of fruit during periods of prolonged wet weather. Fruit infection is prevented by spraying the tree canopy with copper fungicides.

Other important fungus diseases of sweet orange include greasy spot (caused by *Mycosphaerella citri*; Fig. 3), which causes serious defoliation in countries with very hot humid climates; black spot (caused by *Guignardia citricarpa*), which is an important disease in parts of South Africa and Australia; melanose (caused by *Diaporthe citri*), which is prevalent in areas with wet and relatively warm weather during early fruit development; and scab (caused by *Elsinoe australis*), which has been found only in South America. Spray oil or copper fungicides are applied to control greasy spot; benomyl fungicide is used to control black spot; and copper fungicides provide control of melanose and scab. See PLANT PATHOLOGY.

Jack O. Whiteside

Bibliography. A. R. Biggs (ed.), *Handbook of Cytology, Histology, and Histochemistry of Fruit Tree Diseases*, 1992; M. L. Flint, *Integrated Pest Management for Citrus*, 2d ed., 1991; J. Janick and J. N. Moore (eds.), *Advances in Fruit Breeding*, 1975.

## Orbital motion

In astronomy the motion of a material body through space under the influence of its own inertia, a central force, and other forces. Johannes Kepler found empirically that the orbital motions of the planets about the Sun are ellipses. Isaac Newton, starting from his laws of motion, proved that an inverse-square gravitational field of force requires a body to move in an orbit that is a circle, ellipse, parabola, or hyperbola.

**Elliptical orbit.** Two bodies revolving under their mutual gravitational attraction, but otherwise undisturbed, describe orbits of the same shape about a common center of mass. The less massive body has the larger orbit. In the solar system, the Sun and Jupiter have a center of mass just outside the visible disk of the Sun. For each of the other planets, the center of mass of Sun and planet lies within the Sun.

For this reason, it is convenient to consider only the relative motion of a planet of mass  $m$  about the Sun of mass  $M$  as though the planet had no mass

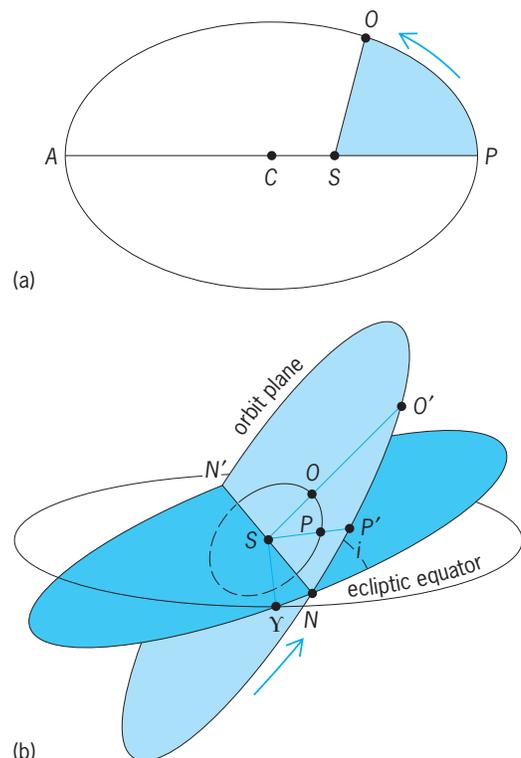
and moved about a center of mass  $M + m$ . The orbit so determined is exactly the same shape as the true orbits of planet and Sun about their common center of mass, but it is enlarged in the ratio  $(M + m)/M$ . See CENTER OF MASS; PLANET.

**Parameters of elliptical orbit.** The illustration shows the elements or parameters of an elliptical orbit. Major axis  $AP$  intersects the ellipse  $AOP$  at the apses; the extension of the major axis is the line of apses. The body is nearest the center of mass at one apside, called perihelion  $P$ , and is farthest away at the other, called aphelion  $A$ . See ELLIPSE.

Shape and size of an orbit are defined by two elements: length of semimajor axis and departure of the orbit from a circle. Semimajor axis  $a$  equals  $CP$ ; this length is expressed in units of the mean distance from the Earth to the Sun. Eccentricity  $e$  equals  $CS/CP$  where  $C$  is the center of the ellipse and  $S$  is a focus. For elliptical orbits  $e$  is always less than unity.

Position of a body in its orbit at time  $t$  can be computed if  $a$ ,  $e$ , and time of perihelion passage  $p$  and period of revolution  $T$  are known. Let  $O$  be the position of a planet at time  $t$  and  $OSP$  be the area swept out in time  $t - p$ . From Kepler's area law, area  $OSP$  equals  $(t - p)/T$  multiplied by the area of the full ellipse.

To describe the orientation of an orbit in space, several other parameters are required. All orbits in the solar system are referred to the plane of the ecliptic, this being the plane of the orbit of Earth about the Sun. The reference point for measurement of celestial longitude in the plane of the ecliptic is the vernal equinox, the first point of Aries. This is the



Parameters, or elements, of an elliptical orbit. (a) Relative orbit. (b) Orbit in space.

point where the apparent path of the Sun crosses the Earth's equator from south to north. The two points of intersection of the orbit plane with the plane of the ecliptic ( $N$  and  $N'$ ) are called the nodes, and the line joining them is the line of nodes. Ascending node  $N$  is the one where the planet crosses the plane of the ecliptic in a northward direction;  $N'$  is the descending node. The angle as seen from the Sun  $S$  measured in the plane of the ecliptic from the vernal equinox to the ascending node is  $\gamma SN$ ; it is termed the longitude of the ascending node  $\Omega$  and fixes the orbit plane with respect to the zero point of longitude. The angle at the ascending node between the plane of the ecliptic and the orbit plane is called the inclination  $i$  and defines the orientation of the orbit plane with respect to the fundamental plane. The angle as seen from the Sun, measured in the orbit plane from the ascending node to perihelion, is  $NSP'$  and is referred to as the argument of perihelion; it defines the orientation of the ellipse within the orbit plane. The angle  $NSP' + \Omega$ , measured in two different planes, is called the longitude of perihelion  $\tilde{\omega}$ . Because dynamically the semimajor axis  $a$  and period  $T$  of a planet of mass  $m$  revolving under the influence of gravitation  $G$  about the Sun of mass  $M$  are related by Eq. (1),

$$\frac{4\pi^2}{T^2} = \frac{G(M + m)}{a^3} \quad (1)$$

only six elements,  $a$ ,  $e$ ,  $i$ ,  $\Omega$ ,  $\tilde{\omega}$ , and  $p$ , are required to fix the orbit and instantaneous position of a planet in space. Instead of these elements, however, a position vector  $x$ ,  $y$ ,  $z$  and the associated velocity vector  $\dot{x}$ ,  $\dot{y}$ ,  $\dot{z}$  at a given instant of time would serve equally well to define the path of a planet in a rectangular coordinate system with origin at the Sun.

**Orbital velocity.** Orbital velocity  $v$  of a planet moving in a relative orbit about the Sun may be expressed by Eq. (2) where  $a$  is the semimajor axis and  $r$  is the

$$v^2 = G(M + m) \left( \frac{2}{r} - \frac{1}{a} \right) \quad (2)$$

distance from the planet to the Sun. In the special case of a circular orbit,  $r = a$ , and the expression becomes Eq. (3). When the eccentricity of an orbit

$$v^2 = \frac{G(M + m)}{a} \quad (3)$$

is exactly unity, the length of the major axis becomes infinite and the ellipse degenerates into a parabola. The expression for the velocity then becomes Eq. (4).

$$v^2 = G(M + m) \left( \frac{2}{r} \right) \quad (4)$$

This parabolic velocity is referred to as the velocity of escape, since it is the minimum velocity required for a particle to escape from the gravitational attraction of its parent body. See ESCAPE VELOCITY.

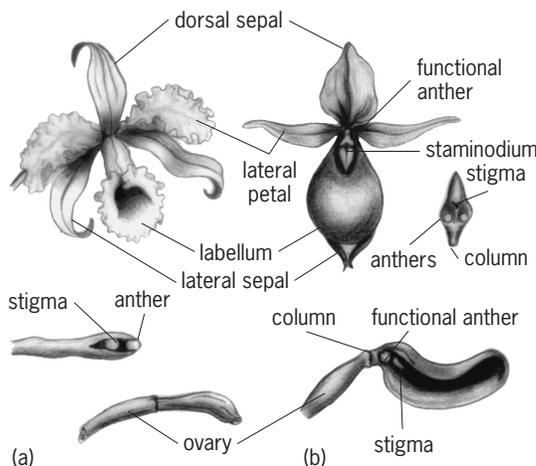
Eccentricities greater than unity occur with hyperbolic orbits. Because in a hyperbola the semimajor axis  $a$  is negative, hyperbolic velocities are greater than the escape velocity.

Parabolic and hyperbolic velocities seem to be observed in the motions of some comets and meteors. Aside from the periodic ones, most comets appear to be visitors from cosmic distances, as do about two-thirds of the fainter meteors. For ease of computation, the short arcs of these orbits that are observed near perihelion are represented by parabolas rather than ellipses. Although the observed deviation from parabolic motion is not sufficient to vitiate this computational procedure, it is possible that many of these "parabolic" comets are actually moving in elliptical orbits of extremely long period. The close approach of one of these visitors to a massive planet, such as Jupiter, could change the velocity from parabolic to elliptical if retarded, or from parabolic to hyperbolic if accelerated. It is possible that many of the periodic comets, especially those with periods under 9 years, have been captured in this way. See CELESTIAL MECHANICS; COMET; GRAVITATION; PERTURBATION (ASTRONOMY); STELLAR ROTATION. Raynor L. Duncombe

Bibliography. J. E. Prussing and B. A. Conway, *Orbital Mechanics*, Oxford University Press, 1993; A. E. Roy, *Orbital Motion*, 4th ed., Institute of Physics Publishing, 2004; V. G. Szebehely and H. Mark, *Adventures in Celestial Mechanics: A First Course in the Theory of Orbits*, 2d ed., Wiley, 1998; B. D. Tapley, B. E. Shultz, and G. H. Born, *Statistical Orbit Determination*, Elsevier Academic Press, 2004.

## Orchid

Any member of the orchid family (Orchidaceae), among the largest families of plants, estimated to contain up to 35,000 species. Orchids are monocots; their flowers have inferior ovaries, three sepals, and three petals. They are distinguished by the differentiation of one petal into a labellum, and the fusion of pistil and stamens into the column (see **illus.**). Pollen is usually contained in pollinia, that is, bundles that are removed intact by pollinators, usually insects or sometimes birds. Self-pollination and asexual reproduction without fertilization also occur. The



Structure of representative orchid flowers. (a) *Cattleya*. (b) *Cypripedium*.

combination of lip and column structure, flower color, fragrance, and other factors may limit the range of pollinators. Differing pollination mechanisms often provide barriers to cross-pollination between related species. Each flower can produce large quantities of seeds, with numbers in the millions in some tropical species. Seeds are minute, with undifferentiated embryo and no endosperm. Germination and establishment depends on symbiotic mycorrhizae that provide nutrients and water.

Orchids occur on all continents except Antarctica; they range from arctic tundra and temperate forest and grassland to tropical rainforest, where as epiphytes they reach their greatest abundance and diversity. Vanilla is obtained from seed pods of some species of *Vanilla*, and the beauty and mystique of many orchids make them important horticultural subjects. See MYCORRHIZAE; ORCHIDALES; VANILLA.

Charles J. Sheviak

Bibliography. R. L. Dressler, *Phylogeny and Classification of the Orchid Family*, 1993.

## Orchidales

An order of flowering plants, division Magnoliophyta (Angiospermae), in the subclass Liliidae of the class Liliopsida (monocotyledons). The order consists of four families: the Orchidaceae, with perhaps 15,000–20,000 species; the Burmanniaceae, with about 130 species; the Cordiaceae, with 9 species; and the Geosiridaceae, with 1 species. The Orchidales are mycotrophic, sometimes nongreen Liliidae with very numerous tiny seeds that have an undifferentiated



Eastern American species of moccasin flower (*Cypripedium acaule*). (U.S. Forest Service photograph by R. Dale Sanders)

embryo and little or no endosperm. The ovary is always inferior and apparently lacks the septal nectaries found in many Liliopsida although other kinds of nectaries are present. The various species of Orchidaceae (orchids) have numerous, highly diverse adaptations to particular insect pollinators. The flowers are highly irregular (see *illus.*) and usually have only one or two stamens, which are adnate to the thickened style. The pollen grains of most genera cohere in large masses called pollinia, so that a single act of pollination can result in the production of many thousands of seeds. Orchids are most abundant and diversified in moist, tropical regions, where they often grow high above the ground on forest trees. They do not parasitize the trees, but merely roost on them, as epiphytes. The most familiar corsage orchids belong to the genus *Cattleya* or hybrids of it. See FLOWER; LILIIDAE; LILIOPSIDA; MAGNOLIOPHYTA; ORCHID; PLANT KINGDOM.

Arthur Cronquist; T. M. Barkley

## Ordovician

The second-oldest period in the Paleozoic Era. The Ordovician is remarkable because not only did one of the most significant Phanerozoic radiations of marine life take place (early Middle Ordovician), but also one of the two or three most severe extinctions of marine life occurred (Late Ordovician). The early Middle Ordovician radiation of life included the initial colonization of land. These first terrestrial organisms were nonvascular plants. Vascular plants appeared in terrestrial settings shortly afterward. See GEOLOGIC TIME SCALE.

The rocks deposited during this time interval (these are termed the Ordovician System) overlie those of the Cambrian and underlie those of the Silurian. The Ordovician Period was about  $7 \times 10^7$  years in duration, and it lasted from about  $5.05 \times 10^8$  to about  $4.35 \times 10^8$  years ago.

The British geologist Charles Lapworth named the Ordovician in 1879, essentially as a resolution to a long-standing argument among British geologists over division of the Lower Paleozoic. Until that time, one school of thought, that of R. I. Murchison and his followers, had maintained that only a Silurian Period encompassed the lower part of the Paleozoic. Adam Sedgwick and his followers advocated that two intervals, the Cambrian and the Silurian, could be recognized in the Lower Paleozoic. By 1879 Lapworth observed that “three distinct faunas” had been recorded from the Lower Paleozoic, and he pointed out that each was as “marked in their characteristic features as any of those typical of the accepted systems of later age.”

To the stratigraphically lowest and oldest of the three, Lapworth suggested in 1879 that the appellation Cambrian be restricted. To the highest and youngest, Lapworth stated that the name Silurian should be applied. To the middle or second of three, Lapworth gave the name Ordovician, taking the

CENOZOIC	QUATERNARY	
	TERTIARY	
MESOZOIC	CRETACEOUS	
	JURASSIC	
	TRIASSIC	
PALEOZOIC	PERMIAN	
	CARBONIFEROUS	PENNSYLVANIAN
		MISSISSIPPIAN
	DEVONIAN	
	SILURIAN	
	ORDOVICIAN	
	CAMBRIAN	
PRECAMBRIAN		

name from an ancient tribe renowned for its resistance to Roman domination.

The type area for the Ordovician System is those parts of Wales and England that include rocks bearing the fossils that composed the second of the three major Lower Paleozoic faunas cited by Lapworth. The Ordovician System in Britain was divided into six major units called series, each distinguished by a unique fossil fauna. The time intervals during which each series formed are epochs. From oldest to youngest, the epochs and series of the British Ordovician are Tremadoc, Arenig, Llanvirn, Llandeilo, Caradoc, and Ashgill. Each of them has a type area in Britain where their characteristic faunas may be collected.

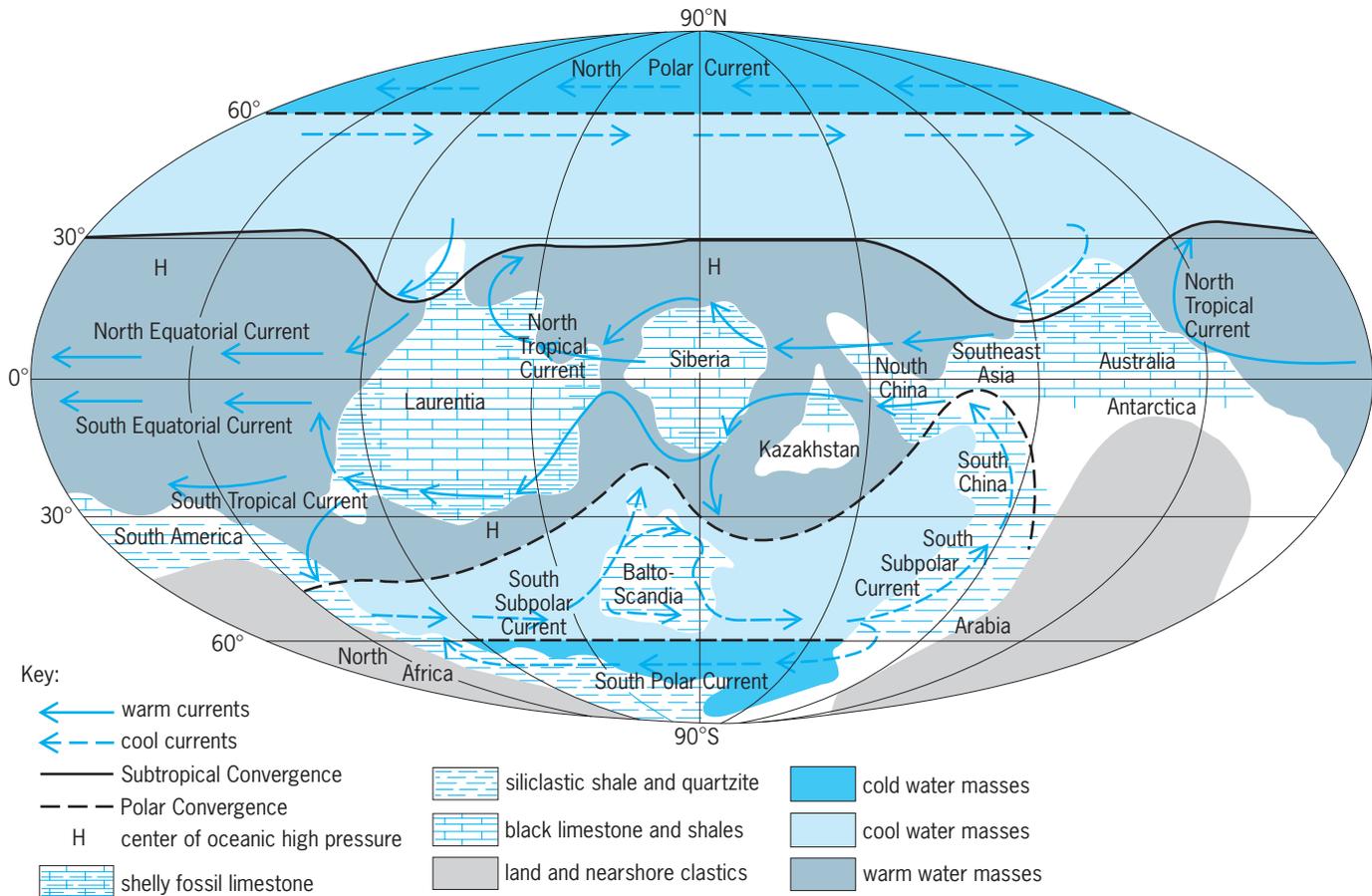
Intervals of shorter duration than those of the epoch are recognized as well in Britain. One set of such intervals is based on the evolutionary development of the fossils called graptolites. These intervals are the graptolite zones recognized by Lapworth and his associates Ethel M. R. Wood and Gertrude Elles. Each graptolite zone was about  $3 \times 10^6$  to  $5 \times 10^6$  years in duration. The boundary between the Ordovician and superjacent Silurian System has been designated as the base of the *Parakidograptus acuminatus* graptolite zone by international agreement. The type locality for that boundary is at Dob's Linn, near Moffat, southern Scotland. Black, graptolite-bearing shales are exposed there.

The Ordovician System is recognized in nearly all parts of the world, including the peak of Mount Everest, because the groups of fossils used to characterize the system are so broadly delineated. The British epochs and zones may not be recognized in all areas

where Ordovician rocks are found because the fossils used to characterize them are limited to certain geographic areas. Biogeographic provinces limited the distribution of organisms in the past to patterns similar to those of modern biogeographic provinces. Three broadly defined areas of latitude—the tropics, the midlatitudes (approximately 30–60°S), and the Southern Hemisphere high latitudes—constitute the biogeographic regions. Provinces may be distinguished within these three regions based upon organismal associations unique to each province. Epochs and zones are limited to a single region, and consequently each region has a unique set of epochs and zones for the Ordovician.

**Dynamic interrelationships.** The Earth's crust is essentially a dynamic system that is ceaselessly in motion. Plate positions and plate motions are linked closely with, and potentially exert a primary driving force that underlies, ocean circulation, ocean-atmosphere interactions, climates and climate change, and expansion and reduction of environments. Life responds to these physical aspects of the Earth's crust.

Several lines of evidence, including remanent magnetism, distributions of reefs and other major accumulations of carbonate rocks, positions of shorelines, and sites of glacial deposits, may be used to deduce many aspects of Ordovician paleogeography and paleogeographic changes. The Ordovician configuration of land and sea was markedly different from today. Much of the Northern Hemisphere above the tropics was ocean. The giant plate Gondwana was the predominant feature of the Southern Hemisphere. Modern Africa and South America were joined and occupied most the Southern Hemisphere high latitudes. The South Pole lay approximately in north-central Africa. A massive lobe of the plate extended northward from eastern Africa into the tropics. The modern Middle East, Turkey, India, Antarctica, and Australia constituted much of that huge lobe. A number of small plates lay on the margins of Gondwana. Certain of them may have been joined to Gondwana, and others were close to it. Some of these high-latitude plates included Perunica (modern Czech Republic); Avalonia (possibly in two parts, an eastern and a western), which included parts of Wales, southern Britain, Newfoundland, and Maritime Canada; and a number of plates that today make up southern Europe, including those found in the Alps, those that constitute the Iberian Peninsula, and those that made up Armorica (much of France). Plates that were within about 30–55°S latitude included Baltica (modern Scandinavia and adjacent eastern Europe east to the Urals), the Argentine Precordillera, South China, Tarim (in Asiatic China), the Exploits, and perhaps similar small plates. Parts of the Andes within modern northern Argentina, Peru, and Bolivia were within this midlatitudinal interval. Plates in tropical latitudes included Laurentia (modern North America, Greenland, Scotland, and some of northern Ireland), North China, Siberia, one or more plates that made up the Kazakh plate, and



**Early Ordovician (Arenig) paleogeography, major rock suites, and potential ocean surface currents; based on remanent magnetism data supplemented by plots of the positions of shelf sea carbonates and reefs.**

several plates that were close to or attached to northern or tropical Gondwana (these make up modern southeastern Asia and southeastern China). See PALEOGEOGRAPHY; PALEOMAGNETISM; PLATE TECTONICS.

During the early part of the Ordovician (Tremadoc-Arenig), prior to a significant number of plate movements, siliclastic materials (sand, silts, muds) spread northward from a Gondwana landmass into river, delta, and nearshore marine environments on the Gondwana plate and on those plates close to it, especially those in high latitudes. Coeval tropical environments were sites of extensive carbonate accumulations. Most midlatitude plates were sites of siliclastic and cool-water carbonate deposition.

Extensive plate motions and major volcanic activity at the margins of many plates characterize the Arenig-Llanvirn boundary interval of the early Middle Ordovician. Many plates on the Gondwanan margins began a northward movement that continued for much of the remainder of the Paleozoic. In addition, Laurentia bulged upward to such an extent that marine environments, which had covered most of the plate early in the Ordovician, were driven to positions on the plate margins. The Avalonian plates joined and moved relatively quickly northward to collide with the eastern side of Laurentia near the end of the Ordovician. Prior to that collision, the Popelogan or Medial New England plate collided with the

Laurentian plate at about the position of modern New England. That collision, which occurred about 455 million years ago, classically has been called the Taconic orogeny. Baltica not only moved northward relatively rapidly, but also rotated about  $90^\circ$  during the latter part of the Ordovician. The Argentine Precordillera plate moved southward across a midlatitude interval of ocean to collide with what is today the western side of Argentina in the Middle Ordovician. Africa shifted northward during the Ordovician with the result that northern Africa and the regions adjacent to the Middle East shifted into cool-temperate conditions.

**Life and environments.** As plate motions took place, environments changed significantly, as well as life in them. Both oceanic and terrestrial settings became the sites of significant radiations.

Early Ordovician (Tremadoc-Arenig) environmental conditions in most areas were similar to those of the Late Cambrian. Accordingly, Early Ordovician life was similar to that of the latter part of the Cambrian. Trilobites were the prominent animal in most shelf sea environments. Long straight-shelled nautiloids, certain snails, a few orthoid brachiopods, sponges, small echinoderms, algae, and bacteria flourished in tropical marine environments. Linguloid brachiopods and certain bivalved mollusks inhabited cool-water, nearshore environments.

Middle Ordovician plate motions were accompanied by significant changes in life. On land, nonvascular, mosslike plants appeared in wetland habitats. Vascular plants appeared slightly later in riverine habitats. The first nonvascular plants occurred in the Middle East on Gondwanan shores. The Middle Ordovician radiation of marine invertebrates is one of the most extensive in the record of Phanerozoic marine life. Corals, bryozoans, several types of brachiopods, a number of crinozoan echinoderms, conodonts, bivalved mollusks, new kinds of ostracodes, new types of trilobites, and new kinds of nautiloids suddenly developed in tropical marine environments. As upwelling conditions formed along the plate margins, oxygen minimum zones—habitats preferred by many graptolites—expanded at numerous new sites. Organic walled microfossils (chitinozoans and acritarchs) radiated in mid- to high-latitude environments. Ostracoderms (jawless, armored fish) radiated in tropical marine shallow-shelf environments. These fish were probably bottom detritus feeders. *See* PALEOECOLOGY.

**Glaciation.** When the Avalon plate collided with the Laurentian, a major mountain chain developed in a tropical setting. Vast quantities of siliclastic materials were shed from that land to form what is called the Queenston delta in the present-day Appalachians. As the Queenston delta grew, glaciation commenced at or near the South Pole. Continental glaciation spread from its North African center for a 1–2-million-year interval late in the Ordovician. Glacially derived materials (including many drop-stones) occurred in the Late Ordovician strata in Morocco, Algeria, southern France, Germany, Spain, Portugal, and the Czech Republic. Sea level dropped by at least 70 m at the glacial maximum. As a result, most shallow to modest-depth marine environments were drained. Karsts formed across many carbonates that had accumulated in the shallow marine settings. Upwelling along many platform margins ceased or became quite limited. As a consequence, former extensive oxygen minimum zones were markedly diminished. The loss of wide expanses of shallow marine environments and extensive oxygen minimum zones led to massive extinctions of benthic marine organisms, as well as those graptolites living near oxygen minimum zones. These extinctions took place over a 1–2-million-year interval as environments shifted, diminished, or eventually were lost. Oxygen isotope studies on brachiopod shells suggest that tropical sea surface temperatures dropped by as much as 4°C. *See* GEOMORPHOLOGY; PALEOCEANOGRAPHY.

The latest Ordovician stratigraphic record suggests that the ice melted relatively quickly, accompanied by a relatively rapid sea-level rise in many areas. Some organisms—certain conodonts, for example—did not endure significant extinctions until sea levels began to rise and shelf sea environments began to expand. *See* STRATIGRAPHY.

**Ocean surface circulation.** Surface circulation in Ordovician seas was controlled in the tropics by the several platforms and in the Southern Hemisphere by Gondwanaland. Equatorial surface currents flowed

east to west, but they were deflected by the shallow shelf environments. The tropical or warm-water faunal provinces were influenced by these deflections. Homogeneity of the tropical faunas was maintained by the surface water currents. Southern Hemisphere currents were influenced by the relatively long west coast of Gondwanaland and of the Baltoscanian Plate. Upwelling conditions would have been generated along these coasts. Location, size, and relief on Gondwanaland probably led to monsoonal seasonal reversals in surface ocean currents near what is today South China. Absence of lands or shallow shelf seas north of the Northern Hemisphere tropics would permit oceanic surface circulation to be zonal; that is, currents flowed from east to west north of 30° north latitude, and they flowed from west to east between 30 and 60° north latitude.

**Economic resources.** Ordovician shelf and shelf margin rock sequences in areas where there has been little post-Ordovician volcanic activity or severe deformation have yielded petroleum and natural gas. Quartzites interbedded with carbonates formed in shelf sea environments have been used as a source of silica for glass manufacture. Ordovician carbonates are hosts for lead-zinc-silver ores mined in the western United States, including Missouri and Washington. Significant quantities of gold were recovered from Ordovician graptolite-bearing strata in eastern Australia in the late 1800s. Gold-bearing Ordovician rocks occur in Nevada (western United States) where they are part of one of the most prolific gold-producing areas in the world.

W. B. N. Berry

**Bibliography.** J. D. Cooper, M. I. Droser, and S. C. Finney (eds.), *Ordovician Odyssey: Short Papers for the 7th International Symposium on the Ordovician System*, Pacific Section, Society for Sedimentary Geology, 1995; P. Kraft and O. Fatka (eds.), *Quo Vadis Ordovician?*, *Acta Universitatis Carolinae Geologica*, vol. 43, no. 1/2, 1999; C. Lapworth, On the tripartite classification of the Lower Palaeozoic rocks, *Geol. Mag.*, 6:1–15, 1879; T. H. Torsvik, Palaeozoic palaeogeography: A North Atlantic viewpoint, *GFF*, 120:109–118, 1998.

## Ore and mineral deposits

Ore deposits are naturally occurring geologic bodies that may be worked for one or more metals. The metals may be present as native elements, or, more commonly, as oxides, sulfides, sulfates, silicates, or other compounds. The term ore is often used loosely to include such nonmetallic minerals as fluorite and gypsum. The broader term, mineral deposits, includes, in addition to metalliferous minerals, any other useful minerals or rocks. Minerals of little or no value which occur with ore minerals are called gangue. Some gangue minerals may not be worthless in that they are used as by-products; for instance, limestone for fertilizer or flux, pyrite for making sulfuric acid, and rock for road material.

**TABLE 1. Elemental composition of Earth's crust based on igneous and sedimentary rocks\***

Element	Weight %	Atomic %	Volume %
Oxygen	46.71	60.5	94.24
Silicon	27.69	20.5	0.51
Titanium	0.62	0.3	0.03
Aluminum	8.07	6.2	0.44
Iron	5.05	1.9	0.37
Magnesium	2.08	1.8	0.28
Calcium	3.65	1.9	1.04
Sodium	2.75	2.5	1.21
Potassium	2.58	1.4	1.88
Hydrogen	0.14	3.0	

\*After T. F. W. Barth, *Theoretical Petrology*, John Wiley and Sons, 1952 (recalculated from F. W. Clarke and H. S. Washington, 1924).

Mineral deposits that are essentially as originally formed are called primary or hypogene. The term hypogene also indicates formation by upward movement of material. Deposits that have been altered by weathering or other superficial processes are secondary or supergene deposits. Mineral deposits that formed at the same time as the enclosing rock are called syngenetic, and those that were introduced into preexisting rocks are called epigenetic.

The distinction between metallic and nonmetallic deposits is at times an arbitrary one since some substances classified as nonmetals, such as lepidolite, spodumene, beryl, and rhodochrosite, are the source of metals. The principal reasons for distinguishing nonmetallic from metallic deposits are practical ones, and include such economic factors as recovery methods and uses.

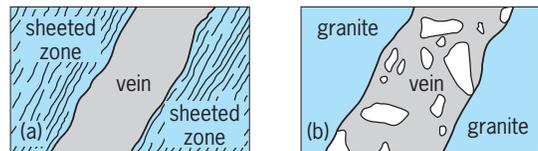
**Concentration.** The Earth's crust consists of igneous, sedimentary, and metamorphic rocks. **Table 1** gives the essential composition of the crust and shows that 10 elements make up more than 99% of the total. Of these, aluminum, iron, and magnesium are industrial metals. The other metals are present in small quantities, mostly in igneous rocks (**Table 2**).

Most mineral deposits are natural enrichments and concentrations of original material produced by different geologic processes. To be of commercial grade, the metals must be present in much higher concentrations than the averages shown in **Table 2**. For example, the following metals must be concentrated in the amounts indicated to be considered ores: aluminum, about 30%; copper, 0.7–10%; lead, 2–4%; zinc, 3–8%; and gold, silver, and uranium, only a small fraction of a percent of metal. Therefore, natural processes of concentration have increased the aluminum content of aluminum ore 3 or 4 times, and even a low-grade gold ore may represent a concentration of 20,000 times. Economic considerations, such as the amount and concentration of metal, the cost of mining and refining, and the market value of the metal, determine whether the ore is of commercial grade.

**Forms of deposits.** Mineral deposits occur in many forms depending upon their origin, later deformation, and changes caused by weathering. Syngenetic

deposits are generally sheetlike, tabular, or lenticular, but may on occasion be irregular or roughly spherical.

Epigenetic deposits exhibit a variety of forms. Veins or lodes are tabular or sheetlike bodies that originate by filling fissures or replacing the country rock along a fissure (**Fig. 1a**). Replacement bodies in limestone may be very irregular. Veins are usually inclined steeply and may either cut across or conform with the bedding or foliation of the enclosing rocks. The inclination is called the dip, and is the angle between the vein and the horizontal. The horizontal trend of the vein is its strike, and the vertical angle between a horizontal plane and the line of maximum elongation of the vein is the plunge. The veins of a mining district commonly occur as systems which have a general strike, and one or more systems may be present at some angle to the main series. In places the mineralization is a network of small, irregular, discontinuous veins called a stockwork.



**Fig. 1. Typical forms of deposits. (a) Vein developed in fissured or sheeted zone. (b) Brecciated vein in granite.**

Mineral deposits are seldom equally rich throughout. The pay ore may occur in streaks, spots, bunches, or bands separated by low-grade material or by gangue. These concentrations of valuable ore are called ore shoots; if roughly horizontal they are called ore horizons, and if steeply inclined they are called chimneys. After formation, mineral deposits may be deformed by folding, faulting, or brecciation (**Fig. 1b**).

**TABLE 2. Abundance of metals in igneous rocks**

Element	%
Aluminum	8.13
Iron	5.00
Magnesium	2.09
Titanium	0.44
Manganese	0.10
Chromium	0.02
Vanadium	0.015
Zinc	0.011
Nickel	0.008
Copper	0.005
Tin	0.004
Cobalt	0.0023
Lead	0.0016
Arsenic	0.0005
Uranium	0.0004
Molybdenum	0.00025
Tungsten	0.00015
Antimony	0.0001
Mercury	0.00005
Silver	0.00001
Gold	0.000005
Platinum	0.000005

**Metasomatism, or replacement.** Metasomatism, or replacement, is the process of essentially simultaneous removal and deposition of chemical matter. A striking manifestation of this process in mineral deposits is the replacement of one mineral by another mineral or mineral aggregate of partly or wholly different composition. A large volume of rock may be transformed in this manner, and the resulting deposit is generally of equal volume. Commonly the original structure and texture of the replaced rock is preserved by the replacing material.

Replacement, evidence for which is found in many mineral deposits, operates at all depths under a wide range of temperature. The evidence indicates that the new minerals formed in response to conditions that were unstable for the preexisting ones.

Usually the replacing material moves to the site of metasomatism along relatively large openings such as faults, fractures, bedding planes, and shear zones. It then penetrates the rock along smaller cracks and finally enters individual mineral grains along cleavage planes, minute fractures, and grain boundaries where substitution may take place on an atomic scale until the entire mass has been transformed (Fig. 2). After gaining access to individual grains, the replacement may proceed by diffusion of ions through the solid, controlled in large part by imperfections in the crystal structure. In many deposits repeated movement has opened and reopened channelways, which would otherwise have become clogged, to permit continued and widespread replacement. The process may take place through the action of gases or solutions or by reactions in the solid state. See METASOMATISM.

**Classification.** Mineral deposits are generally classified on the basis of the geologic processes responsible for their formation as magmatic, contact metamorphic, pegmatitic, hydrothermal, sedimentary, residual, and regional metamorphic deposits.

*Magmatic deposits.* Some mineral deposits originated by cooling and crystallization of magma, and the concentrated minerals form part of the body of the igneous rock. If the magma solidified by simple crystallization, the economically valuable mineral is distributed through the resulting rock; diamond deposits found in peridotite are believed by some geologists to be of this type. However, if the magma has differentiated during crystallization, early formed

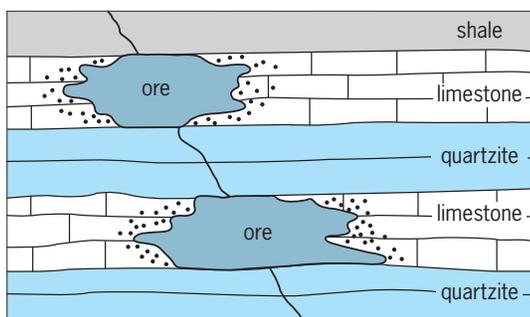


Fig. 2. Replacement of limestone by ore along a fissure. Disseminated ore, indicated by the dots, is forming in advance of the main body.

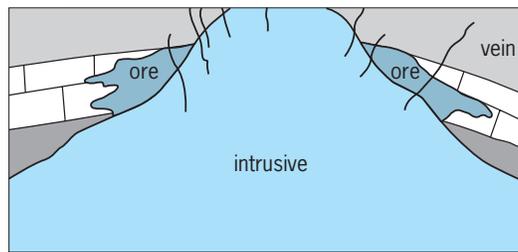


Fig. 3. Association of contact metamorphic and vein deposits with intrusive magmas.

minerals may settle to the bottom of the magma chamber and form segregations such as the chromite deposits of the Bushveld in South Africa. Late-formed minerals may crystallize in the interstices of older minerals and form segregations like the Bushveld platinum deposits. Occasionally, the residual magma becomes enriched in constituents such as iron, and this enriched liquid may form deposits, such as the Taberg titaniferous iron ores of Sweden. It is also possible that during differentiation some of the crystals or liquid may be injected and form sills or dikes. The iron ores of Kiruna, Sweden, have been described as early injections, and certain pegmatites are classed as late magmatic injections. Magmatic deposits are relatively simple in mineral composition and few in number. See DIAMOND; MAGMA; PERIDOTITE.

*Contact metamorphic deposits.* During the crystallization of certain magmas a considerable amount of fluid escapes. This fluid may produce widespread changes near the contacts of magma with the surrounding rocks (Fig. 3). Where such changes are caused by heat effects, without addition of material from the magma, the resulting deposits are called contact metamorphic. If appreciable material is contributed by the magma, the deposits are termed contact metamorphic. The term skarn is applied to the lime-bearing silicates formed by the introduction of Si, Al, Fe, and Mg into a carbonate rock; some skarns contain ore bodies. The magmas that produce these effects are largely silicic in composition and resulting mineral deposits are often irregular in form. See SKARN.

A complicating case exists where little fluid escaped from the magma but the heat of the intrusion was great enough to cause dissolution and movement of certain metals from the surrounding rocks. It is believed by some investigators that solutions formed in this manner may become concentrated in metal content and subsequently deposit these metals near the contact of the intrusion and the surrounding rocks. In this case, the ore minerals were deposited by replacing preexisting rocks but the source of the ore is the surrounding rocks, not the magma. To further complicate matters, the ore in some deposits appears to consist of material derived from both the intrusion and the surrounding rocks. In such deposits the source of the ore is generally controversial, and the size, amount, and composition of the mineralization would depend upon the relative contributions from the intrusion and the associated rocks.

Under contact metasomatic conditions, the (ore-forming) fluids extensively replace the country rock to produce a variety of complex minerals. Contact metasomatic deposits include a number of important deposits, whereas contact metamorphic deposits are rarely of economic value. Many garnet, emery, and graphite deposits are classed as contact metasomatic, as are such metalliferous deposits as the iron ores of Cornwall, Pennsylvania, Iron Springs, Utah, and Banat, Hungary; many copper ores of Utah, Arizona, New Mexico, and Mexico; the zinc ores of Hanover, New Mexico; and various tungsten ores of California and Nevada.

*Pegmatite deposits.* Pegmatites are relatively coarse-grained rocks found in igneous and metamorphic regions. The great majority of them consist of feldspar and quartz, often accompanied by mica, but complex pegmatites contain unusual minerals and rare elements. Many pegmatites are regular tabular bodies; others are highly irregular and grade into the surrounding rocks. In size, pegmatites range from a few inches in length to bodies over 1000 ft (3000 m) long and scores of feet across. Some pegmatites are zoned, commonly with a core of quartz surrounded by zones in which one or two minerals predominate. See FELDSPAR; QUARTZ.

Pegmatites may originate by various igneous and metamorphic processes. Fractional crystallization of a magma results in residual solutions that are generally rich in alkalis, alumina, water, and other volatiles. The volatiles lower the temperature of this liquid and make it unusually fluid; the low viscosity promotes the formation of coarse-grained minerals. The rare elements that were unable by substitution to enter into the crystal structure of earlier-formed minerals, principally because of differences in size of their atomic radii, are concentrated in the residual pegmatite solutions. Late hydrothermal fluids may alter some of the previously formed pegmatite minerals.

Some pegmatites develop by replacement of the country rock and commonly these are isolated bodies with no feeders or channels in depth. They occur in metamorphic regions usually devoid of igneous rocks and contain essentially the same minerals as those in the country rocks. In some regions small pegmatites have grown by forcing apart the surrounding metamorphic rock, and others have formed by filling a fissure or crack from the walls inward. In both cases growth is believed to have taken place by diffusion and consolidation of material in the solid state. See PEGMATITE.

*Hydrothermal deposits.* Most vein and replacement deposits are believed to be the result of precipitation of mineral matter from dilute, hot ascending fluids. As the temperature and pressure decrease, deposition of dissolved material takes place. It is not altogether certain how important the gaseous state is in the transport of ore material. It may be that at relatively shallow depth and high temperature gaseous solutions transport significant amounts of ore-forming material.

W. Lindgren, who developed the hydrothermal

theory, divided these deposits into three groups on the basis of temperature and pressure conditions supposed to exist at the time of formation. Deposits thought to form at temperatures of 50–200°C (120–390°F) at slight depth beneath the surface are called epithermal. Many ores of mercury, antimony, gold, and silver are of this type. Deposits formed at 200–300°C (390–570°F) at moderate depths are known as mesothermal and include ores of gold-quartz, silver-lead, copper, and numerous other types. Hypothermal deposits are those formed at 300–500°C (570–930°F) at high pressures; certain tin, tungsten, and gold-quartz ores belong to this type.

The nature of hydrothermal fluids is inferred by analogy with laboratory experiments, and by investigation of deposits forming around volcanoes and hot springs at the present time. Studies of liquid inclusions in minerals, of mineral textures, and of inversion temperatures of minerals indicate that mineralization takes place at elevated temperatures. Layers of minerals on the walls of open fissures with crystal faces developed toward the openings suggest deposition from solution. In some of these cavities later crystals were deposited on earlier ones in a manner that suggests growth in moving solutions. Certain secondary replacement phenomena, such as weathering and oxidation of mineral deposits, also indicate deposition from liquid solutions. Studies of wall rock alteration where hydrothermal solutions have attacked and replaced rock minerals indicate that these solutions change in character from place to place. Sulfur in such solutions may react with or leach metals from the surrounding rocks or partly solidified magma to form certain kinds of mineral deposits. On the basis of geochemical data it has been estimated that most hydrothermal ore-forming solutions had a temperature in the range 50–600°C (120–1100°F), formed under pressures ranging from atmospheric to several thousand atmospheres, commonly contained high concentrations of NaCl and were saturated with silica but were not highly concentrated in ore metals, were neutral or within about 2 pH units of neutrality; and that the metals probably were transported as complexes.

The principal objections to the hydrothermal theory are the low solubility of sulfides in water and the enormous quantities of water required. W. Lindgren realized this and, for some deposits, favored colloidal solutions as carriers of metals. Laboratory synthesis of sulfide minerals by G. Kullerud shows that some ore-bearing solutions must have been considerably more concentrated than is generally believed.

Two common features of hydrothermal deposits are the zonal arrangement of minerals and alteration of wall rock.

1. Zoning of mineralization. Many ore deposits change in composition with depth, lateral distance, or both, resulting in a zonal arrangement of minerals or elements. This arrangement is generally interpreted as being due to deposition from solution with decreasing temperature and pressure, the solution precipitating minerals in reverse order of their solubilities. Other factors are also involved such

as concentration, relative abundance, decrease in electrode potentials, and reactions within the solutions and with the wall rocks as precipitation progresses.

Zonal distribution of minerals was first noted in mineral deposits associated in space with large igneous bodies, and has since been extended to include zoning related to sedimentary and metamorphic processes in places where no igneous bodies are in evidence. Although many geologists interpret zoning as a result of precipitation from a single ascending solution, others believe deposition is achieved from solutions of different ages and of different compositions.

The distribution of mineral zones is clearly shown at Cornwall, England, and at Butte, Montana. At Cornwall, tin veins in depth pass upward and outward into copper veins, followed by veins of lead-silver, then antimony, and finally iron and manganese carbonates. Such zoning is by no means a universal phenomenon, and, in addition to mines and districts where it is lacking, there are places where reversals of zones occur. Some of these reversals have been explained more or less satisfactorily by telescoping of minerals near the surface, by the effects of structural control or of composition of the host rock in precipitating certain minerals, and by the effects of supergene enrichment on the original zoning, but many discrepancies are not adequately explained.

2. Wall rock alteration. The wall rocks of hydrothermal deposits are generally altered, the most common change being a bleaching and softening. Where alteration has been intense, as in many mesothermal deposits, primary textures may be obliterated by the alteration products. Chemical and mineralogical changes occur as a result of the introduction of some elements and the removal of others; rarely a rearrangement of minerals occurs with no replacement.

Common alteration products of epithermal and mesothermal deposits are quartz, sericite, clay minerals, chlorite, carbonates, and pyrite. Under high-temperature hypogene conditions pyroxene, amphibole, biotite, garnet, topaz, and tourmaline form. In many mines sericite has been developed nearest the vein and gives way outward to clay minerals or chlorite. The nature and intensity of alteration vary with size of the vein, character of the wall rock, and temperature and pressure of hydrothermal fluids. In the large, low-grade porphyry copper and molybdenum deposits associated with stocklike intrusives, alteration is intense and widespread, and two or more stages of alteration may be superimposed.

Under low-intensity conditions, the nature of the wall rock to a large extent determines the alteration product. High-intensity conditions, however, may result in similar alteration products regardless of the nature of the original rock. Exceptions to this are monomineralic rocks such as sandstones and limestones. Wall rock alteration may develop during more than one period by fluids of differing compositions, or it may form during one period of mineralization as the result of the action of hydrothermal fluids that

did not change markedly in composition. Alteration zones have been used as guides to ore and tend to be most useful where they are neither too extensive nor too narrow. Mapping of these zones outlines the mineralized area and may indicate favorable places for exploration.

*Sedimentary and residual deposits.* At the Earth's surface, action of the atmosphere and hydrosphere alters minerals and forms new ones that are more stable under the existing conditions. Sedimentary deposits are bedded deposits derived from preexisting material by weathering, erosion, transportation, deposition, and consolidation. Different source materials and variations in the processes of formation yield different deposits. Changes that take place in a sediment after it has formed and before the succeeding material is laid down are termed diagenetic. They include compaction, solution, recrystallization, and replacement. In general, the sediment is consolidated by compaction and by precipitation of material as a cement between mineral grains. Sedimentation as a process may itself involve the concentration of materials into mineral deposits. *See* DIAGENESIS.

The mineral deposits that form as a result of sedimentary and weathering processes are commonly grouped as follows: (1) sedimentary deposits, not including products of evaporation; (2) sedimentary-exhalative deposits; (3) chemical evaporites; (4) placer deposits; (5) residual deposits; and (6) organic deposits. *See* VOLCANO.

1. Sedimentary deposits. Included in this group are the extensive coal beds of the world, the great petroleum resources, clay deposits, limestone and dolomite beds, sulfur deposits such as those near Kuibyshev, Russia, and the deposits of the Gulf Coast region, and the phosphate of North Africa and Florida. Metalliferous deposits such as the minette iron ores of Lorraine and Luxembourg, and Clinton iron ores of the United States, and the manganese of Tchiaturi, Georgia, and Nikopol in the Ukraine also belong here. There are other deposits of metals in sedimentary rocks whose origin remains an enigma, such as the uranium of the Colorado Plateau, the Witwatersrand in South Africa, and Blind River in Ontario; and the copper deposits of Mansfeld, Germany, and of the Copperbelt of Zambia and the Democratic Republic of the Congo. These deposits have characteristics of both syngenetic and epigenetic types. A controversy centers around the genesis of these and similar deposits of the world. *See* HEAVY MINERALS; SEDIMENTOLOGY.

2. Sedimentary-exhalative deposits. Many large stratiform deposits are found in marine sedimentary rocks associated with volcanic rocks. It is well known that volcanoes and fumaroles carry in their gases a number of metals. On land these gases escape into the atmosphere. Under water the gases, if they carry matter which is insoluble under the existing conditions, will precipitate their metals as oxides, sulfides, or carbonates in the vicinity of the gas emission. If the gases contain matter that is soluble, the metal content of the seawater will increase, and upon reaching saturation level will precipitate an

extensive disseminated ore deposit. Where submarine emissions take place in a large ocean basin, they may be deposited over the floor of the basin as part of the sedimentation process. *See* VOLCANO.

Deposits exemplified by lead-zinc-barite-fluorite mineralization, most commonly found in carbonate rocks, occur in the Mississippi Valley region of North America and also on other continents. These ores are included with the sedimentary-exhalative type, but could also be discussed under several other classes of deposits since they are very difficult to categorize. They have been considered by various geologists to be true sediments, diagenetic deposits, lateral secretion deposits, deposits formed by downward leaching of overlying lean ores, deposits formed by solutions that descended and subsequently ascended, deposits resulting from magmatic-hydrothermal processes, and sea-floor deposits from thermal springs. Most geologists favor either a syngenetic-sedimentary hypothesis or an epigenetic-hypogene one. Some studies hypothesize a source of metal-bearing waters similar to those in deep brines which rise and move through fissures in overlying rocks or are poured out on the sea floor and are added to accumulating sediments. A single generally acceptable hypothesis of origin, if such eventually emerges, must await the accumulation and interpretation of additional geological and geochemical data.

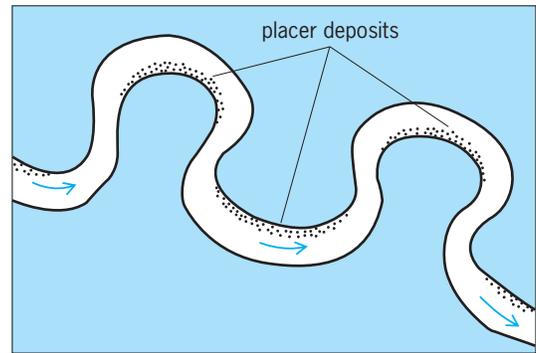
3. Chemical evaporites. These consist of soluble salts formed by evaporation in closed or partly closed shallow basins. Deposits of salt or gypsum that are several hundred feet thick are difficult to explain satisfactorily. Oschsenius suggested that they formed in basins which were separated from the ocean by submerged bars except for a narrow channel (inlet); such barriers are common along coastal areas. Intermittently, seawater flowed over the barrier and was concentrated into saline deposits by evaporation. Modifications of this theory have been proposed to account for the omissions of certain minerals and the interruptions in the succession.

Deposits of gypsum and common salt (halite) are found in many countries, whereas the larger concentrations of potash salts, borates, and nitrates are much more restricted in occurrence. *See* SALINE EVAPORITES.

4. Placer deposits. Placers are the result of mechanical concentration whereby heavy, chemically resistant, tough minerals are separated by gravity from light, friable minerals. Separation and concentration may be accomplished by streams, waves and currents, and air, or by soil and hill creep. The most important economic placer deposits are those formed by stream action (**Fig. 4**).

Stream and beach placers are widespread in occurrence and include the famous gold placers of the world, as well as deposits of magnetite, ilmenite, chromite, wolframite, scheelite, cassiterite, rutile, zircon, monazite, and garnet. Placer deposits of diamond, platinum, and gemstones are less common.

5. Residual deposits. Complete weathering results in distribution of the rock as a unit and the seg-



**Fig. 4.** Deposition of placer by stream action on the inside of meander bends.

regation of its mineral constituents. This is accomplished by oxidation, hydration, and solution, and may be accelerated by the presence of sulfuric acid. Some iron and manganese deposits form by accumulation without change, but certain clay and bauxite deposits are created during the weathering of aluminous rocks. Residual concentrations form where relief is not great and where the crust is stable; this permits the accumulation of material in place without erosion. *See* WEATHERING PROCESSES.

Large residual deposits of clay, bauxite, phosphate, iron, and manganese have been worked in many parts of the world, as have smaller deposits of nickel, ocher, and other minerals.

6. Organic deposits. Plants and animals collect and use various inorganic substances in their life processes, and concentration of certain of these substances upon the death of the organisms may result in the formation of a mineral deposit. Coal and peat form from terrestrial plant remains and represent concentration by plants of carbon from the carbon dioxide of the atmosphere. Petroleum originates by the accumulation of plant and animal remains. Many limestone, phosphate, and silica deposits also form by plant and animal activity. Hydrated ferric oxide and manganese dioxide are precipitated by microorganisms; anaerobic bacteria can reduce sulfates to sulfur and hydrogen sulfide. There is considerable controversy, however, as to whether microorganisms are responsible for the formation of certain iron, manganese, and sulfide deposits. Some uranium, vanadium, copper, and other metalliferous deposits are considered to have formed, in part at least, by the activity of organisms.

*Deposits formed by regional metamorphism.* Regional metamorphism includes the reconstruction that takes place in rocks within orogenic or mountain belts as a result of changes in temperature, pressure, and chemical environment. In these orogenic belts, rocks are intensely folded, faulted, and subjected to increases in temperature. The changes that occur in this environment affect the chemical and physical stability of minerals, and new minerals, textures, and structures are produced, generally accompanied by the introduction of considerable material and the removal of other material.

Some geologists believe that the water and metals released during regional metamorphism can give rise to hydrothermal mineral deposits. Along faults and shear zones movement of fluids could take place by mechanical flow, though elsewhere movement might be by diffusion. The elements released from the minerals would migrate to low-pressure zones such as brecciated or fissured areas and concentrate into mineral deposits. It has been suggested that the subtraction of certain elements during metamorphism also can result in a relative enrichment in the remaining elements; if this process is sufficiently effective, a mineral deposit may result. Certain minerals also may be concentrated during deformation by flow of material to areas of low pressure such as along the crests of folds.

Deposits of magnetite, titaniferous iron, and various sulfides may form in metamorphic rocks, as well as deposits of nonmetallic minerals such as kyanite, corundum, talc, graphite, and garnet.

Opponents of the concept of mineral formation by regional metamorphism believe that a dispersal of minerals, rather than a concentration, would result from the operative processes. However, if movement of material were confined to specific channelways, this objection would not necessarily hold. See METAMORPHISM.

**Oxidation and supergene enrichment.** Many sulfide minerals form at depth under conditions differing markedly from those existing at the surface. When such minerals are exposed by erosion or deformation to surface or near-surface conditions, they become unstable and break down to form new minerals. Essentially all minerals are affected.

The oxidation of mineral deposits is a complex process. Some minerals are dissolved completely or in part, whereas elements of others recombine and form new minerals. The principal chemical processes that take place are oxidation, hydration, and carbonation. The oxidation of pyrite and other sulfides produces sulfuric acid, a strong solvent. Much of the iron in the sulfides is dissolved and reprecipitated as hydroxide to form iron-stained outcrops called gossans. Metal and sulfate ions are leached from sulfides and carried downward to be precipitated by the oxidizing waters as concentrations of oxidized ores above the water table. Oxides and carbonates of copper, lead, and zinc form, as do native copper, silver, and gold. The nature of the ore depends upon the composition of the primary minerals and the extent of oxidation. If the sulfates are carried below the water table, where oxygen is excluded, upon contact with sulfides or other reducing agents they are precipitated as secondary sulfides. The oxidized zone may thus pass downward into the supergene sulfide zone. Where this process has operated extensively, a thick secondary or supergene-enriched sulfide zone is formed. Enrichment may take place by removal of valueless material or by solution of valuable metals which are then transported and reprecipitated. This enrichment process has converted many low-grade ore bodies into workable deposits. Supergene enrichment is characteristic of copper deposits

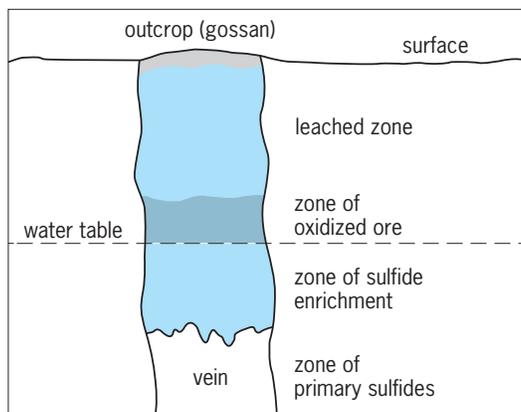


Fig. 5. Vein deposit of sulfide ore, showing changes due to oxidation and supergene enrichment.

but may also take place in deposits of other metals. Beneath the enriched zone is the primary sulfide ore (Fig. 5).

The textures of the gossan minerals may give a clue to the identity of the minerals that existed before oxidation and enrichment took place. These have been used as guides in prospecting for ore.

**Sequence of deposition.** Studies of the relationships of minerals in time and space have shown that a fairly constant sequence of deposition, or paragenesis, is characteristic of many mineral deposits. This sequence has been established largely by microscopic observations of the boundary relationships of the minerals in scores of deposits. Subsequent experimental studies of mineral phases have contributed to the knowledge of paragenesis. In magmatic and contact metasomatic ores, silicates form first, followed by oxides and then sulfides. W. Lindgren presented the paragenesis for hypogene mineral associations, and others have discussed the problems involved. The sequence of common minerals starts with quartz, followed by iron sulfide or arsenide, chalcopyrite, sphalerite, bornite, tetrahedrite, galena, and complex lead and silver sulfosalts. It indicates the existence of some fundamental control but attempts to explain the variations in it have been largely unsuccessful, or are applicable to only part of the series or to specific mineralized areas. Local variations are to be expected since many factors such as replacement, unmixing, superimposed periods of mineralization, structural and stratigraphic factors, and telescoping of minerals may complicate the order of deposition.

Paragenesis is generally thought to be the result of decreasing solubility of minerals with decreasing temperature and pressure. It has also been explained in terms of relative solubilities, pH of the solutions, metal volatilities, decreasing order of potentials of elements, free energies, and changing crystal structure of the minerals as they are deposited. R. L. Stanton has reevaluated paragenetic criteria as applied to certain stratiform sulfide ores in sedimentary and metamorphic rocks. He proposes that the textures of such ores do not represent sequences of deposition but are the result of surface energy requirements

during grain growth, or annealing of deformed minerals. To explain mineral paragenesis more satisfactorily, many additional experiments must be made to determine phase relations at different temperatures and pressures. See DEPOSITIONAL SYSTEMS AND ENVIRONMENTS; MINERAL.

**Mineralogenetic provinces and epochs.** Mineral deposits are not uniformly distributed in the Earth's crust nor did they all form at the same time. In certain regions conditions were favorable for the concentration of useful minerals. These regions are termed mineralogenetic provinces and they contain broadly similar types of deposits, or deposits with different mineral assemblages that appear to be genetically related. The time during which these deposits formed constitutes a mineralogenetic epoch; such epochs differ in duration, but in general they cover a long time interval that is not sharply defined. Certain provinces contain mineral deposits of more than one epoch.

During diastrophic periods in the Earth's history mountain formation was accompanied by plutonic and volcanic activity and by mineralization of magmatic, pegmatitic, hydrothermal and metamorphic types. During the quieter periods, and in regions where diastrophism was milder, deposits formed by processes of sedimentation, weathering, evaporation, supergene enrichment, and mechanical action.

During the 1960s numerous studies were made of the regional distribution of mineral deposits associated with long subsiding belts of sediments, or geosynclines, and with platform areas of relatively thin sediments adjoining the thick geosynclinal wedge. Geosynclinal areas commonly suffer folding and later uplift and become the sites of complex mountain ranges. It has been proposed that the outer troughs and bordering deep-seated faults contain ore deposits of subcrustal origin, the inner uplifts contain deposits of crustal origin, and the platforms contain ores derived from subcrustal and nonmagmatic platform mantle rocks. V. I. Smirnov and others have summarized available information on types of mineral deposits characteristic of the processes most active during the evolutionary stages of geosynclinal and platform regions. In the early pre-folding stage of subsidence, subcrustal juvenile basaltic sources of ore fluids prevail, and the characteristic metals are Cr, titanomagnetite, Pt metals, skarn Fe and Cu, and deposits of pyritic Cu and Fe and Mn. In the folding episode, rocks of the geosyncline are melted to produce magma from which ore components are extracted or leached by postmagmatic fluids. The most typical ores of this stage are Sn, W, Be, Ni, Ta, and various polymetallic deposits. The late stage is characterized by ore deposits associated with igneous rocks and other deposits with no apparent relationship to igneous rocks. Smirnov believes these ores originated by the combined effect of subcrustal, crustal, and nonmagmatic sources of ore material. Typical metals of this stage include Pb, Zn, Cu, Mo, Sn, Bi, Au, Ag, Sb, and Hg. In the tectonically activated platform areas, deposits of Cu-Ni sulfides, diamonds, various magmatic and pegmatitic deposits, and hydrother-

mal ores of nonferrous, precious, and rare metals are found. In addition, there are nonmagmatic deposits of Pb and Zn. Some ore material is believed to be both subcrustal and nonmagmatic in origin. Relative proportions of types of mineralization differ from one region to another.

The relationship between mineral deposition and large-scale crustal movements permits a grouping of mineralogenetic provinces by major tectonic features of the continents such as mountain belts, stable regions, and Precambrian shields. The Precambrian shield areas of the world contain the Lake Superior, Kiruna, and Venezuelan iron provinces, the gold provinces of Kirkland Lake and Porcupine in Canada, the gold-uranium ores of South Africa, the gold deposits of western Australia, and the base metals of central Australia. In the more stable regions are the metalliferous lead-zinc province of the Mississippi Valley and provinces of salt and gypsum, iron, coal, and petroleum in different parts of the world. The mountain belts are the location of many diverse kinds of mineral provinces such as the gold-quartz provinces of the Coast Range and the Sierra Nevadas, various silver-lead-zinc provinces of the western United States, the Andes, and elsewhere, and numerous base-metal provinces in the Americas, Africa, Australia, and Europe.

**Localization of mineral deposits.** The foregoing discussion has shown that mineral deposits are localized by geologic features in various regions and at different times. Major mineralized districts within the shield areas and mountain belts are often localized in the upper parts of elongate plutonic bodies. Specific ores tend to occur in particular kinds of rocks. Thus tin, tungsten, and molybdenum are found in granitic rocks, and nickel, chromite, and platinum occur in basic igneous rocks. In certain regions mineral deposits are concentrated around plateau margins. Tropical climates favor the formation of residual manganese and bauxite deposits, whereas arid and semiarid climates favor the development of thick zones of supergene copper ores. Major mineralized districts are also localized by structural features such as faults, folds, contacts, and intersections of superimposed orogenic belts. The location of individual deposits is commonly controlled by unconformities, structural features, the physical or chemical characteristics of the host rock (Fig. 6), topographic features, basins of deposition, ground-water action, or

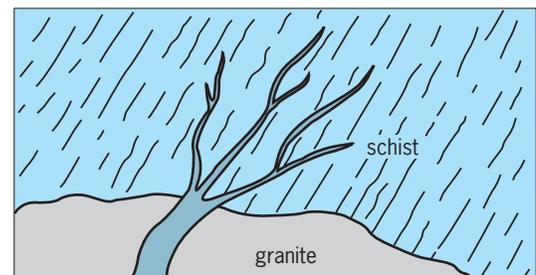


Fig. 6. Strong vein in granite dividing into stringers upon entering schist.

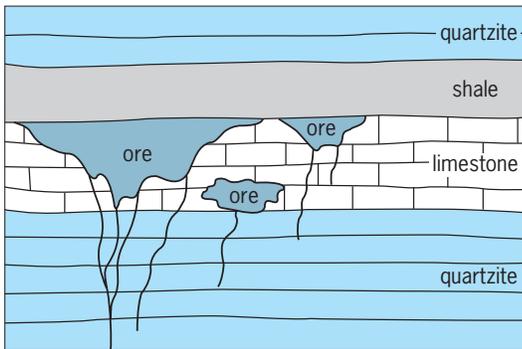


Fig. 7. Ore in limestone beneath impervious shale.

by restriction to certain favorable beds (Fig. 7). See PLUTON.

**Source and nature of ore fluids.** Widely divergent views have been expressed as to the original source and mode of transport of mineral deposits. Each view has certain advantages when applied to specific types of deposits. However, the complex nature of some mineralizations and the highly diverse physicochemical environments in which mineral deposits form make it impossible to select one theory to account for the source of all ore-forming materials.

According to one view, the source of the ore material was a juvenile subcrustal basaltic magma from which mineral deposits crystallized by simple crystallization, or in some cases were concentrated by differentiation. Most of the ore deposits associated with such magmas show a close spatial relationship to the enclosing igneous rocks and are similar in composition from one province to another. The exceptions to this are certain pyritic copper and skarn ores that apparently were introduced into sedimentary rocks and are now removed from the postulated source magma.

Another hypothesis holds that many ore deposits associated with granitic rocks were derived from magmas generated by remelting of deep-seated sedimentary rocks, followed by movement of the magma into higher levels of the Earth's crust. As an end product of crystallization and differentiation, an ore fluid was produced containing concentrations of metals originally present in the magma. Commonly, such deposits are confined to the apical portions of granitic plutons that have been altered by postmagmatic fluids. An increasing number of geologists ascribe to the view that the ore material was removed from the solidified magma by these late-stage fluids. Such ore deposits are complex, and their composition, dependent in part on the composition of the remelted rocks, is diverse and variable from one region to another. For certain ores associated with major deep-seated faults or with intersections of extensive fault or fissure systems, an ore source in deeper, subcrustal, regions has been advocated.

Circulation of surface waters may have removed metals from the host rocks and deposited them in available openings; this is the lateral secretion theory. The metals were carried either by cool surface

waters or by such waters that moved downward, became heated by contact with hot rocks at depth, and then rose and deposited their dissolved material.

As sediments are compacted and lithified, huge volumes of water may be expelled. It has been suggested that the ore-forming fluid in some sedimentary deposits contains metals that were in solution before the sediment was buried, plus metals acquired during diagenesis of the sediments. For certain ores with colloform textures, it is believed that movement took place as a finely divided suspension, and that the ore minerals initially precipitated as gels. The metals could have been held as adsorbates on clays and other colloids and then released for concentration during later crystallization of the colloids. Crystallization would exclude the metals as finely divided material which could combine with released water and move to places favorable for precipitation.

A considerable amount of experimental work has been done on the geochemistry of ore fluids in an attempt to determine the source, nature, solubility, transport, and deposition of these fluids. Studies of metal-bearing saline waters and of thermal waters and brines in igneous, sedimentary, and metamorphic rocks have also contributed to the knowledge of this complex subject. D. E. White has analyzed and summarized these studies, and he stresses that four mutually interdependent factors must be considered: a source of the ore constituents, the dissolving of these constituents in the hydrous phase, migration of the ore-bearing fluid, and the selective precipitation of the ore constituents in favorable environments. The ore-bearing fluids are Na-Ca-Cl brines that may form by magmatic or connate processes, solution of evaporates by dilute water, or membrane concentration of dilute meteoric water.

During regional metamorphism large quantities of hydrothermal fluids may be released from rocks in deep orogenic zones. These fluids remove metals and other minerals from the country rock and redeposit them at higher levels along favorable structures. Elements may also move by diffusion along chemical, thermal, and pressure gradients.

A number of the famous mineralized districts of the world that have characteristics of both epigenetic and syngenetic deposits have been modified by later metamorphism, thereby further obscuring their origin. In some of these districts the fissure and joint systems in the rocks reflect the pattern in deeper-seated rocks. H. Schneiderhohn has suggested that repeated rejuvenation of these systems by tectonic movements, accompanied by the dissolving action of thermal waters on old ore deposits in depth, would result in upward movement and reprecipitation of metals in higher formations; Schneiderhohn calls these deposits secondary hydrothermal ores. Elsewhere old folded rocks and ore deposits have been greatly deformed, and the ores taken into solution and transported to higher and younger strata; such deposits Schneiderhohn terms regenerated ores. Controversy centers around suitable criteria for epigenetic and syngenetic deposits, the problems of solubility of metals in thermal

waters, their transport over long distances, and whether such rejuvenated and regenerated ores would be dispersed or concentrated by the processes envisaged by Schneiderhohn. A. F. Hagner

**Mineral and chemical composition.** The common minerals of hydrothermal deposits (Table 3) are sulfides, sulfo salts, oxides, carbonates, silicates, and native elements, although sulfates, a fluoride, tungstates, arsenides, tellurides, selenides, and others are by no means rare. Many minor elements which seldom occur in sufficient abundance to form discrete minerals of their own may substitute for the major elements of the minerals and thus be recovered as by-products. For example (as shown in Table 3), the ore mineral of cadmium, indium, and gallium is sphalerite; the major ore mineral of silver and thallium is galena; and pyrite is sometimes an ore of cobalt. See ELEMENTS, GEOCHEMICAL DISTRIBUTION OF.

Ore deposits consist, in essence, of exceptional concentrations of given elements over that com-

monly occurring in rocks. The degree of concentration needed to constitute ore varies widely, as shown in Table 4, and is a complex function of many economic and sometimes political variables. The quantity of these elements in the total known or reasonably expected ore bodies in the world is infinitesimal when compared with the total amounts in the crust of the Earth. Thus, each and every cubic mile of ordinary rocks in the crust of the Earth contains enough of each ore element to make large deposits (Table 4). Although there is a large number of geologic situations that are apparently favorable, only a few of them contain significant amounts of ore. Thus, it is evident that the processes leading to concentration must be the exception and not the rule, and obviously any understanding or knowledge of these processes should aid in the discovery of further deposits.

Each step in the process of ore formation must be examined carefully if this sporadic occurrence of ore is to be placed on a rational basis. In order for ores to form, there must be a source for the metal,

TABLE 3. Some common primary minerals of hydrothermal ore deposits

Element	Common minerals	Idealized formulas	Significant minor elements occurring in these minerals, underlined where economically important
Iron	Hematite	Fe <sub>2</sub> O <sub>3</sub>	
	Magnetite	Fe <sub>3</sub> O <sub>4</sub>	<u>Mn</u>
	Pyrite	FeS <sub>2</sub>	<u>Au</u> ,* <u>Co</u> , Ni
	Pyrrhotite	Fe <sub>1-x</sub> S	<u>Ni</u> , Co
	Siderite	FeCO <sub>3</sub>	Mn, Ca, Mg
	Arsenopyrite	FeAsS	Sb, Co, Ni
Copper	Chalcopyrite	CuFeS <sub>2</sub>	<u>Ag</u> , Mn, <u>Se</u>
	Bornite	Cu <sub>5</sub> FeS <sub>4</sub>	
	Chalcocite	Cu <sub>2</sub> S	<u>Ag</u>
	Enargite	Cu <sub>3</sub> AsS <sub>4</sub>	<u>Ag</u> , Sb
	Tetrahedrite	Cu <sub>12</sub> Sb <sub>4</sub> S <sub>13</sub>	<u>Ag</u> , Fe, Zn, Hg, As
Zinc	Sphalerite	ZnS	Fe, Mn, <u>Cd</u> , Cu, <u>Ga</u> , <u>Ge</u> , Sn, <u>In</u>
Lead	Galena	PbS	<u>Ag</u> , Bi, As, <u>Tl</u> , Sn, Se, Sb
Bismuth	Native bismuth	Bi	
	Bismuthinite	Bi <sub>2</sub> S <sub>3</sub>	
Silver	Native silver	Ag	<u>Au</u>
	Argentite	Ag <sub>2</sub> S	
Gold	Various sulfo salts		
	Native gold	Au	<u>Ag</u> , Cu
	Various tellurides of gold and silver		
Mercury	Cinnabar	HgS	
Tin	Cassiterite	SnO <sub>2</sub>	
Uranium	Uraninite	UO <sub>2</sub>	<u>Ra</u> , Th, Pb
Cobalt	Cobaltite	CoAsS	
	Smaltite	CoAs <sub>2</sub>	
Nickel	Pentlandite	(Fe, Ni) <sub>9</sub> S <sub>8</sub>	
Tungsten	Scheelite	CaWO <sub>4</sub>	Mo
	Wolframite	(Fe, Mn)WO <sub>4</sub>	Mo
Molybdenum	Molybdenite	MoS <sub>2</sub>	Re
Manganese	Rhodochrosite	MnCO <sub>3</sub>	Fe, Mg, Ca
	Rhodonite	MnSiO <sub>3</sub>	Ca
Others	Calcite	CaCO <sub>3</sub>	Mn
	Dolomite (and ankerite)	CaCO <sub>3</sub> · MgCO <sub>3</sub>	Fe, Mn
	Barite	BaSO <sub>4</sub>	
	Fluorite	CaF <sub>2</sub>	
	Quartz	SiO <sub>2</sub>	
	Sericite, chlorite, feldspars, clays, and various other silicates		

\*Intimately associated as minute particles of metallic gold, but not in the crystal structure of the pyrite.

TABLE 4. Approximate concentration of ore elements in Earth's crust and in ores

Element	Approximate concentration in average igneous rocks, %	Tons/mi <sup>3</sup> rock (metric ton/km <sup>3</sup> )	Approximate concentration in ores, %	Concentration factor to make ore
Iron	5.0	560,000,000 (120,000,000)	50	10
Copper	0.007	790,000 (170,000)	0.5–5	70–700
Zinc	0.013	1,500,000 (330,000)	1.3–13	100–1000
Lead	0.0016	180,000 (39,000)	1.6–16	1000–10,000
Tin	0.004	450,000 (99,000)	0.01*–1	2.5–250
Silver	0.00001	1,100 (240)	0.05	5000
Gold	0.0000005	56 (12)	0.0000015*–0.01	3–2000
Uranium	0.0002	22,000 (4,800)	0.2	1000
Tungsten	0.003	340,000 (74,000)	0.5	170
Molybdenum	0.001	110,000 (24,000)	0.6	600

\*Placer deposits.

a medium in which it may be transported, a driving force to move this medium, a “plumbing system” through which it may move, and a cause of precipitation of the pre elements as an ore body. These inter-related requirements are discussed below in terms of the origin of the hydrothermal fluid, its chemical properties, and the mechanisms by which it may carry and deposit ore elements.

**Source of metals.** It is not easy to determine the source for the metals in hydrothermal ore deposits because, as shown above, they exist everywhere in such quantities that even highly inefficient processes could be adequate to extract enough material to form large deposits.

*Fluids associated with igneous intrusion.* In many deposits there is evidence that ore formation was related to the intrusion of igneous rocks nearby, but in many other deposits intensive search has failed to reveal any such association. Because the crystal structures of the bulk of the minerals (mostly silicates) crystallizing in igneous rocks are such that the common ore elements, such as copper, lead, and zinc, do not fit readily, these elements are concentrated into the residual liquids, along with H<sub>2</sub>O, CO<sub>2</sub>, H<sub>2</sub>S, and other substances. These hot, water-rich fluids, remaining after the bulk of the magma has crystallized, are the hydrothermal fluids which move outward and upward to areas of lower pressure in the surrounding rocks, where part or all of their contained metals are precipitated as ores. A more detailed discussion of the composition of these fluids is presented below.

*Fluids obtained from diagenetic and metamorphic processes.* Fluids of composition similar to the above also could be obtained from diagenetic and metamorphic processes. When porous, water-saturated sediments containing the usual amounts of hydrous and carbonate minerals are transformed into essentially nonhydrous, nonporous metamorphic rocks, great quantities of water and carbon dioxide must be driven off. Thus, each cubic mile of average shale must lose about  $3 \times 10^9$  tons of water (each cubic kilometer, about  $6.5 \times 10^8$  metric tons) and may lose large amounts of carbon dioxide on metamorphism to gneiss. The great bulk of the water presumably

comes off as connate water (entrapped at time of rock deposition) under conditions of fairly low temperature. In many respects this water has the same seawater composition as it had to start with. However, as metamorphism proceeds, accompanied by slow thermal buildup from heat flow from the Earth's interior and from radioactivity, the last fluids are given off at higher temperatures and are richer in CO<sub>2</sub> and other substances. These fluids would have considerably greater solvent power and can be expected to be similar to those coming from cooling igneous rocks.

*Role of surface and other circulating waters.* It is very likely that the existence of a mass of hot rock under the surface would result in heating and circulation of meteoric water (from rain and snow) and connate water. The possible role of these moving waters in dissolving ore elements from the porous sedimentary country rocks through which they may pass laterally and in later depositing them as ore bodies has been much discussed. The waters may actually contribute ore or gangue minerals in some deposits. The test of this theory of lateral secretion on the basis of precise analyses of the average country rocks around an ore body would involve an exceedingly difficult sampling job. It also would require analytical precision far better than is now feasible for most elements, as each part per million uncertainty in the concentration of an element in a cubic mile of rock represents about 10,000 tons of the element or  $1 \times 10^6$  tons of 1% ore.

*Movement of ore-forming fluids.* In addition to the high vapor pressures of volatile-rich fluids acting as a driving force to push them out into the surrounding country rocks and to the surface, there may well be additional pressures from orogenic or mountain-building forces. When a silicate magma has an appreciable percentage of liquid and is subjected to orogenic forces, it moves en masse to areas of lower pressure (it is intruded into other rocks). But if the magma has crystallized 99% or more of its bulk as solid crystals and has only a very small amount of water-rich fluid present as thin films between the grains, and then is squeezed, this fluid may be the only part sufficiently mobile to move toward regions

of lower pressure. (If the residual fluid, containing the ore elements, stays in the rock, it reacts with the early formed, largely anhydrous minerals of the rock to form new hydrated ones, such as sericite, epidote, amphibole, and chlorite, and its ore elements precipitate as minute disseminated specks and films along the silicate grain boundaries.)

The ore-bearing fluid leaves the source through a system comprising joints, faults, porous volcanic plugs, or other avenues. As the fluid leaves the source, it moves some appreciable but generally unknown distance laterally, vertically, or both, and finally reaches the site of deposition. This system of channels is of utmost importance in the process of ore formation.

*Localization of mineral deposits.* It is stated frequently that ore deposits are geologic accidents; yet there are reasons, however abstruse, for the localization of a mineral deposit in a particular spot. One reason for localization is mere proximity to the source of the ore-forming fluids, as in rocks adjacent to an area of igneous activity or near a major fracture system which may provide plumbing for solutions ascending from unknown depths. Zones of shattering are favored locales for mineralization since these provide plumbing and offer the best possibility for the ore solution to react with wall rock, mix with other waters, and expand and cool, all of which may promote precipitation. Some types of rock, particularly limestone and dolomite, are especially susceptible to replacement and thus often are mineralized preferentially. The chemical or physical properties which cause a rock to be favored by the replacing solutions often are extremely subtle and certainly not understood fully.

*Zoning and paragenesis.* Mineral deposits frequently show evidence of systematic spatial and temporal changes in metal content and mineralogy that are sufficiently consistent from deposit to deposit to warrant special mention under the terms zoning and paragenesis. Zoning may be on any scale, though the range is commonly on the order of a few hundred to a few thousand feet, and may have either lateral or vertical development. In mining districts, such as Butte, Montana, or Cornwall, England, where zoning is unusually well developed, there is a peripheral zone of manganese minerals grading inward through successive, overlapping silver-lead, zinc, and copper zones (and in the case of Cornwall, tungsten, and finally tin). The same sequence of zones appears in many deposits localized about intrusive rocks, suggesting strongly that the tin and tungsten are deposited first from the outward-moving hydrothermal solutions and that the copper, zinc, lead, and silver were deposited successively as the solutions expanded and cooled. In other districts the occurrences of mercury and antimony deposits suggest that their zonal position may be peripheral to that of silver or manganese. The paragenesis, or the sequence of deposition of minerals at a single place, as interpreted from the textural relations of the minerals, follows the same general pattern as the zoning, with the tin and tungsten early and the lead and

silver late. With both zoning and paragenesis there are sometimes reversals in the relative position of adjacent zones, and these are usually explained as successive generations of mineralization. Some metals, such as iron, arsenic, and gold, tend to be distributed through all of the zones, whereas others, such as antimony, tend to be restricted to a single position.

The sequence of sulfide minerals observed in zoning and paragenesis matches in detail the relative abilities of the heavy metals to form complex ions in solution. This observation strongly supports the hypothesis developed later that most ore transport occurs through the mechanism of complex ions, since no other geologically feasible property of the ore metals or minerals can explain the zoning relations.

**Environment of ore deposition.** Important aspects of the environment of ore deposition include the temperature, pressure, nature, and composition of the fluid from which ores were precipitated.

*Temperatures.* Although there is no geological thermometer that is completely unambiguous as to the temperatures of deposition of ores, there is a surprising number of different methods for estimating the temperatures that prevailed during events long since past that have been applied to ores with reasonably consistent results. Those ore deposits which had long been considered to have formed at high temperatures give evidence of formation in the range of 500–600°C (930–1100°F), or possibly even higher. Those that were thought to be low-temperature deposits show temperatures of formation that are in the vicinity of 100°C (212°F) or even less, and the bulk of the deposits lie between these extremes. *See GEOLOGIC THERMOMETRY.*

*Pressures.* It would be useful to know the total hydrostatic pressure of the fluids during ore formation. Most of the phenomena used for determination of the temperatures of ore deposition are also pressure-dependent, and so either an estimate of the correction for pressure must be made, or two independent methods must be used to solve for the two variables.

Pressures vary widely from nearly atmospheric in hot springs to several thousand atmospheres in deposits formed at great depth. Maximum reasonable pressures are considered to be on the order of that provided by the overlying rock; conversely, the minimum reasonable pressures are considered to be about equal to that of a column of fluid open to the surface. Pressures therefore range from approximately 500 to 1500 lb/in.<sup>2</sup> per 1000 ft (10–30 MPa per 1000 m) of depth at the time of mineralization. *See HIGH-PRESSURE MINERAL SYNTHESIS.*

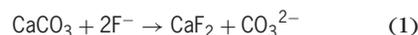
*Evidence of composition.* Geologists generally concede that most ore-forming fluids are essentially hot water or dense supercritical steam in which are dissolved various substances including the ore elements. There are three lines of evidence bearing on the composition of this fluid. These are fluid inclusions in minerals, thermal springs and fumaroles, and the mineral assemblage of the deposit and its associated alteration halos.

1. Fluid inclusions in minerals. Very small amounts of fluid are trapped in minute fluid-filled inclusions during the growth of many ore and gangue minerals in veins, and these inclusions have been studied intensively for evidence of temperature and composition. Although the relative amounts may vary widely, these fluids will have 5–25 or even more weight percent soluble salts, such as chlorides of Na, K, and Ca, plus highly variable amounts of carbonate, sulfate, and other anions. Some show liquid CO<sub>2</sub> or hydrocarbons as separate phases in addition to the aqueous solution. A few show detectable amounts of H<sub>2</sub>S and minor amounts of many other substances. After losing some CO<sub>2</sub> and H<sub>2</sub>S through release of pressure and oxidation when the inclusions are opened, the solutions are within 2 or 3 pH units of neutral. There is little evidence of sizable quantities (>1 g/liter or 0.13 oz/gal) of the ore metals in these solutions, and the evidence indicates that the concentrations of the ore elements must generally be very low (<0.1 g/liter or 0.013 oz/gal). Even if the concentrations were in the range of 0.1 g/liter, there should be analytical evidence in the fluid inclusion studies, but this is lacking. In addition, if fluids of such composition were trapped in fluid inclusions in transparent minerals and on cooling precipitated even a fraction of their metal content as opaque sulfides, these should be visible (under the microscope) within the inclusions, but none are seen. If the concentrations of ore elements are much less than 0.001 g/liter (0.00013 oz/gal), the volume of fluids that must be moved through a vein to form an ore body becomes geologically improbable.

2. Thermal springs and fumaroles. These provide the closest approach to a direct look at the processes of ore deposition as some ore and gangue minerals form within the range of direct observation. The solutions from these springs give diluted and possibly contaminated, partly oxidized and partly devolatilized samples of the sort of fluid that presumably forms ore bodies at greater depths. Isotopic studies show that the solutions have been diluted by local meteoric water until less than 5% (if any) of the fluid emitted at the surface is of deep-seated origin. The compositions of these thermal springs, after correction for such dilution, are in good agreement with the data from fluid inclusions.

3. Mineral assemblage. The assemblage of minerals that occurs within a deposit provides a great deal of information about the chemical nature of the fluid from which the ores were precipitated. There are a great number of stable inorganic compounds of the heavy metals known, yet unaltered ore deposits contain only a relatively small number of minerals. For example, lead fluoride, lead chloride, lead carbonate, lead sulfate, lead oxide, lead sulfide, and many others are known stable compounds of lead, yet of these, primary ore deposits contain only the sulfide (galena). Some elements, such as calcium, which occur in combination with several types of anions, for example, the carbonate, fluoride, sulfate, and numerous silicates, are found with the ore minerals. A quantitative approach to the compositional problem

may be made by considering such reactions as shown in (1). The equilibrium constant for this reaction is

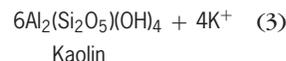
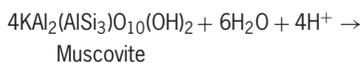


$(\text{CO}_3^{2-})/(\text{F}^-)^2 = 10^{1.4}$  at 25°C (77°F). Thus when calcite and fluorite are in equilibrium, the requirements for the constant are met, and the  $(\text{CO}_3^{2-})/(\text{F}^-)^2$  ratio is known. A large number of such equations can be evaluated and from comparison with the mineral assemblage known to occur in ores, limits on the possible variation of the composition of the ore-forming fluid may be estimated. Unfortunately, calculations of this sort involving ionic equilibria are limited to fairly low temperatures (less than 100–200°C or 212–390°F) since there are few reliable thermodynamic data on ionic species at high temperature. At any temperature, reactions such as shown in (2) can be



used to evaluate or place limits on the possible variation of the chemical potential of some components in the ore-forming fluid. See IONIC EQUILIBRIUM; SULFIDE PHASE EQUILIBRIA.

The composition of the ore fluid tends to become adjusted chemically by interaction with the rocks with which it comes in contact, and these changes may well contribute to the precipitation of the ore minerals. Thus, the K<sup>+</sup>/H<sup>+</sup> ratio may be controlled by such reactions as (3), where the equilibrium constant has the form shown in Eq. (4). Likewise, the



$$K = \frac{(\text{K}^+)^4}{(\text{H}_2\text{O})^6 \cdot (\text{H}^+)^4} \quad (4)$$

quantitatively small but nevertheless important partial pressures of sulfur and oxygen may be governed by such reactions as (5). Such changes in the wall



rock come under the general heading of wall-rock alteration and may be of many types, only a few of the more common of which are mentioned below.

High-temperature alteration of limestones usually results in the formation of water-poor calcium silicates, such as garnet, pyroxenes, idocrase, and tremolite, and the resulting rock is termed skarn. At lower temperatures in the same types of rock, dolomitization and silicification are the predominant forms of alteration, because the partial pressure of CO<sub>2</sub> is too high to permit calcium silicate to form. See SILICATE PHASE EQUILIBRIA.

At high temperatures in igneous and metamorphic rocks near granite in composition, the solutions are approximately in equilibrium with the primary rock-forming minerals, and thus there is little alteration except development of sericite and occasionally topaz and tourmaline. At lower temperatures, the

characteristic sequence of alteration from fresh rock toward the vein is first an argillic zone, then a sericitic zone, and finally a silicified zone bordering the vein.

*Summary.* Summarizing the environment of ore deposition, there are various lines of evidence to show that most hydrothermal ore deposits were formed at temperatures of 100–600°C (212–1100°F) and at pressures ranging from nearly atmospheric to several thousand atmospheres. The solutions were dominantly aqueous and were fairly concentrated in sodium chloride and potassium chloride; however, they were relatively dilute in terms of the ore metals.

**Mechanisms of ore transport and deposition.** The ore minerals, principally the sulfides, are extremely insoluble in pure water at high temperatures as well as low; the solubility products are so low, in fact, that literally oceans of water would be required to transport the metal for even a small ore body. Thus, it is not easy to explain the mechanism whereby the minerals are solubilized to the extent necessary for ore transport.

Crystals of ore and gangue minerals frequently exhibit evidence of repeated partial re-solution (or leaching) and regrowth. This demonstrates that the process of ore formation may, at least in some instances, be reversible. In such cases studies of artificial systems at equilibrium are applicable.

The re-solution of ore minerals is important in another connection. Some geologists have advocated colloidal solutions, or sols, as an alternative to true solutions for ore transport. This was based on the belief, now known to be generally false, that colloform textures in ore minerals are a result of original deposition as a colloidal gel. Colloidal solutions were attractive also because they permitted ore metal concentrations—even in the presence of sulfide—many orders of magnitude higher than true solutions. The re-solution of ore minerals precludes the process of colloidal ore transport, as colloidal solutions are supersaturated and therefore cannot redissolve a crystal of the dispersed phase. *See* COLLOID.

In addition to the fact that the absolute solubilities, calculated from the solubility products, are extremely low, the relative solubilities of the sulfides are radically different. For example, according to the solubility products, FeS is many, many times more soluble than PbS (about  $10^{10}$  times at 25°C or 77°F), yet the two minerals occur together in ore deposits and behave as if galena were slightly more soluble than pyrrhotite. From this and other lines of evidence, it appears necessary to conclude that the solubilities of the various contemporaneous minerals in a given deposit could not have differed among themselves by more than a few orders of magnitude. *See* SOLUBILITY PRODUCT CONSTANT.

The only geologically and chemically feasible mechanism by which these solubilities may be equalized approximately is the formation of complex ions of the heavy metals. Such complexes can increase the solubilities of heavy metals tremendously. As an example, the activity (thermodynamic concentration)

of  $\text{Hg}^{2+}$  in a solution saturated with HgS (cinnabar) and  $\text{H}_2\text{S}$  at 25°C (77°F), 1 atm ( $10^5$  Pa) pressure, and pH 8, is only about  $10^{-47}$  mole/liter, representing a concentration much less than 1 atom of mercury in a volume of water equal to the entire volume of the oceans of the world. However, in the same solution is formed a very stable sulfide complex of mercury,  $\text{HgS}_2^{2-}$ , which increases the total concentration of mercury in solution by the impressive factor of about  $10^{42}$ , giving a concentration on the order of 0.001 g/liter (0.00013 oz/gal). Not only does complex formation provide a means to achieve adequate solubility for ore transport, but the relative tendency for metals to form certain types of complexes matches in detail the commonly observed zoning and paragenetic sequences mentioned previously. The metals whose sulfides are the least soluble tend to form the most stable complexes, and metals whose minerals are comparatively soluble form weaker complexes. *See* COORDINATION COMPLEXES.

There are many kinds of complexing ions or molecules (ligands) of possible geologic importance; a few of the more significant are sulfide ( $\text{S}^{2-}$ ), hydrosulfide ( $\text{HS}^-$ ), chloride ( $\text{Cl}^-$ ), polysulfides ( $\text{S}_x^{2-}$ ), thiosulfate ( $\text{S}_2\text{O}_3^{2-}$ ), sulfate ( $\text{SO}_4^{2-}$ ), and carbonate ( $\text{CO}_3^{2-}$ ), with the first three being most frequently considered. One of the major unsolved problems concerns the behavior of sulfur: What is its oxidation state and concentration relative to metals? If solutions were rich in reduced sulfur species, then the sulfide or hydrosulfide complexes would be dominant. On the other hand, solutions poor in reduced sulfur may transport the metals as chloride complexes.

The precipitation of minerals from complexed solutions takes place either by shifts in equilibrium caused by changing (usually cooling) temperature or by a decrease in the concentration of the ligand, thereby reducing the ability of the solution to carry the metals. This latter alternative can take place in several ways, for example, by reaction with wall rock, by mixing with other solutions, or by formation of a gas phase through the loss of pressure. *See* PRECIPITATION (CHEMISTRY).

**Oxidation and secondary enrichment.** When ore deposits are exposed at the surface, they are placed in an environment quite different from that in which they were formed, and the character of the deposit is changed through the processes of oxidation and weathering. The sulfides give way to oxides, sulfates, carbonates, and other compounds which are more or less soluble and tend to be leached away, leaving a barren gossan of insoluble siliceous iron and manganese oxides. Some minerals, such as cassiterite and native gold, may leach away at a less rapid rate than does the surrounding material; thus they are concentrated as a surficial residuum.

Where the country rock is relatively inert to the acid solutions generated by the oxidizing sulfides, as in the case of quartzites and some hydrothermally altered rocks, copper and especially zinc are leached away readily; lead and silver may be retained temporarily in the oxidized zone as the carbonate or

sulfate, and the chloride or native metal, respectively; but eventually these too are dissolved away. The various metallic ions are carried downward until they reach unoxidized sulfides in the vicinity of the water table, where the solutions interact with these sulfides to form a new series of supergene sulfide minerals. Copper sulfide is the least soluble sulfide of the base and ferrous metals in the solution, and hence the zone of supergene sulfide enrichment is predominantly a copper sulfide zone with occasional rich concentrations of silver. Zinc nearly always remains in solution and is lost in the ground water.

In reactive wall rocks, such as limestones, reaction with the wall rock prevents the solutions from becoming acid enough for large amounts of metal to be removed in solution; the base metals are retained almost in place as carbonates, sulfates, oxides, and halides, and there is no appreciable sulfide enrichment.

The behavior of some elements is governed by the availability of other materials. Thus, for example, uranium is readily leached from the oxidized zone in many deposits; however, when the oxidizing solutions contain even very small amounts of potassium vanadate, the extremely insoluble mineral carnotite precipitates and uranium is immobilized. Highly soluble materials, such as uranium in the absence of chemicals that precipitate it, may be temporarily fixed in the oxidized zone by adsorption on colloidal materials such as freshly precipitated ferric oxides.

**Trends in investigation.** There has been a great increase in the degree to which the experimental methods and principles of physical chemistry have been applied to aid in understanding the processes by which ores have formed, and this approach can be expected to be even more fruitful in the future. Several avenues appear promising and are under active investigation in numerous laboratories. Among these are the following:

1. Phase equilibrium studies of both natural and synthetic ore and gangue minerals.
2. Distribution coefficients for trace elements between coexisting phases, and between various forms on the same crystal.
3. Experimental solubility studies in dominantly aqueous solutions.
4. Studies of the composition and origin of thermal spring waters and fluid inclusions in minerals.
5. Thermodynamic properties of minerals.
6. Isotopic fractionation during transportation and deposition processes.
7. Rate studies on crystal growth, habit, diffusion, reaction, and transformation, as well as studies of sluggish homogeneous reactions, such as the reduction of sulfate.
8. Crystal structure determinations and crystal chemical studies of ore and gangue minerals.
9. Distribution of elements in the Earth's crust and in various rock types.
10. Detailed field studies of the relations between minerals in ore deposits. See HEAVY MINERALS; PETROLEUM GEOLOGY.

For a discussion of sensitive chemical analytical techniques used in the search for ore deposits see GEOCHEMICAL PROSPECTING

For chemical principles involved in ore deposition see GEOLOGIC THERMOMETRY; LEAD ISOTOPES (GEOCHEMISTRY); LITHOSPHERE.

Paul B. Barton, Jr.; Edwin Roedder

**Bibliography.** J. W. Barnes, *Ores and Minerals: Introducing Economic Geology*, 1988; R. L. Bates, *Geology of the Industrial Rocks and Minerals*, 1969; R. L. Bates, *Industrial Minerals: How They Are Found and Used*, 1988; D. Derry, *Concise World Atlas of Geology and Mineral Deposits: Non-metallic Minerals, Metallic Minerals and Energy Minerals*, 1980; A. R. Dutton, *Hydrogeology and Hydrochemical Properties of Salt-Dissolution Zones*, 1987; R. Edwards and K. Atkinson, *Ore Deposit Geology and Its Influence on Mineral Exploration*, 1986; R. M. Garrels and C. L. Christ, *Solutions, Minerals, and Equilibria*, 1965, reprint 1982; M. L. Jensen and A. M. Bateman, *Economic Mineral Deposits*, 3d rev. ed., 1981; K. B. Krauskopf, *Introduction to Geochemistry*, 3d ed., 1994; J. Parnell, H. Kucha, and P. Landais, *Bitumens in Ore Deposits*, 1993; F. Pirajano, *Hydrothermal Mineral Deposits: Principles and Fundamental Concepts for the Exploration Geologist*, 1992.

## Ore dressing

Treatment of ores to concentrate their valuable constituents (minerals) into products (concentrate) of smaller bulk, and simultaneously to collect the worthless material (gangue) into discardable waste (tailing). The fundamental operations of ore-dressing processes are the breaking apart of the associated constituents of the ore by mechanical means (severance) and the separation of the severed components (beneficiation) into concentrate and tailing, using mechanical or physical methods which do not effect substantial chemical changes.

**Severance.** Comminution is a single- or multi-stage process whereby ore is reduced from run-of-mine size to that size needed by the beneficiation process. The process is intended to produce individual particles which are either wholly mineral or wholly gangue, that is, to produce liberation. Since the mechanical forces producing fracture are not susceptible to detailed control, a class of particles containing both mineral and gangue (middling particles) are also produced. The smaller the percentage of middlings the greater the degree of liberation. Comminution is divided into crushing (down to 6- to 14-mesh) and grinding (down to micrometer sizes). Crushing is usually done in three stages: coarse crushing from run-of-mine size to 4-6 in. (10-15 cm) or coarser; intermediate crushing down to about  $\frac{1}{2}$  in. (1.27 cm); and fine crushing to  $\frac{1}{4}$  in. (0.64 cm) or less.

Screening is a method of sizing whereby graded products are produced, the individual particles in each grade being of nearly the same size. In beneficiation, screening is practiced for two reasons: as an

integral part of the separation process, for example, in jigging; and to produce a feed of such size and size range as is compatible with the applicability of the separation process. See SCREENING.

**Beneficiation.** This step consists of two fundamental operations: the determination that an individual particle is either a mineral or a gangue particle (selection); and the movement of selected particles via different paths (separation) into the concentrate and tailing products. When middling particles occur, they will either be selected according to their mineral content and then caused to report as concentrate or tailing, or be separated as a third product (middling). In the latter case, the middling is reground to achieve further liberation, and the product is fed back into the stream of material being treated.

Selection is based upon some physical or chemical property in which the mineral and gangue particles differ in kind or degree or both. Thus, in hand picking, the oldest form of beneficiation, color, luster, and shape are used to decide whether a lump of ore is predominantly mineral or gangue. Use is made of differences in other physical or chemical properties, such as specific gravity, magnetic permeability, inductive charging (electrostatic separation), surface chemical properties, bulk chemical properties, weak planes of fracture (separation by screening), and gamma-ray emission (automatic sorting of radioactive materials). See MECHANICAL SEPARATION TECHNIQUES.

Separation is achieved by subjecting each particle of the mixture to a set of forces which is usually the same irrespective of the nature of the particles excepting for the force based upon the discriminating property. This force may be present for both mineral and gangue particles but differing in magnitude, or it may be present for one type of particle and absent for the other. As a result, separation is possible, and the particles are collected as concentrate or tailing.

**Magnetic separation.** Magnetic separation utilizes the force exerted by a magnetic field upon magnetic materials to counteract partially or wholly the effect of gravity. Thus under the action of these two forces, different paths are produced for the magnetic and nonmagnetic particles. **Figure 1** shows a continuously moving endless belt B onto which are introduced the particles to be separated. The magnetic field is produced by a square-shaped bottom pole B and an upper pole A, so curved as to concentrate the lines of force as shown. When a magnetic particle comes within the magnetic field, it is attracted strongly enough to move upward against the force of gravity to the surface of an endless belt A conformed to the surface of the pole. The movement of belt A perpendicular to belt B carries the particle to a concentrate bin. The unattracted nonmagnetic particle, held by gravity, keeps moving with belt B until it falls into the tailings bin.

All substances placed in a magnetic field acquire magnetic properties. Magnetic permeability is a measure of the ease with which these properties are induced. The ratio of permeabilities may be as low as 5:1 for successful separation. However, for such com-

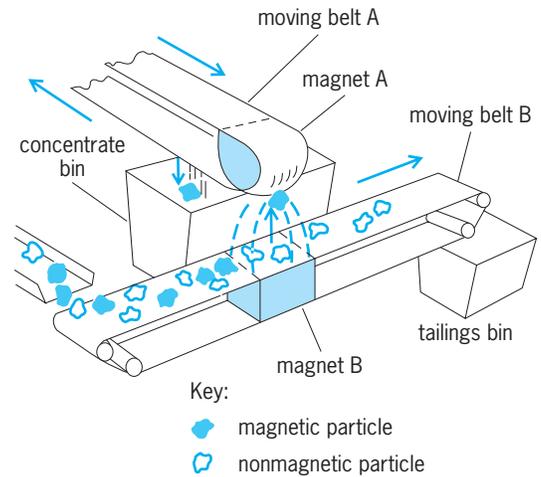


Fig. 1. Magnetic separation.

monly separated materials as magnetite and quartz it is about 110:1.

The most important practical separations are those of the iron ores. In the magnetite ores, magnetite is separated from quartz, feldspars, and the like. In the hematite and limonite ores, the ore is first roasted to convert these iron oxides partly into the magnetic oxide, and this is then separated from gangue. In the preparation of industrial minerals, magnetic separation is used to clean up iron introduced during grinding, as in the preparation of china clay, body slip, or glaze, or to remove trace magnetic minerals, such as biotite, garnet, and tourmaline, from feldspar.

**Gravity methods.** Gravity concentration is based on a discriminating force the magnitude of which varies with specific gravity. The other force that is usually operating in gravity methods is the resistance to relative motion exerted upon the particles by the fluid or semifluid medium in which separation takes place.

Jigging is a gravity method which separates mineral from gangue particles by utilizing an effective difference in settling rate through a periodically dilated bed. In **Fig. 2**, the mixture of particles (feed)

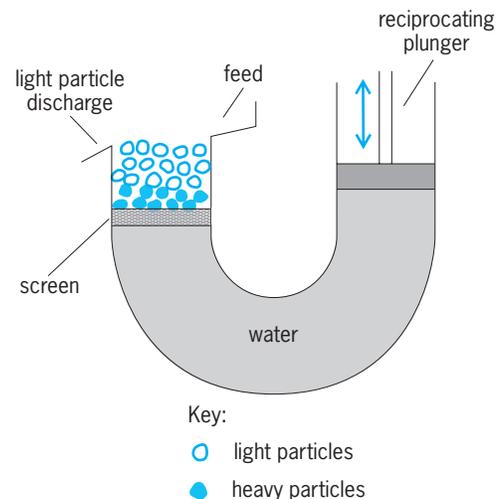
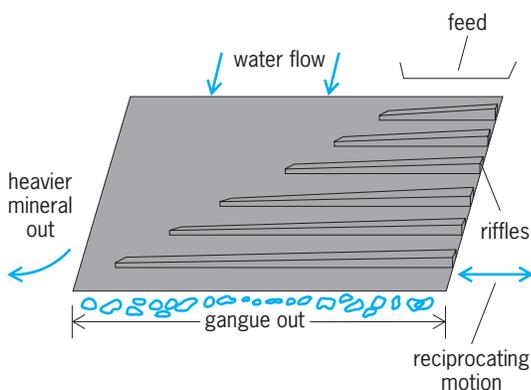


Fig. 2. Diagram of the jigging method.

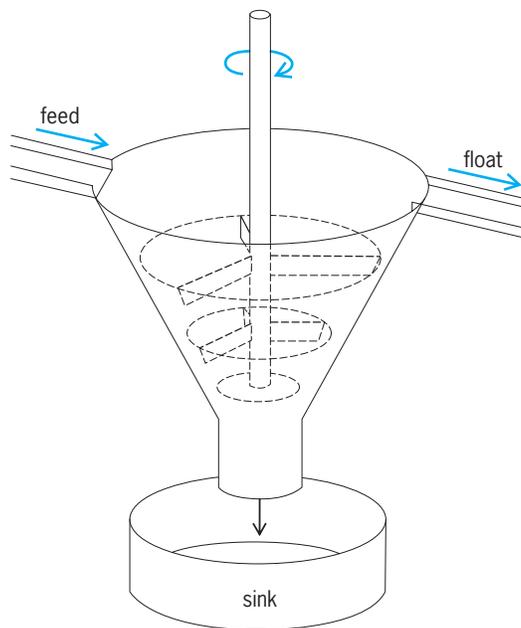
falls into the jig compartment where it is supported by the screen. Reciprocation of the plunger forces water through the screen and causes periodic dilation and contraction of the bed of particles. During the dilation heavier particles work their way to the bottom while the lighter particles remain on top and are discharged over the lip. Jigging is practiced on materials which are liberated upon being reduced to sizes ranging from 1½ in. (3.8 cm) down to several millimeters. It has been used on such diverse ores as coal, iron ores, gold, and lead ores.

Tabling is a gravity method in which the feed, introduced onto an inclined plane and reciprocated deck, moves in the direction of motion while simultaneously being washed by a water film which moves it also at right angles to the motion of the deck. In **Fig. 3**, feed enters at the top of the table, and collects within the valleys formed by the narrow cleats, or riffles, which taper in height from right to left. Under the effect of the reciprocating action, the particles stratify with the heavier particles on the bottom, and they also move from right to left. Owing to decreasing taper of the riffles, the exposed upper surface of the stratified material is acted upon by crosscurrents of water and by the tilt of the deck, as indicated by the arrow, and is moved downhill. The heavier mineral and the lighter gangue are usually collected over the edges of the deck as shown. The boundary between the heavier mineral and lighter gangue particles is roughly a linear diagonal band on the deck of the table. This diagonal band is not stationary; rather it tends to move about a mean position. In practice, therefore, a third product, the middling, is collected between the discharge edges of concentrate and gangue. If the feed to the table has been crushed or ground to produce liberation, then the middling is returned to the feed. If liberation has not been achieved, the middling is returned to the crushing-grinding section of the mill. Tables may be used to treat relatively coarse material (sand tables) or fines (slime tables), with sizes ranging from about 0.08–0.12 in. (2–3 mm) down to 0.003 in. (0.07 mm).

Sink-float separation is the simplest gravity method and is based on existing differences in specific gravity. The feed particles are introduced into a suspension, the specific gravity of which is between that of



**Fig. 3.** Diagram of the tabling method.



**Fig. 4.** Sink-float separation.

the mineral and gangue particles, with the result that particles of higher specific gravity sink while those of lower specific gravity float. In **Fig. 4** the separator is a cone equipped with a slowly operated stirrer which serves to impart a slow rotary motion to the suspension and to prevent the suspension from settling out on the walls. Feed is introduced at one point of the circumference and is slowly moved by the rotating motion of the suspension. By the time this material has reached the discharge point on the circumference, those particles whose specific gravity is greater than that of the suspension have moved down through it so that only float particles are discharged at the top. The sink particles are discharged at the bottom.

The suspension may have a specific gravity ranging from 1.3, using quartz, to 2.4, using galena. Magnetite and ferrosilicon are also used at intermediate densities. Although there is no top limit to the size of feed particles, a lower limit of about 1/8 in. (3 mm) exists with the more standard equipment.

Earliest use of sink-float was in the separation of slate from coal using quartz in suspension. A most important use is to produce coarse-size tailing which can be discarded. In this manner it is possible early in the treatment to reduce the quantity of material handled by the concentrating plant, thereby effecting a saving in the capital investment.

**Filtration.** Filtration is a method of separation based on the differences in size of the things being separated. Water, one of the things being separated, has no lower size, whereas the solid from which it is being separated has a lower size. Consequently if a barrier (filter cloth) having openings which can pass water but not the solids is provided, and if the pulp is placed on one side of the filter, filtration will take place if a pressure is exerted on the pulp. Other methods of concentration are electrostatic

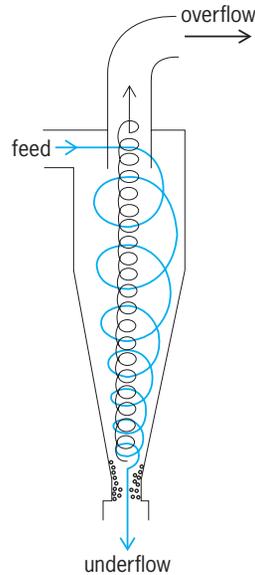


Fig. 5. Separation by cyclone classification.

separation, flotation, and leaching. See FILTRATION; FLOTATION; LEACHING.

**Cyclone classification.** Cyclone classification is a method of separating coarser from finer particles based upon the difference in magnitude of the centrifugal force to which they are subjected. The feed (a mixture of particles and water) is fed tangentially at the top of the cyclone (Fig. 5). The angular motion thus imparted to the pulp causes centrifugal forces to act on all the particles. The magnitude of this force depends upon the angular velocity of the pulp and the weight of the particle which, in turn, depends upon its size and specific gravity. This force is greatest on the heaviest particles and least on the lightest. Thus, the coarser particles move outward from the axis of rotation to the inner surface of the cyclone, whereas the finer particles stay closer to the axis of rotation. Because of the geometry of the cyclone, the pulp follows a downward, inward spiraling path with increasing angular velocity which produces increasing centrifugal forces. Thus, the coarser particles collect near the inner surface of the cyclone and move downward to the underflow discharge. The lighter particles with most of the water are carried in an upward spiraling column near the axis of rotation to the overflow discharge. By controlling the diameter of the underflow apex relative to the diameter of the overflow, it is possible to vary the size of the particles which report in overflow and underflow.

The principal use of the cyclone is to classify the discharge from a grinding mill and return the coarser particles to the feed for the grinding mill. Overflow goes to the concentrating plant. This makes it possible to avoid grinding beyond the desired degree of liberation of the valuable minerals from the gangue. Overgrinding can cause problems in later concentration processes, leading to serious losses in recovery of desirable minerals. See METALLURGY; PYROMETALLURGY, NONFERROUS.

Menelaos D. Hassialis

**Bibliography.** I. I. Inculet, *Electrostatic Mineral Separation*, 1984; A. Lynch, *Mineral Crushing and Grinding Circuits: Their Simulation, Design, and Control*, 1977; A. L. Mular and R. B. Bhappu (eds.), *Mineral Processing Plant Design*, 2d ed., 1980; B. A. Wills, *Mineral Processing Technology: An Introduction to the Practical Aspects of Ore Treatment and Mineral Recovery*, 6th ed., 1997.

## Orectolobiformes

Members of the order Orectolobiformes (carpet sharks), share the following characteristics: two dorsal fins without spines; a short snout that leaves the mouth terminal or almost terminal; and each nostril connected to the mouth by a deep groove, the anterior margin having a well-developed barbel. The classification hierarchy is:

Class Chondrichthyes  
Subclass Elasmobranchii  
Superorder Euselachii  
Order Orectolobiformes

The order consists of about 31 species in 14 genera and 7 families and is represented by at least one family in the seas of the world. Among them is the nurse shark (Family *Gingymostomatidae*). It occurs near shore in the Atlantic, Indian, and Pacific oceans and is often seen in marine aquariums. This order also includes the whale shark (Family *Rhincodontidae*), which has a large, terminal mouth, long gill rakers, and exceptionally large gill openings, characteristics indicative of a filter feeder. The whale shark reaches a length of at least 12 m (40 ft), making it the world's largest cold-blooded animal. See CHONDRICHTHYES; ELASMOBRANCHII.

Herbert Boschung

**Bibliography.** L. Compagno, M. Dando, and S. Fowler, *Sharks of the World*, Princeton Field Guides, 2005; J. S. Nelson, *Fishes of the World*, 3d ed., 1994.

## Oregano

An herb, also known as wild marjoram. The dried leaves of several species of aromatic plants are known as oregano; thus oregano is a common name for a general flavor and aroma rather than the name of a specific plant.

European (*Origanum vulgare*) and Greek (*O. heracleoticum*) oregano are both in the mint family (Lamiaceae). Mexican oregano is obtained primarily from plants of *Lippia graveolens*. These small aromatic shrubs in the verbena family grow wild in Mexico. Origanum oil used in perfumery is steam-distilled primarily from Spanish oregano, *Thymus capitatus*. See LAMIALES.

European oregano, the most cultivated oregano species, is often confused with its close relative marjoram because they look and smell similarly. Marjoram has a bittersweet flavor, while European oregano can be distinguished by its strong piquant character

and tall growth with dark, broad leaves; it is a perennial erect herb 2–3 ft tall (0.6–1 m) with pubescent stems, ovate dark green leaves, and white or purple flowers. Native to southern Europe, southwest Asia, and the Mediterranean countries, European oregano is usually found growing in the dry, rocky, calcareous soils of the mountain regions. Greece, Italy, Spain, Turkey, and the United States are the primary sources of European oregano.

Although gathered from wild plants, European oregano is widely cultivated. It is often established with transplants from stem cuttings due to its slow germination and seedling development. European oregano does well on light, slightly alkaline, and well-drained soils that are kept dry. The plants can be harvested several times a year for 4 or 5 years.

Dried oregano leaves are used as a culinary herb in meat and sausage products, salads, soups, Mexican foods, and barbeque sauces. The essential oil of oregano is used in food products, cosmetics, and liqueurs. *See* MARJORAM; SPICE AND FLAVORING.

Seth Kirby

## Organic chemistry

The study of the structure, preparation, properties, and reactions of carbon compounds. The term organic was early applied to compounds derived from plant and animal sources. These substances from living systems were usually distillable liquids or low-melting solids and were flammable, in contrast to metals, salts, and oxides from mineral sources. Until about 1830 it was held by some that organic compounds contained some special quality, or vital force. This notion was dispelled, but the term organic remained and became broadened to include carbon compounds in general. *See* CARBON.

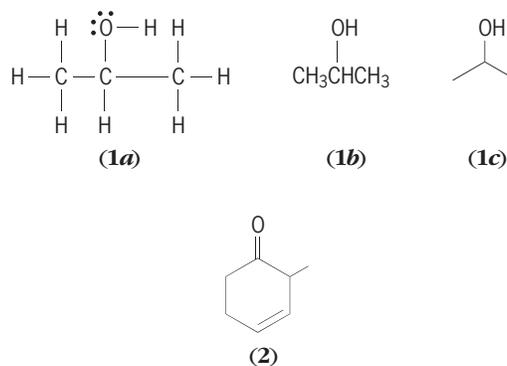
### Structure

Many organic compounds were known at the beginning of the nineteenth century, but a unifying structural principle was not available until about 1850. A crucial key was the concept that a carbon atom in its compounds always formed four linkages or bonds to other atoms. This was followed by rapid development of a structural theory that included the ideas of linear and cyclic carbon chains, multiple bonds, and isomerism. In 1874, to account for the phenomenon of optical rotation, it was proposed that carbon had a tetrahedral structure. These principles, greatly expanded and elaborated by additional concepts such as the electron-pair bond and later the molecular orbital picture of bonding, provide the underpinning for the hundreds of types of compounds and the millions of individual compounds known today.

The structures of organic compounds are described by a molecular framework of carbon atoms on which substituents may be located at various points. The *Beilstein Handbuch* is a multivolume reference that provides a systematic record of compounds organized into three classes: acyclic (aliphatic), carbocyclic, and heterocyclic (with rings

containing atoms other than carbon). Each of these classes is divided into subclasses according to the presence of functional groups.

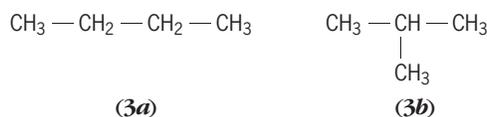
Structures can be represented in several ways, as illustrated for the three-carbon alcohol 2-propanol (**1**) and the cyclic ketone 2-methyl-3-cyclohexenone (**2**).



The expanded structure of 2-propanol (**1a**) shows all bonds and electron pairs, including unshared electrons on oxygen. More compact and convenient is the condensed structure (**1b**) in which the C—C and C—H bonds are implied. In the bond-line convention (**1c**), all C—C bonds are indicated by a line, as shown for 2-propanol. Carbon atoms are not shown explicitly, but rather are implied at the ends of each line segment, together with enough hydrogen atoms to complete the tetravalency at each carbon. The bond-line convention is particularly convenient for cyclic structures such as 2-methyl-3-cyclohexenone; each vertex and the end of each line segment represents a carbon and appropriate number of hydrogens. *See* STRUCTURAL CHEMISTRY; VALENCE.

**Functional groups.** A functional group is an atom other than carbon or a multiple bond, such as the hydroxyl group (OH) of 2-propanol, the double bond (C=C), or the carbonyl group (C=O) of 2-methyl-3-cyclohexenone. The group defines a class of compounds and is the point at which characteristic reactions occur, for example, oxidation, reduction, or addition of an electrophilic or nucleophilic reagent. Some of the principal functional groups are shown in the **table**. *See* ELECTROPHILIC AND NUCLEOPHILIC REAGENTS.

**Isomerism.** The fact that there can be two or more compounds, known as isomers, with the same molecular composition was one of the key points in development of a structural theory. One type of isomerism, structural or constitutional, is illustrated by the two isomers that have the formula C<sub>4</sub>H<sub>10</sub>, butane (**3a**) and isobutane (2-methylpropane; **3b**). The



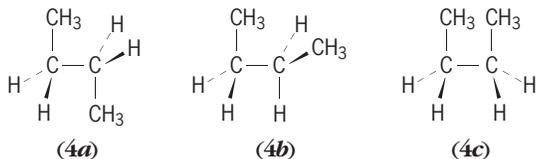
number of possible structural isomers becomes enormous in larger molecules. Examples of the molecular formulas of some hydrocarbons and the number of

possible isomers for each are given below.

$C_3H_8$	1	$C_8H_{18}$	18
$C_4H_{10}$	2	$C_{10}H_{22}$	75
$C_6H_{14}$	5	$C_{20}H_{42}$	366, 319

See MOLECULAR ISOMERISM.

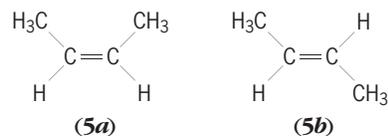
**Conformation.** Several three-dimensional representations of butane, showing the tetrahedral geometry of the carbon atoms, are given in structures (4). As



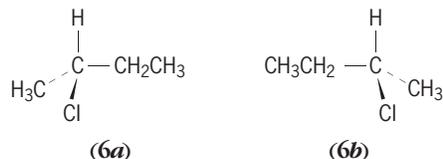
indicated in these structures, butane can exist in several forms, called conformations, which differ in the relative positions of the carbon atoms, and thus the overall shape of the molecule. However, the barrier to rotation around the central C—C bond is so low that these individual conformational isomers are not separable, and butane is thus a single compound. See CONFORMATIONAL ANALYSIS.

In an alkene, rotation around the C=C bond does not occur, and 2-butene, for example, exists as two

isomeric compounds, cis (Z) and trans (E) (5a and 5b, respectively).

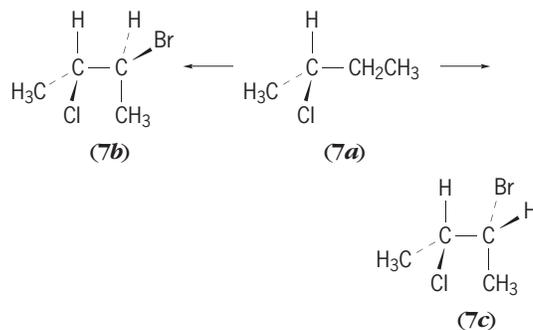


**Stereoisomerism.** Stereoisomers are compounds that have the same bond sequence but differ in the spatial array of the bonds. When a carbon atom is bonded to four unlike atoms or groups, the tetrahedral geometry of carbon causes the atom to be dissymmetric or chiral. A compound, with a chiral atom can exist in two isomeric forms, known as enantiomers. The relative positions of all atoms is identical in the two enantiomers, but they differ in handedness, a characteristic of an asymmetric object and its nonsuperposable mirror image, as in structures (6) of 1-chlorobutane.



See STEREOCHEMISTRY.

When two chiral centers are present, two stereoisomers can arise from each enantiomer. Thus enantiomer (7a) of chlorobutane can lead to isomeric structures (7b) and (7c), in which the relative



positions of the atoms is not identical. In this case, the isomers are known as diastereoisomers. With  $n$  chiral centers, there can be  $2^n$  stereoisomers.

**Acyclic compounds.** The simplest organic compound is methane ( $CH_4$ ). It is the first member of the homologous series of alkanes, in which successive compounds differ by an additional  $-CH_2-$  group ( $CH_3CH_3$ ,  $CH_3CH_2CH_3$ , and so forth). Methane is the major constituent of natural gas, and is the most abundant organic compound in the biosphere. It is also released into the atmosphere by the decay of rotting vegetation and fermentation in the rumens of grazing animals. See ALKANE; METHANE; NATURAL GAS.

Higher alkanes,  $CH_3(CH_2)_nCH_3$  ( $n = 3-20$ ), and also branched isomers and cyclic hydrocarbons are the principal components of petroleum. These compounds have no reactive functional groups.

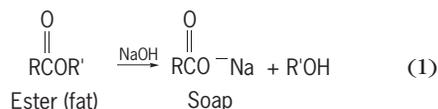
#### Principal organic functional groups

Composed class	Group	Structure
Alkene	Double bond	$>C=C<$
Alkyne	Triple bond	$-C\equiv C-$
Alcohol	Hydroxyl	$-OH$
Amine	Amino	$-NH_2(-NR_2)^*$
Aldehyde	Carbonyl	$\begin{array}{c} O \\    \\ -CH \end{array}$
Ketone	Carbonyl	$\begin{array}{c} O \\    \\ -CR \end{array}$
Acid	Carboxyl	$\begin{array}{c} O \\    \\ -COH \end{array}$
Ester	Alkoxycarbonyl	$\begin{array}{c} O \\    \\ -COR \end{array}$
Amide	Carbamoyl	$\begin{array}{c} O \\    \\ CN< \end{array}$
Nitrile	Cyano	$-C\equiv N$
Azide	Azido	$-N=N=N$
Nitro		$-NO_2$
Sulfide		$-S-$
Sulfoxide		$\begin{array}{c} O \\    \\ -S- \end{array}$
Sulfonic acid		$-SO_3H$

\*R = any carbon group, for example,  $CH_3$ .

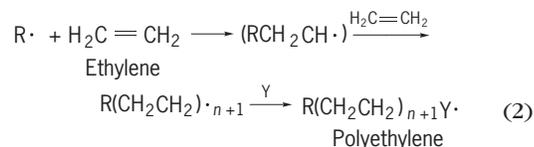
However, several reactions that involve major structural changes, such as cracking into smaller molecules, isomerization of carbon chains, and removal of hydrogen are carried out on these compounds by catalytic processes, often at high temperatures, in petroleum refining. See CRACKING; HYDROCRACKING; ISOMERIZATION.

Both acyclic and cyclic carbon frameworks can contain multiple bonds; oxygen, nitrogen, and sulfur atoms; and other functional groups listed in the table. Alcohols, alkenes, aldehydes, amines, acids, and so on occur in plant and animal tissues and can be isolated from these sources. If they are required for some use they can be prepared as necessary by laboratory methods, or on large scale by the chemical process industry. As an example, long straight molecules, often containing one or more double bonds and terminating in an ester group, are the characteristic structural features of fats and waxes. Hydrolysis of fats by reaction with aqueous alkali gives the salts of fatty acids [soaps; reaction (1)].

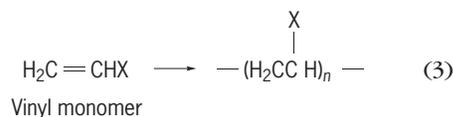


These acids were among the compounds available to organic chemists in early investigations. See FAT AND OIL; HYDROLYSIS; SOAP.

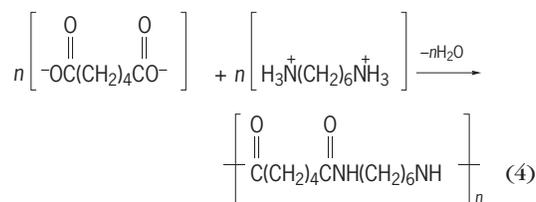
**Polymers.** There is no fixed limit to the length of the carbon chain. Alkenes, with the functional group C=C, can undergo polymerization; successive molecules add to a reactive intermediate (R·). An important example is the formation of polyethylene from ethylene (H<sub>2</sub>C=CH<sub>2</sub>), reaction (2), where R· is



the initiating radical. The value of *n* in polyethylene can be as high as 10,000. Other polymers, each with characteristic properties and specific uses, can be obtained from various vinyl monomers [H<sub>2</sub>C=CHX; reaction (3), where X = alkyl, halogen, or some



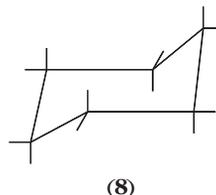
other group]. These macromolecules are manufactured on a scale of millions of tons per year, and are by far the largest-volume product of the chemical process industry. Another major type of polymer is obtained by condensation, in which monomer units are linked by reactions that involve the formation of amide or ester bonds. A typical example is formation of a nylon polyamide [reaction (4)].



See POLYAMIDE RESINS; POLYESTER RESINS; POLYMER; POLYMERIZATION.

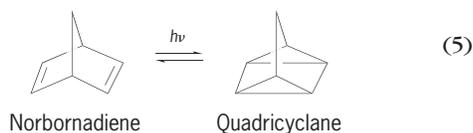
**Carbocyclic compounds.** The two large groups of compounds with rings containing only carbon are alicyclic and aromatic. The parent hydrocarbons in the former series are cycloalkanes and in the latter, benzene.

*Alicyclic compounds.* Cycloalkanes are named simply by adding the prefix cyclo- to the name of the alkane corresponding to the number of CH<sub>2</sub> groups in the ring. In cyclopropanes and cyclobutanes, with three- and four-membered rings, respectively, the normal tetrahedral bond angles of 109° are compressed, and reactions that lead to breaking a C—C bond are facilitated. Rings with five or more carbons are not significantly strained; the rings are puckered, and the chemistry is much like that in acyclic compounds. A cyclic structure restricts the number of conformations available to a molecule. The lowest-energy conformation of a cyclohexane ring is a chairlike structure (8): six bonds (the heavy lines in the structure),



termed axial, project above and below the ring, and six others, termed equatorial, lie in the mean plane of the ring.

Compounds with several six-membered rings fused together are very common among naturally occurring terpenes and steroids. Highly condensed multiring structures have been prepared by various methods, often taking advantage of intramolecular photocyclizations. An example is the reversible isomerization of norbornadiene to quadricyclane [reaction (5)], which has been suggested as a means of



energy storage. Among the interesting compounds that have been prepared are a tetrahedrane, cubane, and dodecahedrane (Fig. 1), with carbon skeletons corresponding to three of the five platonic solids—the tetrahedron, the cube, and the dodecahedron.

*Aromatic compounds.* Benzene, a hydrocarbon with composition C<sub>6</sub>H<sub>6</sub>, played a major role in the concept of cyclic structures and later in the development of bonding theory. The structure of benzene (9) is a planar six-membered ring with six electrons in a

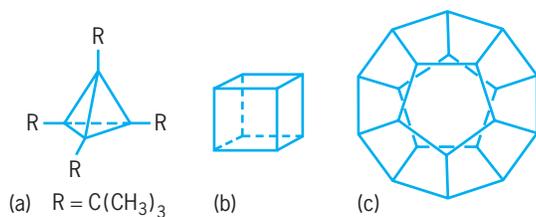
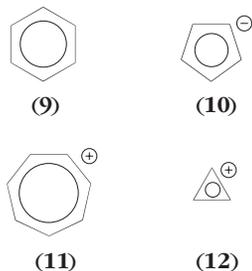


Fig. 1. Highly condensed multiring structures. (a) Tetrahedrane. (b) Cubane. (c) Dodecahedrane.

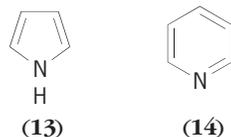
delocalized array. Since benzene does not show the reactions expected for a highly unsaturated compound, the details of the bonding were obscure for many decades. The Hückel formulation of an aromatic ring as one containing  $4n + 2 \pi$  electrons in a cyclic conjugated system provides a unifying basis for the structure of benzene and also nonbenzenoid aromatic compounds. Thus, the five- and seven-membered rings in cyclopentadiene anion (10) and cycloheptatriene cation (11), respectively, with six  $\pi$  electrons, and the cyclopropenium cation (12), with two electrons ( $4n + 2$  where  $n = 0$ ), have properties that qualify these species as aromatic. An aromatic system can be symbolized by a circle inscribed within the appropriate ring.



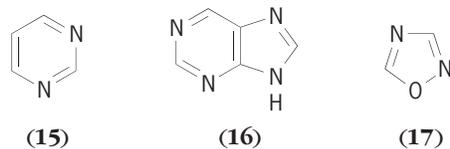
See AROMATIC HYDROCARBON; BENZENE; RESONANCE (MOLECULAR STRUCTURE).

Benzene, its methyl derivatives toluene and xylenes, and a number of polycyclic aromatic hydrocarbons such as naphthalene and anthracene can be isolated from coal tar distillates, although the major source of benzene and toluene is now cyclization and dehydrogenation of alkanes from petroleum. Benzene was available in quantity in the mid-nineteenth century, and the first extensive studies of the reactions of organic chemistry were those of aromatic compounds. Among the major developments were synthetic dyes based on aniline, which were the foundation of several large German chemical firms. See AROMATIZATION; COAL CHEMICALS; DEHYDROGENATION; DESTRUCTIVE DISTILLATION; DYE.

**Heterocyclic compounds.** A nitrogen, oxygen, or sulfur atom can take the place of carbon in either alicyclic or aromatic rings. The most numerous and important heterocyclic compounds are those with nitrogen in a five- or six-membered aromatic system. Pyrrole (13) and pyridine (14) are the simplest com-



pounds, and one or several other heteroatoms can be present at various positions as well, giving rise to thousands of heterocyclic systems with one or two rings. Examples are the pyrimidines (15) and purines (16), which make up the backbone of the



nucleic acids. All of the B vitamins, chlorophyll, heme pigments, and alkaloids have heterocyclic structures, for example, 1,2,4-oxadiazole (17). See CHLOROPHYLL; PURINE; PYRIDINE; PYRIMIDINE; PYRROLE; VITAMIN.

Two major areas of contemporary organic chemistry are the development of new drugs, and also chemical agents for use as selective herbicides, pesticides, and plant growth regulators. The primary thrust in both areas is the preparation and pharmacological or agricultural evaluation of many candidate compounds. These possess, almost without exception, heterocyclic structures. See HETEROCYCLIC COMPOUNDS.

**Naturally occurring compounds.** All living organisms, from prokaryotes to mammals and higher plants, use the same basic groups of compounds in their metabolism. These components are lipids, carbohydrates, nucleic acids, and proteins. The major lipids are fats, which are triesters of glycerol with three long-chain acids, phospholipids with one carboxylic group replaced by a phosphoric acid derivative, and nonsaponifiable compounds such as cholesterol. Carbohydrates consist of sugars, which are five- and six-carbon polyhydroxyaldehydes and their polymeric acetals starch and cellulose. Nucleic acids are very large polymers built up from four different nucleotides by phosphate-ester bonds. Proteins are smaller polymers of  $\alpha$ -amino acids linked by peptide (amide) bonds; a total of 20 different amino acids are used in mammalian proteins.

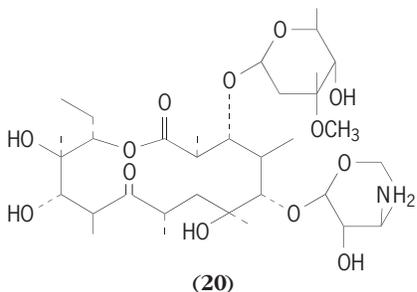
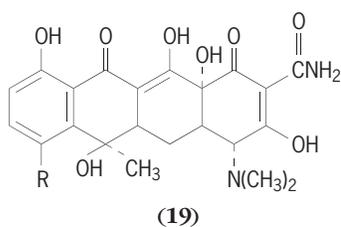
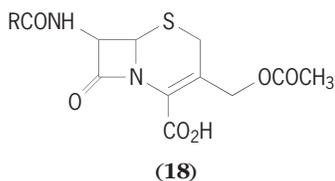
The chemistry of each of these groups of compounds has become a major area of itself, with distinctive experimental methods. Investigations are concerned primarily with the formation, functions, and interactions of these molecules in living cells. Although the original structural studies on these compounds were carried out by organic chemists in the nineteenth century and have added immensely to organic chemistry, contemporary work is usually considered the province of biochemistry. See BIOCHEMISTRY; CARBOHYDRATE; LIPID; NUCLEIC ACID; PROTEIN.

**Secondary metabolites.** An area of natural-products chemistry that was pursued at a very early period was the isolation of active principles of medicinal plants. One group of such compounds comprises the alkaloids. These are cyclic amines with basic properties, and include some compounds with extremely complex structures such as morphine and strychnine. See ALKALOID; AMINE; MORPHINE ALKALOIDS; STRYCHNINE ALKALOIDS.

Another large group of plant products comprises the terpenes. These are hydrocarbons, alcohols, and ketones with 10, 15, 20, or 30 carbon atoms. They are present, sometimes in substantial quantities, in many essential oils. Some terpenes from marine plants contain several halogen atoms. Work on the elucidation of the structures of terpenes and their chemistry led to the first clear recognition that the carbon skeleton of a molecule can undergo rearrangement, or scrambling, during a reaction. *See* ESSENTIAL OILS; TERPENE.

Alkaloids, terpenes, and another group of compounds with a phenylpropanoid structure are referred to collectively as secondary metabolites. Their biochemical origin in plant cells has been studied extensively and is understood in some detail, but their functions in the plants and their evolutionary significance are obscure.

*Antibiotics.* Much attention in contemporary work on naturally occurring compounds has been directed to antibiotic substances produced by various microorganisms. Antibiotics are a very diverse class of compounds, but some of the clinically useful compounds can be grouped into a few general structural types:  $\beta$ -lactam, such as cephalosporin (**18**); tetracycline (**19**), with R = N, Cl, or OH; and macrolide, such as erythromycin (**20**). Each type has several representatives.



*See* ANTIBIOTIC.

**Separation and structure determination.** Progress in every phase of organic chemistry has depended on the methods available for separation, purification, and structure determination of compounds from sources such as plant extracts, coal tar, petroleum, or complex reactions. The separation of mixtures of compounds with very similar properties has required increasingly refined experimental methods

and has provided a constant impetus for the development of new methods.

Early separations depended mainly on fractional crystallization and distillation; both methods require substantial amounts of material. Chromatographic separations based on adsorption and selective elution were first applied to pigments and then to sugars, which were difficult to crystallize and impossible to distill. A major advance was partition chromatography, in which a solution of a mixture is passed over a column of cellulose or strip of paper. The components of the mixture are separated by differential extraction between a stationary phase, for example, water absorbed on the cellulose, and a mobile solvent or vapor phase. This principle is employed in gas chromatography and high-pressure liquid chromatography (HPLC). Modern refinements permit the analysis and separation of extremely complex mixtures on a nanogram scale. A more recent application is the use of a stationary phase with a chiral coating on which enantiomers can be separated by diastereoisomeric complexation, permitting accurate determination of enantiomeric purity. *See* CHEMICAL SEPARATION TECHNIQUES; CHROMATOGRAPHY; CRYSTALLIZATION; DISTILLATION; GAS CHROMATOGRAPHY; LIQUID CHROMATOGRAPHY.

Until about 1940, methods for establishing the structure of a new compound depended entirely on chemical characterization of various features and functional groups, systematic degradation to smaller molecules, and a great deal of insight. This was an arduous task; the final structures of some of the alkaloids that had been isolated in the 1820s were not solved until 1950.

The development of spectroscopic and x-ray diffraction methods revolutionized elucidation of structure in the span of one generation. Ultraviolet, visible, and infrared spectroscopy reveal the presence or absence of certain functional groups and bonds; such information was the key to determination of the structure of penicillin G. A far more powerful tool is nuclear magnetic resonance (NMR). By this method, a signal for nearly every carbon, hydrogen, and nitrogen atom, as well as the environment of the atoms, can be seen in the spectrum. An even more direct view of a molecule is provided by x-ray crystallography, which shows the exact location of atoms in the crystal lattice. With these methods, only microgram-milligram samples are needed for a complete determination of structure, and they have enormously speeded the pace of research. *See* INFRARED SPECTROSCOPY; NUCLEAR MAGNETIC RESONANCE (NMR); X-RAY CRYSTALLOGRAPHY.

### Synthesis Reactions

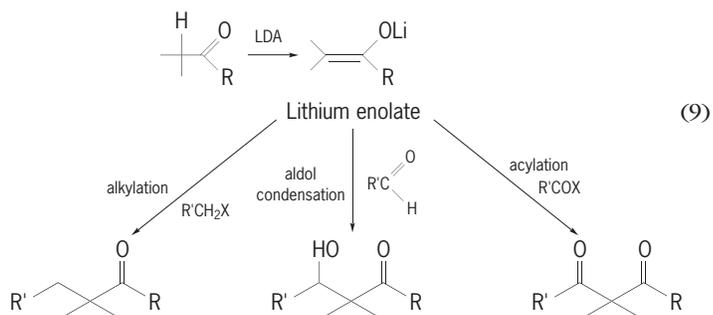
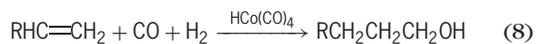
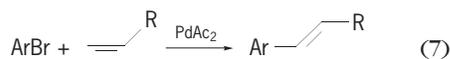
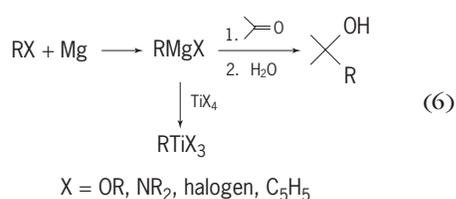
The preparation of compounds occupies much of the effort of organic chemistry, and is the principal business of the chemical industry. The manufacture of drugs, pigments, and polymers entails the preparation of organic compounds on a scale of thousands to billions of kilograms per year, and there is constant research to develop new products and processes. Synthesis of new substances is carried out for many purposes beyond the goal of a commercial

product. A compound of a specified structure may be needed to test a mechanistic proposal or to evaluate a biochemical response such as inhibition of an enzyme. Synthesis may provide a more dependable and less expensive source of a naturally occurring compound; moreover, a synthetic approach permits variations in the structure that may lead to enhanced biological activity.

The term synthesis usually implies a planned sequence of steps leading from simple starting compounds to a desired end product. Each of these steps involves a reaction that may lead to formation of a C—C bond or to the introduction, alteration, or removal of a functional group. Progress in synthesis depends on the availability of a wide range of reactions that bring about these changes in good yield, with a minimum of interfering by-products. An integral part of synthesis is the development of new methods and reagents that are selective for a desired transformation, and, very importantly, proceed with control of the stereochemistry. See ASYMMETRIC SYNTHESIS; ORGANIC SYNTHESIS.

**Carbon-carbon bond formation.** Most syntheses involve the creation of one or more C—C bonds, and there are many ways to accomplish this. Two of the more important general approaches are the reactions of organometallic compounds and the reactions of enolates.

*Organometallic reactions.* The first generally useful organometallic compounds were organomagnesium halides, discovered by V. Grignard. These compounds are readily prepared from the metal and an organic halide, and react with any carbonyl group. Grignard reagents are effective for the preparation of simple alcohols, for example, but because of their high reactivity they are not selective, and other functional groups interfere. However, conversion of the

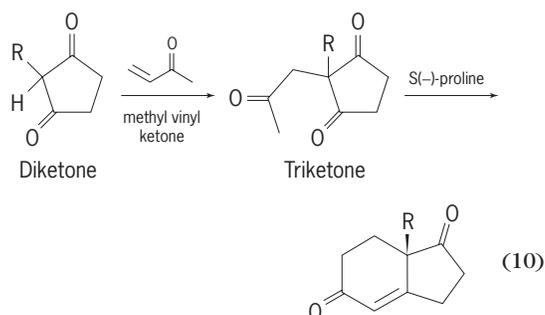


magnesium compound to another organometallic derivative such as a titanium complex leads to reagents with a range of reactivity and high stereoselectivity [reactions (6)]. See GRIGNARD REACTION; ORGANOMETALLIC COMPOUND.

Several other transition metals have important uses in synthesis. In these processes, the reactants are brought together and react in the coordination shell of the metal. An organometallic compound is formed, but in most of these reactions this intermediate is not stable and a catalytic cycle can be maintained. Examples are the coupling of a halide and an alkene [reaction (7)] and hydroformylation [reaction (8)]; the latter is a major industrial process. See HYDROFORMYLATION; TRANSITION ELEMENTS.

*Enolate reactions.* Removal of an  $\alpha$ -hydrogen from a carbonyl compound gives an enolate derivative, which can then serve as a nucleophile, that is, an electron pair donor, for alkylation, aldol addition, or acylation [reaction scheme (9); X = any leaving group]. The original carbonyl compound can be an aldehyde, ketone, or ester, and the products can be reduced, dehydrated, or transformed in other ways. Use of the very strong base lithium diethylamide (LDA) permits complete conversion of the carbonyl compound to the lithium enolate at  $-70^\circ\text{C}$ , greatly facilitating control of the subsequent steps. These reactions provide a very general and flexible approach for the assembly of carbon skeletons with functional groups at specific locations. See ACYLATION; ALDEHYDE; ESTER; KETONE.

An illustration of these processes is shown in reaction (10). Alkylation of the diketone by 1,4-addition

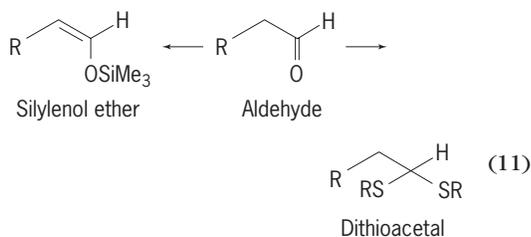


to methyl vinyl ketone gives the triketone. Aldol condensation of the latter in the presence of S(-)-proline as a chiral catalyst gives a single enantiomer of the bicyclic product, which is an important precursor for the synthesis of steroid hormones and analogs.

**Functional group transformations.** Many thousands of reactions and reagents have been discovered or invented to introduce, modify, and remove functional groups. These reactions involve all of the general processes of organic chemistry such as substitution, elimination, addition, oxidation, and reduction, and involve multiple bonds, oxygen, nitrogen, sulfur, and numerous other elements. Most frequently, the transformation is the end objective, as in the introduction and reduction of a nitro group, ( $-\text{NO}_2$ ), to prepare an amine ( $\text{RNH}_2$ ).

On other occasions a functional group may be modified to obtain a synthon (reagent) to be used

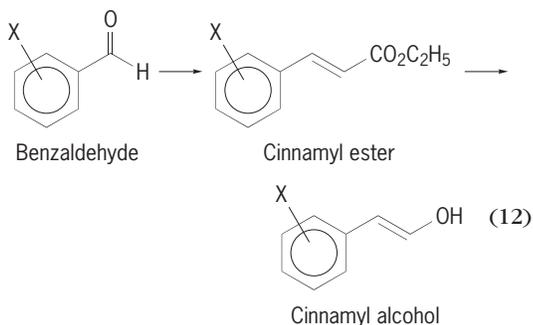
in the formation of a new C—C bond. Examples are conversion of an alcohol to an iodide or trifluoromethylsulfonyl ester ( $\text{CF}_3\text{SO}_2\text{R}$ ) for alkylation, or an aldehyde to the silyl enol ether or dithioacetal [reactions (11)].



Similarly, a carboxylic acid can be converted to the more reactive acid chloride ( $\text{RCOCl}$ ) or to a mixed anhydride or activated ester for peptide synthesis. See ACETAL; ALCOHOL; CARBOXYLIC ACID; ETHER.

Another objective is the introduction and removal of protective groups. In complex syntheses it is often necessary to block reactions of a functional group until a later stage of the sequence, when it is then regenerated under mild, selective conditions. An example is use of the trichloroethyl ether or ester group to protect an alcohol or acid, respectively, against undesired reactions. The group is removed under nonhydrolytic conditions by reductive elimination with zinc.

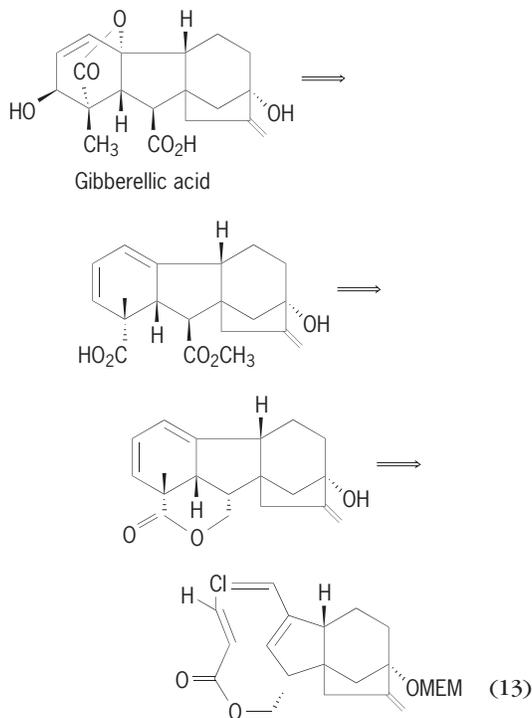
**Logical design in synthesis.** Relatively simple syntheses can be based on an obvious relationship between the desired product and readily available precursors. For example, a cinnamyl alcohol can be obtained by reduction of the corresponding ester, and the latter can be prepared by condensation of the appropriate benzaldehyde [reaction (12)]. The X



term shows that this is a general reaction (any group X) at any position.

In the synthesis of more complex structures, such as those of naturally occurring compounds, many different reaction sequences are possible. In this situation, it is desirable to recognize many alternatives and to decide which will be the most practical and efficient. This problem can be facilitated by a method known as retrosynthetic analysis, in which the structure of the target compound is examined in order to discover any bonds whose disconnection will lead to simpler molecules or fragments. An example of this approach is the synthesis of the plant regulator gibberellic acid. Working backward, retrosynthetic steps (indicated by  $\Rightarrow$ ) sug-

gested that the precursors shown in sequence (13)



[OMEM = methoxyethoxymethyl] would provide a suitable route, and the synthesis was accomplished in this way. Computer programs with files of information on methods for bond formation, functional group manipulation, and protective groups have been developed to aid in this type of analysis. See COMPUTATIONAL CHEMISTRY.

### Theory and Mechanisms

As knowledge of structures and reactions of organic compounds has accumulated, understanding of reaction mechanisms and correlation of structure and reactivity have advanced by application of the methods of physical chemistry and the principles of molecular orbital theory.

**Bonding.** In 1916, G. N. Lewis suggested that covalent bonds could be described as electron pairs shared by two atoms. A more precise formulation followed from the development of quantum theory, and the concept of atomic orbitals, which are functions of the electron density at various points of space surrounding the atomic nucleus. The orbitals of carbon that are important in bonding are designated  $2s$ ,  $2p_x$ ,  $2p_y$ ,  $2p_z$  or hybrids of these,  $sp$ ,  $sp^2$ , and  $sp^3$ , in which one, two, or three of the  $p$  atomic orbitals are "mixed" with the  $s$  atomic orbital. When the atomic orbitals of two atoms combine, two molecular orbitals, one bonding and one antibonding ( $\text{MO}^*$ ) at a higher energy level, are formed. Overlap along the axis between the nuclei leads to a  $\sigma$  molecular orbital, as shown in Fig. 2a for the combination of two  $sp^3$  methyl radicals to form the  $\sigma$  C—C bond in ethane.

Another type of bond is possible with  $p$  orbitals, which can overlap sidewise to give a  $\pi$  molecular orbital, as illustrated in the double bond of an alkene

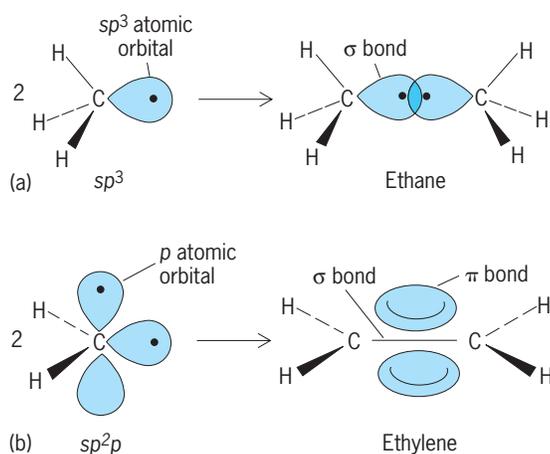


Fig. 2. Bonding orbitals in a (a)  $\sigma$  bond and (b)  $\pi$  bond.

(Fig. 2b). This additional  $\pi$  bond has a nodal plane passing through the atoms, with electron density above and below, and it contributes only 40% of the total bonding energy. A triple bond can be represented as a  $\sigma$  and two orthogonal  $\pi$  bonds.

In Fig. 2 only bonding orbitals are depicted. Treatment of conjugated systems, and the analysis of reactions in molecular orbital terms, requires consideration of the energy levels of both bonding and antibonding molecular orbitals. Diagrams of the  $\pi$ -electron systems of ethylene and 1,3-butadiene are shown in Fig 3. The size and shading of the orbital at each position indicates the magnitude and algebraic sign; there is a node, and thus no bonding, between lobes of opposite sign. The highest-energy molecular orbital that is occupied by electrons (HOMO)

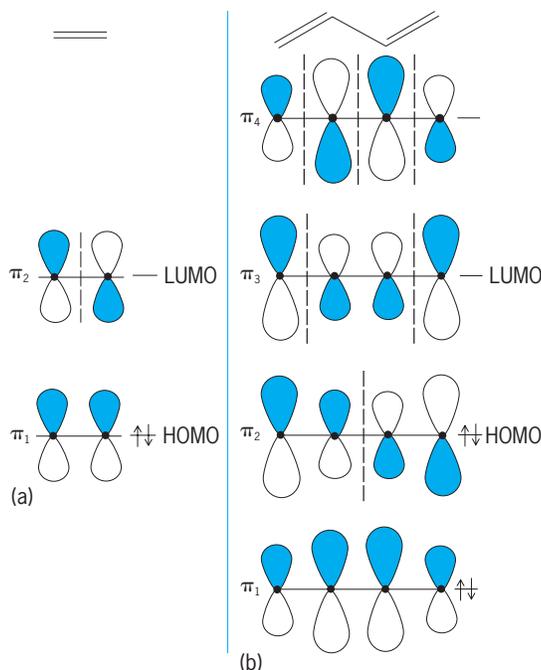


Fig. 3. Molecular orbitals of (a) ethylene and (b) 1,3-butadiene. (Broken vertical line designates node.)

and the lowest-energy unoccupied orbital (LUMO) are designated the frontier orbitals.

In the concept of frontier molecular orbital theory, interactions of the HOMO of one reactant with the LUMO of another produces a mutual perturbation of molecular orbitals as bond changes begin. When a reactant has two or more reactive sites, or when two reactants are compared, the product is generally the one that results from the most effective of the possible HOMO-LUMO interactions. Applications of this principle have provided a unifying basis for understanding reactions and observations that are not otherwise accounted for. See CHEMICAL BONDING; MOLECULAR ORBITAL THEORY; WOODWARD-HOFFMANN RULE.

**Reactions.** The course of any reaction is governed by thermodynamics. In the reaction  $A + B \rightleftharpoons C + D$ , the extent to which the reaction proceeds,  $K_{eq}$ , depends on the free-energy change ( $\Delta G$ ), which in turn is determined by the changes in enthalpy ( $\Delta H$ ) and entropy ( $\Delta S$ ) as expressed by Eq. (14).

$$\Delta G^\circ = \Delta H^\circ - T\Delta S^\circ = -RT \ln K_{eq} \quad (14)$$

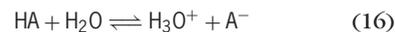
In reactions at relatively low temperatures, the entropy term is usually small, and the major factor is  $\Delta H$ , which represents the difference between the energies of the bonds made in the products and that lost in the reactants. See CHEMICAL EQUILIBRIUM; CHEMICAL THERMODYNAMICS; ENTHALPY; ENTROPY; FREE ENERGY; PHYSICAL ORGANIC CHEMISTRY; THERMOCHEMISTRY.

A separate consideration is the reaction rate ( $k_r$ ) and the kinetic parameters. A widely used concept is the activation energy ( $E_A$ ) of a transition barrier that must be overcome in the rate-determining step of the reaction. The measured rate and  $E_A$  are related by the Arrhenius equation (15), where  $A$  is a term known

$$k_r = Ae^{-E_A/RT} \quad (15)$$

as the preexponential factor,  $R$  is the gas constant, and  $T$  is the absolute temperature. Measurement of the rate at several temperatures permits evaluation of  $\Delta H^\ddagger$  and  $\Delta S^\ddagger$  in reaching the transition state. See CHEMICAL DYNAMICS; GAS CONSTANT.

**Acidity.** An important property of organic compounds is proton acidity, as expressed by the equilibrium constant ( $K_A$ ) for dissociation in water ( $H_2O$ ), reaction (16), where HA represents the acid and  $H_3O^+$  is the hydronium ion.



$$K_A = \frac{[H_3O^+][A^-]}{[HA]}$$

$$pK_A = -\log K_A$$

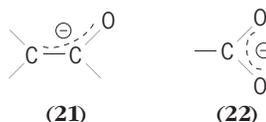
See HYDROGEN ION.

For compounds that are weaker acids than water, the equilibrium is determined in a nonaqueous solvent and related to the  $K_A$  in water. The acid dissociation constant  $K_A$  is usually expressed as the negative logarithm  $pK_A$ . In this way, acid strengths ranging

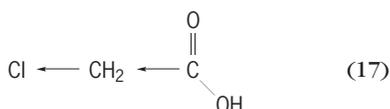
over  $10^{50}$  from alkanes ( $pK_A \approx 50$ ) to strong acids ( $pK_A \approx 0-2$ ) can be compared on a common scale of  $pK_A$ . Some representative  $pK_A$  values are given below:

$\text{CH}_3\text{CH}_3$	50		
$\text{CH}_2=\text{CH}_2$	44	$\text{CH}_3\text{CH}_2\text{OH}$	15.9
$\text{C}_6\text{H}_5\text{CH}_3$	41	$\text{C}_6\text{H}_5\text{OH}$	10.0
$\text{CH}_3\text{CH}_2\text{NH}_2$	35	$\text{O}$	
$\text{HC}\equiv\text{CH}$	25	$\parallel$	
		$\text{CH}_3\text{COH}$	4.7
$\text{O}$		$\parallel$	
$\parallel$		$\text{CH}_2\text{ClCOH}$	2.8
$\text{CH}_3\text{CCH}_3$	20		

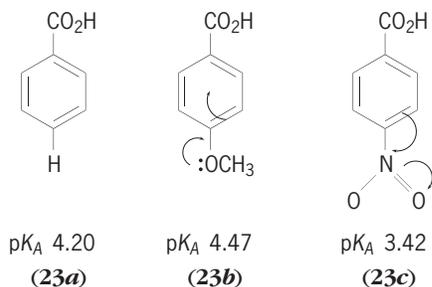
The  $pK_A$  of an acid is a sensitive probe of structural effects. The acidity of a proton bonded to carbon is extremely low, but it increases significantly when the carbon is doubly or triply bonded ( $sp^2$  or  $sp$  hybridized) because the electrons of the C—H bond are brought closer to the atom nucleus. An adjacent carbonyl group greatly increases the acid strength of C—H, N—H, or O—H bonds, reflecting the dispersal and stabilization of the charge on the anions, as in the enolate (21) or carboxylate (22). The enhanced acid-



ity of an  $\alpha$ -chloroacid results from the inductive effect of the electronegative halogen atoms, as shown in notation



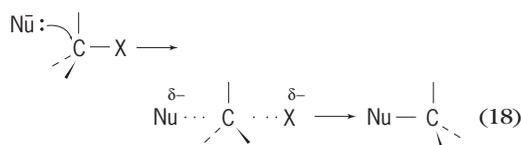
The influence of electronic factors on  $pK_A$  can be seen in the substituted benzoic acids (23). An



electron-releasing group,  $-\text{OCH}_3$ , reduces acid strength, and an electron-withdrawing group,  $-\text{NO}_2$ , increases it. The effects of substituents on the  $pK_A$  of benzoic acids are the basis of a more general correlation termed a linear free-energy relationship. By this correlation, the effect of a given *meta* or *para* substituent on the  $pK_A$  of a benzoic acid can be compared with the effect of the same substituent on some other process. This approach

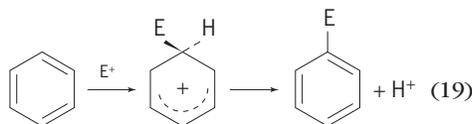
provides a powerful tool in studies of mechanism. See ACID AND BASE; PK; SOLUTION.

**Aliphatic nucleophilic substitution.** Substitution of a group X on a  $sp^3$  carbon by a nucleophile that supplies an electron pair is a very general and extensively studied reaction. In a typical case, the leaving group X is a halide or sulfonate anion. Two mechanisms for nucleophilic substitution are designated  $S_N1$  and  $S_N2$ , referring to the kinetic order. The  $S_N2$  process depends on both substrate and nucleophile (Nu) concentrations; the transition state involves attack of nucleophile from the side opposite X, with inversion of stereochemical configuration [reaction (18)]. The reaction rate is greatly retarded by the presence of bulky substituents.

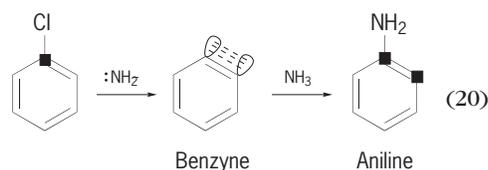


In the  $S_N1$  mechanism, the group X departs in an initial step with formation of a carbenium ion,  $\text{R}_3\text{C}^+$ , which then reacts with a nucleophile, often a solvent molecule. The rate depends strongly on the stability of the electron-deficient intermediate, which may undergo skeletal arrangement during the process. See HALOGENATED HYDROCARBON; STERIC EFFECT (CHEMISTRY); SUBSTITUTION REACTION.

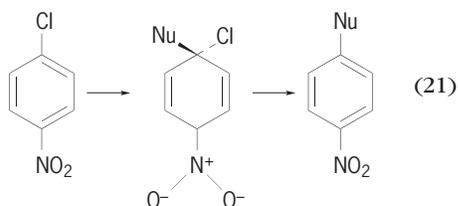
**Aromatic substitution.** Aromatic compounds undergo several types of substitution. With strongly electrophilic reagents ( $\text{E}^+$ ) such as bromonium ( $\text{Br}^+$ ), nitronium ( $\text{NO}_2^+$ ), acylium ( $\text{RCO}^+$ ), or sulfur trioxide ( $\text{SO}_3$ ), substitution occurs by formation of a cationic intermediate, which then loses a proton [reaction (19)].



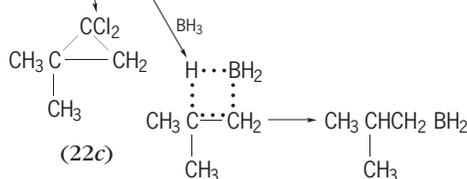
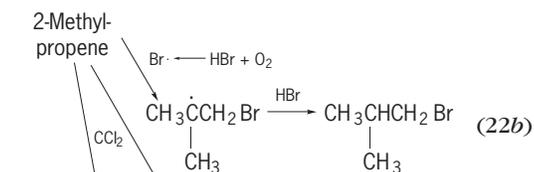
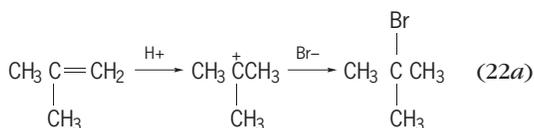
Nucleophilic substitution of halogen on an aromatic ring with strongly basic reagents can occur by an elimination-addition process, as in the formation of aniline via benzyne intermediate. This mechanism was established by isotopic labeling (the black square represents  $^{14}\text{C}$ ), which showed that two adjacent positions in the ring became labeled [reaction (20)].



If an electron-withdrawing group is present, substitution can occur by an addition-elimination sequence [reaction (21)].

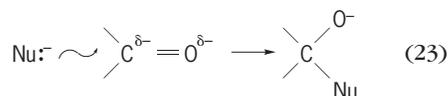


**Addition reactions.** A multiple bond can undergo addition, with formation of two new  $\sigma$  bonds and loss of a  $\pi$  bond, by several mechanisms. Addition reactions of alkenes occur by attack of an electrophilic reagent on the electron-rich  $\pi$  bond. With 2-methylpropene, for example, reaction with acids leads to the more stable carbenium ion, which then reacts with a nucleophile such as bromide ion ( $\text{Br}^-$ ) to give the bromide [reaction (22a)]. In the presence



of air, the electrophile is a bromine radical ( $\text{Br}\cdot$ ), and the isomeric product is produced [reaction (22b)]. Other additions occur by essentially simultaneous formation of both bonds as in the reaction with carbenes to give cyclopropanes [reaction (22c)], or by a four-center transition state with borane to give an alkylborane [reaction (22d)]. See HYDROBORATION; REACTIVE INTERMEDIATES.

Additions to the  $\text{C}=\text{O}$  group of carbonyl compounds are among the most important organic reactions. In this case, addition occurs by attack of a nucleophile at the carbon atom of the polarized  $\text{C}=\text{O}$  bond. A diverse range of products results, depending on the nature of the carbonyl compound and reagent, but the first step in nearly all these reactions is to form the tetrahedral intermediate [reaction (23)].

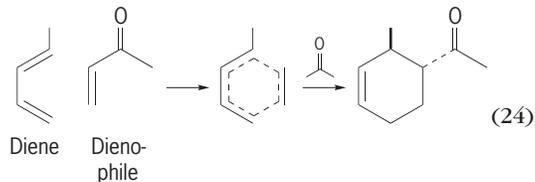


See ACID HALIDE.

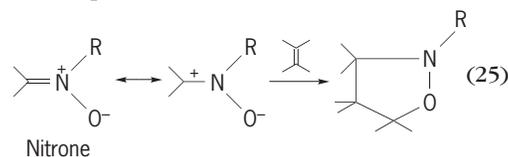
**Pericyclic reactions.** Several reactions, termed pericyclic, occur by a concerted reorganization of bonds, with no intermediate. In these reactions, the transition state between reactants and products is an aro-

matic system. Two of these processes, cycloaddition and sigmatropic rearrangement, are very useful synthetic reactions.

Cycloadditions are exemplified by the Diels-Alder reaction, in which a diene combines with a dienophile to form a six-membered ring [reaction (24)]. Another type of cycloaddition involves a dipo-

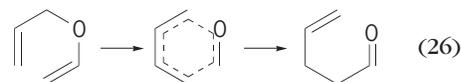


lar molecule such as a nitrone rather than a diene; the product in this case is a five-membered heterocyclic compound [reaction (25)].



See DIELS-ALDER REACTION; PERICYCLIC REACTION.

In a sigmatropic rearrangement, a bond in a diene chain is broken and a new bond is formed. The chain may contain heteroatoms and may be part of a cyclic structure [reaction (26)].



James A. Moore

Bibliography. F. A. Carey and R. J. Sundberg, *Advanced Organic Chemistry* pts. A and B, 4th ed., 2000; E. J. Corey and X. M. Cheng, *The Logic of Chemical Synthesis*, 1989; T. L. Gilchrist, *Heterocyclic Chemistry*, 3d ed., 1997; H. Hart, *Organic Chemistry: A Short Course*, 9th ed., 1994; J. Lambert et al., *Organic Structural Analysis*, 1976; T. H. Lowry and K. S. Richardson, *Mechanism and Theory in Organic Chemistry*, 3d ed., 1987; J. March, *Advanced Organic Chemistry: Reactions, Mechanisms and Structure*, 5th ed., 2001; A. Streitwieser, Jr., C. H. Heathcock, and E. M. Kosower, *Introduction to Organic Chemistry*, 4th ed., 1992; K. P. C. Vollhardt and N. E. Schore, *Organic Chemistry*, 3d ed., 1998; K. Weissmermel and H.-J. Arpe, *Industrial Organic Chemistry*, 3d ed., 1997.

## Organic conductor

An organic substance with low electrical resistance. Two major classes of organic conductors are charge-transfer compounds and conducting polymers.

### Charge-Transfer Compounds

The charge transfer compounds, a class of well-ordered molecular crystals, display metallic-like electrical conduction and many other unusual phenomena, such as the stabilization of charge-density waves

and spin-density waves, unusual mechanisms for electronic transport, organic superconductivity, and unusual states of matter produced under strong magnetic fields. Most of these phenomena are due to the low dimensionality (one or two dimensions) of the electron gas in these compounds. Superconductivity is observed at temperatures as high as 10 K ( $-442^{\circ}\text{F}$ ). In addition, experiment and theory suggest a Cooper-pair binding mechanism for electrons in these conducting compounds which, unlike that of normal inorganic superconductors, does not require lattice vibrations but may be mediated by magnetic interactions between electrons. See SUPERCONDUCTIVITY.

**Molecular structure.** Charge-transfer compounds are two-component materials containing anionic and cationic species originating by charge transfer between donor and acceptor entities; these may be two organic molecules or an organic molecule with an inorganic ion. Representative systems are planar molecules with  $\pi$ -orbitals around carbon and heteroatoms with  $p_z$  atomic orbitals (Fig. 1). Tetrathiafulvalene-tetracyanoquinodimethane (TTF-TCNQ) is the prototype of charge-transfer organic crystals. Its crystal structure exhibits piled-up segregated columns of donor TTF and acceptor TCNQ molecules (Fig. 2a). See MOLECULAR ORBITAL THEORY.

Another class of organic conductors is exemplified by radical cation salts such as  $(\text{TMTSF})_2\text{X}$  where the organic molecule is tetramethyltetraselenafulvalene and X is an inorganic anion such as phosphorus hexafluoride ( $\text{PF}_6^-$ ), perchlorate ( $\text{ClO}_4^-$ ), or nitrate ( $\text{NO}_3^-$ ). In this class of materials (Bechgaard salts) the

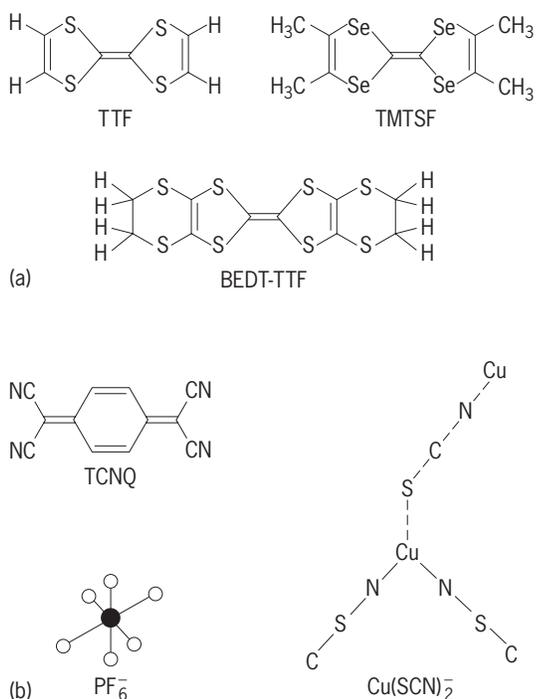


Fig. 1. Structures of organic conductors. (a) Electron donors. (b) Acceptor molecule (TCNQ) and two monovalent anions. Abbreviations are explained in the text.

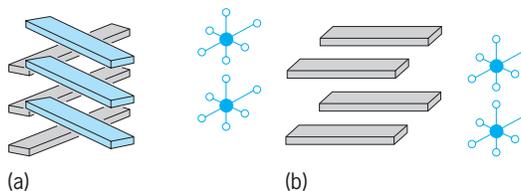


Fig. 2. Stacking of charge-transfer compounds. (a) Segregated stacking of TTF-TCNQ-like materials. (b) Typical zigzag stacking of the  $(\text{TMTSF})_2\text{X}$  series.

molecules display a zigzag packing along the stacking axis (Fig. 2b), where one positive charge (hole) is shared between two organic molecules.

**One-dimensional conduction.** The strong overlap between electron clouds of neighboring molecules along the stacks spreads the partially filled molecular electronic states into an energy band 0.5–1 eV wide. This bandwidth is large enough to allow electron delocalization among all molecules on a given stack and to promote electrical conduction similar to that in metal crystals. The resistivity of organic conductors is between  $2 \times 10^{-2}$  and  $1 \times 10^{-3} \Omega \cdot \text{cm}$  at ambient temperature and shows a large decrease on cooling, much as with pure metals. A peculiarity of most organic crystals is strong directionality: the conductivity is 100 to 1000 times lower along the directions perpendicular to the packing axis. See BAND THEORY OF SOLIDS; DELOCALIZATION; ELECTRICAL CONDUCTIVITY OF METALS.

**Charge- and spin-density waves.** Unlike ordinary metals, which remain metallic (highly conducting) down to millikelvin temperatures, organic conductors often undergo low-temperature phase transitions toward insulating states. Monochromatic x-ray diffraction of the insulating state of TTF-TCNQ revealed weak satellite diffraction peaks in the vicinity of each Bragg peak related to the crystal structure. These satellites point to the existence of a lattice-distorted Peierls state (charge-density wave) in TTF-TCNQ, following the argument that a uniform conducting one-dimensional electron gas is unstable at low temperature against a lattice distortion. See CHARGE-DENSITY WAVE; PHASE TRANSITIONS; X-RAY DIFFRACTION.

In  $(\text{TMTSF})_2\text{PF}_6$ , however, no such satellites of Bragg peaks are found at low temperature. Instead, magnetic data reveal the onset of an internal magnetic modulation (spin-density wave) occurring together with the phase transition at 12 K. In this case, repulsive electrostatic interactions between electrons and low dimensionality are responsible for the existence of the magnetic ground (Overhauser) state. See SPIN-DENSITY WAVE.

**Superconductivity.** The compound  $(\text{TMTSF})_2\text{PF}_6$  exhibits unusual behavior under pressure. The magnetic and insulating ground state is rapidly suppressed. Above 9 kilobars (0.9 gigapascal), superconductivity accompanied by zero resistance and magnetic flux expulsion (the Meissner effect) is observed below the critical temperature  $T_c \approx 1 \text{ K}$  ( $-458^{\circ}\text{F}$ ). Other investigations of the  $(\text{TMTSF})_2\text{X}$

series have shown that with the proper anion (in particular  $\text{ClO}_4^-$ , which is smaller than  $\text{PF}_6^-$ ) superconductivity can be stabilized at ambient pressure below 1.2 K ( $-457.5^\circ\text{F}$ ). Significantly higher critical temperatures have been obtained in the quasi-two-dimensional compound  $(\text{BEDT-TTF})_2\text{Cu}(\text{SCN})_2$ , where the organic molecules bis(ethylenedithio)tetrathiafulvalene (BEDT-TTF; also called the ET molecule) are arranged in two-dimensional conducting sheets sandwiched between insulating copper thiocyanate  $[\text{Cu}(\text{SCN})_2]$  layers.

**High-magnetic-field phenomena.** Apart from superconductivity, the response of the electron gas in  $(\text{TMTSF})_2\text{ClO}_4$  to an applied magnetic field is highly unusual. Under high magnetic fields the conducting state undergoes a transition to a new state above a threshold field that is accompanied by a drastic reduction in the density of carriers from 1 to about  $10^{-2}$  per unit cell. In addition the high-field state exhibits an antiferromagnetic modulation. The interpretation of these phenomena lies in the reinforcement of the one-dimensional character of the quasi-one-dimensional electron gas by the applied magnetic field. Furthermore, the field-induced state reveals a sequence of subphases for which the Hall coefficient remains independent of the magnetic field. This effect clearly resembles the quantum Hall effect in a two-dimensional electron gas. See HALL EFFECT. D. Jérôme

### Conducting Polymers

Polymeric materials are typically considered as insulators, and in fact important applications rely on this property. However, research since the late 1970s has led to the discovery of polymeric materials with extremely high conductivity, approaching that of copper. The prospect of materials combining the properties of plastics and metals or semiconductors has led to the pursuit of applications. Improved polymers, particularly polypyrrole and polyaniline (see table), no longer suffer from such drawbacks as low stability, processing difficulties, and brittleness. Most conducting polymers can be switched reversibly between conductive and nonconductive states, with

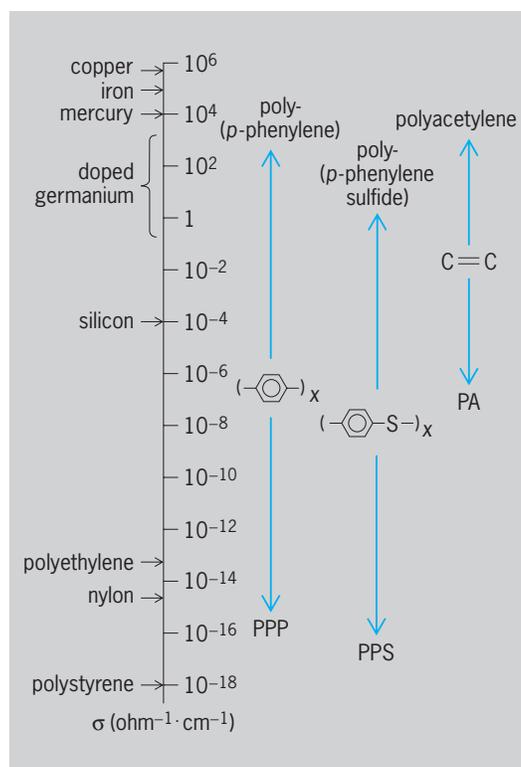


Fig. 3. Comparison of the range of electrical conductivities of some  $\text{AsF}_5$ -doped conducting polymers with some metals, semiconductors, and nonconducting polymers.

the result that their conductivities can span an enormous range (Fig. 3). This switching is accomplished through redox chemistry, the conductivity being sensitive to the degree of oxidation of the polymer backbone. This property distinguishes conducting polymers from metals and semiconductors and is the basis of many existing and potential applications. Also, certain polymers become conducting upon oxidation or reduction and thus can exhibit either *p*- or *n*-type conduction. See OXIDATION-REDUCTION; POLYMER; SEMICONDUCTOR.

**Structure and properties.** An oxidant is used to create carriers, and the resulting room-temperature

#### Examples of dopole polymers, oxidants, dopant ions, and resulting conductivities

Polymer	Formula	Oxidant	Dopant	Conductivity, $(\Omega \cdot \text{cm})^{-1}$
Polyacetylene	$(-\text{CH}=\text{CH}-)_n$	$\text{I}_2$	Triiodide ( $\text{I}_3^-$ )	10,000
Polypyrrole		Electrode (anode)	Tetrafluoroborate ( $\text{BF}_4^-$ )	200
Poly(3-hexylthiophene)		Electrode (anode)	Perchlorate ( $\text{ClO}_4^-$ )	30
Polyaniline (emeraldine form)		None (protonation leads to internal redox reaction)	$\text{Cl}^-$ , $\text{CH}_3-\text{C}_6\text{H}_4-\text{SO}_3^-$	$\sim 200$

conductivities, in conducting polymers (see table). The oxidant removes electrons from the  $\pi$ -electron system of the polymer, creating radical cations that, at high concentrations, dimerize to form cation pairs known as bipolarons. Charge-balancing counterions are concomitantly incorporated between polymer chains. The overall process is referred to as doping, and the counteranion (or counteranion in the case of reduction) is the dopant. Doping can be carried out by exposing the polymer to vapors or solutions of oxidant or reductant. The polymer can also be oxidized or reduced electrochemically, incorporating dopant ions from the electrolyte. This method is frequently preferred, since dopant ions of precise structure can be introduced and the degree of oxidation or reduction controlled by the electrochemical potential.

Conducting polymers have delocalized  $\pi$ -electrons at least over a few structural units. Such unsaturated polymers favor carrier generation because of the possibility of resonance delocalization of the resulting radical ions; good intramolecular carrier mobility may also result. In addition, the geometry of  $\pi$ -orbitals allows for good orbital overlap and encourages intermolecular carrier transport. The fact that the polymer chains are much shorter than typical sample dimensions indicates that intermolecular transport is dominant, especially in view of the disorder observed in most conducting polymer systems. However, as the number of defects (cross-links or twists that inhibit conjugation) decreases, it might be anticipated that conductivity would be higher because carriers can travel greater distances along a chain before an intermolecular electron transfer of higher activation energy becomes necessary. Indeed, certain stretch-oriented polymers afford significantly greater conductivities than their unoriented counterparts. See CONJUGATION AND HYPERCONJUGATION.

Considerable progress has been made on the theory of important parameters such as oxidation potentials, band gaps, and band widths, often with good agreement with experiment. Such work is important for the design of new conductive polymers with specific properties. The elucidation of conduction mechanisms has also received much attention. To a first approximation, carrier transport may be viewed as a combination of intrachain resonance delocalization processes and interchain redox processes involving electron transfer between segments of neutral and oxidized (or reduced) polymer. The interchain processes are reminiscent of intrastack conduction in small-molecule organic radical ion salts (discussed above). See ELECTRON-TRANSFER REACTION.

**Processibility.** Conducting polymers are frequently intractable to processing, although advances in this area have alleviated this problem. For example, alkyl-substituted polythiophenes, unlike the parent polythiophene, are soluble in common solvents and yet have nearly comparable conductivities when doped. Polyacetylene and poly(*p*-phenylene vinylene) can be fabricated into

films and fibers from soluble precursor polymers. Doped poly(*p*-phenylene sulfide) can be solubilized if the doping is carried out in solvents such as arsenic fluoride (AsF<sub>3</sub>). Many approaches are available for direct synthesis of conducting polymers as films. See FILM (CHEMISTRY).

Melt-blending of conducting polymer powders with thermoplastic polymers offers the opportunity to combine conductivity of the conducting polymer filler with desirable mechanical properties of the thermoplastic matrix. Blends of doped polyaniline and poly(vinyl chloride) are available (with the dopants for polyaniline, such as *p* toluenesulfonate, selected to afford maximum thermal stability).

**Applications.** The utility of most conducting polymers lies in their unique properties or functions, rather than replacing existing materials. Of particular interest is the fact that many of these materials can be oxidized and reduced reversibly, with concomitant changes in conductivity, optical absorption, and wettability.

Conductive polymers appear to be promising materials as battery cathodes or anodes, and may even function as both anode and cathode in a cell (the all-polymer battery). Electrochemical doping of polyacetylene at an anode or cathode is a charging process. Placement of a load across the polymer electrodes results in discharge as the doped polymers neutralize each other, the oxidized polymer (anode) becoming reduced and vice versa. The high surface areas of many conducting polymer films and powders provide for useful power densities. See BATTERY.

Other applications include corrosion-inhibiting coatings on steel, biosensors, and matrices for the timed release of chemicals. In the last case, advantage is taken of the fact that the counteranions in *p*-type (oxidized) conducting polymers are released into solution when the polymer is reduced electrochemically. However, it may be desirable for cations rather than anions to be released. This can be done with a polypyrrole-poly(styrene sulfonate) composite. The appropriate cation is bound initially to the poly(styrene sulfonate), although upon electrochemical oxidation of polypyrrole the sulfonate moieties of the poly(styrene sulfonate) become the counteranions for the conductive polymer, releasing their previously bound cations into solution. Cations can also be bound with self-doped polyheterocyclic compounds containing pendant sulfonates. The fact that neutral (undoped) polymers are typically hydrophobic whereas their doped forms are polar salts has led to the construction of ion-gate membranes with variable ion permeabilities, depending on the electrochemically controllable oxidation state of the polymer. Doping also changes optical absorption properties, and thus many conducting polymers exhibit electrochromism with possible utility in display technology and so-called smart windows. Microelectrochemical transistors are also fabricated from conducting polymers by using techniques common in the semiconductor industry. See COMPOSITE

MATERIAL; ELECTROCHEMISTRY; ELECTROCHROMIC DEVICES; TRANSISTOR.

Poly(phenylene vinylene) and related conjugated polymers (but without doping) exhibit electroluminescence, wherein carriers of opposite sign generated at attached electrodes recombine in the polymer with emission of light. The color of the light can be controlled by tailoring the polymer structure, opening the prospect of flexible, color-tuned displays. See ELECTROLUMINESCENCE; SOLID-STATE CHEMISTRY; SOLID-STATE PHYSICS. Gary E. Wnek

Bibliography. P. Chandrasekhar, *Conducting Polymers, Fundamentals and Applications: A Practical Approach*, 1999; J.-P. Farges (ed.), *Organic Conductors, Fundamentals and Applications*, 1994; R. B. Kaner and A. G. MacDiarmid, Plastics that conduct electricity, *Sci. Amer.*, 258(2):106-108, February 1988; V. Z. Kresin and W. A. Little (eds.), *Organic Superconductivity*, 1990; L. Ouahab and E. Yagubskii (eds.), *Organic Conductors, Superconductors and Magnets: From Synthesis to Molecular Electronics*, 2004; T. A. Skotheim, R. L. Elsenbaumer, and J. R. Reynolds (eds.), *Handbook of Conducting Polymers*, 2d ed., 1997.

## Organic evolution

Organic, or biological, evolution is the modification of living organisms during their descent, generation by generation, from common ancestors. It is to be distinguished from other phenomena to which the term evolution is often applied, such as chemical evolution, cultural evolution, or the origin of life from non-living matter. Organic evolution includes two major processes: anagenesis, the alteration of the genetic properties of a single lineage over time; and cladogenesis, or branching, whereby a single lineage splits into two or more distinct lineages that continue to change anagenetically.

**Study of evolution.** The subject matter of evolutionary biology may be roughly divided into the analysis of the history of evolutionary events and the analysis of the mechanisms, or processes, of evolutionary change. The study of evolutionary history attempts to determine the ancestry of and genealogical relationships among different kinds of organisms, the pathways by which their morphological, biochemical, and other features have become modified, the history by which they arrived at their present geographical distributions, and the changes in the diversity and number of species throughout geological time. The methods by which such inferences are made include analysis of the fossil record and the phylogenetic analysis of living taxa, many having an inadequate fossil record. Phylogenetic analysis, using data on the comparative anatomy, molecular characteristics [for example, protein and deoxyribonucleic acid (DNA) sequences], and geographical distributions of organisms, is part of the province of biological systematics.

The analysis of the mechanisms of evolutionary change addresses primarily the factors that cause

changes in the genetic composition of populations and species, and those that influence diversification and extinction of species. The mathematical theory of population genetics is important to this enterprise. Experimental and observational testing of the theory includes molecular, genetic, and developmental analysis of genetic variation and the mechanisms by which it arises; ecological genetics, the study of the impact of ecological factors on genetic change of populations; studies in functional morphology, physiology, behavior, and ecology that address the adaptive value of genetically different traits; and taxonomic and phylogenetic analyses that shed light on processes such as cladogenesis. Thus the study of evolution embraces all of biology.

**History.** Although some ancient Greek philosophers had vague, often mythological, intimations of evolution, Platonic and Aristotelian philosophy, in which variation represented imperfect reflection of eternal, unchanging essences or "ideas," was antithetical to evolution. The adoption of this framework by Christian theology, and the literal interpretation of the first chapters of Genesis, led to the belief that all living things had been directly created in their current form (special creation) only a few thousand years ago. The first challenges to this view did not arise until the eighteenth century, when speculations on cosmic change, on the antiquity and dynamic nature of the Earth, and on human progress led naturally to the idea that living things might change as well. The French biologist G. de Buffon (1707-1788) was one of the first to hint at evolution, but his countryman J. de Lamarck (1744-1829) was the first to argue forcefully for evolution and to propose a mechanism by which it might occur. Lamarck supposed that new forms arise continually by spontaneous generation, and then progress toward greater complexity and perfection because of "powers conferred by the supreme author of all things" and because their behavioral responses to the environment cause changes in their structure. Lamarckism was soon discredited by the implausibility and vagueness of the mechanisms he postulated, and by forceful arguments against evolution by leading French biologists of the day. Nevertheless, the idea of evolutionary change was "in the air" in the early nineteenth century.

Charles Darwin (1809-1882), son of an English physician, apparently came to think of the possibility of evolution toward the end of his 5-year (1831-1836) voyage as naturalist on the H.M.S. *Beagle*. He conceived of the theory of natural selection in 1838, and spent the next 20 years synthesizing and amassing evidence and refining his ideas until, faced with the possibility that the young naturalist Alfred Russel Wallace, who had independently thought of natural selection, might gain priority for the idea, he published an "abstract" of the massive book which he had been preparing. The "abstract" was a 490-page book, *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life*, published on November 24, 1859. Darwin spent the rest of his life conducting

research and writing books on an extraordinary range of subjects, from the power of movement in plants to the “descent of man,” all of which are related directly or indirectly to the themes in his most famous book. But *The Origin of Species*, as it is generally known, is his triumph, one of the most important works in the history of western civilization.

*The Origin of Species* accomplishes two things. Darwin marshals evidence from every quarter of biology and geology that evolution has in fact occurred: that living things are descended with modification from common ancestors. Second, he presents an explicit, purely mechanistic theory of the causes of evolution. Every species, Darwin points out, has hereditary variation in numerous characteristics. Some variants will be better suited to the exigencies of life than others, and so will survive better or reproduce more prolifically than the inferior variants. Since descendants inherit their superior properties, the proportion of individuals in the population that bear superior characteristics will increase, and the proportion with inferior traits will decrease from generation to generation, until the species has been transformed. The new character itself is subject to further variation and to further alteration by this process of natural selection, so that in the vastness of time the feature comes to differ extremely from the original form—but it is a great change accomplished in small steps. Because different populations experience different environments and adapt to different resources, numerous forms may diverge from an original stock, each adapted to a different environment or way of life. This branching process, continued through the immensity of geological time, gives rise to the great “tree of life.”

Darwin's ideas at first met with much opposition, but within 20 years after the *Origin* appeared, most scientists had been convinced of the reality of evolution, and many were stimulated to work in the areas of paleontology, comparative anatomy, and comparative embryology that provide evidence of the historical (phylogenetic) relationships among organisms. But Darwin's theory of the cause of evolution, natural selection, was not widely accepted for lack of sufficient evidence. It fell even deeper into disrepute in the early twentieth century, when the new science of genetics (developing after the rediscovery in 1900 of Mendel's work) seemed to provide alternative mechanisms for evolution, such as mutation.

Modern evolutionary theory began in the late 1920s and early 1930s, when naturalists and systematists amassed evidence on adaptive variation and the nature of species, and when the mathematical geneticists R. A. Fisher and J. B. S. Haldane in England and S. Wright in the United States developed equations showing that natural selection and mendelian genetics (including mutation) are not alternatives, but rather work in combination just as Darwin had proposed. Their theories, which include other causes of evolution in addition to the coaction of natural selection and mutation, are the foundation of mathematical population genetics. See POPULATION GENETICS.

During the 1930s through the early 1950s, the theory of population genetics, together with the ideas of experimental workers and taxonomists concerned with evolution, congealed into the “modern synthesis” commonly known as neo-Darwinism. The experimental geneticists S. S. Chetverikov, Th. Dobzhansky, and E. B. Ford, the zoologists E. Mayr, B. Rensch, and J. S. Huxley, the paleontologist G. G. Simpson, and the botanist G. L. Stebbins, were among the chief figures who, together with Wright, Fisher, and Haldane, formulated a coherent, comprehensive theory of evolution and showed that it was consistent with data from diverse fields of study. Since the modern synthesis, new data and theories concerning especially evolution at the molecular level, evolution of behavior and of ecological interactions among species, and long-term patterns of evolution over geological time have been added to the neo-Darwinian framework, and some neo-Darwinian ideas have been called into question. Thus, although most of the neo-Darwinian theory is accepted by most evolutionary biologists, there remain controversies, as in any other field of science, about some aspects of the theory. There is no controversy, however, about the reality of evolution as a historical event. That organisms have descended from common ancestors is accepted by knowledgeable biologists as fact. Molecular and other similarities imply that all living things are related to each other by common ancestry.

**Mechanisms of species transformation.** Anagenesis consists of change in the genetic basis of the features of the organisms that constitute a single species. Populations in different geographic localities are commonly considered members of the same species if they exchange members at some rate and interbreed with each other (or are thought to be potentially able to interbreed); but unless the level of interchange (gene flow) is very high, some degree of genetic difference among different populations is likely to develop. The changes that transpire in a single population may be spread to other populations of the species by gene flow. See SPECIES CONCEPT.

**Genetic variation.** Almost every population harbors several different alleles (forms of a gene) at each of a great many of the gene loci; hence many characteristics of a species are genetically variable. This is evident by the existence of morphological and physiological variations that are shown to be inheritable, and by recent study of genes themselves (deoxyribonucleic acid, or DNA, sequences) and of their products (proteins). Studies of protein variation on *Drosophila*, humans, and numerous other species suggest that 40% or more of the gene loci may be variable within a population, and analyses of DNA suggest that almost all of a species' genes vary. Thus, to evolve in response to many environmental changes, a population need not wait for new mutations to occur; genetic variations that happen to be suitable for new conditions may already be present. For example, genes for insecticide resistance are present in low frequency in populations of insects that have never been exposed to insecticides.

**Sources of genetic variation.** All genetic variations ultimately arise by mutation of the genetic material. Broadly defined, mutations include changes in the number or structure of the chromosomes and changes in individual genes, including substitutions of individual nucleotide pairs, insertion and deletion of nucleotides, and duplication of genes. Many such mutations alter the properties of the gene products (RNA and proteins) or the timing or tissue localization of gene action, and consequently affect various aspects of the phenotype (that is, the morphological and physiological characteristics of an organism). Whether and how a mutation is phenotypically expressed often depends on developmental (epigenetic) events, some of which may “canalize” the phenotype to reduce phenotypic variability even if genetic variation exists. The phenotypic effect of mutations ranges from extremely subtle to drastic; many drastic mutations alter development so greatly that the organism’s viability is severely reduced. However, some large mutations, as well as more subtle mutations which slightly change body size, the form of an appendage, the activity of an enzyme, or other features, are deleterious, neutral, or beneficial, depending on the organism’s environment. Because a gene typically affects several aspects of the phenotype, it may have both harmful and beneficial effects (pleiotropy), the relative magnitude of which determines the net effect of the mutation on survival and reproduction. Because different genes interact in determining phenotypes, the positive or negative value of a mutation may depend on which alleles are prevalent at other gene loci.

The frequency with which a given gene mutates to a recognizably different allele is fairly low (often about 1 mutation per gene per 100,000 gametes), but can be considerably higher when mutation is caused by factors such as transposable segments of DNA that become inserted into a gene and affect its function. Although the mutation rate of any given gene is low, the total flux of mutations is appreciable (at least one new mutation somewhere in the genome of each gamete) because the number of genes is high. Mutation occurs spontaneously because of molecular “noise,” but can be enhanced by factors such as heat, radiation, and chemical mutagens. *See* MUTATION.

Mutations are generally believed to occur at random, not in the sense that all genes mutate at the same rate or that all possible mutations of a gene are equally likely, but rather in the sense that the likelihood of a particular mutation is not influenced by whether it would be advantageous to the organism in its prevailing environment. Genes cannot respond to the environment by mutating in an appropriate direction; nor can the experiences of an individual organism enable it to change in an adaptive manner the genetic basis for the characteristics that the organism uses in its relations to the environment. This is the principal point of difference between the neo-Darwinian and the Lamarckian theories of evolution.

Recombination, arising from sexual reproduction and crossing-over, combines the mutant variants at various loci into a vast number of possible combi-

nations; thus mutation and recombination together generate immense genetic variability among the members of a population. When, as is usually the case, many different loci each contribute to a particular trait such as body size, the result is “continuous” variation, often in the form of a bell-shaped (normal) distribution of the trait in the population. *See* CROSSING-OVER (GENETICS); DISTRIBUTION (PROBABILITY).

**Natural selection.** The fundamental event in evolution is a change in the frequency of an allele in a population. In its full form, this entails the spread through a population of an allele that, having just come into existence by mutation, is very rare, but which ultimately comprises 100% of the gene copies at that locus (in a population of  $N$  diploid organisms, there are  $2N$  gene copies at a locus that is not sex-linked). The allele is then said to have been fixed in the population. One of the factors that causes this process is natural selection.

Natural selection is a consistent difference in the average rate at which genetically different entities leave descendants to subsequent generations; such a difference arises from differences in fitness (that is, in the rate of survival, reproduction, or both). In fact, a good approximate measure of the strength of natural selection is the difference between two such entities in their per capita rate of increase  $r$ . The entities referred to are usually different alleles at a locus, or phenotypically different classes of individuals in the population that differ in genotype. Thus selection may occur at the level of the gene, as in the phenomenon of meiotic drive, whereby one allele predominates among the gametes produced by a heterozygote. Selection at the level of the individual organism, the more usual case, entails a difference in the survival and reproductive success of phenotypes that may differ (in body size, for instance) at one locus (for example, genotypes  $AA$  and  $aa$  differ in size) or at more than one locus (for example,  $AABBCC$  versus  $aabbcc$ , where different letters denote different loci and upper- and lowercase letters denote different alleles at a locus). As a consequence of the difference in fitness, the proportion of one or the other allele increases in subsequent generations (**Fig. 1**). Numerous cases of such differences in fitness, often of very considerable magnitude, have been documented in both laboratory-maintained and natural populations of many species. *See* POPULATION ECOLOGY.

The relative fitness of different genotypes usually depends on environmental conditions. Thus, for example, brown shell color in terrestrial snails is advantageous in forests, providing cryptic protection against predators, whereas yellow coloration is advantageous in open fields. Relative fitnesses of genotypes often switch as the environment changes from season to season or year to year, and the frequencies of genotypes often fluctuate in consequence. Thus changes in genetic composition are the passive consequence of changes in environment; a population cannot alter its genetic composition in anticipation of some future environmental change. This implies that populations cannot adapt so as to avoid future

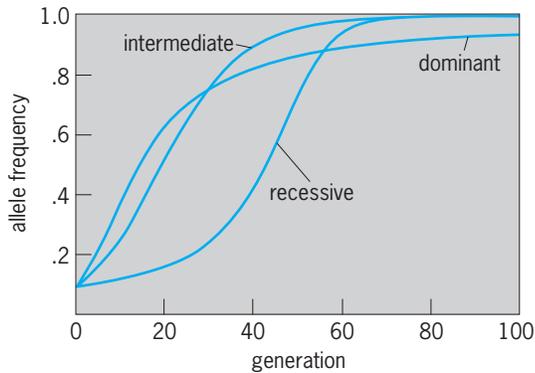


Fig. 1. Increase in the frequency of an advantageous allele within a population, from an initial frequency of 0.1. The fitness, or relative rate of increase, of the fittest genotype is 20% greater than that of the least fit genotype in each of three cases, in which the fittest genotype is dominant, recessive, or partially dominant (intermediate). (After D. J. Futuyma, *Evolutionary Biology*, Sinauer Associates, 1979)

extinction. Natural selection does not act for the benefit of the species as a whole; it is a purely mechanical, mathematical phenomenon: the difference in reproductive success among individual members of the species.

Selection is sometimes directional, meaning that one extreme state of a phenotypic feature is most fit and ultimately becomes fixed if the environmental conditions to which it is adapted prevail long enough (Fig. 2). Often, however, selection is stabilizing, meaning that an intermediate phenotype is most fit. If this phenotype is produced by a heterozygote at a single locus, selection maintains both alleles in the population. Several cases of stabilizing selection are known in which the heterozygote is most fit. Genetic variation can also be maintained by diversifying selection, whereby several different phenotypes are favored. Diversifying (disruptive) selection includes the phenomenon of frequency-dependent selection: the rarer a genotype is, the higher its fitness becomes. This phenomenon often arises because of interactions among the members of a population, each of which constitutes part of the environment of every other member. For example, genotypes often differ in their relative ability to use one or another resource, such as different kinds of food. If there is competition for limited food, a rare genotype experiences less competition for its particular food type than the common genotype does for its particular food type; thus the per capita rate of increase of the rarer genotype is relatively greater. When it attains high frequency, the tables are turned, so an equilibrium is attained at which both genotypes persist.

Under directional selection, new mutations that enhance a particular trait may be consistently advantageous over the long run. For example, all other things being equal, greater height is advantageous for a tree because shorter trees are overtopped by their taller neighbors, and so do not receive as much light. But there is a limit to the height that natural selection favors, because countervailing selection pressures come into play: an excessively tall tree will be

topped by wind unless it has a sufficiently strong trunk and root system. Because wind conditions differ from one place to another, and the trunk and root structure differs among species, the relative strength of conflicting selection pressures differs among populations and species, which therefore evolve along divergent lines. In addition to divergent evolution, however, evolution is often convergent: different organisms sometimes experience similar selection pressures and so evolve at least superficially similar characteristics.

**Random genetic drift.** Different alleles of a gene that provides an important function do not necessarily differ in their effect on survival and reproduction; such alleles are said to be neutral. The proportion of two neutral alleles in a population fluctuates randomly from generation to generation by chance, because not all individuals in the population have the same number of surviving offspring. Random fluctuations of this kind are termed random genetic drift. If the process continues long enough in the absence of countervailing factors, one or another allele will ultimately fluctuate all the way to fixation, and the other alleles will be lost from the population by chance (Fig. 3). This process occurs more rapidly, the smaller the population. Many natural populations are quite small; in fact, their effective size, which can be thought of as the number of individuals that actually succeed in reproducing, is considerably smaller than the total number of individuals. Thus all populations are susceptible to genetic drift, and the process

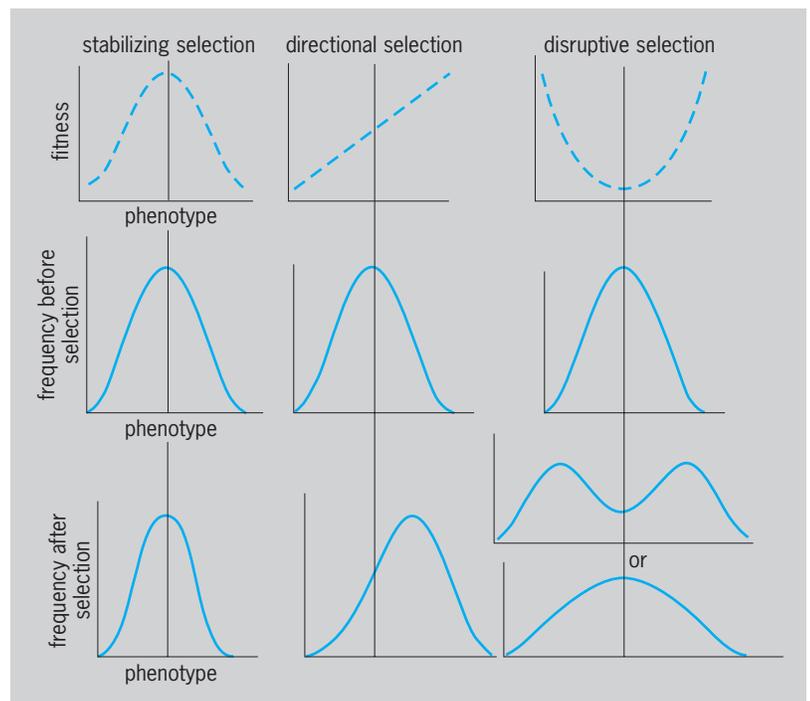
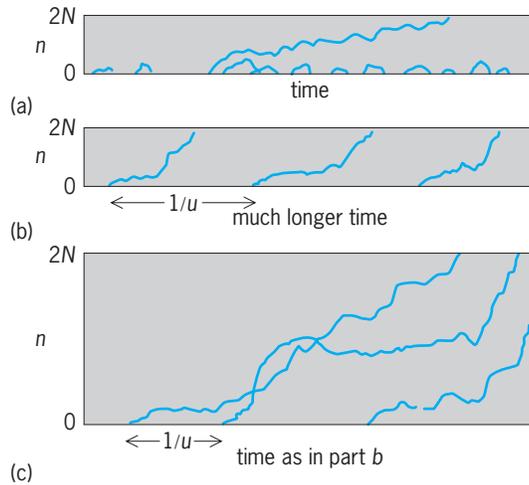
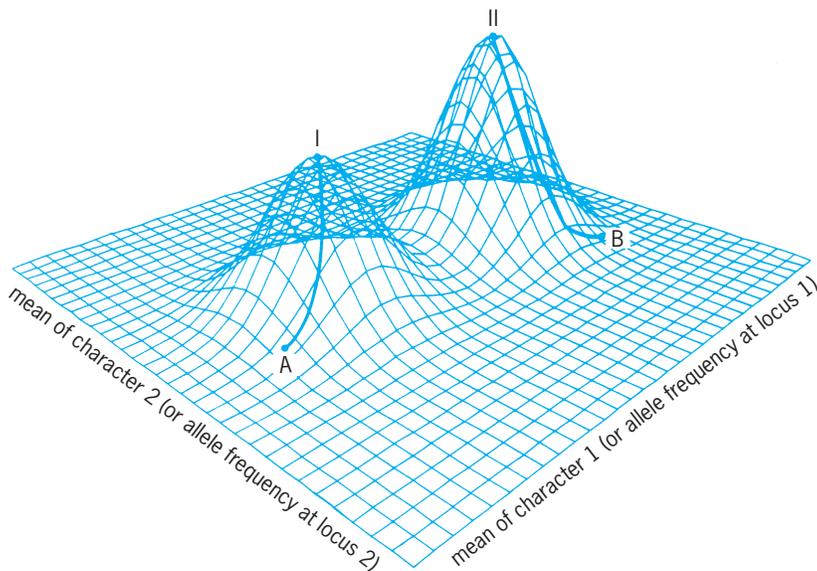


Fig. 2. Three modes of selection on a quantitatively varying phenotypic character. Because of the relation between fitness and phenotype, portrayed in the upper panels, the frequency distribution of the trait in the population changes during one generation of selection from the patterns shown in the middle to those shown in the lower panels. (After L. L. Cavalli-Sforza and W. F. Bodmer, *The Genetics of Human Populations*, W. H. Freeman, 1971)



**Fig. 3. Gene substitution by genetic drift.** A population of  $N$  diploid individuals has  $2N$  genes at any locus. The number of copies  $n$  of a mutation fluctuates by chance. (a) Numerous mutations arise, of which one is fixed while others increase and then decrease in frequency. (b, c) Only those mutations at various loci which are fixed are shown. The average number of generations between the origin of successive mutations that are ultimately fixed by genetic drift is  $1/u$ , where  $u$  is the mutation rate. In a large population, shown in c, several loci are polymorphic at any time, as it takes longer for mutations to reach fixation. (After J. F. Crow and M. Kimura, *An Introduction to Population Genetics Theory*, Burgess, 1970)

is likely to proceed quite rapidly in many species. Because the identity of the fixed allele is only a matter of chance, the allele that becomes fixed will differ from one population to another, so that the genetic composition of different populations or of different species diverges over time, at a rate inversely propor-



**Fig. 4. Natural selection may drive a population to different genetic equilibria depending on initial allele frequencies.** The elevation of a point on this landscape represents the mean fitness of a population that has a particular average value for each of two traits (or that has a particular gene frequency at each of two loci). Character combinations I or II represent high fitness, whereas intermediate phenotypes have low fitness. A population beginning with average phenotypes represented by point B will evolve to equilibrium II, whereas one beginning at point A will become stabilized by natural selection at equilibrium I, even though this represents an inferior condition relative to II. (After D. J. Futuyma and M. Slatkin, *Coevolution*, Sinauer Associates, 1983)

tional to their effective population size. The theory of genetic drift has been validated by laboratory experiments and by patterns of genetic difference among natural populations of species.

**Drift versus selection.** If different alleles do indeed differ in their effects on fitness, both genetic drift and natural selection operate simultaneously. The deterministic force of natural selection drives allele frequencies toward an equilibrium, while the stochastic (random) force of genetic drift brings them away from that equilibrium. The outcome for any given population depends on the relative strength of natural selection (the magnitude of differences in fitness) and of genetic drift (which depends on population size), just as the trajectory of a dust particle in still air depends on the relative power of gravitation and of brownian motion. Thus the fate of an allele that differs in fitness only slightly from other alleles may be dominated by selection if the population is large, but by genetic drift if the population is small.

The relative importance of genetic drift versus natural selection is the subject of considerable controversy and research, the chief focus of which is on genetic differences at the molecular level, such as different forms of an enzyme or slight differences in nucleotide sequences in DNA. There is reason to believe that many such differences have only a slight impact on fitness and that considerable evolution at the molecular level occurs primarily by genetic drift. The best evidence is provided by synonymous codons of DNA and by the nucleotide sequences of DNA that are not transcribed into RNA and protein. Many such changes at the DNA level are unlikely to affect the organism's fitness, yet comparison among species shows that these sequences have evolved at a high rate—higher than that of DNA sequences that are functional. Fixation of purely neutral alleles, according to mathematical theory, should occur at a constant rate per generation, and there is some evidence that the rate at which species have diverged at the molecular level has been moderately constant. Thus some evolutionary change has certainly transpired by genetic drift, although divergence in clearly adaptive features of morphology, physiology, and behavior has certainly come about primarily by natural selection.

In theory, genetic drift can act together with natural selection to enhance adaptation of a species to its environment. Often the pressure of natural selection can drive the population toward any one of several different genetic constitutions: alternative genetic "solutions" to the same "problem" (Fig. 4). Which genetic constitution is achieved depends on which alleles happen to be most prevalent when the environmental challenge arises. The genetic "solution" actually achieved may be inferior to another genetic constitution that might have been achieved had the initial genetic composition of the population been different; yet the equations of gene frequency change show that natural selection alone cannot move the population from one stable genetic equilibrium to a different, superior, genetic equilibrium. If the population is small enough, however,

genetic drift can destabilize the population, carrying the gene frequencies away from the inferior equilibrium enough for natural selection to bring them to a new, superior, genetic configuration. This theory is part of Sewall Wright's shifting-balance theory of genetic change.

**Speciation and cladogenesis.** The great diversity of organisms has come about because individual lineages (species) branch into separate species, which continue to diverge by the processes described above. This splitting process, speciation, occurs when genetic differences develop between two populations that prevent them from interbreeding and forming a common gene pool. The genetically based characteristics that cause such reproductive isolation are usually termed isolating mechanisms, but there is little reason to believe that they evolve specifically to prevent interbreeding, as the unfortunate term mechanism implies. Rather, reproductive isolation seems to develop usually as a fortuitous by-product of genetic divergence that occurs for other reasons (either by natural selection or by genetic drift).

The most common mode of speciation is undoubtedly genetic divergence among populations that are sufficiently spatially isolated that their gene pools are not homogenized by gene flow. This allopatric mode of speciation may occur when two widespread populations are separated by unsuitable habitat (for example, European and American populations), but is probably more frequent and more rapid when a population in a restricted locality is cut off (for example, by colonization across a habitat barrier) from the main body of the species, and undergoes rapid divergence because of genetic drift and different selection pressures. If sufficient genetic divergence transpires before these populations expand and encounter each other, they will not exchange genes when they meet; if divergence has been insufficient, they interbreed and speciation has not been completed.

This genetic theory of speciation is well supported by evidence from taxonomists' analyses of the relationships among populations, which run the gamut from slight genetic differentiation to complete reproductive isolation. The genetic divergence that provides reproductive isolation sometimes consists of changes in chromosome structure that cause infertility of hybrids, or of genic alterations of development, especially in gametogenesis, that cause hybrids to be inviable or infertile. Hybrid inviability and infertility constitute postmating isolating mechanisms. Many species, however, are capable of forming fertile, viable hybrids, but do not do so in nature because they do not mate with each other; differences in time of reproduction, courtship behavior, or (in plants) flower form are among the premating isolating mechanisms that maintain the distinction between the species. Such differences can arise as pleiotropic by-products of the genetic changes that developed by genetic drift or as responses to different environmental factors. In many groups of animals, courtship signals and responses can diverge rapidly because of sexual selection, in which a characteristic (such as a

peacock's train) becomes exaggerated to enable its bearer (often the male) to compete more successfully for mates. Analogous phenomena may occur in some plants, such as orchids. See REPRODUCTIVE BEHAVIOR.

At least in plants, speciation can occur sympatrically, without initial geographic isolation, by polyploidy. Under special circumstances, sympatric speciation can theoretically also occur in animals, but the frequency with which this occurs is strongly debated. Ernst Mayr has developed cogent arguments for supposing that it is rare. See SPECIATION.

**Adaptations.** A frequent consequence of natural selection is that a species comes to be dominated by individuals whose features equip them better for the environment or way of life of the species. Such features are termed adaptations. Thus in England in the nineteenth century, dark variants of the peppered moth (*Biston betularia*) replaced the light genotype because they were less evident to predatory birds; their dark coloration may be considered an adaptation to predation. It is sometimes difficult to determine the adaptive significance of a trait, or whether it is an adaptation at all. However, this can often be accomplished by experiment or observation. For example, it has been shown that the color patterns of certain butterflies that resemble distasteful species provide protection against predation, by experimentally altering their color patterns and finding that the altered individuals suffer greater predation. Convergent evolution is often evidence of adaptation: for example, leaflessness has evolved in many groups of desert plants as a mechanism of reducing water loss.

Most adaptive features benefit the individuals that bear them, rather than the population as a whole. Certain features, though, appear at first sight to be altruistic, such as the warning calls that some birds emit when they see a predator. Other members of the flock benefit from being alerted, but the call seems likely to place the warner in jeopardy, so that alleles for such behavior would appear unlikely to increase in the population. However, features like this can evolve because selection can act not only at the level of individual organisms but at other levels. One possibility, the prevalence of which is strongly disputed, is group selection, whereby the genetic composition of the species changes by the differential survival (or proliferation) of whole populations that differ in gene frequency. Thus populations that happen to have a high frequency of genes for warning behavior might survive best, so that these genes increase in the species as a whole, even if they are selected against within populations. A more likely explanation for apparently altruistic traits is a form of selection at the level of the gene known as kin selection. In developing this theory, William Hamilton argued that because relatives share alleles by common descent, an allele may increase in frequency if it causes its bearer to help its relatives survive and reproduce, even if the fitness of the bearer suffers. Kin selection, the cornerstone of the study of sociobiology, is the most likely explanation for the sterility of workers in social insects, and for warning calls and

many other aspects of social behavior. *See* SOCIOBIOLOGY.

Although many features of organisms are adaptive, not all are, and it is a serious error to suppose that species are capable of attaining ideal states of adaptation. Some characteristics are likely to have developed by genetic drift rather than natural selection, and so are not adaptations; others are side effects of adaptive features, which exist because of pleiotropy or developmental correlations. The absence of appropriate genetic variations and the constraints imposed by processes of development limit the variety of adaptive responses of which a species is capable, so that the path of ideal adaptation may be closed to it. The phylogenetic history of a species determines its current state, and thus the kinds of variations that may be available to natural selection. Most species are not capable of adapting ideally to all environmental changes: more than 99% of all the species that have ever lived are extinct.

**Origin of higher taxa.** Higher taxa are those above the species level, such as genera and families. A taxon such as a genus is typically a group of species, derived from a common ancestor, that share one or more features so distinctive that they merit recognition as a separate taxon. The degree of difference necessary for such recognition, however, is entirely arbitrary: there are often no sharp limits between related genera, families, or other higher taxa, and very often the diagnostic character exists in graded steps among a group of species that may be arbitrarily divided into different higher taxa. Moreover, a character that in some groups is used to distinguish higher taxa (such as the number of cotyledons, one of the features that divides flowering plants into Monocotyledonae and Dicotyledonae) sometimes varies among closely related species or even within species (as does the number of cotyledons in a few flowering plants). In addition, the fossil record of many groups shows that a trait that takes on very different forms in two living taxa has developed by intermediate steps along divergent lines from their common ancestor; thus the inner ear bones of mammals may be traced to jaw elements in reptiles that in turn are homologous to gill arch elements in Paleozoic fishes. Thus evolution of differences of great enough magnitude to define higher taxa appears to proceed usually by gradual incremental changes corresponding to the variations evident within species and among closely related species. However, it is possible that in some instances slight differences at the genetic level have altered developmental patterns so as to yield large discrete changes in a suite of developmentally correlated traits. Neoteny in salamanders, for example, whereby changes in hormone levels cause the retention of a complex of larval characteristics into the reproductive ages, appears to have a simple genetic (or in some cases a purely environmental) basis. The frequency with which such discontinuous, or saltational, changes in phenotype have occurred in evolution is a matter of some controversy. *See* ANIMAL SYSTEMATICS; MACROEVOLUTION; NEOTENY.

Among the primary reasons for changes in the

form of a structure is a change in its function. For example, the sting of a wasp is a morphologically and functionally altered ovipositor, the structure used by females of more primitive forms to insert eggs into the plants or animals in which the larvae develop. (Incidentally, this explains why male wasps and bees lack stings.) In many desert plants such as cacti, leaves or branches have taken on a defensive function and are modified into spines.

To a certain degree, simple evolutionary changes in structure or physiology can be reversed during evolution; but complex features, once lost or highly modified, are seldom regained in their original form, for natural selection and genetic drift can act only on the variations of whatever "raw materials" are available. Thus numerous groups of plants, especially those that are wind-pollinated, have lost their petals; in some of these groups, insect pollination has secondarily evolved, but the structures that are colored and otherwise modified to attract insects are leaves, sepals, or other parts, the petals having been lost. *See* PHYLOGENY.

**Rates of evolution.** The characteristics of a species evolve individually or in concert with certain other traits that are developmentally or functionally correlated. Because of this mosaic pattern of evolution, it is meaningful to speak of the rate of evolution of characters, but not of species or lineages as total entities. Thus in some lineages, such as the so-called living fossils, many aspects of morphology have evolved slowly since the groups first came into existence, but evolution of their DNA and amino acid sequences has proceeded at much the same rate as in other lineages. Every species, including the living fossils, is a mixture of traits that have changed little since the species' remote ancestors, and traits that have undergone some evolutionary change in the recent past.

Rates of evolution are highly variable: whereas many characteristics have changed hardly at all for many millions of years in some lineages, others have responded rapidly to changes in selection pressures. For example, within decades, numerous species of insects have evolved resistance to insecticides, and some plants have become adapted to soils impregnated with toxic metals from mine works. Geographical variation in morphological traits has evolved within a century in house sparrows introduced into North America from Europe. Based on the geological history of their habitats, speciation in some groups of fresh-water fishes is believed to have transpired in less than 10,000 years, whereas in some groups of plants populations isolated for several million years have not become different species. In contrast to the highly variable rate of evolution of morphological and physiological characters, the rates at which nucleotide sequences of DNA and amino acid sequences of proteins have changed appears to be considerably more uniform; Motoo Kimura, Allan Wilson, and some other authors have in fact claimed that the rate of evolution at the molecular level is nearly constant, providing a "molecular clock" that may be used to estimate the time since species diverged from their common ancestors.

The fossil record of certain organisms, especially marine invertebrates, seems to indicate that species often change rather abruptly in morphology after being virtually static for up to a million years; the geological record, however, is usually so coarse that the “abrupt” changes may well have proceeded gradually over a period of many thousands of years. This pattern, of stasis punctuated by brief periods of change, has been termed punctuated equilibrium. Niles Eldredge, Stephen Gould, and Steven Stanley have suggested that the observed changes represent the origin of new species in localized populations, which become evident in the fossil record only after they expand from their site of origin. They also propose that after speciation the species becomes incapable of substantial further change, so that evolutionary change is virtually restricted to divergence during speciation. There is little direct evidence for this notion of genetic paralysis, which is contested by most evolutionary geneticists. See MACROEVOLUTION.

There is abundant evidence that rates of evolution are greatest when a lineage adapts to new ecological opportunities—to vacant ecological niches. Rapid divergent evolution is common, for example, when species colonize islands that harbor few competitors; similarly, the rate of evolution is high in lineages that have survived mass extinction events. The usual pattern in such instances is one of adaptive radiation: the origin, by speciation, of numerous descendant lineages that become adapted in different ways to a variety of available resources (Fig. 5). A famous example is the radiation of Darwin’s finches in the Galápagos Islands, where related species have diverged in beak morphology and have become specialized for feeding on resources that on continents are typically preoccupied by unrelated families of birds. The major adaptive radiation of mammals occurred soon after the demise of the last dinosaurs, leading many authors to suspect that the mammalian radiation was possible only because competition had been alleviated. Thus extinction has played an important role in the history of life, making possible the subsequent diversification of groups that otherwise might not have flourished.

**Evolutionary trends.** The history of life is not one of progress in any one direction, but of adaptive radiation on a grand scale: the descendants of any one lineage diverge as they adapt to different resources, habitats, or ways of life, acquiring their own specialized features as they do so. There is no evidence that evolution has any goal, nor does the mechanistic theory of evolutionary processes admit of any way in which genetic change can have a goal or be directed toward the future.

Nevertheless, in examining the history of a major group of organisms, it is sometimes possible to discern in retrospect a trend in one or more characters. For example, in one group of horses that became increasingly adapted to grazing and running, body size, tooth size, and the number of the toes changed more or less monotonically (although not at a constant rate) during the Tertiary Period. Once

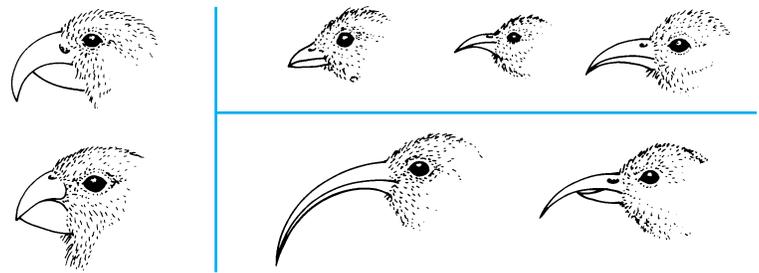


Fig. 5. Some members of an adaptive radiation, the Hawaiian honeycreepers. Although descended from finches, these birds have become adapted to a variety of ecological roles in the Hawaiian archipelago; some feed primarily on seeds (thick beaks), others on insects (short, thin beaks), others primarily on nectar (long, thin beaks). (After D. J. Futuyma, *Evolutionary Biology*, Sinauer Associates, 1979)

embarked on a way of life, a lineage experiences selection for improvement in the features that adapt it to that particular way of life, and the more specialized such features become, the less capable they are of being modified in other directions. A very common trend, for example, is reduction and simplification of parts, as in the reduction of limbs, lungs, and other features of snakes, which are now so morphologically and developmentally committed to their ways of life that they would be incapable of reevolving the lizardlike form of their ancestors. But although such trends can be discerned retrospectively, no one can say that a group such as the snakes was destined from their beginning to take the evolutionary path that they in fact followed: other roads may have been open to them at one time, but are open no longer. All history, including evolutionary history, is contingent on antecedent events.

For particular groups, one can document trends of increasing reduction and simplification; of the converse, increasing structural or behavioral complexity; of increased ecological specialization; of its converse, increased homeostatic freedom from the environment; and of increased mechanical or functional efficiency of morphological and physiological features. Few such trends continue unabated throughout evolutionary time; most are terminated by complete extinction, the fate of most of the higher taxa of organisms that have ever existed.

For life taken as a whole, there is some very tentative evidence that species (at least of marine invertebrates) have on average become more resistant to extinction, so that Cenozoic species have persisted longer than Mesozoic or Paleozoic species; but any such trend has been interrupted repeatedly by mass extinction events (the causes of which are actively disputed) that have eliminated much of the Earth’s biota. The total number of species has increased after each such extinction event, and has been amplified in the last 100 million years or so by continental drift, which has accentuated differences in species composition in different parts of the world. More than 1½ million living species have been described (of the 5 to 10 million that probably exist); and even though many major taxa have become extinct, the diversity of species now and in the recent past is higher than ever before in Earth’s history. For life taken as

a whole, the only clearly discernible trend is toward ever-increasing diversity. Douglas J. Futuyma

Bibliography. J. C. Avise, *Molecular Markers, Natural History, and Evolution*, 1994; D. J. Futuyma, *Evolutionary Biology*, 3d ed., 1997; D. J. Futuyma, *Science on Trial: The Case for Evolution*, 1983; D. L. Hartl and A. G. Clark, *Principles of Population Genetics*, 3d ed., 1997; J. S. Levinton, *Genetics, Paleontology, and Macroevolution*, 2d ed., 2001; W.-H. Li and D. Graur, *Fundamentals of Molecular Evolution*, 2d ed., 2000; J. Maynard Smith, *Evolutionary Genetics*, 2d ed., 1998; E. Mayr, *The Growth of Biological Thought: Diversity, Evolution, and Inheritance*, 1985; E. Mayr, *Populations, Species, and Evolution*, 1970.

### Organic geochemistry

The study of the abundance and composition of naturally occurring organic substances, their origins and fate, and the processes that affect their distributions on Earth and in extraterrestrial materials.

Organic geochemistry was born from a curiosity about the organic pigments extractable from petroleum and black shales. It developed with extensive investigations of the chemical characteristics of petroleum and petroleum source rocks as clues to their occurrence and formation, and now encompasses a broad scope of activities within interdisciplinary areas of earth and environmental science. This range of studies recognizes the potential of geological records of organic matter to help characterize sedimentary depositional environments and to provide evidence of ancient life and indications of evolutionary developments through the Earth's history. Organic geochemistry includes determinations of anthropogenic contaminants amid the natural background of organic molecules and the assessment of their environmental impact and fate. Marine organic geochemistry addresses and interprets aquatic processes involving carbon species. It involves investigations of the chemical character of particulate and dissolved organic matter, evaluation of oceanic primary production including the factors (light, temperature, nutrient availability) that influence the uptake of carbon dioxide (CO<sub>2</sub>), the composition of marine organisms, and the subsequent processing of organic constituents through the food web. Organic geochemistry extends to broader biogeochemical issues, such as the carbon cycle, and the effects of changing carbon dioxide levels, especially efforts to use geochemical data and proxies to help constrain global climate models. Examination of the organic chemistry of meteorites and lunar materials also falls within its compass, and as a critical part of the quest for remnants of life on Mars, such extraterrestrial studies are now regaining the prominence they held in the 1970s during lunar exploration. See COSMOCHEMISTRY; GEOCHEMISTRY.

These activities share the common need for identification, measurement, and assessment of organic matter in its myriad forms. Hence, many advances

within this area have paralleled developments in analytical methodology or instrumentation that permit new approaches to the determination of the abundance, structure, and composition of organic matter in geological materials.

**Global inventories of carbon.** Carbon naturally exists as oxidized and reduced forms in carbonate carbon and organic matter. The major reservoir of both forms of carbon on Earth is the geosphere. It contains carbonate minerals deposited as sediments and organic matter accumulated from the remains of dead organisms. Estimates of the size of the geological reservoir of carbon vary within the range of  $5$  to  $7 \times 10^{22}$  g, of which 75% is carbonate carbon and 25% is organic carbon. The amounts of carbon contained in living biota ( $5 \times 10^{17}$  g), dissolved in the ocean ( $4 \times 10^{19}$  g), and present in atmospheric gases ( $7 \times 10^{17}$  g) are miniscule compared to the quantity of organic carbon buried in the rock record. Most of it occurs in finely disseminated form within sedimentary rocks, especially shales. The importance of buried organic matter extends beyond its sheer magnitude; it includes the fossil fuels—coal, natural gas, and petroleum—that supply 85% of the world's energy. Global annual consumption of these three fuels represent only approximately 0.4%, 1.5%, and 2.3%, respectively, of their proven economic reserves, which in turn collectively account for only about 0.12% of the total accumulation of organic matter. These figures illustrate the immense quantities of buried organic carbon, which represent a geological legacy inherited from ancient life. They also show how the release of carbon dioxide from burning of these fossil fuels can readily increase the comparatively small atmospheric reservoir of carbon, perturbing the natural balance of the global carbon cycle. See BIOGEOCHEMISTRY; CARBON; CARBON DIOXIDE; CARBONATE MINERALS; FOSSIL FUEL; LIMESTONE; SEDIMENTARY ROCKS; SHALE.

**Sedimentary organic matter.** The vast amounts of organic matter contained in geological materials represent the accumulated vestiges of organisms amassed over the expanse of geological time. Yet, survival of organic cellular constituents of biota into the rock record is the exception rather than the norm. Only a small portion of the carbon fixed by organisms during primary production, especially by photosynthesis, escapes degradation as it settles through the water column and eludes microbial alteration during subsequent incorporation and assimilation into sedimentary detritus. Most biological debris is rapidly broken down or remineralized by surficial processes within the oceans, in lakes, in wetlands, and in soils and other terrigenous environments. Atmospheric and aqueous oxidation reactions contribute to this recycling process, which is also enhanced by consumption associated with food web processes and facilitated by sediment bioturbation (movement of sediment by animals). Protective packaging of organic matter can improve its chance of survival. Rapidly sinking zooplankton fecal pellets contain materials derived from phytoplankton that have survived ingestion and digestion and descend in discrete

capsules, thereby evading degradative influences. Hierarchical communities of aerobic and anaerobic microbes act as the predominant agents for degradation and alteration of organic matter in both the water column and sediments. They use it as their carbon source during a wide range of distinct metabolic processes, including heterotrophy, fermentation, sulfate reduction, methanogenesis, and methylotrophy. Microbial consortia selectively remove target substrates, and augment the surviving organic materials with secondary biomass, by adding distinctive components and isotopic signatures of their activity. The application of various microbiological procedures in examination of sedimentary materials, including incubation tests for specific metabolic processes and characterization of DNA isolates, is rapidly expanding the known distribution of microbial activity in the geosphere. These approaches demonstrate that the highly complex character of sedimentary organic matter often includes diagnostic evidence of its diverse biological sources and of the cumulative effects of alteration processes. *See* BIODEGRADATION; PETROLEUM MICROBIOLOGY.

Two distinct but related phenomena can enhance sequestration of organic matter in sedimentary environments. First, high levels of primary production can generate a supply of organic matter that exceeds the capacity of microbial flora to effectively degrade it. Second, depositional conditions, especially anoxia, can restrict the active populations of microbial communities engaged in the degradative process. Both circumstances can lead to exceptional preservation of organic matter, although their comparative importance in the process continues to be debated. Depositional conditions that favor survival of greater quantities of carbon, in terms of its elemental abundance, also affect the molecular and isotopic compositions of the organic matter by increasing survival of labile components. For example, hypersaline environments tend to facilitate the process of sulfur incorporation, which aids preservation of less stable structural features.

**Kerogen and bitumen.** Sedimentary organic matter can be divided operationally into solvent-extractable bitumen and insoluble kerogen. Bitumens contain a myriad of structurally distinct molecules, especially hydrocarbons, which can be individually identified (such as by gas chromatography-mass spectrometry) although they may be present in only minute quantities (nanograms or picograms). The range of components includes many biomarkers that retain structural remnants inherited from their source organisms, which attest to their biological origins and subsequent geological fate. *See* BITUMEN; KEROGEN.

For many years, kerogen was thought to form in sediments by a series of condensation reactions that combined the building blocks of life (amino acids, carbohydrates, lipids) into larger insoluble components. Undoubtedly such reactions do occur, aided by microbial alteration processes, but it is now apparent that a significant portion of kerogen represents macromolecular remnants, such as fragments

of cell walls, inherited directly from living organisms or partially modified by removal of labile structural elements by degradative processes. The biological sources of organic matter, their survival during sediment deposition, and the extent of alteration effectively establish the elemental composition of kerogen that determines its character. Kerogen is broadly classified into three main types based on the proportions of hydrogen and oxygen relative to carbon (H/C and O/C ratios). Type I kerogen possesses high H/C and low O/C ratios, and is comparatively rare. It arises from contributions of organic matter that are dominated by phytoplankton and bacteria and usually occur in lacustrine (lake) settings. Type III kerogen exhibits low H/C and high O/C ratios, which result from a predominance of terrestrial organic matter that is typical in continental margin settings. It tends to be gas-prone, generating methane (CH<sub>4</sub>). Type II kerogen displays intermediate H/C and O/C ratios, which reflects the mixed sources of organic matter that typify marine settings. Most oil source rocks contain this kerogen type, although some enriched in sulfur are labeled type II-S. The process of burial modifies all kerogen types, reducing both H/C and O/C ratios as organic matter experiences thermal alteration. Ultimately, the distinctive characteristics of the separate kerogen types and evidence of their sources are obscured by the burial processes that constitute diagenesis.

**Diagenetic alteration of organic matter.** Few organic constituents synthesized by living organisms remain stable in the sedimentary domain, although diagnostic features of their structures can survive for eons. To persist in sediments, organic matter must first survive the processes of recycling modification by microbial action during burial and consolidation. Subsequently, the organic matter experiences progressive transformation caused by thermal processes as its host rock undergoes compaction, dewatering, and perhaps remineralization. Modifications of organic matter proceed in a sequential manner and ultimately cause dramatic changes in its character. Hydrogen and oxygen atoms are lost incrementally, and eventually carbon-carbon bonds are broken, liberating mobile hydrocarbons, which is a crucial step in petroleum generation known as catagenesis. Continued alteration during burial at elevated temperature and pressure further depletes the hydrogen in residual organic material and ultimately produces graphite, one of the pure forms of carbon. *See* DIAGENESIS; GRAPHITE.

**Petroleum formation.** A number of specific geological conditions must be met for generation of recoverable petroleum, which is an inherently inefficient process. First, there is a need for the raw materials that can generate hydrocarbons, which must be provided by a source rock containing appreciable quantities (typically >0.5%) of preserved organic matter of appropriate character. The nature of the organic matter determines the potential yield of petroleum and its composition. Second, the source rock must be buried at sufficient depth that kerogen cracking liberates hydrocarbons. This process is dependent

on a combination of time and temperature; petroleum generation can occur rapidly in regions where the Earth's heat flow is high, or it can proceed slowly over  $10^8$ – $10^9$  years where thermal gradients are low. Third, the hydrocarbons released must migrate from the source rock to a suitable trap, often in folded or faulted rock strata, where they can accumulate in a reservoir. Primary migration occurs as generated oil moves out of the source rock aided by the thermal expansion and fracturing of kerogen. Subsequently, driven by pressure gradients, it moves through porous conduits displacing formation waters or along planes of weaknesses during the secondary migration to reach the reservoir. Finally, the integrity of the petroleum within its reservoir must be preserved for tens or hundreds of millions of years. Tectonic forces may fracture the reservoir allowing petroleum to escape, or the reservoir may be uplifted and eroded at the surface. Contact with meteoric waters percolating through the petroleum reservoir can lead to water washing, which selectively extracts the soluble petroleum constituents. It may also introduce microbial populations, which can sequentially remove components amenable to biodegradation, leaving residues that are increasingly resistant to alteration. Alternatively, petroleum may seep upward through fractures, to emanate as surface or submarine seeps, where it can evaporate or be washed or degraded by microbes. All of these processes tend to reduce the quality and value of petroleum. *See* PETROLEUM; PETROLEUM GEOLOGY; SEDIMENTOLOGY.

The composition of petroleum reflects its source materials, thermal and migration history, and extent of biodegradation. The combined effects and disparate interactions between these influences produce oils that vary widely in their chemistry and physical properties. Thus, investigation of the compositional characteristics of petroleum enables the cumulative history of these processes, and their effects, to be interpreted. Oil generation is an essential part of a continuum of changes that affect source rocks during their progressive burial, and that lead to successive liberation of petroleum and gases, contingent on the composition of the source material. Thus, the thermal processes that transform kerogen into petroleum continue beyond oil generation, and fragment organic matter into progressively smaller molecules, which leads to production of hydrocarbon gases. The volume and composition of the gases generated is dependent on whether the source rock is oil- or gas-prone, which in turn is a function of its composition, especially its dominant kerogen type. However, as thermal maturity increases, there is a trend in composition from wet gas to dry gas, as the proportion of hydrocarbon gases that condense as liquids (for example butane and propane) decreases while methane becomes predominant. *See* METHANE; NATURAL GAS.

Methane can also be produced in shallow sediments by the microbial activity of methanogenic bacteria. Such bacterial methane forms many recoverable gas accumulations at comparatively shallow

depths and represents the source of methane in gas hydrates. Hydrates occur in sediments where the combination of low temperature and high pressure stabilizes the ice cages that can entrap methane. The worldwide frequency of hydrate occurrences supports estimates that they may greatly exceed the combined resources of oil and gas (0.5% versus 0.017% of global carbon). Moreover, massive release of methane from hydrates is a direct consequence of global warming and is now recognized as the climatic perturbation that caused significant changes in Earth's flora and fauna associated with the late Paleocene thermal maximum, 55 million years ago. *See* HYDRATE; METHANOGENESIS (BACTERIA).

**Coal.** Peat deposits contain an abundance of terrestrial organic matter that is progressively transformed by burial compaction and thermal alteration into lignite, bituminous coal, and ultimately graphite. Coals vary in type and chemical characteristics, largely dependent on the composition of their organic matter (typically type III kerogen), its extent of bacterial degradation, and its thermal history. *See* LIGNITE; PEAT.

Coals contain many plant fragments, which are often altered but can be identified under the microscope and classified according to their origins. These integral parts of coal are called macerals, which are individually described and named according to their source, for example, sporinite (spores), resinite (resin), and vitrinite (woody). Macerals can be recognized in other sediments, where they provide evidence of contributions of organic matter from terrestrial plants and aid assessment of depositional environments. The reflectance of the maceral vitrinite changes systematically with increasing burial and thermal maturity, whether in coals or other rocks. Thus, determination of vitrinite reflectance ( $R_o$ ) by microscopic examination provides a direct measure of its thermal history, and is widely used in assessing the maturity of petroleum source rocks. *See* COAL.

**Biomarkers.** Biomarkers are individual compounds whose chemical structures carry evidence of their origins and history. Recognition of the specificity of biomarker structures initially helped confirm that petroleum was derived from organic matter produced by biological processes. For example, the structural and stereochemical characteristics of the hydrocarbon cholestane that occurs in petroleum can be directly and unambiguously linked to an origin from the sterol cholesterol, which is a widespread constituent of plants and animals. Of the thousands of individual petroleum components, hundreds reflect precise biological sources of organic matter, which distinguish and differentiate their disparate origins. The diagnostic suites of components may derive from individual families of organisms, but contributions at a species level can occasionally be recognized. Biomarker abundances and distributions help to elucidate sedimentary environments, providing evidence of depositional settings and conditions. They also reflect sediment maturity, attesting to the progress of the successive,

sequential transformations that convert biological precursors into geologically occurring products. Thus, specific biomarker characteristics permit assessment of the thermal history of individual rocks or entire sedimentary basins. *See* BASIN.

**Carbon isotopes.** Carbon naturally occurs as three isotopes: carbon-12 ( $^{12}\text{C}$ ), carbon-13 ( $^{13}\text{C}$ ), and radiocarbon ( $^{14}\text{C}$ ). The proportion of  $^{13}\text{C}$  in living organic matter is controlled by the organism's source of carbon, fractionation during carbon fixation, and both biosynthetic and metabolic processes. Measurements of  $^{13}\text{C}$  are expressed as a ratio, which compares the amount of  $^{13}\text{C}$  in the sample to a standard (carbonate in the Peedee belemnite) and provides a value in parts per thousand (‰), defined as  $^{13}\text{C} = \{[(^{13}\text{C}/^{12}\text{C})_{\text{sample}} / (^{13}\text{C}/^{12}\text{C})_{\text{standard}}] - 1\} \times 1000$ . Organic carbon tends to be depleted of  $^{13}\text{C}$  (negative  $^{13}\text{C}$  values) relative to inorganic carbon (carbon dioxide and carbonate) as organisms preferentially uptake and use  $^{12}\text{C}$  because its reactivity is greater than that of  $^{13}\text{C}$ . However, the extent of  $^{13}\text{C}$  depletion is dependent on biosynthetic pathways. For example, photosynthetic carbon fixation of carbon dioxide by plants that employ the  $\text{C}_3$  and  $\text{C}_4$  pathways produces organic matter averaging  $-26\text{‰}$  and  $-12\text{‰}$ , respectively. The significant difference is caused by the ability of the latter group, mainly comprising grasses, to store carbon dioxide before fixation. Methylophilic bacteria use methane, which is highly depleted in  $^{13}\text{C}$ , rather than carbon dioxide as their carbon source. Consequently, they produce organic compounds that possess highly negative  $^{13}\text{C}$  values ( $-55\text{‰}$  to  $-120\text{‰}$ ). By contrast, the products synthesized by bacteria that use the reverse-TCA (tricarboxylic acid, or Krebs) cycle exhibit  $^{13}\text{C}$  values typically less than  $-10\text{‰}$ . Thus, the carbon isotopic composition of many individual components found in sediments attests to the diversity of their origins and demonstrates the variety of microbial processes active in the shallow subsurface. *See* ISOTOPE; PHOTOSYNTHESIS; PLANT RESPIRATION.

Temporal excursions in the  $^{13}\text{C}$  values of sediment sequences can reflect perturbations of the global carbon cycle. For example, episodes of enhanced burial of organic matter that occurred during widespread oceanic black shale deposition in the Cretaceous effectively depleted the global  $^{12}\text{C}$  reservoir, leading to an enrichment of  $^{13}\text{C}$  in sedimentary organic matter. Also, global shifts in carbon isotopic signals provide critical evidence that the late Paleocene thermal maximum was associated with a release of methane from hydrates. Similarly, the remarkable variations in  $^{13}\text{C}$  values of organic matter in the Late Proterozoic are consistent with the "snowball Earth" hypothesis that ice sheets extended to the Equator, creating long-term episodes of glaciation. These events ended catastrophically when greenhouse conditions gradually created by a build-up of atmospheric carbon dioxide became sufficient to melt the ice, leading to extraordinarily rapid weathering of continents as their ice burden dissipated, and dramatic deposition of marine carbonates as the ocean again

became a sink for carbon dioxide. *See* BLACK SHALE; CRETACEOUS; MARINE SEDIMENTS; PALEOCEANOGRAPHY.

Radiocarbon is widely employed to date archeological artifacts, but the sensitivity of its measurement also permits its use in exploration of the rates of biogeochemical cycling in the oceans. This approach permits assessment of the ages of components in sediments, demonstrating that bacterial organic matter is of greater antiquity than components derived from phytoplankton sources. *See* RADIOCARBON DATING.

**Applications.** A major impetus for the development of organic geochemistry has been its use as a tool in petroleum exploration. Determination of the quantity, nature, and degree of thermal alteration of organic matter from kerogen analysis provides evidence of the richness and maturity of source rocks and the timing of oil generation. Such information is critical in evaluation of prospects for petroleum formation and accumulation. Similarly, chemical clues provided by the molecular and isotopic composition of petroleum enable the determination of the biological sources of organic matter (marine or terrestrial) and its environment of deposition. Such information can also place constraints on the age of petroleum based on the timing of the appearance of specific biomarkers, based on evolutionary trends in their source organisms. Molecular distributions also provide valuable fingerprints for the correlation of source rocks and oils, and the  $^{13}\text{C}$  values of hydrocarbon gases help distinguish their origins from bacterial or thermogenic sources, especially when coupled with the determination of their hydrogen isotope compositions (deuterium values). *See* DEUTERIUM; HYDROGEN.

Molecular and isotopic tools are gaining acceptance as paleoclimate proxies. The proportion of unsaturation (number of carbon-carbon double bonds) in a group of compounds known as alkenones is biosynthetically controlled by the growth temperature of the phytoplankton that synthesize them. The survival of this temperature signal in marine sediment sequences provides a temporal record of sea surface temperatures that reflect past climates, including evidence of the timing and extent of global cooling during ice ages, and the warming associated with El Niño events. *See* CHEMOSTRATIGRAPHY; EL NIÑO.

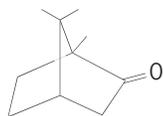
Also, differences between the carbon isotopic composition of the alkenones and co-occurring carbonate can be interpreted in terms of ancient atmospheric levels of carbon dioxide, when the effect of the growth rates of the alkenone-producing phytoplankton is excluded. *See* PHYTOPLANKTON.

Simon C. Brassell

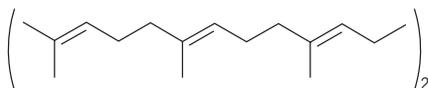
**Bibliography.** M. Engel and S. A. Macko, *Organic Geochemistry*, 1993; J. M. Hunt, *Petroleum Geochemistry and Geology*, 2d ed., 1996; S. Killips and V. J. Killips, *An Introduction to Organic Geochemistry*, 1993; L. M. Pratt, J. B. Comer, and S. C. Brassell (eds.), *Geochemistry of Organic Matter in Sediments and Sedimentary Rocks*, 1992.

## Organic nomenclature

A system by which a unique and unambiguous name or other designation is assigned to a given organic molecular structure. A set of rules adopted by the International Union of Pure and Applied Chemistry (IUPAC) is the basis for a standardized name for any organic compound. Common or nonsystematic names are used for many compounds that have been known for a long time. The latter names have the advantage of being short and easily recognized, as in the examples below. However, in contrast to system-



Common name: camphor  
Systematic name:  
1,7,7-trimethyl-  
bicyclo[2.2.1]heptan-6-one

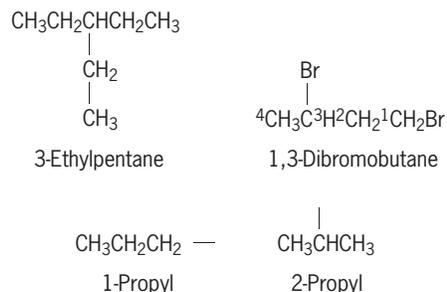


Common name: squalene  
Systematic name:  
2,6,10,15,19,23-hexamethyl-  
2,6,10,14,18,22-tetracosahexaene

atic names, common or trivial names do not convey information from which the structure can be written by reference to prescribed rules.

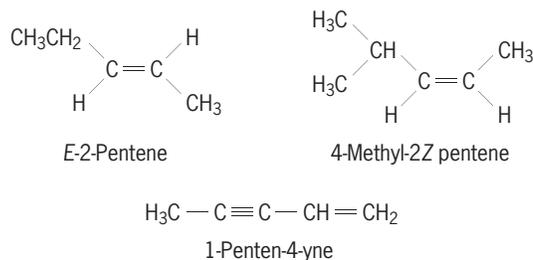
**Aliphatic compounds.** In the IUPAC system, a name is formed by combination of a parent alkyl chain or ring with prefixes and suffixes to denote substituents. For aliphatic compounds, the parent name is a stem that denotes the longest straight chain in the structure with the ending -ane. The first four members of the alkane series are methane, ethane, propane, and butane. From five carbons on, the names follow the Greek numerical roots, for exam-

ple, pentane and hexane. Changing the ending -ane to -yl gives the name of the corresponding radical or group; for example,  $\text{CH}_3\text{CH}_2-$  is the ethyl radical or ethyl group. Branches on an alkyl chain are indicated by the name of a radical or group as a prefix. The location of a branch or other substituent is indicated by number; the chain is numbered from whichever end results in the lowest numbering. Examples are shown below.



See ALKANE.

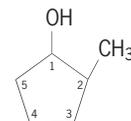
Double or triple bonds are indicated by the endings -ene or -yne, respectively. The configuration of the chain at a double bond is denoted by *E*- when two similar groups are on opposite sides of the plane bisecting the bond, and *Z*- when they are on the same side. This is illustrated by the structures below.



See ALKENE; ALKYNE.

A compound containing a functional group is named by adding to the parent name a suffix characteristic of the group. If there are two or more groups, a principal group is designated by suffix and the other(s) by prefix. Examples are given in **Table 1**, in which groups are listed in descending order of priority as the suffix.

**Carbocyclic compounds.** The same general principles apply to cyclic compounds. For alicyclic rings, the prefix cyclo- is followed by a stem indicating the number of carbon atoms in the ring, as is illustrated in the structure below. In bicyclic compounds, the



2-Methylcyclopentanol

total number of carbons in the ring system is prefixed by bicyclo- and numbers in brackets which indicate the number of atoms in each connecting chain. The atoms in a bridged bicyclic ring system are numbered by commencing at a bridgehead atom and

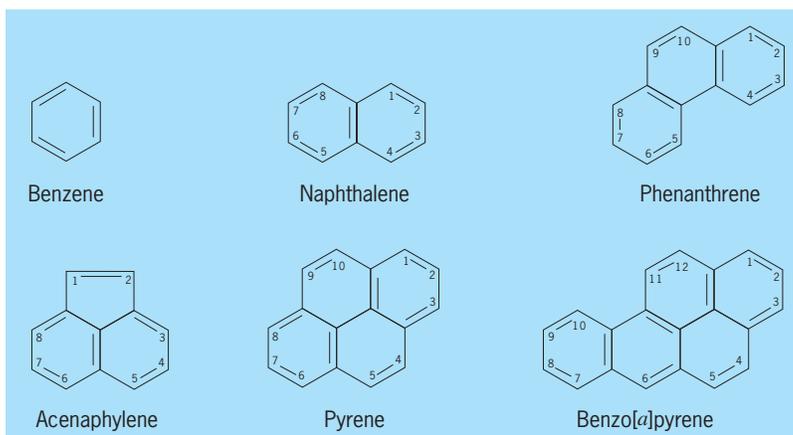
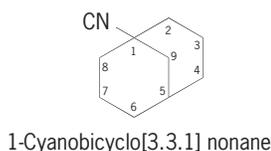


Fig. 1. Structures and position numbering of benzene and some representative polycyclic aromatic structures.

TABLE 1. Names of compounds containing functional groups				
Group*	Suffix	Prefix	Structure	Name
$\begin{array}{c} \text{O} \\ \parallel \\ \text{—COH} \end{array}$	-oic acid	carboxy-	$\text{CH}_3\text{CH}_2\text{COH}$	Propanoic acid
$\begin{array}{c} \text{O} \\ \parallel \\ \text{—COR} \end{array}$	alkyl -oate	alkoxycarbonyl-	$\text{CH}_3\text{CH}_2\text{COCH}_3$	Methyl propanoate
			$\begin{array}{c} \text{O} \quad \text{O} \\ \parallel \quad \parallel \\ \text{CH}_3\text{OCCH}_2\text{CH}_2\text{COH} \end{array}$	3-Methoxycarbonyl propanoic acid
$\text{—C}\equiv\text{N}$	-nitrile	cyano-	$\text{CH}_3\text{CH}_2\text{C}\equiv\text{N}$	Propanenitrile
			$\begin{array}{c} \text{O} \\ \parallel \\ \text{CH}_3\text{CHCNCOH} \end{array}$	2-Cyanopropanoic acid
$\begin{array}{c} \text{O} \\ \parallel \\ \text{—CH} \end{array}$	-al	formyl-	$\text{CH}_3\text{CH}_2\text{CH}$	Propanal
$\begin{array}{c} \text{O} \\ \parallel \\ \text{—CR} \end{array}$	-one	oxo-	$\text{CH}_3\text{CH}_2\text{CCH}_3$	Butanone
			$\begin{array}{c} \text{O} \quad \text{O} \\ \parallel \quad \parallel \\ \text{CH}_3\text{CCH}_2\text{COH} \end{array}$	3-Oxobutanoic acid
$\text{—OH}$	-ol	hydroxy-	$\text{CH}_3\text{CH}_2\text{CH}_2\text{OH}$	1-Propanol
			$\begin{array}{c} \text{O} \\ \parallel \\ \text{HOCH}_2\text{CH}_2\text{CH} \end{array}$	3-Hydroxypropanal
$\text{—NH}_2$	-amine	amino-	$\text{CH}_3\text{CH}_2\text{CH}_2\text{NH}_2$	1-Propanamine
$\text{—OR}$	—	alkoxy-	$\text{NH}_2\text{CH}_2\text{CH}_2\text{CH}_2\text{OCH}_3$	3-Methoxy-1-propanamine

\*R = any alkyl group.

numbering by the longest path to the other bridgehead, and then the shorter bridge, as is illustrated in the structure below.



See ALICYCLIC HYDROCARBON.

The names of aromatic hydrocarbons have the ending -ene, which denotes a ring system with the maximum number of noncumulative double bonds. Each of the simpler polycyclic hydrocarbons has a different parent name; a few of these, with the ring numbering, are shown in Fig. 1. Other fused ring systems can be named by adding the prefix benzo-, which denotes attachment of an additional  $\text{—CH=CH—CH=CH—}$  chain at two adjacent carbons.

To number the positions in a polycyclic aromatic ring system, the structure must first be oriented with the maximum number of rings arranged horizontally and to the right. The system is then numbered in clockwise sequence starting with the atom in the most counterclockwise position of the upper right-

hand ring, omitting atoms that are part of a ring fusion. See AROMATIC HYDROCARBON.

**Heterocyclic compounds.** Systematic names for rings containing a heteroatom (O, S, N) are based on a combination of a prefix denoting the heteroatom(s) and a suffix denoting the ring size, as indicated in Table 2. Examples are shown in Fig. 2; as is often the case, the simplest rings have names that depart from this system. See CHEMICAL SYMBOLS AND

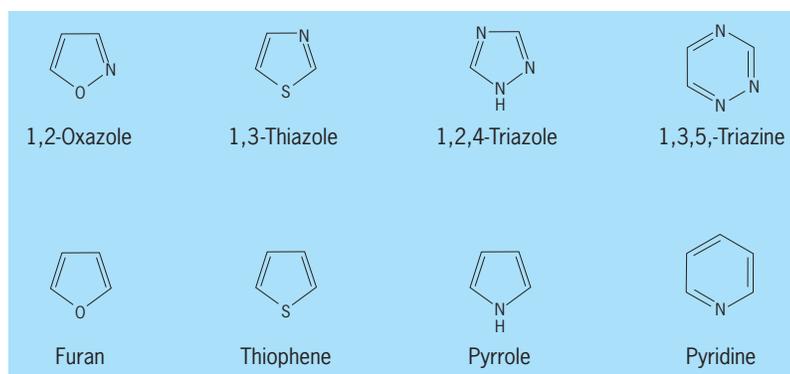


Fig. 2. Examples of structures and names of some representative heterocyclic compounds.

**TABLE 2. Basis of nomenclature of heterocyclic compounds**

Heteroatom	Prefix	Ring size	Suffix
O	ox(a)-	4	-ete
S	thi(a)-	5	-ole
N	az(a)-	6	-ine
		7	-epine

FORMULAS; ORGANIC CHEMISTRY; STRUCTURAL CHEMISTRY. James A. Moore

Bibliography. F. A. Carey, *Organic Chemistry*, 2d ed., 1992; International Union of Pure and Applied Chemistry, *Nomenclature of Organic Chemistry*, 1979.

### Organic photochemistry

A branch of chemistry that deals with light-induced changes of organic material. Because it studies the interaction of electromagnetic radiation and matter, photochemistry is concerned with both chemistry and physics. In considering photochemical processes, therefore, it is also necessary to consider physical phenomena that do not involve strictly chemical changes, for example, absorption and emission of light, and electronic energy transfer.

Because several natural photochemical processes were known to play important roles (namely, photosynthesis in plants, process of vision, and phototropism), the study of organic photochemistry started very early in the twentieth century. However, the breakthrough occurred only after 1950 with the availability of commercial ultraviolet radiation sources and modern analytical instruments for nuclear magnetic resonance (NMR) spectroscopy and (gas) chromatography. See GAS CHROMATOGRAPHY; NUCLEAR MAGNETIC RESONANCE (NMR).

**Photo products.** In general, most organic compounds consist of rapidly interconvertible, nonsepa-

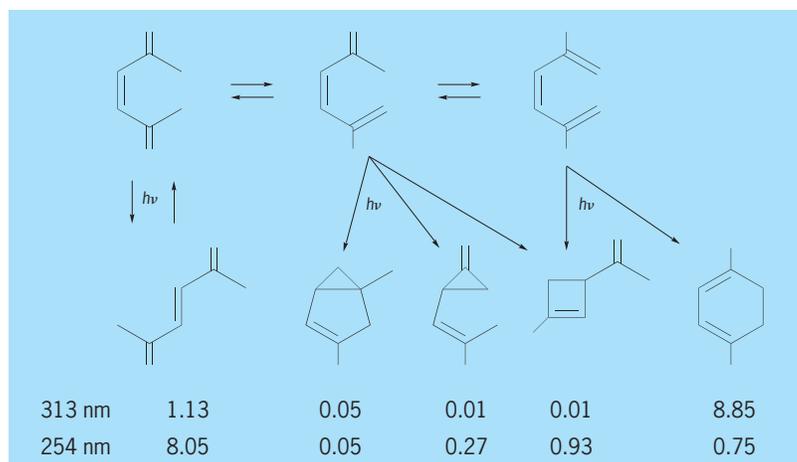
rable conformers, because of the free rotation about single bonds. Each conformer has a certain energy associated with it, and its own electronic absorption spectrum. This is one cause of the broad absorption bands produced by organic compounds in solution. The equilibrium of the conformers may be influenced by the solvent and the temperature. According to the Franck-Condon principle, which states that promotion of an electron by the absorption of a photon is much faster than a single vibration, each conformer will have its own excited-state configuration. See CONFORMATIONAL ANALYSIS; FRANCK-CONDON PRINCIPLE; MOLECULAR STRUCTURE AND SPECTRA.

Because of the change in the  $\pi$ -bond order upon excitation of the substrate and the short lifetime of the first excited state, the excited conformers are not in equilibrium and each yields its own specific photoproduct (**Fig. 1**). Though different conformers may lead to the same photoproduct and one excited conformer may lead to several photoproducts, a change in solvent, temperature, or wavelength of excitation influences the photoproduct composition. This is especially true with small molecules; larger molecules with aromatic groups are less sensitive for small wavelength differences.

The influence of wavelength is also of importance when the primary photoproduct also absorbs light and then gives rise to another photoreaction. Excitation with selected wavelengths or monochromatic light by use of light filters or a monochromator, respectively, may then be profitable for the selective production of the primary product. Similarly, irradiation at a low temperature is helpful in detecting a primary photoproduct that is unstable when heated (thermolabile). See COLOR FILTER.

In addition to photoreactions and fluorescence from the first excited singlet state ( $S_1$ ) of an organic compound, the intersystem crossing (isc) to the triplet state ( $T_1$ ) can also occur. In principle, this state cannot be attained by direct irradiation because of the Pauli exclusion principle. The quantum yield of intersystem crossing depends on the difference in energy between the singlet and triplet state. See EXCLUSION PRINCIPLE; FLUORESCENCE; PHOTOCHEMISTRY; TRIPLET STATE.

Compounds containing heteroatoms [other than carbon (C) or hydrogen (H)], especially those with heavy groups such as bromine (Br) or nitro ( $\text{NO}_2$ ), have a high quantum yield of intersystem crossing caused by spin-orbit coupling. Theoretically the conversion of the triplet state to the ground state ( $S_0$ ) is not allowed, and therefore the lifetime of the triplet state is much longer than that of the first excited singlet state. For this reason, photoreactions are much more probable from the triplet state than from the first excited singlet state. The physical properties of the excited triplet state are different from those of the first excited singlet state, and the unpaired electrons of the triplet state cause the compound to have a biradical character. The photoproducts derived from the two excited states will therefore be different, although a given product may be derived from both states. See QUANTUM CHEMISTRY.



**Fig. 1.** Reaction scheme of three rapidly interconvertible, nonseparable conformers and their yields of specific photoproducts, showing the composition of the photoproduct mixture (in percent) of the Z isomer of 2,5-dimethylhexa-1,3,5-triene after 10% conversion at two wavelengths.

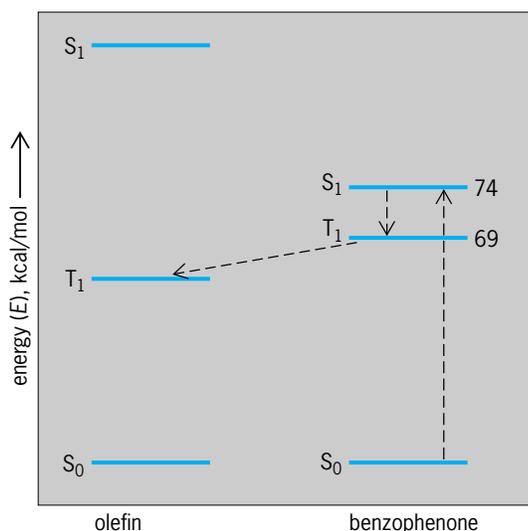


Fig. 2. Benzophenone as a triplet sensitizer for an olefin. Symbols are explained in the text.

**Sensitization.** To investigate which product is derived from the first excited singlet state, a triplet quencher can be added to a compound. The triplet energy of the quencher must be lower than that of the compound. However, to identify the products from the triplet state or to obtain only products from this state, the triplet state has to be populated without formation of the first excited singlet state. This can be done by a method known as intermolecular energy transfer. This involves adding a sensitizer to the compound. The sensitizer compound has to possess a high quantum yield of intersystem crossing, and a singlet energy lower and a triplet energy higher than that of the original compound. (Fig. 2). Frequently, aromatic ketones and certain dyes that fulfill the electronic requirements are used as sensitizers.

**Energy surface description.** It is assumed that a potential energy curve controls nuclear motion, except for situations where two surfaces come close together. Each point on such a curve represents a specific geometry (horizontal axis) and a specific energy (vertical axis) [Fig. 3]. The energy barrier between the reactant and the product on the ground-state surface may be too high to be overcome by a thermal reaction. Irradiation of the reactant gives rise to the excited species, usually with enough vibrational energy to overcome small barriers in the first excited singlet state surface and arrive at a point that is close to the ground state for a given geometry. By a jump from the first excited singlet state to the ground state at this position, the product and/or the reactant can be attained. This is an example of a nonadiabatic or diabatic photoreaction, which is the most usual type of photoreaction. Thus a small change in the mutual positions of the minimum and maximum of the curves of the first excited singlet state and the ground state due to a substituent, solvent, or another factor will affect the balance of the formation of product and reactant. While organic molecules, consisting of

several atoms, possess much more complicated energy hyperfaces, this kind of approximation is found to be useful.

**Photoreactions.** Organic photochemistry involves all classes of organic compounds. Some examples include the photochemistry of saturated hydrocarbons, unsaturated hydrocarbons, aromatic compounds, and carbonyl compounds.

*Saturated hydrocarbons.* These compounds have only  $\sigma$  bonds and absorb in the vacuum ultraviolet part of the spectrum. Their photochemistry has not been extensively investigated. Substituted alkanes that absorb above about 200 nanometers are photolyzed on irradiation and form alkyl radicals, which react in the usual way. See CHEMICAL BONDING; REACTIVE INTERMEDIATES.

*Unsaturated hydrocarbons.* These compounds possess one or more double bonds. The second bond of a double bond is known as a  $\pi$  bond; it has a lower energy than the  $\sigma$  bond. With light, the excited  $\pi\pi^*$  state is populated.

Because of the large distance between the excited singlet and triplet states of these compounds, the intersystem crossing is inefficient and most photoreactions on direct excitation occur from the  $^1\pi\pi^*$  state. Usually the *trans,cis*-isomerization is the most efficient reaction (Fig. 1). Other common intramolecular photoreactions are sigmatropic shifts, electrocyclicizations, and cycloadditions.

Not uncommonly in photochemistry, in a transition state or in the primary photoproduct the aromaticity is lost, because of the high energy of the excited state.

*Aromatic compounds.* Contrary to thermal reactions of aromatic compounds, which occur with retention of aromaticity, photochemical reactions of these compounds give a variety of nonaromatic products. For example, prismane can be derived as a photoproduct of butylbenzene, in more than 70% yield. In general, the yield of the photoproducts depends on the type and position of the substituents in the benzene ring.

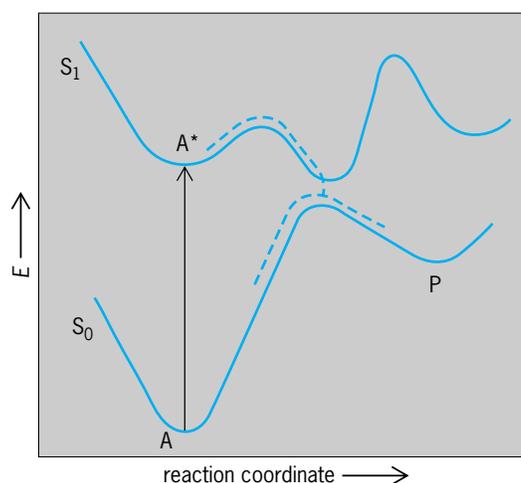


Fig. 3. Energy curves of the ground state ( $S_0$ ) and the first excited singlet state ( $S_1$ ) of the photoreaction. By a jump from the first excited singlet state to the ground state, the product (P) or reactant (A) can be attained

Although the energy content of benzene is lower than that of prismane, the latter molecule is stable at room temperature as the thermal ring-opening is not allowed theoretically, because of orbital symmetry considerations. Photochemical isomerizations of benzene derivatives occur via benzvalene or prismane intermediates and not by shifts of the substituents.

Irradiation of benzene and its derivatives in the presence of olefins usually leads to cycloaddition products. The preferred photoreaction with olefins bearing an electron-withdrawing group is a 1,2-(ortho) cycloaddition. The more interesting 1,3-(meta) cycloaddition is generally observed when nonpolar alkenes are used. Not only intermolecular reactions take place, but also intramolecular cycloadditions are found when the double bond is at a suitable position to react. Substitution reactions of aromatic compounds induced by light are usually nucleophilic, again in contrast to those in the ground state, which are electrophilic in nature. See SUBSTITUTION REACTION.

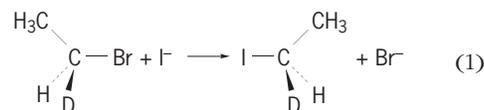
**Carbonyl compounds.** Carbonyl compounds contain a carbonyl group (C=O). In the ground state the C-O bond is polarized, with a partial negative charge on the oxygen atom. Excitation by light causes an  $n-\pi^*$  transition, with the oxygen becoming electron deficient and the carbon more electron rich. The ( $n,\pi^*$ ) triplet state of the carbonyl group has a diradical character. These properties determine the reactivities of this class of compounds and make the photochemistry of these compounds rather complicated. In aliphatic carbonyl compounds the intersystem crossing from the first excited singlet state to the triplet state is relatively slow, so that reactions are probable from both states. However, in conjugated aryl ketones and  $\alpha,\beta$ -unsaturated carbonyl compounds, the intersystem crossing is very fast and only the triplet state is the reactive state for reactions. The most important types of photoreaction are  $\alpha$  cleavage of the molecule, intramolecular hydrogen abstraction and photoelimination, photoreduction, and cycloaddition to olefins. See ORGANIC CHEMISTRY; ORGANIC SYNTHESIS. Wim. H. Laarhoven Bibliography. A. M. Braun, E. Oliveres, and M. T. Maurette, *Photochemical Technology*, 1991; A. Gilbert and J. Baggott, *Essentials of Molecular Photochemistry*, 1991; W. M. Horspool (ed.), *Handbook of Organic Photochemistry and Photobiology*, 1994; J. Mattay and A. G. Griesbeck (eds.), *Photochemical Key Steps in Organic Syntheses*, 1994; S. A. L. Murov, *Handbook of Photochemistry*, 2d ed., 1993.

### Organic reaction mechanism

A complete, step-by-step account of how a reaction of organic compounds takes place. A fully detailed mechanism would correlate the original structure of the reactants with the final structure of the products and would account for changes in structure and energy throughout the progress of the reaction. A com-

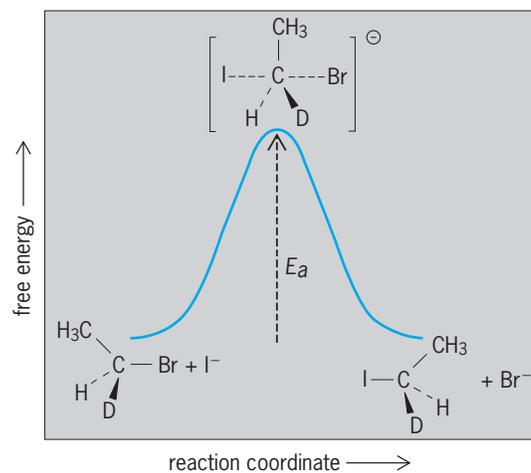
plete mechanism would also account for the formation of any intermediates and the rates of interconversions of all of the various species. Because it is not possible to detect directly all of these details of a reaction, evidence for a reaction mechanism is always indirect. Experiments are designed to produce results that provide logical evidence for (but can never unequivocally prove) a mechanism. For most organic reactions, there are mechanisms that are considered to be well established (that is, plausible) based on bodies of experimental evidence. Nevertheless, new data often become available that provide further insight into new details of a mechanism or that occasionally require a complete revision of an accepted mechanism.

**Classification of organic reactions.** The description of an organic reaction mechanism typically includes designation of the overall reaction (for example, substitution, addition, elimination, oxidation, reduction, or rearrangement), the presence of any reactive intermediates (that is, carbocations, carbanions, free radicals, radical ions, carbenes, or excited states), the nature of the reagent that initiates the reaction (such as electrophilic or nucleophilic), the presence of any catalysis (such as acid or base), and any specific stereochemistry. For example, reaction (1) would



be described as a concerted nucleophilic substitution of an alkyl halide that proceeds with inversion of stereochemistry. A reaction that proceeds in a single step, without intermediates, is described as concerted or synchronous. Reaction (1) is an example of the  $S_N2$  mechanism (substitution, nucleophilic, bimolecular).

**Potential energy diagrams.** A common method for illustrating the progress of a reaction is the potential energy diagram, in which the free energy of the system is plotted as a function of the completion of the reaction (see *illus.*).

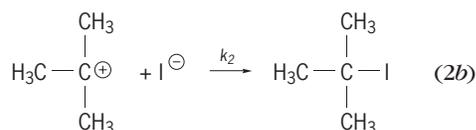
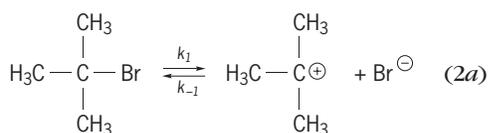


Potential energy–reaction coordinate diagram for a typical nucleophilic substitution reaction that proceeds by the  $S_N2$  mechanism.  $E_a$  = activation energy.

The reaction coordinate is intended to represent the progress of the reaction, and it may or may not correlate with an easily observed or measurable feature. In reaction (1), the reaction coordinate could be considered to be the increasing bond length of the carbon-bromine (C-Br) bond as it is broken, or the decreasing separation of C and iodine (I) as they come together to form a bond. In fact, a complete potential energy diagram should illustrate the variation in energy as a function of both of these (and perhaps several other relevant structural features), but this would require a three-dimensional (or higher) plot.

Besides identifying the energy levels of the original reactants and the final products, the potential energy diagram indicates the energy level of the highest point along the reaction pathway, called the transition state. It is important to recognize that the reaction pathway actually taken in a reaction mechanism is chosen because it represents the lowest-energy pathway. The analogy often used is a mountain pass between two valleys, in which the top of the pass represents the transition state; the top of the pass represents the highest point traveled, but not the highest point in the vicinity. Because the transition state represents the highest energy that the molecules must attain as they proceed along the reaction pathway, the energy level of the transition state is a key indication of how easily the reaction can occur. Features that tend to make the transition state more stable (lower in energy) make the reaction more favorable. Such stabilizing features could be intramolecular, such as electron donation or withdrawal by substituents, or intermolecular, such as stabilization by solvent. See CHEMICAL BONDING; ENERGY.

**Kinetics.** Another way to illustrate the various steps involved in a reaction mechanism is as a kinetic scheme that shows all of the individual steps and their rate constants. The  $S_N2$  mechanism is a single step, so the kinetics must represent that step; the rate is observed to depend on the concentrations of both the organic substrate and the nucleophile. However, for multistep mechanisms the kinetics can be a powerful tool for distinguishing the presence of alternative pathways. For example, when more highly substituted alkyl halides undergo nucleophilic substitution, the rate is independent of the concentration of the nucleophile. This evidence suggests a two-step mechanism, called the  $S_N1$  mechanism, as shown in reaction scheme (2), where the  $k$  terms represent rate constants.



The  $S_N1$  mechanism accomplishes the same overall nucleophilic substitution of an alkyl halide, but does so by initial dissociation of the leaving group ( $\text{Br}^-$ ) to form a carbocation, step (2a). The nucleophile then attaches to the carbocation to form the final product, step (2b). Alkyl halides that have bulky groups around the carbon to be substituted are less likely to be substituted by the direct  $S_N2$  mechanism, because the nucleophile encounters difficulty in making the bond to the inaccessible site (called steric hindrance). If those alkyl groups have substituents that can support a carbocation structure, generally by electron donation, then the  $S_N1$  mechanism becomes preferable.

A crucial feature of a multistep reaction mechanism is the identification of the rate-determining step. The overall rate of reaction can be no faster than its slowest step. In the  $S_N1$  mechanism, the bond-breaking reaction (2a) is typically much slower than the bond-forming reaction (2b). Hence, the observed rate is the rate of the first step only. Thus, kinetics can distinguish the  $S_N1$  and  $S_N2$  mechanisms, as shown in Eqs. (3) and (4), where R is an alkyl group, X is a halo-

$$\text{Rate} = k [\text{RX}] [\text{Nu}] \quad \text{for an } S_N2 \text{ mechanism} \quad (3)$$

$$\text{Rate} = k [\text{RX}] \quad \text{for an } S_N1 \text{ mechanism} \quad (4)$$

gen or other leaving group, Nu is a nucleophile, and the terms in the brackets represent concentrations.

A more complete description of the  $S_N1$  mechanism was recognized when it was observed that the presence of excess leaving group [for example,  $\text{Br}^-$  in reaction (2)] can affect the rate (called the common ion rate depression). This indicated that the mechanism should include a reverse step [ $k_{-1}$  in reaction step (2a)] in which the leaving group returns to the cation, regenerating starting material. In this case, the rate depends in a complex manner on the competition of nucleophile and leaving group for reaction with the carbocation. See CHEMICAL DYNAMICS; REACTIVE INTERMEDIATES; STERIC EFFECT (CHEMISTRY).

**Activation parameters.** The temperature dependence of the rate constant provides significant information about the transition state of the rate-determining step. The Arrhenius equation (5)

$$k = A e^{-E_a/RT} \quad (5)$$

expresses that dependence in term of an exponential function of temperature and an activation energy,  $E_a$ ;  $A$  is called the Arrhenius or preexponential factor, and  $R$  is the gas constant. See GAS; LOGARITHM.

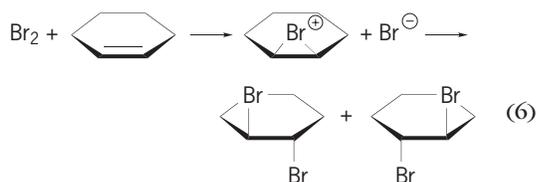
The activation energy represents the energy difference between the reactants and the transition state, that is, the amount of energy that must be provided in order to proceed along the reaction pathway successfully from reactant to product (see illus.).

A more complete description of transition-state theory expresses the rate constant in terms of contributions from both an enthalpy of activation,  $\Delta H^\ddagger$ , and an entropy of activation,  $\Delta S^\ddagger$ . In this case, enthalpy of activation represents the additional

enthalpy content of the transition state compared to reactants;  $\Delta H^\ddagger$  can be rather directly correlated with the energy of activation. The entropy of activation,  $\Delta S^\ddagger$ , represents the additional entropy content of the transition state compared to reactants;  $\Delta S^\ddagger$  is highly informative with respect to the reaction mechanism, because it indicates whether the overall structure becomes more ordered or more disordered as the reaction proceeds from reactants to transition state. For example, the transition state for the  $S_N2$  reaction (see illus.) requires the reactants to be organized in a very specific manner, with the nucleophile approaching the opposite side of the leaving group; the  $S_N2$  reaction typically shows a negative value for  $\Delta S^\ddagger$ . See ENTHALPY; ENTROPY.

**Stereochemistry.** Careful attention to the stereochemistry of a reaction often provides crucial insight into the specific orientation of the molecules as they proceed through the reaction mechanism. The complete inversion of stereochemistry observed in the  $S_N2$  mechanism provides evidence for the backside attack of the nucleophile. Alkyl halides that undergo substitution by the  $S_N1$  mechanism do not show specific stereochemistry, since the loss of the leaving group is completely uncorrelated with the bonding of the nucleophile.

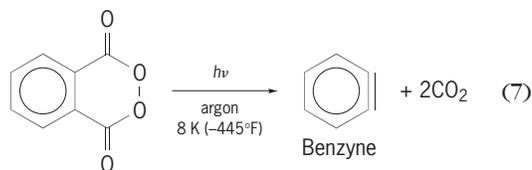
In addition reactions, possible stereochemical outcomes are addition of the new bonds to the same or opposite sides of the original pi bond, called syn and anti addition, respectively. The anti addition of bromine to double bonds provides evidence for the intermediacy of a bridged bromonium ion, as shown in reaction (6).



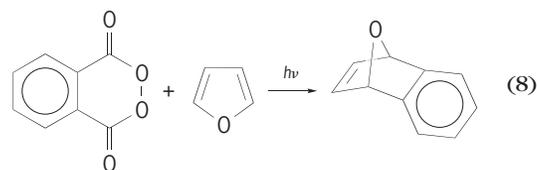
**Experimental probes of mechanisms.** Chemists use a variety of techniques to determine the mechanistic course of reactions. The most direct is actual observation of the reactants, intermediates, and products in real time during the reaction. Fast spectroscopic techniques are becoming ever faster and more sensitive, so that direct evidence for short-lived intermediates is often obtainable. A representative fast technique is flash photolysis, in which a compound is rapidly decomposed by an intense pulse of light. Laser light pulses less than a picosecond ( $10^{-12}$  s) are available; they, combined with detection methods that can record data on a time scale of picoseconds, allow for direct observation of the spectroscopic or other properties of reaction intermediates. See LASER SPECTROSCOPY; ULTRAFast MOLECULAR PROCESSES.

Another approach is to create an environment that is stabilizing for the reaction intermediates to allow study of their properties by standard techniques. A common approach called matrix isolation creates the intermediate in a frozen matrix, so that it is unable to diffuse and react with other species. In this case, the

reactive intermediate must be formed by a unimolecular reaction, since it cannot diffuse to meet another reactant, and typically light-induced photodecompositions are used. An example of the successful use of matrix isolation to generate and observe benzyne is shown in reaction (7).



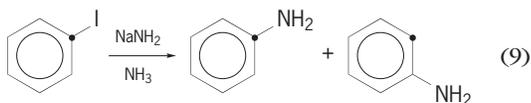
Isolating benzyne and collecting spectroscopic evidence for its structure under matrix isolation conditions provides evidence that it could be involved as an intermediate in the same reaction under other conditions. At room temperature in solution, benzyne reacts with a variety of substrates to form cycloaddition products, such as in reaction (8). The



observation of these same products upon warming a frozen matrix containing benzyne plus the added substrate helps to verify its reactivity patterns and strengthens its implication as a reaction intermediate. See MATRIX ISOLATION.

**Substituent effects and Hammett equation.** By far the major methods for determining mechanisms of reactions as they occur under normal laboratory conditions utilize kinetics and stereochemistry, as already illustrated for the differentiation of  $S_N1$  and  $S_N2$  substitution mechanisms. Variations in the observed kinetics as the structure of the reactants are systematically varied are the basis of one of the most powerful methods of detecting details of a reaction mechanism. Correlation of substituent effects by the Hammett equation is a standard approach that allows the prediction of the extent of electron shifting in a reaction mechanism. Substituents are assigned substituent constants ( $\sigma$ ) that indicate the degree to which they are electron-donating (negative values) or electron-withdrawing (positive values) in a standard reaction (the dissociation of benzoic acid). Reactions then display a dependence on the substituent constants, a reaction constant ( $\rho$ ) that indicates the degree to which the reaction is favored by electron donation (negative values) or favored by electron withdrawal (positive values). The negative  $\rho$  value indicates that the reaction is favored by substituents that donate electron density, a point that would not have been obvious from looking at the reactants and products, all of which are uncharged. This evidence suggests that the transition state is polarized. In that transition state, electron donation provides a stabilizing effect, lowers the energy of the transition state, and thereby increases the rate. See SUBSTITUTION REACTION.

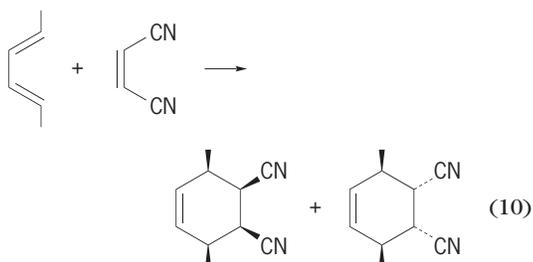
**Isotope effects and isotopic labeling.** The use of isotopes is an important tool in mechanistic studies. Kinetic isotope effects are often subtle but useful methods to distinguish details of the transition state. The magnitude of the kinetic isotope effect provides a measure of the amount of bond breaking at the transition state. Isotopes are also used simply as labels to distinguish otherwise identical atoms. For example, the use of  $^{13}\text{C}$  isotopic labeling in reaction (9), where



the heavy dot represents the position of a  $^{13}\text{C}$  atom, provided evidence that the mechanism was not a simple nucleophilic substitution. The appearance of  $^{13}\text{C}$  equally to two different positions of the product is evidence for a symmetrical intermediate such as benzene. See DEUTERIUM; ISOTOPE.

**Theoretical correlations.** The theory underlying organic chemistry has developed to the stage that both qualitative and quantitative approaches often provide excellent insight into the workings of a reaction mechanism. The principle of conservation of orbital symmetry, developed by R. B. Woodward and R. Hoffmann and often called the Woodward-Hoffmann rules, provides a simple yet powerful method for predicting the stereochemistry of concerted reactions. The principle states that the formation of new bonds and the breaking of old bonds will be carried out preferentially in a manner that maximizes bonding at all times. See PERICYCLIC REACTION; WOODWARD-HOFFMANN RULE.

By knowing the symmetry properties of the molecular orbitals of reactants and products, the preferred (or allowed) pathway can be predicted. The approach can be greatly simplified by recognizing the cyclic delocalized transition state of the reaction and counting the number of electrons involved in the cyclic transition state. If there are  $(4n + 2)$  electrons (that is, 2, 6, 10, ... electrons), the reaction is considered favorable, because all of the bonding can occur on the same sides of the molecules. Thus cycloadditions such as the Diels-Alder reaction (10) are



favorable according to the Woodward-Hoffmann rules, because they involve six electrons in the delocalized transition state.

Reactions involving  $4n$  electrons, such as the cycloaddition of two double bonds to form a four-membered ring, are unfavorable by these rules and are not observed except in photochemical reactions, for which the Woodward-Hoffmann rules are ex-

actly reversed. See DELOCALIZATION; DIELS-ALDER REACTION; MOLECULAR ORBITAL THEORY; PHOTOCHEMISTRY.

The improvement of computing capabilities has allowed quantitative calculations to become accurate enough to predict energies and structures for simple molecules. Quantum-mechanics calculations may be performed either from first principles, that is, without simplifying assumptions, or semiempirically, that is, with some standardizing parameters. The desired result is a potential energy map that correlates the energy levels of the possible structures of the reactants as they transform to products. Such approaches can help to rule out possible pathways as too high in energy or can suggest alternative pathways that appear feasible based on calculated energies. See COMPUTATIONAL CHEMISTRY; MOLECULAR MECHANICS; QUANTUM CHEMISTRY.

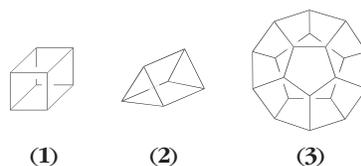
Carl C. Wamser

Bibliography. B. K. Carpenter, *Determination of Organic Reaction Mechanisms*, 1984; R. D. Guthrie and W. P. Jencks, IUPAC recommendations for the representation of reaction mechanisms, *Acc. Chem. Res.*, 22:343-349, 1989; T. H. Lowry and K. S. Richardson, *Mechanism and Theory in Organic Chemistry*, 3d ed., 1987; J. March, *Advanced Organic Chemistry*, 5th ed., 2001.

## Organic synthesis

The making of an organic compound from simpler starting materials.

**Role in chemistry.** Organic synthesis plays an important role in chemistry, biochemistry, medicine, agriculture, molecular biology, physics, materials science, electronics, and engineering by allowing for the creation of specific molecules for scientific and technological investigations. In some cases the target molecule has an unusual structure whose characterization may advance understanding of various theoretical aspects of chemistry. Such a molecule may possess particularly unusual patterns of bonding, such as a strained ring system or unique symmetry; examples are cubane (1), prismane (2), and dodecahedrane (3). Once a structurally unusual molecule



has been synthesized, its properties can be studied to elucidate concepts of chemical bonding, spectroscopy, and reactivity. See CHEMICAL BONDING; ORGANIC CHEMISTRY; STERIC EFFECT (CHEMISTRY).

There are other reasons for synthesizing a particular molecule. For example, a molecule may be isolated as a natural product from some obscure organism, and a synthesis of the molecule in the laboratory can provide larger quantities of it for testing for properties that might indicate its usefulness, such as in medicine or agriculture. Another reason for

synthesizing a molecule is because its structure is different from any known compound and the molecule is predicted to possess some useful property, such as selective ion-binding ability or biological activity such as antibiotic, herbicidal, insecticidal, anticancer, or antiviral activity. In the study of polymers, a particular monomeric subunit might be chosen for synthesis because, upon polymerization, it might yield a polymer having useful physical properties. By hypothesizing that a certain molecule will have a particular property, then obtaining that molecule by synthesis, a chemist can test that hypothesis and thus obtain a better understanding of the relationship between molecular structure and function, and at the same time producing—if the synthesized molecule does possess the desired property—a novel chemical that can be used to better the human condition. *See* POLYMER; POLYMERIC COMPOSITE; POLYMERIZATION.

Finally, a chemist might devise and carry out the synthesis of an already-known compound in order to introduce one or more radioactive isotopes of some of the constituent elements into the product. Such labeled molecules are useful in biology for studying metabolic processes and for isolating enzymes and proteins that bind specifically to that molecule. *See* RADIOISOTOPE (BIOLOGY).

**Synthetic strategy.** The heart of organic synthesis is designing synthetic routes to a molecule. Organic synthesis can be compared with architecture and construction, where the chemist must devise a synthetic route to a target molecule (a “blueprint”), then utilize a repertoire of organic reactions (the “tools”) to complete the “construction project.” Along the way, the synthetic chemist must make extensive use of analytical techniques for purifying and characterizing intermediate products as well as the final product. *See* CHROMATOGRAPHY; CRYSTALLIZATION; DISTILLATION; INFRARED SPECTROSCOPY; MASS SPECTROMETRY; NUCLEAR MAGNETIC RESONANCE (NMR).

The simplest synthesis of a molecule is one in which the target molecule can be obtained by submitting a readily available starting material to a single reaction that converts it to the desired target molecule. However, in most cases the synthesis is not that straightforward; in order to convert a chosen starting material to the target molecule, numerous steps that add, change, or remove functional groups, and steps that build up the carbon atom framework of the target molecule may need to be done.

*Retrosynthetic analysis.* A systematic approach for designing a synthetic route to a molecule is to subject the target molecule to an intellectual exercise called a retrosynthetic analysis. This involves an assessment of each functional group in the target molecule and the overall carbon atom framework in it; a determination of what known reactions form each of those functional groups or that build up the necessary carbon framework as a product; and a determination of what starting materials (synthetic precursors or synthetic intermediates) for each such reaction are required. The resulting starting materials are then subjected to the same retrosynthetic analysis, thus working backward from the target molecule until

starting materials that are commercially available (or available by synthesis following an already published procedure) are derived.

The retrosynthetic analysis of a target molecule usually results in more than one possible synthetic route. It is therefore necessary to critically assess each derived route in order to choose the single route that is most feasible (most likely to proceed as written, with as few unwanted side reactions as possible) and most economical (involving the fewest steps and using the least expensive starting materials). The safety of each possible synthetic route (the toxicity and reactivity hazards associated with the reactions involved) is also considered when assessing alternative synthetic routes to a molecule.

*Stereoselectivity.* Selectivity is an important consideration in the determination of a synthetic route to a target molecule. Stereoselectivity refers to the selectivity of a reaction for forming one stereoisomer of a product in preference to another stereoisomer. Stereoselectivity cannot be achieved for all organic reactions; the nature of the mechanism of some reactions may not allow for the formation of one particular configuration of a chiral (stereogenic) carbon center or one particular geometry (cis versus trans) for a double bond or ring. When stereoselectivity can be achieved in a reaction, it requires that the reaction proceed via a geometrically defined transition state and that one or both of the reactants possess a particular geometrical shape during the reaction. For example, if one or both of the reactants is chiral, the absolute configuration of the newly formed stereogenic carbon center can be selected for in many reactions. Nucleophilic substitution is an example of a reaction that can proceed stereoselectively when the starting material is chiral. Pericyclic reactions also proceed stereoselectively, because they involve transition states that have well-defined geometries. The achievement of stereoselectivity is an important aspect of organic synthesis, because usually a single stereoisomer of a target molecule is the desired goal of a synthesis. Sometimes the target molecule contains a chiral (stereogenic) carbon center; that is, it can exist as either of two possible enantiomers. The possible synthetic routes to the target molecule may not be selective for forming a single enantiomer of the target molecule; each would form a racemic mixture. In some cases, such nonstereoselective synthetic routes to a molecule are acceptable. However, if a synthesis of a single stereoisomer of a target molecule is required, the stereoselectivity of the reactions derived during the retrosynthetic analysis would need to be considered. The development of stereoselective reactions is an active area of research in organic synthesis. *See* ASYMMETRIC SYNTHESIS; ORGANIC REACTION MECHANISM; PERICYCLIC REACTION; RACEMIZATION; STEREOCHEMISTRY.

*Chemoselectivity.* This term refers to the ability of a reagent to react selectively with one functional group in the presence of another similar functional group. An example of a chemoselective reagent is a reducing agent that can reduce an aldehyde and not a ketone. In cases where chemoselectivity

cannot be achieved, the functional group that should be prevented from participating in the reaction can be protected by converting it to a derivative that is unreactive to the reagent involved. The usual strategy employed to allow for such selective differentiation of the same or similar groups is to convert each group to a masked (protected) form which is not reactive but which can be unmasked (deprotected) to yield the group when necessary. The development and usage of protecting groups is an important aspect of organic synthesis.

**Other strategies.** Most target molecules for synthesis are complicated, and their syntheses require many steps. An important aspect of the strategy for synthesizing complex molecules is to devise a convergent synthesis, where relatively large subunits of the target molecule are synthesized individually and then attached together to form the complete or nearly complete target. A convergent synthesis strategy is more economical than the alternative linear synthesis strategy, where the target molecule is built up step by step, one group at a time.

**Synthetic reactions.** A retrosynthetic analysis can be done by using sophisticated computer programs in order to derive as comprehensive a list of possible synthetic routes to a target molecule as possible. One important aspect of synthetic planning is that it depends upon a knowledge (by a chemist, or programmed into a computer program) of known synthetic transformations, reactions that build up carbon skeletons or introduce functional groups or interconvert functional groups. A large variety of organic reactions that can be used in syntheses are known. They can be categorized according to whether they feature a functional group interconversion or a carbon-carbon bond formation.

Functional group interconversions (**Table 1**) are reactions that change one functional group into another functional group. A functional group is a nonhydrogen, non-all-singly-bonded carbon atom or group of atoms. Included in functional group interconversions are nucleophilic substitution reactions, electrophilic additions, oxidations, and reductions. See COMPUTATIONAL CHEMISTRY; ELECTROPHILIC AND NUCLEOPHILIC REAGENTS; OXIDATION-REDUCTION; OXIDIZING AGENT; SUBSTITUTION REACTION.

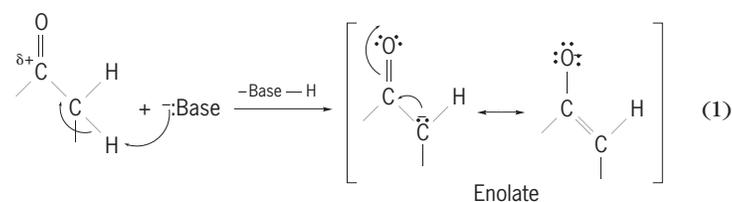
Carbon-carbon bond-forming reactions (**Table 2**) feature the formation of a single bond or double bond between two carbon atoms. This is a particularly important class of reactions, as the basic strategy of synthesis—to assemble the target molecule from simpler, hence usually smaller, starting materials—implies that most complex molecules must be synthesized by a process that builds up the carbon skeleton of the target by using one or more carbon-carbon bond-forming reactions.

**Nucleophilic reactions.** An important feature of many common carbon-carbon bond-forming reactions is the intermediacy of a carbanionic intermediate, a molecule bearing a carbon-metal bond formed by deprotonation, by a strong base, of a carbon-hydrogen bond that is relatively acidic because of a nearby

electron-withdrawing group or because of the insertion of a metal into a carbon-halogen bond. Such carbanionic intermediates are good nucleophiles (electron-rich, partly negatively charged centers) and thus react readily with added aldehydes, alkyl halides, esters, or other electrophilic (electron-poor, partly positively charged) carbon centers to form the carbon-carbon bond. For example, Grignard reagents are formed by the reaction of an organohalide with magnesium metal in ether solvents in the absence of water. Grignard reagents react with aldehydes or ketones by adding the nucleophilic carbon bound to the magnesium to the electrophilic carbonyl carbon of the aldehyde functional group (**Table 2**). The resulting magnesium alkoxide intermediates will, upon subsequent addition of an acid (usually dilute aqueous hydrochloric acid), undergo protonation to yield neutral alcohol products whose carbon frameworks bear both the carbons of the aldehyde or ketone and the carbons of the organohalide. See REACTIVE INTERMEDIATES.

By changing the metal associated with a carbanionic reagent, the reactivity of the reagent can be altered for synthesis. For example, an organocupper (Gilman) reagent will react more readily in nucleophilic substitution reactions with alkyl halides than will organomagnesium (Grignard) reagents (**Table 2**). Gilman reagents, which are lithium diorganocupper salts, can be formed by the reaction of organohalides, in the absence of water and oxygen, with lithium metal to form the organolithium reagent (RLi), followed by the addition of copper(I) iodide to form the lithium diorganocupper reagent. Subsequent addition of an alkyl halide results in a nucleophilic substitution reaction, where the carbon bound to the copper displaces a halide ion to form a coupled product consisting of the carbon framework of the organohalide precursor to the Gilman reagent connected to the carbon framework of the alkyl halide by a carbon-carbon single bond. See ORGANOMETALLIC COMPOUND.

**Addition reactions.** The Grignard reaction is an example of an addition reaction, where the carbanionic reagent adds itself to the carbon-oxygen double bond of the aldehyde or ketone. A similar addition reaction is the aldol addition reaction (**Table 2**), where the carbanionic reagent is an enolate formed by the removal, by a strong base, of a hydrogen from a carbon that is bound to a carbonyl group. The electropositive nature of the carbonyl group enhances the acidity of the carbon-hydrogen (C-H) bond next to it, thus allowing such a deprotonation (removal of the hydrogen) to occur, as shown in reaction (1).



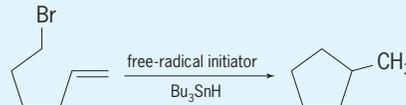
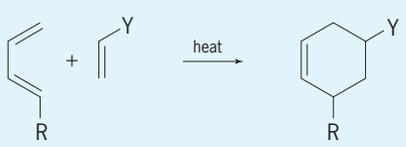
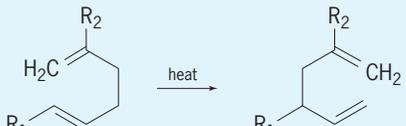
See GRIGNARD REACTION.

TABLE 1. Examples of some functional-group interconversions

General equation for the reaction*	Net transformation (name)
$\begin{array}{c} R_1 \\   \\ R_3 - C - X \\   \\ R_2 \end{array} + Nu^- \longrightarrow Nu - \begin{array}{c} R_1 \\   \\ C \\   \\ R_2 \\   \\ R_3 \end{array} + X^-$ <p>(X = Cl, Br, I, or OSO<sub>2</sub>R; Nu = OH, OR, CN, NR<sub>2</sub>, others)</p>	Alkyl halide to various functional groups (alcohols, ethers, nitriles, amines, others)
$\begin{array}{c} R_1 \\   \\ R_2 - CH - CH - X \\   \quad   \\ H \quad R_3 \end{array} + \text{base} \longrightarrow \begin{array}{c} R_2 \quad R_4 \\ \backslash \quad / \\ C = C \\ / \quad \backslash \\ R_1 \quad R_3 \end{array}$ <p>(such as CH<sub>3</sub>O<sup>-</sup>)</p>	Alkyl halide to alkene (elimination)
$ROH + HX \longrightarrow RX$ <p>(X = Cl, Br, I)</p>	Alcohol to alkyl halide
$\begin{array}{c} R_1 \\   \\ R_2 - CH - CH - OH \\   \quad   \\ H \quad R_3 \end{array} + H_2SO_4 \longrightarrow \begin{array}{c} R_2 \quad R_4 \\ \backslash \quad / \\ C = C \\ / \quad \backslash \\ R_1 \quad R_3 \end{array}$	Alcohol to alkene (dehydration)
$\begin{array}{c} OH \\   \\ R_1 - CH - R_2 \end{array} \xrightarrow{CrO_3\text{-pyridine}} \begin{array}{c} O \\    \\ R_1 - C - R_2 \end{array}$	Oxidation of alcohol to ketone or aldehyde
$R_1YH + \begin{array}{c} O \\    \\ R_2 - C - X \end{array} \longrightarrow \begin{array}{c} O \\    \\ R_2 - C - OR_1 \end{array}$ <p>(Y = O or N; X = OH, Cl, others)</p>	Alcohol and carboxylic acid derivative to ester (esterification); amine and carboxylic acid derivative to amide
$\begin{array}{c} R_2 \quad R_4 \\ \backslash \quad / \\ C = C \\ / \quad \backslash \\ R_1 \quad R_3 \end{array} + RCO_3H \longrightarrow \begin{array}{c} R_2 \quad R_4 \\ \backslash \quad / \\ C - O - C \\ / \quad \backslash \\ R_1 \quad R_3 \end{array}$	Alkene to epoxide (epoxidation)
$\begin{array}{c} R_2 \quad R_4 \\ \backslash \quad / \\ C = C \\ / \quad \backslash \\ R_1 \quad R_3 \end{array} + H_2 \xrightarrow[\text{(or other catalyst)}]{Pd} \begin{array}{c} R_2 \quad R_4 \\   \quad   \\ H - C - C - H \\   \quad   \\ R_1 \quad R_3 \end{array}$	Alkene to alkane (hydrogenation)
$\begin{array}{c} O \\    \\ R_1 - C - R_2 \end{array} + NaBH_4 \longrightarrow \begin{array}{c} OH \\   \\ R_1 - CH - R_2 \end{array}$	Reduction of ketone or aldehyde to alcohol
$R_1COOR_2 \xrightarrow[2. H_2O \text{ workup}]{1. LiAlH_4} R_1CH_2OH + R_2OH$	Reduction of ester to two alcohols
$R_1COOR_2 \xrightarrow{H_2O, \text{ acid or base}} R_1COOH + R_2OH$	Ester to carboxylic acid and alcohol (ester hydrolysis)
$R - CN \xrightarrow[2. H_2O \text{ workup}]{1. LiAlH_4} RCH_2NH_2$	Reduction of nitrile to amine
$\text{C}_6\text{H}_6 + E^+ \longrightarrow \text{C}_6\text{H}_5E$ <p>(E = Br, NO<sub>2</sub>, R, RCO, others)</p>	Benzene to substituted benzene (electrophilic aromatic substitution)

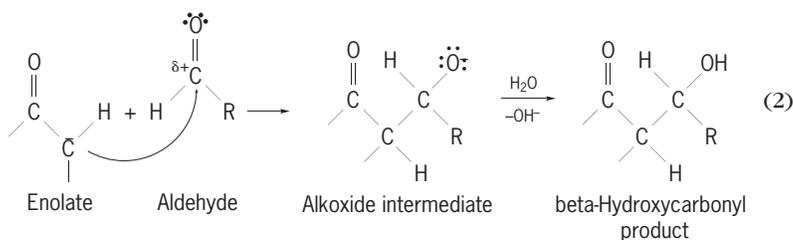
\*R = any organic group (alkyl, aryl, alkenyl) or a hydrogen atom. Nu = nucleophile.

TABLE 2. Examples of some carbon-carbon bond-forming reactions

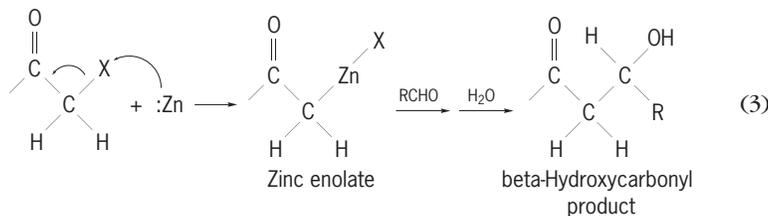
General equation for the reaction	Name of reaction
$R_1X + Mg \longrightarrow R_1MgX \xrightarrow[2. H^+]{1. R_2COR_3} R_1-\overset{\overset{OH}{ }}{C}-R_3$ <p>(X = Cl, Br, I)</p>	Grignard
$R_1X \xrightarrow[2. CuI]{1. Li} (R_1)_2CuLi \xrightarrow{R_2X} R_1-R_2$ <p>(X = Cl, Br, I)</p>	Gilman
$X-\overset{\overset{O}{  }}{C}-CH_2R_1 \xrightarrow[3. H^+]{1. (C_3H_7)_2NLi, 2. R_2CHO} X-\overset{\overset{O}{  }}{C}-\overset{\overset{OH}{ }}{C}-R_1-R_2$ <p>(X = R, RO, NR<sub>2</sub>, others)</p>	Aldol addition
$XCH_2CO-Y + RCHO \xrightarrow[2. H_2O \text{ workup}]{1. Zn} \overset{\overset{OH}{ }}{RCH}-CH_2CO-Y$ <p>(X = Cl, Br, or I; Y = R, RO, NR<sub>2</sub>, others)</p>	Reformatsky
$X-\overset{\overset{O}{  }}{C}-\overset{\overset{R_3}{ }}{C}=\overset{\overset{R_2}{ }}{C}-R_1 + H_2C(COOR)_2 \xrightarrow{NaOR} X-\overset{\overset{O}{  }}{C}-\overset{\overset{R_1}{ }}{C}-\overset{\overset{R_3}{ }}{C}-CH(COOR)_2$ <p>(X = R, RO, NR<sub>2</sub>, others)</p>	Michael addition
$X-\overset{\overset{O}{  }}{C}-CH_2R_1 + R_2CHO \xrightarrow[2. H_3O^+, \text{ heat}]{1. NaOCH_3} X-\overset{\overset{O}{  }}{C}-\overset{\overset{R_1}{ }}{C}=\overset{\overset{R_2}{ }}{C}-R_1$ <p>(X = R, RO, NR<sub>2</sub>, others)</p>	Aldol condensation
$R_1COOR_2 + R_3CH_2COOR_2 \xrightarrow{NaOR_2} R_1-\overset{\overset{O}{  }}{C}-\overset{\overset{R_1}{ }}{C}-\overset{\overset{O}{  }}{C}-OR_2$	Claisen condensation
$R_1CH_2Br \xrightarrow[2. BuLi]{1. PPh_3} R_1-CH=PPh_3 \xrightarrow{3. R_2CHO} R_1CH=CHR_2$	Wittig
$2RCHO \xrightarrow[2. H_2O \text{ workup}]{1. Ti(I)} \overset{\overset{HO}{ }}{R}-\overset{\overset{OH}{ }}{CH}-R$	Pinacol coupling
	Free-radical cyclization
 <p>(Y = COOR, COR, CN, others)</p>	Diels-Alder
	Cope rearrangement

Enolates can be formed by using lithium amide bases (such as lithium diisopropylamide) or, in equilibrium with the carbonyl precursor (most often a ketone, ester, or amide), by using alkoxide bases (such as sodium methoxide). Their structures are resonance hybrids between two resonance structures, one having the negative charge of the enolate centered on the oxygen and the other having the negative charge centered on the carbon. Most of the reactions of enolates used in synthesis involve the anionic carbon atom of the enolate acting as a nucleophile. Enolates can be formed by using weaker bases when the enolate precursor bears two carbonyl groups. See RESONANCE (MOLECULAR STRUCTURE).

In the case of the aldol addition reaction, the enolate formed by deprotonation of a carbonyl precursor is allowed to react with an aldehyde; the enolate reacts like a Grignard reagent, adding to the carbonyl carbon atom of the aldehyde to form an alkoxide intermediate. Subsequent addition of water results in the protonation of the alkoxide, thus producing a beta-hydroxycarbonyl product, as shown in reaction (2).

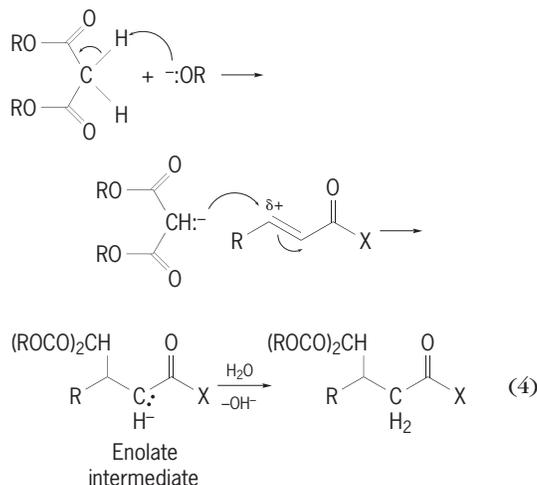


A similar addition reaction is the Reformatsky reaction (Table 2), where a zinc enolate is formed by the insertion of a zinc metal atom into the carbon-halogen bond of an alpha-halocarbonyl precursor. The zinc enolate is then allowed to react with an aldehyde, adding to it in a manner analogous to the Grignard reaction and aldol addition reaction to yield a zinc alkoxide intermediate which, upon subsequent addition of water, undergoes protonation (addition of a hydrogen) to yield a beta-hydroxycarbonyl product, as shown in reaction (3).

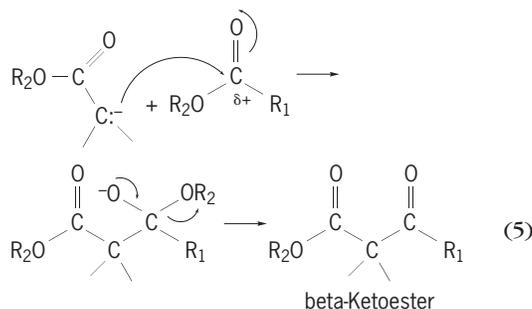


The Michael addition reaction (Table 2) is similar to the aldol addition in that it involves the formation of an enolate, which then acts as a nucleophile. In the Michael reaction, however, the enolate is formed by deprotonation of a carbon next to two carbonyl groups (usually a diester), and the reaction is typified by the addition of the enolate to a carbon-carbon double bond that is activated by being attached to a carbonyl group (an alpha, beta-unsaturated system). The double bond is polarized by the carbonyl group so that it is electrophilic enough to react

with the nucleophilic enolate, thus forming an intermediate enolate which is subsequently protonated to yield an addition product, as shown in reaction scheme (4).



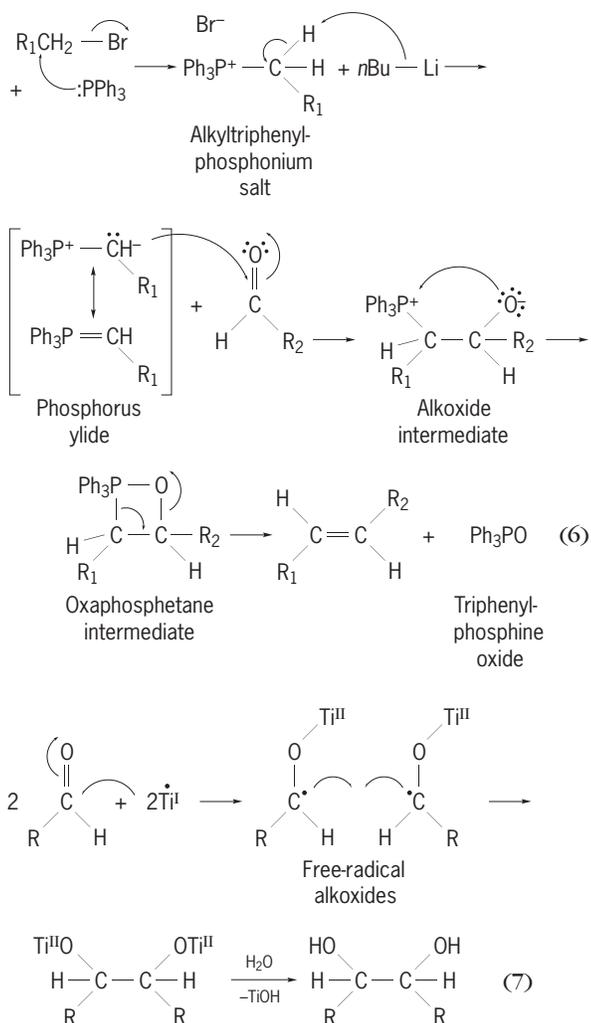
**Condensation reactions.** Numerous carbon-carbon bond-forming reactions are classified as condensation reactions, referring to the fact that the reaction combines the two reactants with a net loss of a small fragment, usually water or an alcohol by-product. Condensation reactions typically proceed in a manner similar to that of the Grignard reaction, the aldol reaction, and the Michael addition reaction, involving the initial addition of a carbanionic intermediate to a carbonyl group. The addition step is followed by the loss of the water (as hydroxide) or alcohol (as an alkoxide). For example, the aldol condensation (Table 2) proceeds initially in the same manner as the aldol addition reaction, an enolate adding to an aldehyde to form a beta-hydroxycarbonyl intermediate, but subsequent addition of an acid to the reaction mixture causes the beta-hydroxyl group to eliminate (a dehydration reaction; Table 1) to form an alkene product (specifically, an alpha, beta-unsaturated carbonyl compound). The Claisen condensation (Table 2) involves the addition of an enolate to an ester to form initially an alkoxide intermediate similar to the intermediates formed by enolates adding to an aldehyde. However, this alkoxide intermediate subsequently undergoes an eliminationlike reaction where the alkoxy group of the ester is displaced by the negative charge of the alkoxide to form a ketone carbonyl group; thus the final product of the Claisen condensation is a beta-ketoester, as shown in reaction (5).



The Wittig reaction (Table 2) also proceeds by way of a net condensation process. It starts with an alkyl halide, which is first treated with triphenylphosphine to form, by a nucleophilic substitution reaction, an alkyltriphenylphosphonium salt. This salt is then treated, in the absence of water, with a strong base such as *n*-butyllithium, which removes a proton from the carbon attached to the phosphorus atom. As in the case with the formation of enolates, the electropositive nature of the positively charged phosphorus atom enhances the acidity of the C-H bond next to it, allowing the deprotonation to occur. The resulting deprotonated intermediate, which is called a phosphorus ylide, has an anionic carbon center that can add to an electrophile. Thus, when an aldehyde is subsequently added to the reaction mixture, the nucleophilic carbon atom of the ylide adds to the carbonyl carbon to form an alkoxide intermediate. This intermediate then cyclizes to a four-membered oxaphosphetane intermediate, which then undergoes an elimination reaction to form the alkene product along with triphenylphosphine oxide as a by-product, as shown in reaction scheme (6). See YLIDE.

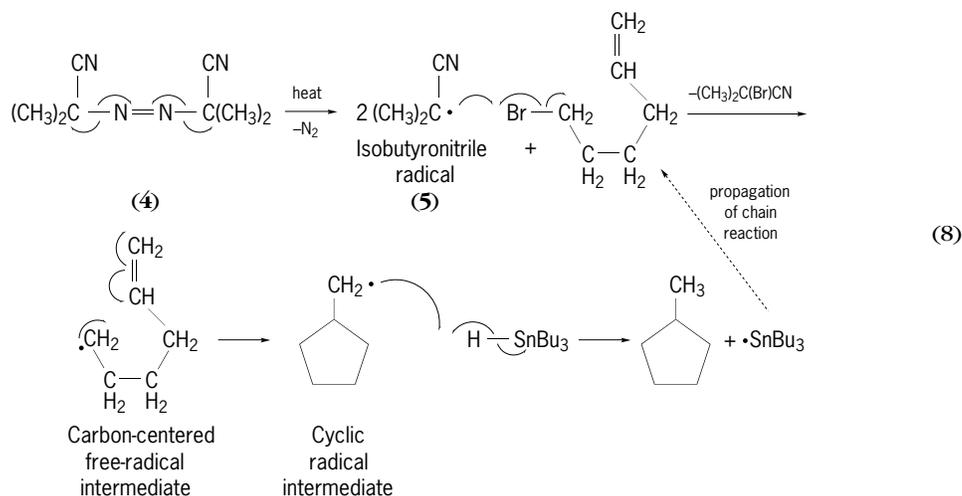
*Free-radical intermediates.* Some carbon-carbon bond-forming reactions involve free-radical intermediates, where a carbon-centered free radical is an intermediate. The free radicals involved in such reactions are molecules that bear a neutral carbon atom that has three bonds to other atoms and a single, unpaired electron. For example, in the pinacol coupling (Table 2) two aldehyde groups are coupled to each other to form, upon subsequent treatment with water as a proton source, a 1,2-diol product. This reaction requires a reduced metal reagent such as titanium(I), and this metal reagent acts to add an electron to the aldehyde group to form a carbon-centered free-radical metal alkoxide (the metal undergoes oxidation as a result of this electron transfer). The free-radical metal alkoxides then couple to each other (a typical reaction of free radicals) to form a coupled dialkoxide product. Subsequent addition of water results in protonation of the dialkoxide to yield the 1,2-diol product, as shown in reaction scheme (7), where Ph represents the phenyl group (C<sub>6</sub>H<sub>5</sub>). See FREE RADICAL.

Another common carbon-carbon bond-forming reaction that involves free radical intermediates is the addition of a carbon-centered free radical to an alkene. The form of this reaction that has found greatest utility for organic synthesis is the intramolecular addition of the radical to an alkene group in the presence of tributyltin hydride to form a reduced cyclic product, a cyclization that proceeds most efficiently when it forms a five-membered ring (Table 2). In these cyclizations, the carbon-centered free radical is usually formed by the cleavage of a carbon-bromine bond by a free-radical initiator [usually the isobutyronitrile radical (5), which is formed by heating azobisisobutyronitrile, or AIBN (4), a commonly used free-radical precursor]. The free-radical initiator abstracts the bromine from the carbon-bromine bond to form the carbon-centered

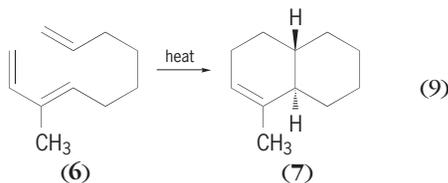


radical, which then attacks the alkene group tethered to it to form a cyclic radical intermediate. This intermediate then abstracts a hydrogen atom from the tin hydride reagent to form the reduced cyclic product and a tin-centered free-radical, which then acts as a radical initiator by abstracting (withdrawing) a bromine atom from another molecule of the starting material. Thus the reaction is a free-radical chain reaction, where free radicals (alternatively, carbon-centered and tin-centered) are continuously produced in a chain, which continues until either the bromine-containing starting material or the tin hydride reagent is consumed, as in reaction scheme (8) where the symbol · identifies a free radical. See CHAIN REACTION (CHEMISTRY).

*Other reactions.* Other carbon-carbon bond-forming reactions proceed in a concerted fashion and do not involve charged or free-radical intermediates. For example, the Diels-Alder reaction (Table 2) is a reaction between a 1,3-diene and an isolated alkene that simultaneously forms two new carbon-carbon bonds and changes the bond order (double bond to single bond, single bond to double bond) of three other carbon-carbon bonds. The Cope rearrangement (Table 2) is a reaction that simultaneously forms and breaks carbon-carbon single bonds while shifting the positions of a carbon-carbon double bond. These



concerted reactions are examples of pericyclic reactions, a class of reactions that are notable for their stereoselectivity. When a diene is tethered to an alkene group in a starting material, an intramolecular Diels-Alder reaction can proceed, thus forming several rings in a single step. For example, by heating the triene (6), the bicyclic product (7) can be formed as a single stereoisomer, as shown in reaction (9). Such intramolecular processes are very use-



ful for the synthesis of complex organic molecules. See DIELS-ALDER REACTION; WOODWARD-HOFFMANN RULE. Robert D. Walkup

**Protecting groups.** A successful chemical synthesis results in a product with a specified structure and set of properties. In designing a chemical synthesis, it is necessary to develop a strategy that will result in a product with these characteristics. In the course of a chemical synthesis, the usual practice is to use compounds with several functional groups. While one of these serves as the reactive center for one reaction, others are saved for later use in the sequence. It is imperative that the conditions required for the desired reaction do not lead to reactions at other groups. Thus, some groups must be protected by first converting them to unreactive derivatives, accomplished with the use of protecting (protective) groups.

A protecting group is a functional group that is attached to a second functional group during a synthesis to prevent interference by or reaction at that site during subsequent chemical transformations. The group should have certain specific properties: (1) It must be possible to install and remove the group in high yield so as not to compromise the efficiency of a synthesis. Every time a protective group is used, it adds two steps to a synthesis. Even if both reac-

tions proceed in 95% yield, the total yield is only 90%. (2) The group should not introduce new stereocenters so that diastereomers are not produced that would complicate analysis by nuclear magnetic resonance and purification. (3) The reagents for its introduction should be low in cost since these ultimately do not add value or carbon to a synthesis. (4) The reagents used for removal of the group should have a high degree of specificity for that group.

*Protection of alcohols, phenols, and thiols.* Among the protective groups, those for alcohols and amines provide the greatest diversity and will serve to illustrate some more general principles involved in protection and deprotection. Protection of an alcohol usually is accomplished by some modification of a reaction known as the Williamson ether synthesis; that is, the alcohol is treated with a base and a suitable alkylating agent to form an ether. Methods that rely on carbonium ion chemistry are also employed, but they are restricted to substrates that stabilize cations. See ALCOHOL.

The greatest variability in protective group chemistry occurs in the methods for deprotection, so that any group in a given molecule can be removed in the presence of the others. This concept is known as orthogonality. Simple ethers or acetals are used as protective groups for alcohols (Table 3). The simple ethers are considered the most robust with respect to most synthetic reagents. The methylene acetals (Table 3) offer a large measure of variability in deprotection methods, and consequently greater discrimination is possible when choosing conditions for carbon-carbon bond formation in a synthesis when these are used. See ACETAL; ETHER.

A third group consists of silyl ethers ( $R_1R_2R_3SiOR$ ). These have the advantage that they can be introduced and removed with great facility under very mild conditions while maintaining excellent stability to a wide array of synthetic reagents. See ORGANOSILICON COMPOUND.

Esters ( $RCO_2R'$ ) are also extensively used to protect alcohols, and they are usually introduced through the acid chloride or anhydride in the presence of a base. Cleavage is normally effected by basic

TABLE 3. Protection for alcohols

Protection group	Deprotection method
<b>Ethers</b>	
-CH <sub>3</sub>	HI or BBr <sub>3</sub>
-C(CH <sub>3</sub> ) <sub>3</sub>	Solvolysis in strong acid (CF <sub>3</sub> CO <sub>2</sub> H)
-CH <sub>2</sub> Ph	Hydrogenolysis (H <sub>2</sub> , Pd-C)
-CH <sub>2</sub> C <sub>6</sub> H <sub>4</sub> OCH <sub>3</sub>	Oxidation (dichlorodicyanoquinone, DDQ)
-CH <sub>2</sub> C <sub>6</sub> H <sub>4</sub> -2-NO <sub>2</sub>	Photolysis
-CH <sub>2</sub> CH=CH <sub>2</sub>	Isomerization-hydrolysis [Rh(Ph <sub>3</sub> P) <sub>3</sub> Cl; HgCl <sub>2</sub> , H <sub>2</sub> O]
<b>Acetals</b>	
-CH <sub>2</sub> OCH <sub>3</sub>	Strong acid
-CH <sub>2</sub> OCH <sub>2</sub> CH <sub>2</sub> OCH <sub>3</sub>	Lewis acid (ZnCl <sub>2</sub> )
-CH <sub>2</sub> OCH <sub>2</sub> CH <sub>2</sub> TMS	Fluoride ion ( <i>n</i> -Bu) <sub>4</sub> NF
-CH <sub>2</sub> OCH <sub>2</sub> CCl <sub>3</sub>	Reduction with zinc
-CH <sub>2</sub> OCH <sub>2</sub> Ph	Hydrogenolysis (H <sub>2</sub> , P-C)
-CH <sub>2</sub> OCH <sub>2</sub> SiPh(CH <sub>3</sub> ) <sub>2</sub>	Oxidation (AcOOH, KBr)
-CH <sub>2</sub> OCH <sub>2</sub> C <sub>6</sub> H <sub>4</sub> OCH <sub>3</sub>	Oxidation (DDQ)
-CH <sub>2</sub> OCH <sub>2</sub> CH <sub>2</sub> CH <sub>2</sub> CH=CH <sub>2</sub>	<i>N</i> -Bromosuccinimide
-CH <sub>2</sub> OC <sub>6</sub> H <sub>4</sub> OCH <sub>3</sub>	Oxidation (DDQ)
-CH <sub>2</sub> SCH <sub>3</sub>	HgCl <sub>2</sub> , CaCO <sub>3</sub> , H <sub>2</sub> O

hydrolysis, and the rate of hydrolysis is dependent upon both electronic and steric effects. By introducing electron-withdrawing groups on the carbon in the alpha position relative to the carbonyl, the rate of base hydrolysis is increased, and this increase can be used to advantage to obtain good cleavage selectivity between several esters. As the crowding or steric demand increases around a particular ester, the rate of hydrolysis decreases because the accessibility to the carbonyl is greatly reduced. For example, the large tertiary butyl ester is much more difficult to hydrolyze with sodium hydroxide than the much smaller methyl ester. As a result, in the case of a substance that contains both these esters only the methyl ester will be hydrolyzed with sodium hydroxide. One advantage of increased steric demand is that it greatly improves the selectivity of protection in polyfunctional substrates. See ACID ANHYDRIDE; ESTER.

For alcohol protection, some carbonates have proven to be useful since they are introduced with the ease of an esterification, but can be cleaved by some of the methods used to cleave ethers, thus avoiding the use of basic conditions. Carbonates are generally more stable to base hydrolysis than esters, and thus they can also provide a good level of selectivity in the protection of multiple hydroxyl groups.

Thiol protection is analogous to alcohol protection. The major differences are that with the increased nucleophilicity of a thiol many of the ether-type derivatives are more easily formed, but they are sometimes more difficult to cleave. This is especially true for situations that employ the noble-metal catalysts which are often poisoned with sulfur compounds. As with the methylthio methyl ether, many of the thioether-based protective groups are cleaved with mercury salts. Strong acids such as trifluoroacetic or hydrobromic acid are also frequently

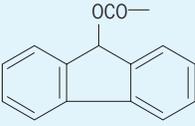
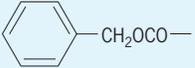
used. Protection of a thiol through disulfide formation is an option not available to alcohols. Disulfides are easily prepared by oxidation and cleaved by reduction. See CATALYSIS; ORGANOSULFUR COMPOUND.

*Protection of amines.* Amine protection has its origin in peptide synthesis, where the nucleophilic amine must be rendered temporarily nonnucleophilic in order to activate the carboxylic acid end for coupling with the amine of another amino acid. An amino acid attached to a polymer is coupled with an amino acid protected at the amine and activated at the carboxylate end to form a dipeptide. Deprotection then leads to a new amine, which can then be reacted with a different activated and protected amino acid to form a tripeptide. After the desired number of iterations of the process, the polypeptide is cleaved from the polymer; this step is also a deprotection. See PEPTIDE.

The most useful protecting groups for amines are based on carbamates [R'RNC(=O)OR'']. Numerous methods exist for their preparation, with the most common method being through reaction of an amine with a chloroformate or activated carbonate in the presence of a base. The carbamate protective groups (Table 4) exhibit a high degree of orthogonality in that most can be deprotected in the presence of the others. It should be noted that there is a strong correlation between the amine protective groups and the simple ethers used to protect alcohols. Both are often cleaved under the same conditions.

Although it is easy to convert an amine to an amide, most are not useful as protective groups because they are too difficult to hydrolyze. The exception is the trifluoroacetamide, which because of its three strongly electron withdrawing fluorine atoms is relatively easy to hydrolyze. A large number of other methods exist for amine protection, but many of these are of a very specialized nature and are used infrequently. In thinking about amine protection, it is necessary to consider the difference between normal primary and secondary amines and heterocyclic amines such as imidazole and tryptophan. Because of their increased acidity, these are often more easily deprotected. This is especially true of acyl

TABLE 4. Carbamate protective groups

Abbreviation	Structure	Deprotection method
BOC	t-BuOCO—	Strong acid
Fmoc		Base
Alloc	H <sub>2</sub> C=CHCH <sub>2</sub> OCO—	Pd (Ph <sub>3</sub> P) <sub>4</sub> , nucleophile
Cbz		Hydrogenolysis
Troc	CCl <sub>3</sub> CH <sub>2</sub> OCO—	Reduction with zinc

derivatives which can be cleaved with hydroxide under very mild conditions in contrast to normal amides which are very difficult to hydrolyze. See AMIDE; AMINE.

**Protection of carbonyls.** The most commonly used method for the protection of the carbonyl is by ketal formation. The most common derivatives are the dimethylacetals, 1,3-dioxolanes and 1,3-dioxanes. These are easily prepared by treating a ketone or aldehyde with methanol, ethylene glycol, or 1,3-propane diol respectively and an acid catalyst, while scavenging water either azeotropically with benzene or toluene or chemically with a suitable drying agent. The most frequently employed ketal is the 1,3-dioxolane, for which numerous methods exist for both its introduction and cleavage. Acid-catalyzed hydrolysis is the most commonly employed cleavage method, but many methods that avoid the use of aqueous conditions are also available. For cases where a very robust protective group is required, the carbonyl can be protected as the 1,3-dithiolane or the 1,3-dithiane. These are very resistant to acid cleavage and are usually cleaved with mercury or silver salts, but many other methods have also been developed, the best of which is probably by visible-light photolysis in the presence of the dye methylene green. See ALDEHYDE; AZEOTROPIC DISTILLATION; KETONE; PHOTOLYSIS.

**Protection of acids.** Carboxylic acids are usually protected as an ester ( $\text{RCO}_2\text{R}'$ ) which serves to tie up the acidic proton, thus avoiding unwanted acid-base reactions. Most commonly the methyl or ethyl ester is used in this capacity. These are easily hydrolyzed with hydroxide. In the event that basic conditions cannot be used to cleave the ester, several other options are available. The allyl and a variety of benzyl esters are often used. These may be cleaved by using conditions similar to the related ethers and carbamates previously discussed. To prevent nucleophilic addition to the carbonyl, sterically hindered esters (that is, esters having bulky substituents near the carbonyl group) such as the *t*-butyl ester are normally used, because they are easily prepared and conveniently cleaved with strong acid. Since hindered esters are not always effective in preventing nucleophilic additions, a second alternative exists: an orthoester can be prepared that completely blocks both the acidity and the carbonyl's susceptibility to nucleophilic attack.

**Phosphate protection.** This is an exercise in selective deprotection, since phosphoric acid is a trivalent acid. The importance of phosphates has centered on the synthesis of oligonucleotides, that is, gene synthesis. Some of the methods used to protect alcohols and acids such as the use of allyl and various benzyl esters are quite effective. Unlike carboxylic acids, simple phosphate esters are not easily cleaved with mild acid and base. They are often cleaved with trimethylsilylbromide. As the leaving group ability (that is, the ability to stabilize a negative charge) of the ester increases, nucleophilic reagents become much more effective. See OLIGONUCLEOTIDE; ORGANOPHOSPHORUS COMPOUND.

Peter G. M. Wuts

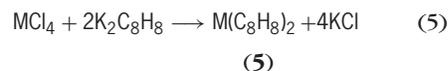
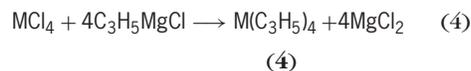
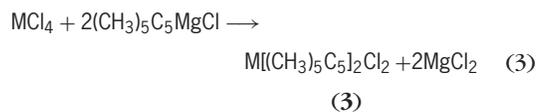
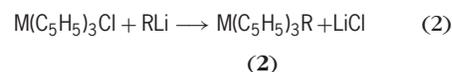
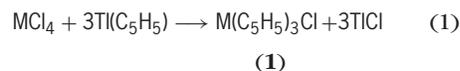
**Bibliography.** F. A. Carey and R. J. Sundberg, *Advanced Organic Chemistry, part B: Reactions and Synthesis*, 4th ed., 2000; W. Carruthers, *Some Modern Methods of Organic Synthesis*, 3d ed., 1986; E. J. Corey and X.-M. Cheng, *The Logic of Chemical Synthesis*, 1989; G. C. Crockett, The chemical synthesis of DNA, *Aldrich Chimica Acta*, 16:47-55, 1983; J. Fuhrhop and G. Penzlin, *Organic Synthesis: Concepts, Methods, Starting Materials*, 2d ed., 1993; T. W. Greene and P. G. M. Wuts, *Protective Groups in Organic Synthesis*, 3d ed., 1999; R. C. Larock, *Comprehensive Organic Transformations: A Guide to Functional Group Preparations*, 2d ed., 1999; R. O. C. Norman and J. M. Coxon, *Principles of Organic Synthesis*, 3d ed., 1993; M. B. Smith, *Organic Synthesis*, 1994; S. Warren, *Organic Synthesis: The Disconnection Approach*, 1982.

## Organoactinides

Organometallic compounds of the actinides—elements 90 and beyond in the periodic table. Both the large sizes of actinide ions and the presence of 5*f* valence orbitals are unique features which differ distinctly from most, if not all, other metal ions.

Organometallic compounds have been prepared for all actinides through curium (element 96), although most investigations have been conducted with readily available and more easily handled natural isotopes of thorium (Th) and uranium (U). Organic groups (ligands) which bind to actinide ions include both  $\pi$ - and  $\sigma$ -bonding functionalities. The importance of this type of compound reflects the ubiquity of metal-carbon two-electron sigma bonds in both synthesis and catalysis. See CATALYSIS; STRUCTURAL CHEMISTRY.

**Synthesis.** The most general preparative route to organoactinides involves the displacement of halide by anionic organic reagents in nonprotic solvents. Some examples of the preparation of cyclopentadienyl,  $\text{C}_5\text{H}_5$  (structure **1**) hydrocarbyl ( $\text{R} = \text{CH}_3$ ,  $\text{C}_6\text{H}_5$ , and so forth;), (**2**), pentamethylcyclopentadienyl,  $(\text{CH}_3)_5\text{C}_5$ , (**3**), allyl,  $\text{C}_3\text{H}_5$ , (**4**), cyclooctatetraenyl,  $\text{C}_8\text{H}_8$  (**5**), derivatives are shown in reactions (1)–(5), where  $\text{M} = \text{Th}$  or  $\text{U}$ . These compounds are



all exceedingly reactive and must be rigorously protected from air at all times.

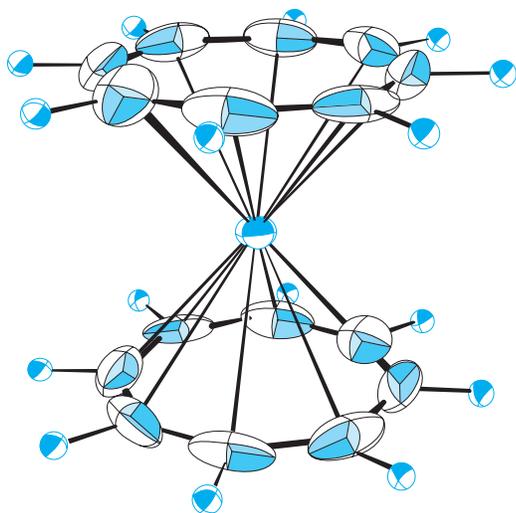


Fig. 1. Molecular structure of  $U(C_8H_8)_2$  determined by single-crystal x-ray diffraction. (After K. O. Hodgson and K. N. Raymond, *Inorg. Chem.*, 12:458, 1973)

**Structures.** The molecular structures of a number of organoactinides have been determined by single-crystal x-ray and neutron-diffraction techniques. In almost all cases the large size of the metal ion gives rise to unusually high (as compared to a transition-metal compound) coordination numbers. That is, a greater number of ligands or ligands with greater spatial requirements can be accommodated within the actinide coordination sphere. The sandwich complex bis(cyclooctatetraenyl)uranium (uranocene) is an example of this latter type (Fig. 1). In most organoactinides, metal-to-ligand atom bond distances closely approximate the sums of the individual ionic radii. This result argues that bonding forces within organoactinides have a large ionic component. Still, a number of spectral (optical, photoelectron, magnetic-resonance, Mössbauer) and magnetic studies reveal that the covalent character of the bonding cannot be ignored, and that there is considerable overlap of metal and ligand orbitals. See METALLOCENES; NEUTRON DIFFRACTION; X-RAY DIFFRACTION.

**Chemical properties.** Studies of organoactinide reactivity have concentrated on understanding the relationship between the identity of the metal, the degree to which the ligands congest the coordination sphere (coordinative saturation), and the types of chemical reactions which the complex undergoes. Employing the methodology of reactions (2) and (3), bis(pentamethylcyclopentadienyl) actinide hydrocarbyl chlorides (6) and bishydrocarbyls (7) have

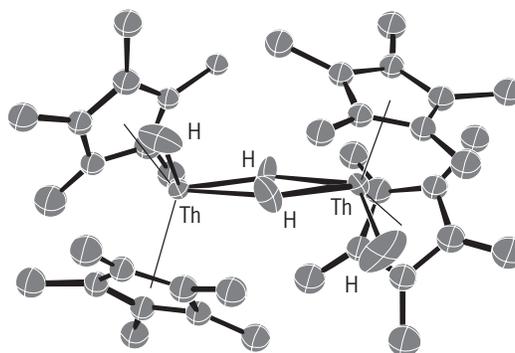
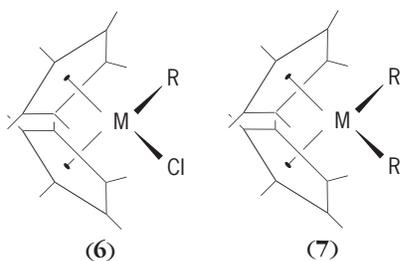
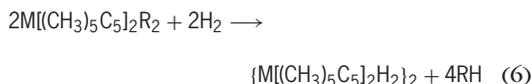


Fig. 2. Molecular structure of dimeric thorium hydride  $\{Th[(CH_3)_5C_5]_2H_2\}_2$  as determined by single-crystal neutron diffraction. (After R. W. Broach et al., *Science*, 203:172, American Association for the Advancement of Science, 1979)

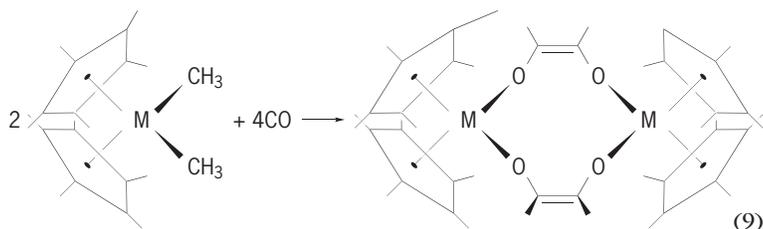
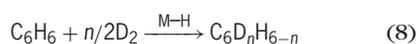
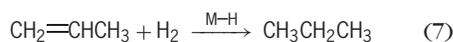
been synthesized. These classes of compounds have proved to be some of the most reactive organoactinides, and some of the most informative in terms of revealing new types of reactivity and bonding.

In an atmosphere of hydrogen, thorium-to-carbon and uranium-to-carbon sigma bonds are rapidly cleaved to yield hydrocarbons and the first known organoactinide hydrides, reaction (6). For the dimeric

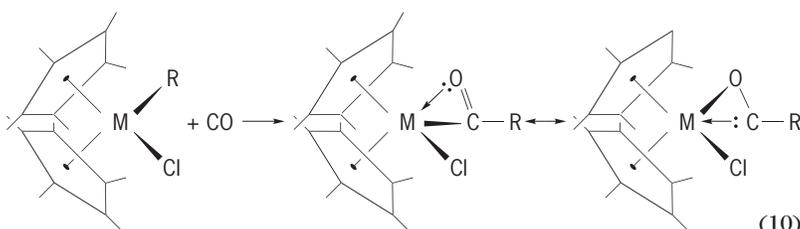


thorium hydride (Fig. 2), hydrogen atoms form both terminal (two-center, two-electron) and bridging (three-center, two-electron) bonds to thorium. In solution the organoactinide hydrides are active catalysts for olefin hydrogenation, reaction (7), and for the conversion of C—H bonds in the presence of  $D_2$  to C—D bonds [reaction (8), where  $M = Th$  or  $U$ ].

The high oxygen affinity of actinide ions and the unsaturation of the bis(pentamethylcyclopentadienyl) ligation environment give rise to several unusual new types of carbonylation reactions. The bis(pentamethylcyclopentadienyl) dimethyl derivatives (8) react rapidly with CO to produce insertion products in which coupling of four carbon monoxide molecules has occurred to produce two carbon-carbon double bonds and four actinide-oxygen bonds, reaction (9), where  $M = Th$  or  $U$ . The alkyl chlorides (9) undergo an insertion reaction to yield acyl  $[MC(O)R]$  complexes (10) in which a very strong metal-oxygen interaction takes place, reaction (10). The marked distortion of the bonding in these complexes, away from a classical acyl in which only the carbon atom is bound to the metal, is evident in structural, spectral, and chemical properties. Thus, the metal-oxygen distances are invariably shorter than the metal-carbon distances, the C-O stretching frequencies (about  $1460\text{ cm}^{-1}$ ) are anomalously low, and the chemistry is decidedly carbenelike. For example, upon heating in solution, the  $R = CH_2C(CH_3)_3$  derivative rearranges to a hydrogen atom migration product (11), and the  $R = CH_2Si(CH_3)_3$  derivative to a trimethylsilyl migration product (12). The

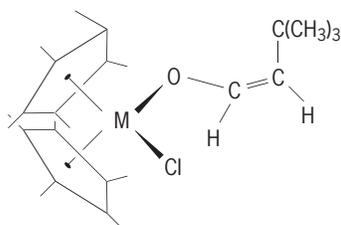


(8)

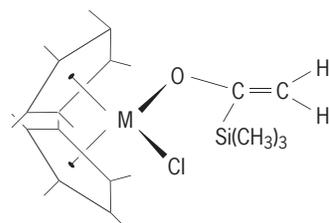


(9)

(10)



(11)



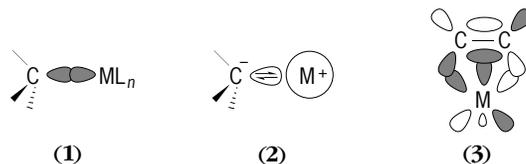
(12)

pronounced oxygen affinity and coordinative unsaturation of the actinide ions in these environments may serve as models for the active surface sites on heterogeneous CO reduction catalysts. See ACTINIDE ELEMENTS; COORDINATION CHEMISTRY; HETEROGENEOUS CATALYSIS; HOMOGENEOUS CATALYSIS; ORGANOMETALLIC COMPOUND. Tobin J. Marks Bibliography. J. Chaiken (ed.), *Laser Chemistry of Organometallics*, 1993; C. Elschenbroich and A. Salzer, *Organometallics: A Concise Introduction*, 2d ed., 1992; T. J. Marks and R. D. Fischer, *Organometallics of the f-Elements*, 1979.

### Organometallic compound

A member of a broad class of compounds whose structures contain both carbon (C) and a metal (M). Although not a required characteristic of

organometallic compounds, the nature of the formal carbon-metal bond can be of the covalent (1; L = ligand), ionic (2), or  $\pi$ -bound (3) type.

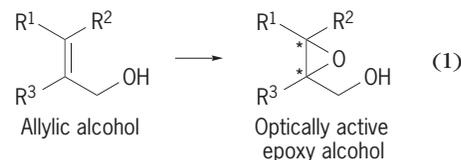


The term organometallic chemistry is essentially synonymous with organotransition-metal chemistry; it is associated with a specific portion of the periodic table ranging from groups 3 through 11, and also includes the lanthanides. See CHEMICAL BONDING; LIGAND; PERIODIC TABLE; TRANSITION ELEMENTS.

This particular area has experienced exceptional growth since the mid 1970s largely because of the continuous discovery of novel structures as elucidated mainly by x-ray crystallographic analyses; the importance of catalytic processes in the chemical industry; and the development of synthetic methods based on transition metals that focus on carbon-carbon bond constructions. From the perspective of inorganic chemistry, organometallics afford seemingly endless opportunities for structural variations due to changes in the metal coordination number, alterations in ligand-metal attachments, mixed-metal cluster formation, and so forth. From the viewpoint of organic chemistry, organometallics allow for manipulations in the functional groups that in unique ways often result in rapid and efficient elaborations of carbon frameworks for which no comparable direct pathway using nontransition organometallic compounds exists.

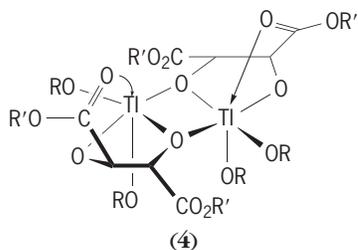
**Early transition metals.** In moving across the periodic table, the early transition metals have seen relatively limited use in synthesis, with two exceptions: titanium (Ti) and zirconium (Zr).

*Titanium.* This transition metal has an important role in a reaction known as the Sharpless asymmetric epoxidation, where an allylic alcohol is converted into a chiral, nonracemic epoxy alcohol with excellent and predictable control of the stereochemistry [reaction (1)]. The significance of the nonracemic



product is that the reaction yields a single enantiomer of high purity.

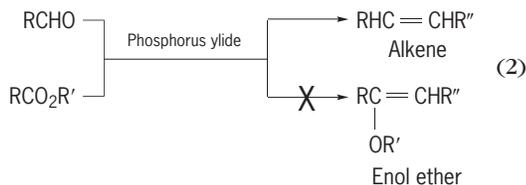
An oxophilic element has a strong affinity for oxygen. The ability of oxophilic titanium(IV) to accommodate the necessary chiral ligand [that is, a (+)- or (-)-tartrate ester], a hydroperoxide, and the educt (that is, the starting material) in an octahedral configuration is crucial to chirality transfer that occurs within the dimeric cluster (4). The ability to induce



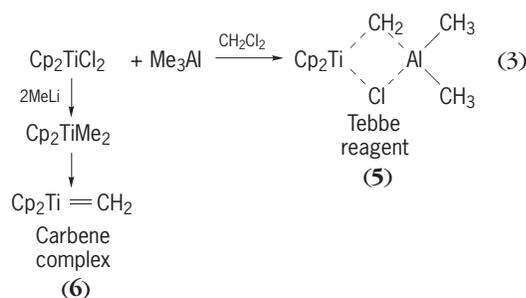
asymmetry by way of reagents, rather than via existing chirality within a substrate, is extremely powerful, especially if both antipodal forms of the source of chirality are available (as is true with tartaric acid esters). See RACEMIZATION.

There are many applications utilizing the Sharpless asymmetric synthesis. Examples of synthetic targets that have relied on this chemistry include riboflavin (vitamin B<sub>2</sub>) and a potent inhibitor of cellular signal transduction known as FK-506.

One classic reaction in organic chemistry is the Wittig olefination, where an aldehyde or ketone, upon treatment with a phosphorus ylide, is converted regioselectively (that is, favoring structural isomer) to an alkene. Such a process, however, is not applicable to a less electrophilic ester carbonyl, the product from which would be an enol ether functional group [reaction (2)]. The oxophilic nature of



titanium again has led to a convenient solution to this synthetic problem, in this case coming in the form of the Tebbe reagent (5). This organometallic compound can be readily prepared from Me<sub>3</sub>Al (Me = CH<sub>3</sub>) and Cp<sub>2</sub>TiCl<sub>2</sub> (Cp = cyclopentadienyl) in CH<sub>2</sub>Cl<sub>2</sub>, although in pyridine the starting material can be viewed as the carbene complex (6) [reaction (3)]. Either form is an effective methylene

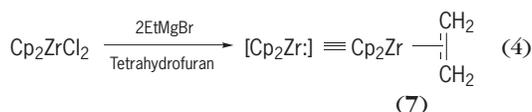


transfer reagent (that is, it donates a CH<sub>2</sub> group) to a variety of carboxylic acid derivatives (RCO-X). See ALDEHYDE; ASYMMETRIC SYNTHESIS; TITANIUM; YLIDE.

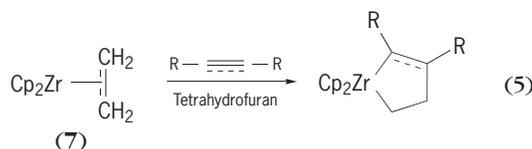
**Zirconium.** Below titanium in group 4 in the periodic table lies zirconium, which is even more oxophilic

than titanium. Most modern organozirconium chemistry concerns zirconium is ready formation of the carbenelike complex [Cp<sub>2</sub>Zr:] that because of its mode of preparation, is more accurately thought of as a π complex (7). Also important are reactions of the zirconium chloride hydride, Cp<sub>2</sub>Zr(H)Cl, commonly referred to as Schwartz's reagent, with alkenes and alkynes, and the subsequent chemistry of the intermediate zirconocenes. See METALLOGENES.

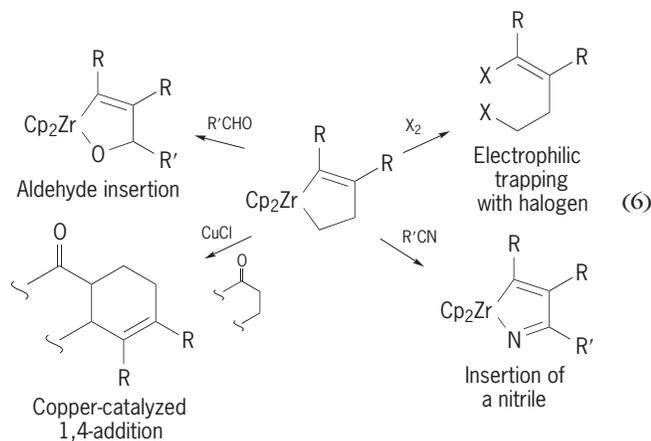
When the complex Cp<sub>2</sub>ZrCl<sub>2</sub> is exposed to two equivalents of ethyl magnesium bromide EtMgBr, the initially formed Cp<sub>2</sub>ZrEt<sub>2</sub> loses a molecule of ethane (C<sub>2</sub>H<sub>6</sub>) to produce the complexed zirconocene [Cp<sub>2</sub>Zr:] (7), as shown in reaction (4). The term



β-hydride elimination implies a process whereby a carbon-bound hydrogen with its two electrons (and thus, a hydride) migrates to a metal located on an adjacent carbon, with incipient generation of a π bond between these adjacent carbon atoms. Upon introduction of another alkene or alkyne, a zirconacyclopentane or zirconacyclopentene is formed, respectively [reaction (5)], where the structure above

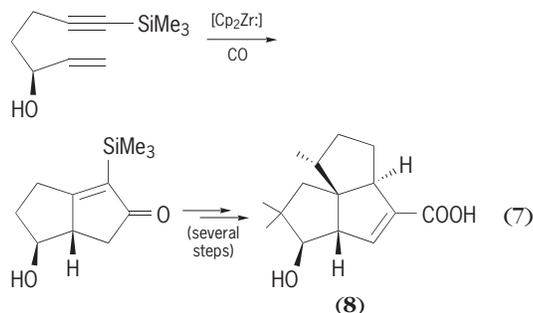


the arrow indicates that the chemistry applies to either alkenes (no third bond) or alkynes (with third bond). These are reactive species that can be converted to many useful derivatives resulting from reactions such as insertions, halogenations, and transmetalation/quenching [reaction scheme (6)]. When the

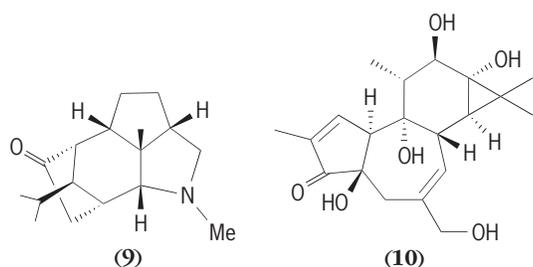


preformed complex (7) is treated with a substrate containing both an alkene and alkyne, a bicyclic zirconocene results that can ultimately yield polycyclic

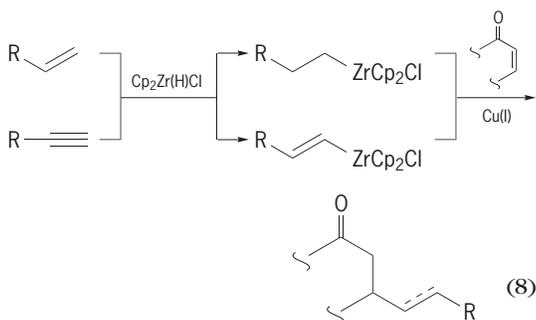
products as in reaction (7), [for example, pental-



enic acid (8), a likely intermediate in the biosynthesis of the antibiotic pentalenolactone]. Such an approach has also been used to synthesize dendrobine (9), a physiologically active component of the Chinese tonic Chin Shih Hu, and the powerful tumor-promoting agent phorbol (10). See REACTIVE INTERMEDIATES.



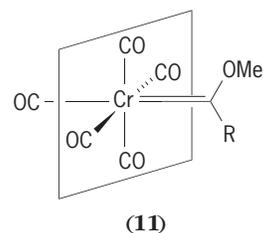
Tetrahedral  $\text{Cp}_2\text{Zr}(\text{H})\text{Cl}$  readily adds across saturated C—C bonds to give alkyl or vinylic zirconocenes, each containing a carbon-zirconium bond. Both are subject to replacement by electrophilic sources of halogen, although perhaps more valuable are their transmetalations to copper (I) and subsequent additions to the  $\beta$  position of an  $\alpha,\beta$ -unsaturated ketone [reaction (8)], or C—C bond-



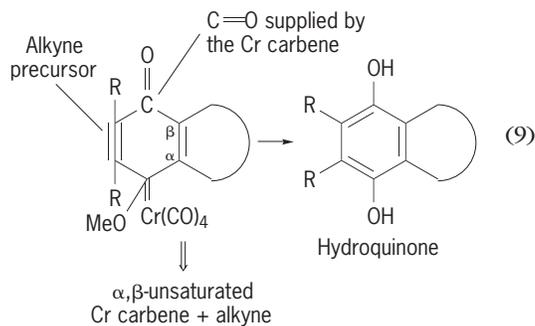
forming reactions, catalyzed by Pd(0), using either vinylic or aryl halides as coupling partners. See ZIRCONIUM.

**Mid transition metals.** Among the group 6–8 metals, chromium (Cr), molybdenum (Mo), and tungsten (W) have been extensively utilized in the synthesis of complex organic molecules in the form of their electrophilic Fischer carbene complexes, which are species having, formally, a double bond between carbon and a metal. They are normally generated as heteroatom-stabilized species bearing a “wall” of car-

bon monoxide ligands (11).



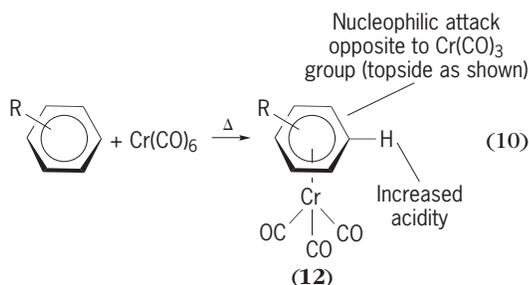
**Chromium.** Most of the synthetic chemistry has been performed with chromium derivatives, which are highly electrophilic at the carbene center because of the strongly electron-withdrawing carbonyl (CO) ligands on the metal. Many different types of reactions are characteristic of these complexes, such as  $\alpha$  alkylation, Diels-Alder cycloadditions of  $\alpha,\beta$ -unsaturated systems, cyclopropanation with electron-deficient olefins, and photochemical extrusions/cycloadditions. The most heavily studied and applied in synthesis, however, is the Dötz reaction. In this reaction, an unsaturated alkoxy-carbene is treated with an alkyne thermochemically yielding a hydroquinone following removal of the arene-bound  $\text{Cr}(\text{CO})_3$  ligand [reaction (9)]. These are



remarkable transformations in light of the level of molecular complexity that results from this one-vessel process. Of the many successful applications of this chemistry, those that relate to the aglycones (that is, non-sugar-containing portions) of the anti-tumor antibiotics in the anthracycline and aureolic acid families have been found to be ideally suited for this chemistry. See DIELS-ALDER REACTION; METAL CARBONYL.

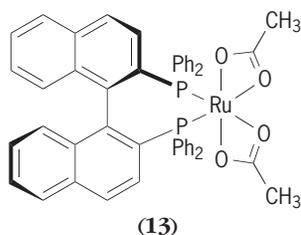
Another valuable synthetic method involves arenechromium tricarbonyl complexes (12), formed from a substituted benzene derivative and  $\text{Cr}(\text{CO})_6$  or its equivalent. The highly electron-withdrawing tricarbonyl chromium ligand inverts the tendency of the normally electron-rich aromatic ring toward attack by electrophiles, and thus, renders it susceptible to nucleophilic attack. The presence of ring hydrogens results in far greater acidity. As an important element of stereochemical control, the bulk of the  $\text{Cr}(\text{CO})_3$  residue shields one face of the planar aromatic ring, thereby directing nucleophiles

to approach from the face opposite the metal [reaction (10)].

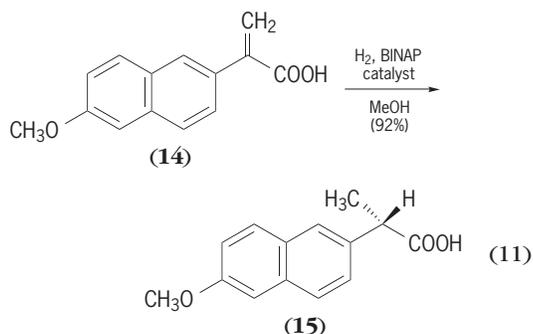


See CHROMIUM; ELECTROPHILIC AND NUCLEOPHILIC REAGENTS.

**Ruthenium.** In its 2+ oxidation state (and hence, a  $d^6$  metal), ruthenium (Ru) forms a highly effective catalyst for asymmetric hydrogenations. The combination of a ruthenium salt such as ruthenium diacetate  $[\text{Ru}(\text{OAc})_2]$  and the bidentate ligand (*R*)- or (*S*)-2,2'-bis(diphenylphosphine)-1,1'-binaphthyl (BINAP) leads to octahedral catalysts such as structure (13). In the presence of a hydrogen atmosphere and



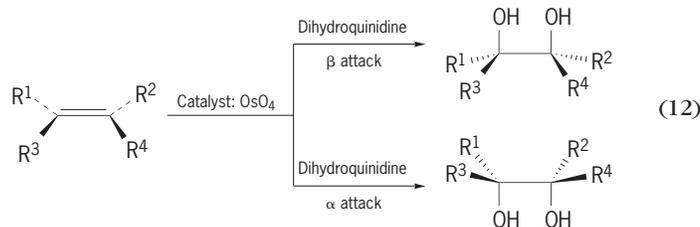
an alcoholic solvent the combination delivers hydrogen to an olefin with excellent facial selectivity. One case in point is that of an acid [structure (14)], where hydrogenation affords the anti-inflammatory agent naproxen (15) with 97% optical purity or enantiomeric excess [reaction (11)]. Tolerance to many



different functional groups exists for example,  $\alpha,\beta$ -unsaturated acids and esters, enamides, allylic alcohols, and unlike most homogeneous catalyst, which are specific for alkene reductions, the octahedral catalyst can likewise reduce a carbonyl moiety as long as a heteroatom is in the  $\alpha$ ,  $\beta$ , or  $\gamma$  location. See CATALYSIS; ELECTRON CONFIGURATION; RUTHENIUM.

**Osmium.** The  $\text{OsO}_4$ -catalyzed *cis*-dihydroxylation of an olefin is a well-established means of forming a 1,2-diol. Controlling the facial delivery of the two oxygen atoms to the (prochiral) alkene in an absolute sense, however, has been difficult to achieve. By using both

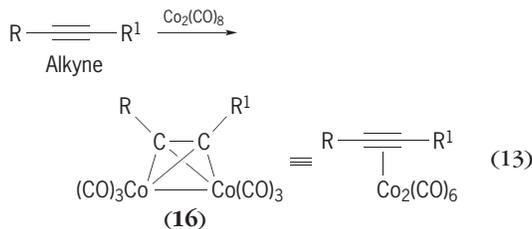
naturally occurring and synthetically modified alkaloid ligands (For example, dihydroquinidine derivatives) as the source of chirality, and potassium ferricyanide  $[\text{K}_3\text{Fe}(\text{CN})_6]$  as stoichiometric reoxidant, alkenes of most substitution patterns participate in this Sharpless asymmetric dihydroxylation, shown in reaction scheme (12).



Many tough mechanistic questions still remain unanswered regarding this remarkably useful process. Nonetheless, as with the Sharpless asymmetric epoxidation, this methodology is so powerful that it was quickly embraced by the synthetic community. See PROCHIRALITY; STOICHIOMETRY.

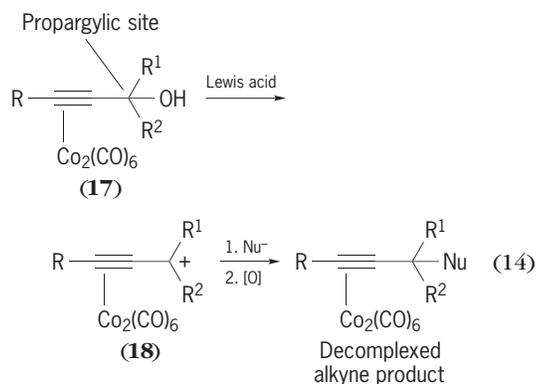
**Late transition metals.** Groups 9–11 contain transition metals that have been the most widely used not only in terms of their abilities to effect C—C bond formations but also organometallic catalysts for some of the most important industrial processes. These include cobalt (Co), rhodium (Rh), palladium (Pd), and copper (Cu).

**Cobalt.** When an alkyne is treated with dicobalt octacarbonyl in an ethereal or hydrocarbon medium, two CO ligands are lost and a stable adduct (16) is formed [reaction (13)]. The acetylene has become



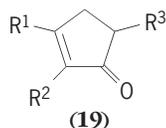
a bridging ligand by donating  $4\pi$  electrons, the hybridization at each carbon now approximating  $sp^2$ -like character.

These alkyne derivatives are, in effect, protected from many of the standard reactions of acetylenes, such as hydroboration. Therefore, they possess some unique features that impart synthetic utility. For example, such a complex imparts special stability to a carbocation located on an adjacent carbon (known as the propargylic site), which undergoes coupling with nucleophiles (Nu) at this site. Therefore, in the Nicholas reaction, a propargylic alcohol or ether can be complexed by cobalt to give structure (17), then treated with a Lewis acid to generate the corresponding cation (18), which is followed by attack of the nucleophile to yield the cobalt-complexed product. Mild oxidation of the metal (for example, with ferric salts) returns the alkyne moiety [reactions (14)]. This type of sequence has proven to be especially valuable en route to the core structure of the enediyne antitumor antibiotics, such as the esperamycins and



calicheamycins, regarded as among the most potent naturally occurring antitumor agents known. *See* ACID AND BASE; ANTIBIOTIC.

Another characteristic feature of dicobalt complexes is their participation in ring forming reactions when heated in the presence of an alkene, leading to cyclopentenones (19).



Hydroformylation (the oxo reaction) was discovered in 1938 by O. Roelen. In this type of reaction the components hydrogen (H) and the formyl group (CHO) are formally added across an alkene to arrive at an aldehyde. Indeed, the reaction for the conversion of ethylene to propionaldehyde, as applied commercially to prepare up to C<sub>15</sub> aldehydes, is the largest-scale industrial process involving a homogeneous catalyst. *See* COBALT; HOMOGENEOUS CATALYSIS.

**Rhodium.** Just below cobalt in group 9, rhodium is also an extremely effective metal for catalyzing various reactions such as hydrogenations by activation of molecular hydrogen. The species (Ph<sub>3</sub>P)<sub>3</sub>RhCl (where Ph = C<sub>6</sub>H<sub>5</sub>), a rhodium(I) derivative referred to as Wilkinson's catalyst, hydrogenates unsaturated C—C bonds, as well as other functional groups, at ambient temperatures and pressures. When L<sub>2</sub>Rh(S)Cl (L = a ligand; S = a solvent molecule) is combined with a chiral, nonracemic *bis*-phosphine, the ligands L<sub>2</sub> on Rh are replaced to form a chiral catalyst. Uptake of hydrogen and a reactant such as an alkene into the coordination sphere of the metal leads to an efficient delivery of hydrogen from one preferred face of the olefin, resulting in high levels of stereoselection; that is, a particular stereochemistry at carbon that was induced or created (in this case) during the hydrogenation of the alkene in the presence of a chiral catalyst. This approach has provided a means of manufacturing the chiral amino acid L-dopa, a clinically useful compound for treatment of Parkinson's disease. *See* PARKINSON'S DISEASE; RHODIUM.

**Palladium.** There are basically two stable oxidation levels of palladium, Pd(0) and Pd(II). It is the facile transition between these two states that accounts for the C—C bond-forming cross-couplings that have

been extensively used to great advantage in organic synthesis. Virtually all reactions that are practical involve catalytic quantities, since palladium metal as well as its numerous salts are relatively expensive. Noteworthy features of palladium chemistry are the remarkable tolerance to functional groups, the breadth of the transformations promoted, and the ready availability and ease of handling of palladium complexes.

The most common source of Pd(0) (a d<sup>10</sup> metal) is its *tetrakis*(triphenylphosphine) derivative, Pd(PPh<sub>3</sub>)<sub>4</sub>, a yellow crystalline solid with slight sensitivity to air. The species [L<sub>4</sub>Pd(0)] at this oxidation level are electron-rich and hence nucleophilic in nature. They readily undergo insertion (that is, oxidative addition) into C<sub>sp</sub><sup>2</sup>-halogen bonds, the ease of which follows the order iodine > bromine ≫ chlorine. After oxidative addition, which produces a Pd(II) intermediate, another organometallic (R-M) present can transmetallate, that is, exchange its ligand R for the halide on Pd. The initially *trans* disposition of the two organic groups on the metal can readily isomerize to *cis*, and then couple (reductively eliminate) to produce the cross-coupled organic product while regenerating the Pd(0) catalyst. There are a number of organometallic nucleophiles that are of general utility in this sequence. An impressive demonstration of the power of this methodology is found in the coupling between C-75 and C-76 present in palytoxin, the toxic principle found in marine soft corals, which is the most poisonous nonpeptidic substance known.

Another common mode of reaction of Pd(0) complexes is their displacement of allylic derivatives such as acetates and carbonates to give rise to an initially  $\sigma$ -bound Pd(II) compound. Isomerization to a  $\pi$ -allyl palladium intermediate occurs that is subject to attack by soft nucleophiles (that is, carbanionic species derived from relatively strong carbon acids), ultimately affording the product of net substitution. Use of this scenario in an intramolecular sense where the  $\pi$ -allyl palladium intermediate is prepared within the same molecule that bears the nucleophile leads to ring-forming reactions. Thus, for allylic carbonate, cyclization gives a key intermediate in the synthesis of (–)-aspochalasin B, a representative of the cytochalasins, which have extraordinary effects on transport across membranes in mammals.

Palladium(II) intermediates R-Pd-X also coordinate alkenes, leading to new  $\sigma$ -Pd-bound intermediates that ultimately undergo  $\beta$ -hydride elimination of H-Pd-X. These Pd(II) species then fall apart to H-X and Pd(0). Thousands of examples are known for this chemistry, usually called the Heck reaction, with many electron-rich or electron-poor olefinic substituents being acceptable.

Olefin insertion into PdCl<sub>2</sub> forms the basis of the Wacker process, the industrial route to acetaldehyde from ethylene. Carried out in aqueous media, the alkene is converted to an intermediate that reads to produce the aldehyde and Pd(0). The CuCl<sub>2</sub> in the reaction vessel reoxidizes Pd(0) to Pd(II), which reenters the cycle. When applied to substituted alkenes,

ketones are formed in good yields. Alteration of the reaction solvent changes the outcome, thereby leading to many key industrial building blocks. See ALKENE; PALLADIUM.

**Copper.** As a group 11 metal, copper has a valence shell corresponding to  $3d^{10}4s^1$ . Since virtually all of the organometallic chemistry of copper as it relates to organic synthesis takes place from the 1+ oxidation state [that is, Cu(I)], and hence of  $d^{10}$  configuration, there are chemists who would claim that it does not therefore qualify as a transition metal. When copper is considered from the mechanistic perspective, however, where reactions of Cu(I) must be undergoing a transition from Cu(I) to Cu(II) [a 1-electron transfer process involving carbon radical intermediates] or via energetically unfavorable Cu(III) state (implying carbanionlike intermediates), it surely belongs in this classification. From the standpoint of synthetic utility, few would argue with the notion that organocopper complexes are among the most useful transition metal reagents. See REAGENT CHEMICALS.

The popularity of organocopper reagents stems from their diversity in reagent makeup and the types of coupling reactions that they facilitate. The combination of two equivalents of an organolithium (RLi) with a copper (I) halide in (ethyloxy) or (tetrahydrofuran) forms cuprates ( $R_2CuLi$ ), commonly called Gilman reagents. These are Cu(I) monoanionic complexes that are regarded as symmetrical dimers, albeit based on a limited number of solid-state x-ray analyses. Nuclear magnetic resonance spectroscopic studies in solution, however, support this view. The hallmark of their chemistry is the ease with which they undergo 1,4 or conjugate addition reactions with  $\alpha,\beta$ -unsaturated carbonyl compounds rather than competitive 1,2-additions to the carbonyl carbon, as do more basic reagents such as the organolithiums and Grignards reagents. See NUCLEAR MAGNETIC RESONANCE (NMR).

Many variations on the  $nRLi + CuX$  theme have appeared since the initial report by H. Gilman in 1952. When  $n = 1$ , the products is  $RCu (+LiX)$ , which is a relatively unreactive, polymeric material. It can be activated with a Lewis acid to give a reagent also prone toward conjugate addition. With CuCN as the Cu(I) source, RLi + CuCN combine to generate the mixed cyanocuprate  $RCu(CN)Li$ , rather than  $RCu + LiCN$ . Because of the electron-withdrawing nitrile ligand, which is tightly bound to copper, these mixed cyanocuprates are much less reactive than are Gilman cuprates. The  $2RLi + CuCN$  combination, commonly referred to as higher-order cyanocuprates  $R_2Cu(CN)Li_2$ , is yet another variation on this theme. These latter species appear to be very reactive toward the usual sorts of electrophilic partners, but they also appear to possess greater thermal stability relative to Gilman reagents. All four reagent types, summarized in reactions (15), have virtues that are routinely called upon in numerous synthetic situations.

Another especially valuable reaction of organocuprates is their cross-couplings with alkyl,allylic,



vinyllic, and aryl halides, as well as with epoxides. These net substitutions occur via different pathways, and notwithstanding the intense usage of this chemistry, definitive mechanistic schemes remain to be elucidated. Most prevalent are displacements of primary alkyl halides, the order of leaving group ability being  $RCH_2I > RCH_2Br > RCH_2Cl$ . Other leaving groups based on oxygen are also common, with sulfonates being roughly equal to iodides in reactivity. Oxirane cleavage is particularly valuable, as the alcohol formed on workup remains in the product for further elaboration. Allylic epoxides are highly activated substrates, in this case the R group on copper being transferred to the olefinic carbon with concomitant opening of the oxirane ring. This arrangement of functionality in cyclopentene epoxide allowed for a cyanocuprate to deliver the  $\alpha$  chain ultimately needed to complete a synthesis of prostaglandins  $E_1$  and  $F_{1a}$ , which are 20-carbon local hormones derived in the organism from the arachidonic acid cascade. See EICOSANOIDS; HORMONE.

Much of organocopper chemistry relies on Grignard reagents as an alternative to organolithiums, leading to magnesio cuprates ( $R_2CuMgX$ ) based on the same stoichiometry:  $2RMgX + CuX$  going to  $R_2CuMgX + MgX_2$ . These are attractive reagents, given the availability of most Grignard reagents (from the precursor halide  $R-X + Mg$  metal). Considering that Grignard reagents readily add to carbonyl groups under similar conditions, implied in these reactions is the in place generation of a highly reactive cuprate (perhaps  $R_2CuMgX$ ) that undergoes 1,4-addition faster than the  $RMgX$  alone can add in a 1,2 fashion to the ketone.

Copper-catalyzed reactions of Grignard reagents are often preferred relative to stoichiometric cuprates as a less costly procedure. The expense factor is even more significant with regard to chemical waste disposal, which became a major issue in the 1990s. Sources of catalytic Cu(I) or Cu(II) for Grignard couplings include not only the usual assortment of salts that have been mentioned for lithio cuprate formation but others such as  $Li_2CuCl_3$  and  $Li_2CuCl_4$ , the latter otherwise known as Kochi's catalyst; this is a Cu(II) species that is rapidly reduced to Cu(I) in the presence of  $RMgX$ .

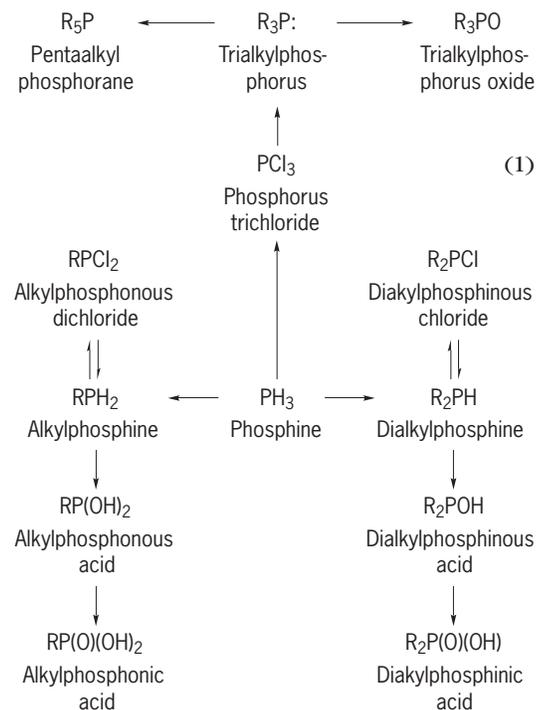
As with lithio cuprates, myriad examples exist describing implementation of this coupling approach in synthetic situations. One well-known target is the anti-inflammatory steroid cortisone. An example of a copper catalyzed alkylation of a Grignard reagent can be found in a synthesis of the side chain of vitamin E ( $\alpha$ -tocopherol), where isopentyl Grignard displacement of a primary tosylate efficiently produced the desired segment of this

coenzyme. See COORDINATION COMPLEXES; COPPER; ORGANIC CHEMISTRY; ORGANIC SYNTHESIS; STEREOCHEMISTRY. Bruce H. Lipshutz

Bibliography. E. W. Abel, F. G. Stone, and G. Wilkinson (eds.), *Comprehensive Organometallic Chemistry II: A Review of the Literature*, 1982–1994, 2d ed., 1995; J. P. Collman et al., *Principles and Applications of Organotransition Metal Chemistry*, 2d ed., 1987; A. DeMeijere and H. T. Dieck, *Organometallics in Organic Synthesis*, 1988; C. Elschenbroich and A. Salzer, *Organometallics*, 2d ed., 1992; L. S. Hegedus, *Transition Metals in the Synthesis of Complex Organic Molecules*, 2d ed., 1999; L. S. Liebeskind (ed.), *Advances in Metal-Organic Chemistry*, 1991.

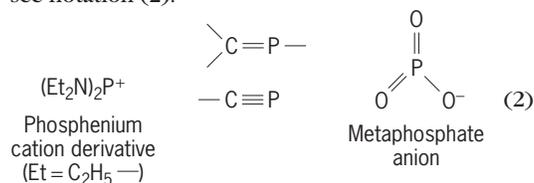
### Organophosphorus compound

One of a series of derivatives of phosphorus that have at least one organic (alkyl or aryl) group attached to the phosphorus atom linked either directly to a carbon atom or indirectly by means of another element (for example, oxygen). The mono-, di-, and trialkylphosphines (and their aryl counterparts) can be regarded formally as the parent compounds of all organophosphorus compounds; see notation (1). Formal substitution of the hydrogen



in phosphine ( $PH_3$ ) by monovalent groups or atoms leads to a number of basic structures; formal addition of bivalent oxygen leads from the mono- and dialkylphosphines to organophosphorus acids and from the trialkylphosphines to their oxides. Many of these organophosphorus molecules have been synthesized in handed (chiral) form, thus serving as stereochemical probes for the study of reaction mechanisms. Formal replacement of the nonbonding electron pair in  $R_3P:$  by two substituents gives  $R_5P$ .

Pentaphenylphosphorane is a stable example in this class. Even hexacoordinate phosphorus (as anionic species) materials are known. There has been considerable research activity devoted to organophosphorus intermediates in atypically bonded structures; see notation (2).



See ORGANIC REACTION MECHANISM; REACTIVE INTERMEDIATES.

Considering the large number of organic groups that may be joined to phosphorus as well as the incorporation of other elements in these materials, the number of combinations is practically unlimited. A vast family in itself is composed of the heterocyclic phosphorus molecules, in which phosphorus is one of a group of atoms in a ring system. See HETEROCYCLIC COMPOUNDS.

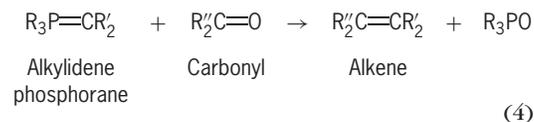
**Preparation.** Organophosphorus compounds may be prepared by a variety of methods. Trialkyl esters of phosphorus acid (trialkylphosphites) react with alkyl halides to form dialkyl alkylphosphonates, in the Michaelis-Arbuzov reaction (3). Since



the preparation involves alkyl halides and depends on the reactivity of the aliphatic halide, aromatic organophosphorus compounds can not be obtained in an analogous manner from aryl halides. The preferred method for preparing aromatic compounds is the introduction of a  $-PCl_2$  group into aromatic hydrocarbons by means of  $PCl_3$ , with anhydrous aluminum chloride as a catalyst. The resulting arylphosphorus dichlorides ( $\text{ArPCl}_2$ ) can react further with the aromatic hydrocarbon to yield the diarylphosphinous chlorides,  $\text{Ar}_2\text{PCl}$ , which are always obtained in addition to the phosphonous dichlorides, but can be made the main product of the reaction by properly choosing reaction conditions.

The reaction of phosphorus halides with organometallic compounds is useful for preparation of trialkyl- or triarylphosphines but is less applicable for making compounds with only one  $P-H$  bond. Organophosphorus derivatives with one or more  $P-H$  bonds can be added across activated alkene or acetylene bonds to form a  $P-C$  linkage. The  $P-H$  group can also react with carbonyl compounds to give  $\alpha$ -hydroxy derivatives. See ORGANOMETALLIC COMPOUND.

**Reactions.** Organophosphorus compounds frequently serve as valuable synthetic reagents. The Wittig reaction (4) is one of the most useful prepar-



ative methods known. The alkylidene phosphorane

is usually prepared from a phosphonium salt ( $R_3P^+CHR'_2Br^-$ ) and an inorganic or organic base (for example,  $C_6H_5Li$ ); subsequent treatment with ketones or aldehydes yield alkenes. The thermodynamic driving force for this and many reactions involving organophosphorus is formation of the relatively strong  $P=O$  bonds. Polyacetylenes, carotenoids (vitamin A), and steroid derivatives are more synthetically accessible by employment of the Wittig reaction. Both the structure of the phosphorane and reactive substrate have been varied over a wide range of combinations. See ORGANIC SYNTHESIS.

**Uses.** Some organophosphorus compounds have been used as polymerization catalysts, lubricant additives, flameproofing agents, plant growth regulators, surfactants, herbicides, fungicides, and insecticides. Hexamethylphosphoramide,  $[(CH_3)_2N]_3PO$ , is a remarkable polar solvent used in organic syntheses and capable of forming complexes with organic, inorganic, and organometallic compounds. Cyclophosphamide (CPA) is a phosphorus heterocycle which has been used in the treatment of cancer; it also is an antiinflammatory agent in diseases. Naturally occurring products with a  $C-P$  bond have been found (for example, 2-aminoethylphosphinic acid) in protozoa, in marine life, and in certain terrestrial animals. On the other hand, the high mammalian toxicity shown by some methylphosphonic acid derivatives (inhibitors of the enzyme cholinesterase) limits the usefulness of a number of related though much less toxic materials because of potential health hazards. The obvious importance of phosphorus in adenosine triphosphate (ATP) and deoxyribonucleic acid (DNA) in biology is also noteworthy. See ADENOSINE TRIPHOSPHATE (ATP); DEOXYRIBONUCLEIC ACID (DNA).

Organophosphorus compounds were made during World War II for use as chemical warfare agents in the form of nerve gases (Sarin, Trilon 46, Soman, and Tabun). See INSECTICIDE; PHOSPHORUS.

Sheldon E. Cremer

Bibliography. D. W. Allen and B. J. Walker, *Organophosphorus Chemistry*, vol. 24, 1993; J. I. Cadogan (ed.), *Organophosphorus Reagents in Organic Synthesis*, 1980; R. S. Edmunson (ed.), *Dictionary of Organophosphorus Compounds*, 1988; G. M. Kosolapoff and L. Maier (eds.), *Organic Phosphorus Compounds*, vols. 1-6, 1970-1975; A. D. F. Toy and E. N. Walsh, *Phosphorus Chemistry in Everyday Living*, 1987; S. Trippett (ed.), *Organophosphorus Chemistry* vols. 1-19, 1972-1988.

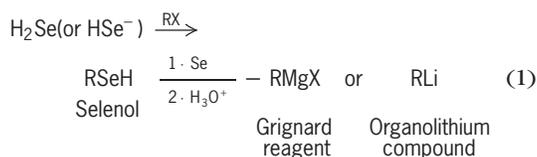
## Organoselenium compound

One of a group of compounds that contain both selenium (Se) and carbon (C) and frequently other elements as well, for example, halogen, oxygen (O), sulfur (S), or nitrogen (N). The first examples of organoselenium compounds were reported shortly after the discovery of elemental selenium by J. J. Berzelius in 1818. Their exceptionally unpleasant and persistent odors may have discouraged early in-

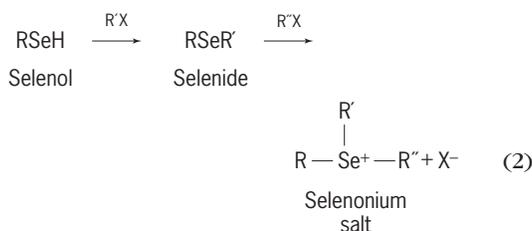
vestigators. However, less volatile derivatives containing aryl substituents or selenium in higher oxidation states are more easily handled.

Organoselenium compounds have become common in organic chemistry laboratories, where they have numerous applications, particularly in the area of organic synthesis. Organoselenium compounds formally resemble their sulfur analogs and may be classified similarly. For instance, selenols, selenides, and selenoxides are clearly related to thiols, sulfides, and sulfoxides. Despite structural similarities, however, sulfur and selenium compounds are often strikingly different with respect to their stability, properties, and ease of formation. See ORGANOSULFUR COMPOUND; SELENIUM.

**Selenols and selenides.** Selenols have the general formula  $RSeH$ , where R represents either an aryl or alkyl group. They are prepared by the alkylation of hydrogen selenide ( $H_2Se$ ) or selenide salts, as well as by the reaction of Grignard reagents or organolithium compounds with selenium, as in reactions (1). Since they are air-sensitive, they must



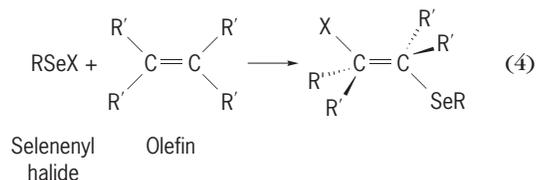
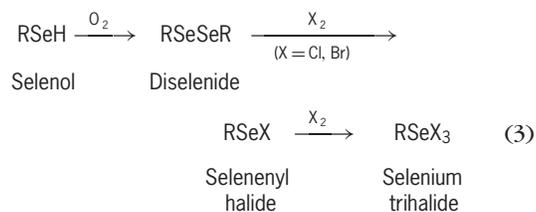
be stored under an inert atmosphere. Selenols are stronger acids than thiols. They and their conjugate bases are powerful nucleophiles that react with alkyl halides ( $R'X$ ;  $R'$  = alkyl group) or similar electrophiles to produce selenides ( $RSeR'$ ); further alkylation yields selenonium salts, as shown in reactions (2), where  $R''$  represents an alkyl group that may be



different from the alkyl group  $R'$ . Both acyclic and cyclic selenides are known, but episelenides (three-membered cyclic selenides) are unstable, losing selenium to produce olefins. Hydrogen ions (protons) can be removed from the carbon atom adjacent to the selenium atom of a selenide by treatment with a suitable base, because the selenium atom stabilizes the resulting carbanion. See ELECTROPHILIC AND NUCLEOPHILIC REAGENTS; GRIGNARD REACTION; REACTIVE INTERMEDIATES.

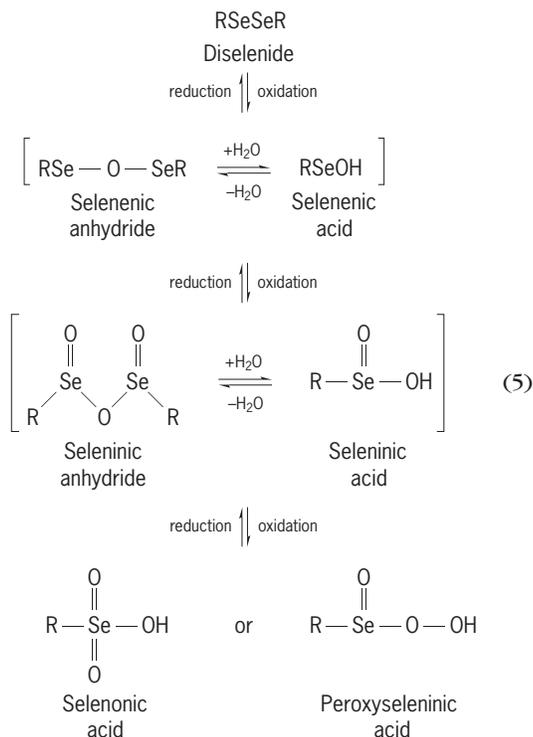
**Diselenides and selenenyl halides.** Diselenides ( $RSeR$ ) are usually produced by the aerial oxidation of selenols; they react in turn with chlorine or bromine, yielding selenenyl halides or selenium trihalides, reactions (3). Selenenyl halides add to olefins [reaction (4)] and undergo halide substitution with various nucleophiles. For example, reactions with amines ( $R'NH_2$  or  $R'_2NH$ ), cyanide ion ( $CN^-$ ), and sulfinates ( $R'SO_2^-$ ) produce selene-namides

(RSeNR'<sub>2</sub>), selenocyanates (RSeCN), and selenosul-



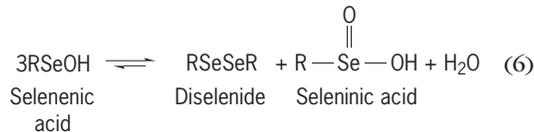
fonates, (RSeSO<sub>2</sub>R'), respectively. Triselenides and polyselenides are also known.

**Selenenic, seleninic, and selenonic acids.** Reaction scheme (5) shows the pathways in which diselenides can be oxidized to selenonic acids. Diselenides are



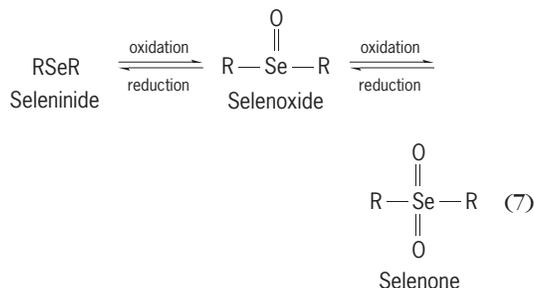
easily oxidized to seleninic acids or anhydrides by reagents such as hydrogen peroxide or nitric acid; further oxidation to selenonic acids is difficult, and the products are relatively unstable. This contrasts with the analogous sulfur compounds where sulfonic acids are stable and easily prepared. The in-place formation of peroxyseleninic acids has also been reported from seleninic acids and hydrogen peroxide. The preparation of selenenic acids (or their anhydrides) from the partial oxidation of diselenides is generally not possible as they disproportionate to diselenides and selenenic acids, as shown in reaction (6). This process is reversible, but the equilibrium favors the products on the right. When generated in the presence of olefins, selenenic acids

undergo addition reactions similar to selenenyl chlo-

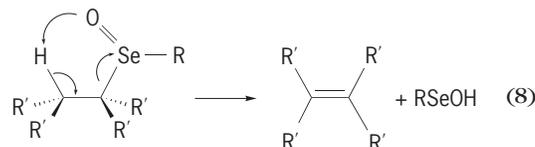


rides [such as reaction (4), except that X = OH].

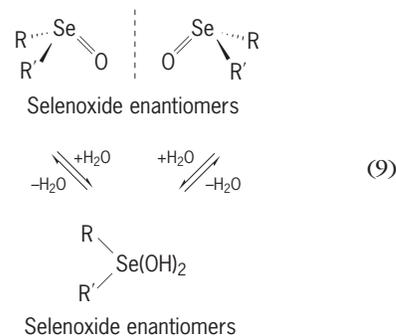
**Selenoxides and selenones.** Selenoxides are readily obtained from the oxidation of selenides with hydrogen peroxide (H<sub>2</sub>O<sub>2</sub>) or similar oxidants [reactions (7)]. Further oxidation to selenones is far more



difficult, again in contrast to the corresponding sulfur analogs (sulfones), which are easily prepared from sulfides or sulfoxides. Selenoxides undergo facile elimination to produce olefins and selenenic acids [reaction (8)]. The process is known as syn-



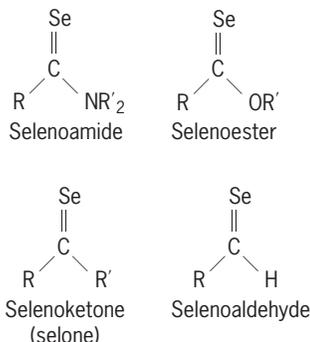
elimination; it requires the selenoxide oxygen atom and the abstracted hydrogen to lie on the same side of the carbon-carbon bond, with all five participating atoms (O, Se, C, C, H) coplanar. The selenium atom of an unsymmetrical selenoxide is a chiral center, and such compounds exist as pairs of enantiomers. They are configurationally stable under anhydrous conditions but racemize rapidly in the presence of water because of the reversible formation of achiral hydrates [reactions (9)].



See OXIDATION-REDUCTION; RACEMIZATION; STEREO-CHEMISTRY.

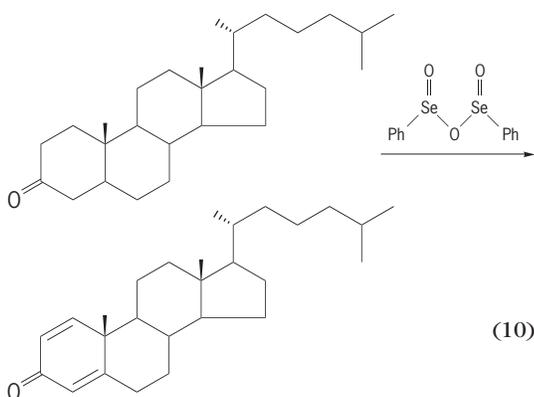
**Selenocarbonyl compounds.** Selenocarbonyl compounds tend to be considerably less stable than their thiocarbonyl or carbonyl counterparts because of the weaker double bond between the carbon and selenium atoms. Selenoamides, selenoesters and related compounds can be isolated, but selenoketones

(selones) and selenoaldehydes are highly unstable unless the selenocarbonyl group is sterically protected by two bulky substituents (for example *t*-butyl). The compounds carbon diselenide (CSe<sub>2</sub>) and carbon oxyseleone (COSe) are also known. The general structures of typical selenocarbonyl compounds are shown below.



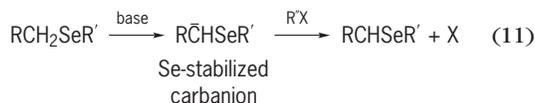
See ALDEHYDE; AMIDE; CHEMICAL BONDING; ESTER; KETONE.

**Applications in synthesis.** The most widely used procedure employing organoselenium compounds is the selenoxide syn-elimination shown in reaction (8), which is an exceptionally mild olefin-forming reaction. Since it often proceeds readily at room temperature, it obviates the need for the forcing pyrolytic condition associated with other types of syn-eliminations. Scavengers for the by-product selenenic acid are sometimes necessary to prevent its readdition to the product olefin. Seleninic acids and anhydrides, as well as the inorganic reagent selenium dioxide (SeO<sub>2</sub>), are useful oxidizing agents. For example, they can be employed in the dehydrogenation of carbonyl compounds, including steroid derivatives [reaction (10); Ph=C<sub>6</sub>H<sub>5</sub>; phenyl]. Perox-



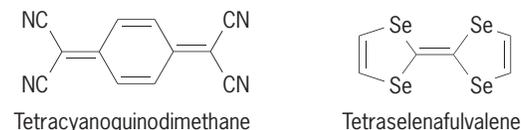
yseleninic acids are efficient reagents for the epoxidation of olefins or the conversion of ketones to esters or lactones. The additions of selenenyl halides, selenenic acids, and related derivatives to olefins [as reaction (4)] provides a useful means for introducing halide, hydroxyl (OH), and other functional groups into the products. (A functional group is an atom or group of atoms considered to have replaced a hydrogen in a hydrocarbon compound.) Selenium-stabilized carbonions derived from selenides react with alkyl halides and other electrophiles, resulting in the formation of new carbon—carbon bonds [reac-

tion (11)], a vital operation in the synthesis of organic



molecules. See ORGANIC REACTION MECHANISM; ORGANIC SYNTHESIS.

**Organic conductors.** The charge-transfer complexes formed between certain organoselenium donor molecules and appropriate acceptors such as tetracyanoquinodimethane are capable of conducting electric current. Tetraselenafulvalene, its sulfur



analog, and their various derivatives are particularly effective donors. Selenium-containing polymers are also of interest as organic conductors. See COORDINATION COMPLEXES; ORGANIC CONDUCTOR.

**Biological importance.** Organoselenium compounds are generally quite toxic and must be handled with appropriate care. However, selenium is also an essential trace element, and its complete absence in the diet is severely detrimental to human and animal health. The element is incorporated into selenoproteins such as glutathione peroxidase, which acts as a natural antioxidant. Harmful peroxides formed as by-products during normal oxidative metabolism are rapidly reduced by glutathione in a process that is catalyzed by the selenoprotein. This protects cells from oxidative damage caused by the peroxides. See BIOINORGANIC CHEMISTRY; COORDINATION CHEMISTRY; PROTEIN.

Thomas G. Back

**Bibliography.** R. F. Burk, *Selenium in Biology and Human Health*, 1993; D. Liotta (ed.), *Organoselenium Chemistry*, 1987; K. C. Nicolaou and N. A. Petasis, *Selenium in Natural Products Synthesis*, 1984; S. Patai and Z. Rappoport (eds.), *The Chemistry of Organic Selenium and Tellurium Compounds*, vols. 1–2, 1986–1987; C. Paulmier, *Selenium Reagents and Intermediates in Organic Synthesis*, 1986; R. J. Shamberger, *Biochemistry of Selenium*, 1983.

## Organosilicon compound

One of a group of compounds in which silicon (Si) is bonded to an organic functional group (R) either directly or indirectly via another atom. Formally, all organosilanes can be viewed as derivatives of silane (SiH<sub>4</sub>) by the substitution of hydrogen (H) atoms. The most common substituents are methyl (CH<sub>3</sub>; Me) and phenyl (C<sub>6</sub>H<sub>5</sub>; Ph) groups. However, tremendous diversity results with cyclic structures and the introduction of heteroatoms. Some representative examples and their nomenclature are given in the **table**. See SILICON.

Organosilicon compounds are not found in nature and must be prepared in the laboratory. The

Representative organosilicon compounds		
Structure		Nomenclature
Single bonds $\begin{array}{c}   \\ -\text{Si}-\text{X} \\   \end{array}$	X = H X = Me X = Ph X = Cl X = O — CR <sub>3</sub> X = O — SiR <sub>3</sub> X = O — H X = NH <sub>2</sub>	Silane Methylsilane Phenylsilane Chlorosilane Alkoxysilane Siloxane Silanol Silylamine
Multiple bonds $\begin{array}{c} \diagup \\ \text{Si} = \text{X} \\ \diagdown \end{array}$ $-\text{Si} \equiv \text{X}$	X = CR <sub>2</sub> X = SiR <sub>2</sub> X = O  X = CR X = SiR	Silene (or silaethylene) Disilene Silanone  Silyne (or silaacetylene) Disilyne
Polymers $(-\text{SiR}_2-)_n$ $(-\text{SiR}_2-\text{O}-)_n$		Polysilane Polysiloxane (or silicone)
Cyclic structures $\square (\text{SiR}_2-)_n$ $\square (\text{SiR}_2-\text{O}-)_n$		Cyclosilane Cyclosiloxane

ultimate starting material is sand (silicon dioxide, SiO<sub>2</sub>) or other inorganic silicates, which make up over 75% of the Earth's crust. This transformation was first accomplished in 1863 by C. Friedel and J. Crafts, who prepared the first organosilicon compound, tetraethylsilane, by the reaction of silicon tetrachloride (SiCl<sub>4</sub>) with diethylzinc [Zn(C<sub>2</sub>H<sub>5</sub>)<sub>2</sub>]. In the following years, many other derivatives were synthesized, but it was not until the useful properties of silicone polymers were identified in the 1940s that widespread interest in organosilicon chemistry appeared.

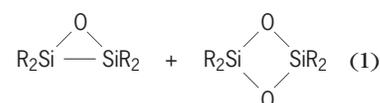
**Bonding and reactivity.** The chemistry of organosilanes can be explained in terms of the fundamental electronic structure of silicon and the polar nature of its bonds to other elements. Silicon appears in the third row of the periodic table, immediately below carbon (C) in group 14 and has many similarities with carbon. However, it is the fundamental differences between carbon and silicon that make silicon so useful in organic synthesis and of such great theoretical interest. *See* CARBON.

In the majority of organosilicon compounds, silicon follows the octet rule and is 4-coordinate. This trend can be explained by the electronic configuration of atomic silicon (3s<sup>2</sup>3p<sup>2</sup>3d<sup>0</sup>) and the formation of sp<sup>3</sup>-hybrid orbitals for bonding. Unlike carbon (2s<sup>2</sup>2p<sup>2</sup>), however, silicon is capable of expanding its octet and can form 5- and 6-coordinate species with electronegative substituents, such as the 6-coordinate octahedral dianion, [Me<sub>2</sub>SiF<sub>4</sub>]<sup>2-</sup>. Also, 5-coordinate species with trigonal bipyramidal geometries are proposed intermediates in nucleophilic displacement reactions at silicon. *See* COORDINATION CHEMISTRY; ELECTRON CONFIGURATION; VALENCE.

Silicon is a relatively electropositive element that forms polar covalent bonds (Si<sup>δ+</sup>—X<sup>δ-</sup>) with carbon

and other elements, including the halogens, nitrogen, and oxygen. The strength and reactivity of silicon bonds depend on the relative electronegativities of the two elements. For example, the strongly electronegative elements fluorine (F) and oxygen (O) form bonds that have tremendous thermodynamic stabilities, 807 and 531 kJ/mol, respectively. For the silicon-carbon bond, where the electronegativity difference is smaller, the bond dissociation energy is 318 kJ/mol and has stability comparable to the carbon-carbon bond, 334 kJ/mol. *See* CHEMICAL THERMODYNAMICS; ELECTRONEGATIVITY.

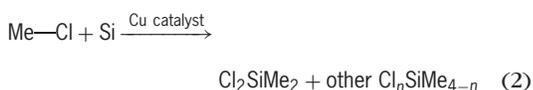
Before 1981, multiply bonded silicon compounds were identified only as transient species both in solution and in the gas phase. However, when tetramesityldisilene (Ar<sub>2</sub>Si=SiAr<sub>2</sub>, where Ar = 2,4,6-trimethylphenyl) was isolated, it was found to be a crystalline, high-melting-point solid having excellent stability in the absence of oxygen and moisture. The key factor for success was the presence of bulky substituents on silicon that impart kinetic stabilization to the double bond by preventing the approach of potential reactants. The bonding is explained in classical terms as a (3p-3p) pi double bond. Investigations into the reactivity of the silicon-silicon (Si=Si) double bond have revealed a rich chemistry, and many of the reactions of disilenes (compounds with the Si=Si group) are without precedent in ethylene (H<sub>2</sub>C=CH<sub>2</sub>) chemistry. For example, disilenes react with molecular oxygen (O<sub>2</sub>) to give unusual cyclic siloxanes [reaction (1), where R = organic group].



Other multiply bonded silicon species have also been isolated; examples are  $\text{Si}=\text{C}$  and  $\text{Si}=\text{N}$ .

Numerous reactive species have been generated and characterized in organosilicon chemistry. Silylenes are divalent silicon species ( $\text{SiR}_2$ , where  $\text{R} = \text{alkyl, aryl, or hydrogen}$ ), analogous to carbenes in carbon chemistry. The dimerization of two silylene units gives a disilene. Silicon-based anions, cations, and radicals are important intermediates in the reactions of silicon. See CHEMICAL BONDING; FREE RADICAL; REACTIVE INTERMEDIATES.

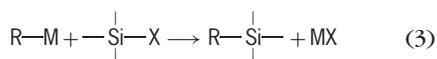
**Preparation.** The direct process for the large-scale preparation of organosilanes has provided a convenient source of raw materials for the development of the silicone industry. The process produces a mixture of chloromethylsilanes from elemental silicon and methyl chloride in the presence of a copper (Cu) catalyst [reaction (2)]. In a variation on the basic re-



action, methyl chloride can be replaced by other organic halides, including chlorobenzene which yields a mixture of chlorophenylsilanes.

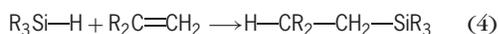
In spite of concentrated research efforts in this area, the synthetic methods available for the controlled formation of silicon-carbon bonds remain limited to only a few general reaction types. These include the reaction of organometallic reagents with silanes catalytic hydrosilylation of multiple bonds, and reductive silylation.

*Organometallic reagents with silanes.* A versatile approach to the synthesis of organosilanes is the reaction of compounds containing Si-X bonds (where X = halogen, OR, or H), with organometallic reagents (RM; M = metal), such as organolithium and organomagnesium compounds. These reactions can be viewed as nucleophilic displacements at the silicon atom [reaction (3)]. They are facilitated by the presence



of electron-withdrawing substituents on silicon. See ORGANOMETALLIC COMPOUND.

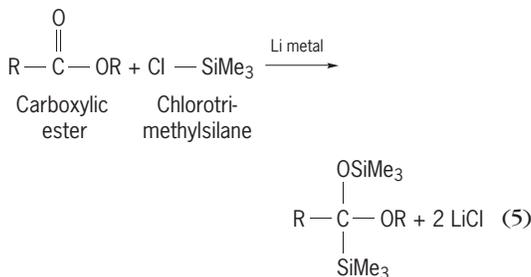
*Catalytic hydrosilylation.* This is used for the preparation of organosilanes from multiply bonded carbon compounds. The reaction effects the addition of a silicon-hydrogen bond across an alkene or an alkyne. It is often catalyzed by peroxides or transition metals, as well as by light. The mechanism is thought to involve silyl radicals ( $\text{R}_3\text{Si}\cdot$ ) as the reactive species [reaction (4)].



See ALKENE; ALKYNE; ORGANIC REACTION MECHANISM.

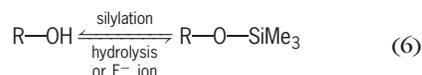
*Reductive silylation.* A less common reaction for the formation of silicon-carbon bonds is reductive silylation. An example is the silylation of a carboxylic ester with lithium (Li) metal in the presence of

chlorotrimethylsilane [reaction (5)].



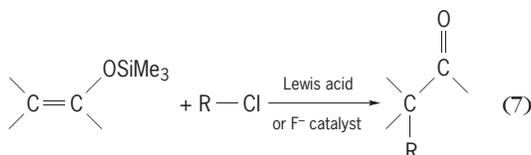
**Reagents in organic synthesis.** The role of silicon in organic synthesis is quite extensive, and chemists exploit the unique reactivity of organosilanes to accomplish a wide variety of transformations. Silicon is usually introduced into a molecule to perform a specific function and is then removed under controlled conditions.

A common use of silanes is as a protecting group to mask an active hydrogen functionality (for example, an alcohol, amine, or thiol) so that it is neither altered by, nor does it interfere with, reagents targeted at other functional groups [reaction (6)]. After the

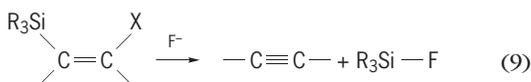
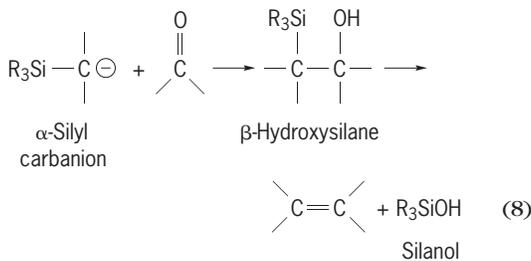


transformation is completed, the silane is removed by hydrolysis or by treatment with fluoride ion ( $\text{F}^-$ ).

In organic synthesis, the conditions chosen to effect a given chemical transformation at a targeted site often are not compatible with another functionality in molecule. As such, silyl enol ethers are powerful synthetic tools, providing an alternative to the nucleophilic alkylation of carbonyl compounds [reaction (7)]. The two approaches are complementary.



In certain synthetic reactions, the formation of energetically favorable silicon-fluorine and silicon-oxygen bonds provides a thermodynamic driving force. Two representative examples of this type are shown in reactions (8) and (9). Reaction (8) is a



valuable synthesis for the preparation of carbon-carbon double bonds from carbonyl compounds. An  $\alpha$ -silylcarbanion is reacted with an aldehyde or

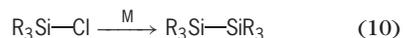
ketone to form a  $\beta$ -hydroxysilane, and subsequent elimination of a silanol gives the olefin. See ALDEHYDE; KETONE.

Reaction (9) is a fluoride ion-assisted elimination of halosilanes to form multiple bonds. The reaction, because it occurs at low temperature under mild conditions, has been used to great advantage in the synthesis of highly strained molecules, such as benzyne.

**Polysilanes.** These are organosilicon compounds that contain either cyclic arrays or linear chains of silicon atoms. The isolation and characterization of both cyclosilanes (with up to 35 silane units) and high-molecular-weight linear polysilanes (with up to 3000 silane units) have demonstrated that silicon is capable of extended chain formation (catenation). Polysilanes contain only silicon-silicon bonds in their backbone, which differentiates them from polysiloxanes (silicones), which contain alternating silicon and oxygen repeat units. See SILICONE RESINS.

The synthetic methods available for the formation of silicon-silicon bonds are extremely limited. The most general procedure for the preparation of polysilanes is via the coupling of halosilanes in the presence of an alkali metal.

The degree of functionalization of the halosilane and the reaction conditions determine the product type. For example, the coupling of monohalosilanes gives disilane products [reaction (10); M = alkali



metal]. If dichlorosilanes are used, the result is the formation of cyclosilanes, under mild conditions, or linear polymers, under more extreme conditions [reaction (11); M = alkali metal]. The introduction



of small amounts of trifunctional silanes leads to branching points, and in the extreme case where only trihalosilane is used, a highly cross-linked polymer network is formed.

The polysilanes have properties that make them attractive materials for many applications. Perhaps the most interesting is that these polymers exhibit extended electron delocalization in the sigma-bonded framework. They are also photosensitive, and the absorption of ultraviolet light results in degradation of the polymers. See DELOCALIZATION; PHOTODEGRADATION; POLYMER.

Polysilanes have found applications as photoresists in microlithography, charge carriers in electrophotography, and photoinitiators for vinyl polymerization. Polysilanes also function as preceramic polymers for the manufacture of silicon carbide fibers. See ORGANIC SYNTHESIS. Howard Yokelson

Bibliography. E. Colvin, *Silicon in Organic Synthesis*, 1981; S. Patai and Z. Rappoport (eds.), *The Chemistry of Organic Silicon Compounds*, 1989; R. West and T. J. Barton, *Organosilicon chemistry*, *J. Chem. Educ.*, 57:165-169, 334-340, 1980; G. Wilkinson, F. G. A. Stone, and E. W. Abel (eds.), *Comprehensive Organometallic Chemistry*, vol. 2, 1982.

## Organosulfur compound

A member of a class of organic compounds with any of several dozen functional groups containing sulfur (S) and often also oxygen (O), nitrogen (N), hydrogen (H), as well as other elements.

Sulfur is an element of the third row of the periodic table; it is larger and less electronegative than oxygen, which lies above it in the second row. Compounds with an expanded valence shell, that is, compounds bonding to as many as six ligands around sulfur, are therefore possible, and a broad range of compounds can be formed. Moreover, sulfur has a much greater tendency than oxygen to undergo catenation to give chains with several atoms linked together through S—S bonds. See CHEMICAL BONDING; PERIODIC TABLE; STRUCTURAL CHEMISTRY; VALENCE.

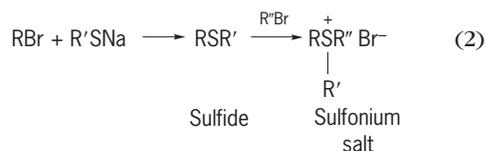
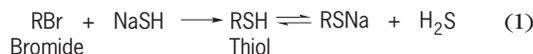
The structures and names of representative types of organosulfur compounds are shown in the **table**. Some compounds and groups are named by using the prefix thio to denote replacement of oxygen by sulfur. The prefix thia can be used to indicate that one or more —CH<sub>2</sub>— groups have been replaced by sulfur, as in 2,7-dithianonane [CH<sub>3</sub>S(CH<sub>2</sub>)<sub>4</sub>SCH<sub>2</sub>CH<sub>3</sub>].

**Thiols and sulfides.** Thiols and sulfides are sulfur counterparts of alcohols and ethers, respectively, and can be prepared by substitution reactions analogous to those used for the oxygen compounds actions [reactions (1) and (2)]. Sulfonium salts are obtained by further alkylation of sulfides. A better

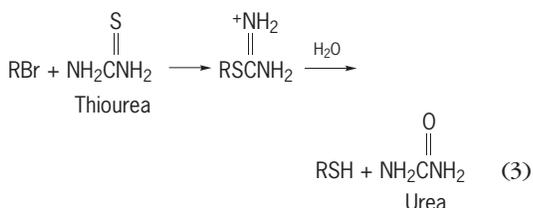
Some types of organosulfur compounds and groups

Structure	Name
RSH	Thiol (mercaptan)
RSR	Sulfide (thioether)
RSSR	Disulfide
RSSSR	Trisulfide (trisulfane)
$\overset{+}{\text{R}}\text{S}^-\text{R} \text{X}^-$	Sulfonium salt
R	
$\text{R}_2\text{C}=\text{S}$	Thio ketone
$\text{R}\overset{\text{N}}{\text{C}}=\text{S}$	Isothiocyanate
O	
RCSR	Thiolate ester (thioic acid S-ester)
S	
RCOR	Thionoate ester
RCS <sub>2</sub> R	Dithioate ester
RSOH	Sulfenic acid
RSCI	Sulfonyl chloride
RSOR	Sulfoxide
NR	
RSR	Sulfimide
RSO <sub>2</sub> H	Sulfonic acid
RSO <sub>2</sub> R	Sulfinate ester
R <sub>2</sub> S=R <sub>2</sub>	Sulfonium ylide (sulfurane)
RSO <sub>2</sub> R	Sulfone
RSO <sub>3</sub> H	Sulfonic acid
RSO <sub>2</sub> NH <sub>2</sub>	Sulfonamide
RSO <sub>2</sub> Cl	Sulfonyl chloride
ROSO <sub>3</sub> R	Sulfate ester

method for preparing thiols is formation of the isoth-



ouronium salt by reaction with thioureas, followed by hydrolysis [reaction (3)]. In these substitutions,



sulfur is a better nucleophile than oxygen. The rate of SN2 reactions with RSNa (Na = sodium) is higher than that with the corresponding oxygen nucleophile (RONa), and the reaction of an ambident nucleophile such as thiosulfate ion gives the *S*-alkyl product exclusively [reaction (4)].



See ALCOHOL; ETHER; ORGANIC REACTION MECHANISM; SUBSTITUTION REACTION.

Although thiols and alcohols are structurally analogous, there are significant differences in the properties of these two groups. Hydrogen bonding (denoted by the broken line) of the type —S—H—S— is very weak compared to —O—H—O—, and thiols are thus more volatile and have lower boiling points than the corresponding alcohols; for example, methanethiol (CH<sub>3</sub>SH) has a boiling point of 5.8°C (42.4°F) compared to 65.7°C (150.3°F) for methanol (CH<sub>3</sub>OH).

Thiols form insoluble precipitates with heavy-metal ions such as lead or mercury; the older name mercaptan, a synonym for thiol, reflects this property. Both thiols and sulfides are extremely malodorous compounds, recalling the stench of rotten eggs (hydrogen sulfide). However, traces of these sulfur compounds are an essential component of the distinctive flavors and aromas of many vegetables, coffee, and roast meat. See MAILLARD REACTION; MERCAPTAN; SPICE AND FLAVORING.

Because of the larger atomic radius and greater polarizability of sulfur, the —SH group is much more acidic than —OH (for example, ethanethiol, p*K*<sub>a</sub> 10.6; ethanol, p*K*<sub>a</sub> 15.9; the lower the value of p*K*<sub>a</sub>, the stronger the acid). For the same reasons, an acyl thiol (thiolester; RS—COR') is more reactive than an ester (RO—COR') toward attack of a nucleophile such as hydroxide ion or a carbanion, and an  $\alpha$ -hydrogen in CH<sub>3</sub>CO—SR is more acidic than that in CH<sub>3</sub>CO—OR. See PK; REACTIVE INTERMEDIATES.

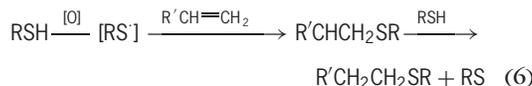
Thiols readily undergo one-electron oxidation to thiyl radicals, which dimerize to disulfides [reac-

tion (5)]. This reaction and the reverse reduction



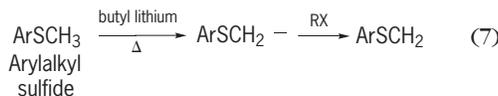
can occur under physiological conditions; this redox system has an important role in biochemistry. See OXIDATION-REDUCTION.

Thiyl radicals are intermediates in the free-radical addition of thiols to alkenes; this reaction is a useful preparative method for sulfides [reaction (6)].

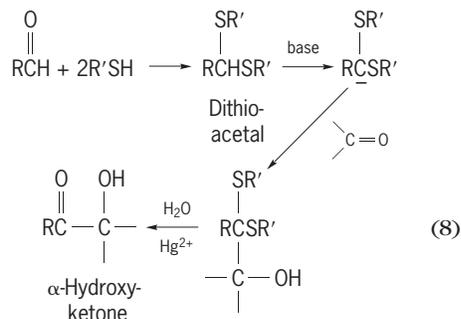


Thiyl radicals are also involved in the vulcanization of rubber, in which natural or synthetic polymers are converted to an elastomeric substance by cross-linking the chains with sulfur. This process is controlled by use of free-radical accelerators such as mercaptobenzimidazole or tetraethylthiuram disulfide. The latter compound has also been used in the treatment of alcoholism. See FREE RADICAL; RUBBER.

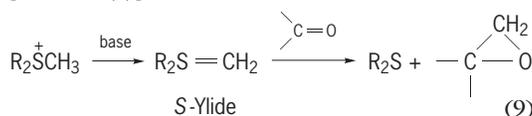
**Sulfur-stabilized carbanions and ylides.** Another important consequence of the high polarizability of sulfur is stabilization of negative charge on carbon by an adjacent sulfur atom. Aryl alkyl sulfides are converted by strong bases to the anion, which can react as a nucleophile with alkylating agents [reaction (7)]. This



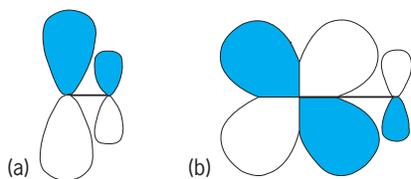
effect is particularly useful with dithioacetals, which are readily deprotonated to the anion; the latter can serve as an acyl anion equivalent, in which the polarity of the original C=O group is reversed, a process known as umpolung. For example, reaction of the dithioanion with a carbonyl compound followed by hydrolysis with a thiophilic metal ion leads to an  $\alpha$ -hydroxy carbonyl [reaction (8)].



Removal of a proton (H<sup>+</sup>) from a sulfonium salt leads to a sulfur-ylide, analogous to the phosphorus ylides obtained from a phosphonium salt. The *S*-ylide reacts with a carbonyl compound to give an epoxide [reaction (9)].

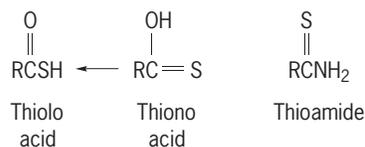


**Thiocarbonyl compounds.** These contain a carbon-sulfur double bond (C=S). Thiocarbonyl



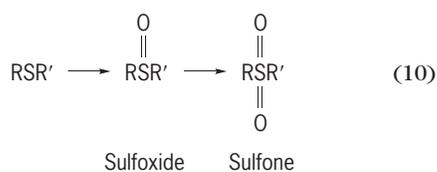
Carbon-sulfur bonding in  $\pi$ -bonds described as orbitals. (a)  $3p$ - $2p$  orbitals. (b)  $3d$ - $2p$  orbitals.

compounds (thiones) are much less common than carbonyl compounds ( $C=O$  bond). Simple thioaldehydes or thioketones have a strong tendency to form cyclic trimers, polymers, or other products. Thiocarboxylic acids have the thiol structure rather than the thione, although the latter is the stable form in thioamides, as in the structures below.



Carbon-sulfur double bonds are weak because the overlap is poor between the small carbon  $2p$  orbital and the relatively large and diffuse sulfur  $3p$  orbital (illus. a). Another type of multiple bonding in organosulfur compounds utilizes  $3d$  sulfur orbitals or hybrids thereof; it is found in compounds in which the sulfur is in a higher oxidation state than dicovalent sulfur, as in sulfoxides, sulfones, and sulfuranes. The sulfur-oxygen or sulfur-carbon double bonds in these compounds can be described as  $3d$ - $2p$   $\pi$  bonds (illus. b). See MOLECULAR ORBITAL THEORY.

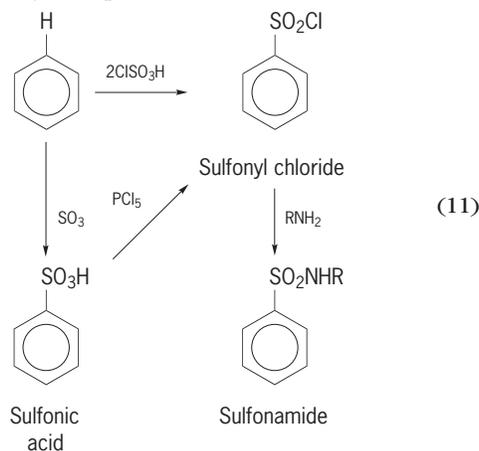
**Compounds with higher oxidation states of sulfur.** Sulfides can be oxidized sequentially to sulfoxides and sulfones, containing the sulfinyl ( $-\text{SO}-$ ) and sulfonyl ( $-\text{SO}_2-$ ) groups, respectively [reaction (10)].



Dimethyl sulfoxide (DMSO) is available in large quantities as a by-product of the Kraft sulfite paper process. It is useful as a polar solvent with a high boiling point and as a selective oxidant and reagent in organic synthesis. Sulfoxides with unlike groups attached to the sulfur atoms are chiral compounds, and thus the enantiomers can be resolved. See DIMETHYL SULFOXIDE; STEREOCHEMISTRY.

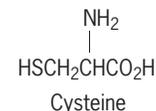
Compounds containing the sulfonyl group include sulfones, sulfonyl chlorides, sulfonic acids, and sulfonamides. Sulfones can be obtained by oxidation of sulfoxides or by sulfonation of an aromatic ring with a sulfonyl chloride, as shown in reaction scheme (11). The sulfonyl group resembles a carbonyl in the acidifying effect on an  $\alpha$ -hydrogen. The diaryl sulfone unit is the central feature of polysulfone resins, used in some high-performance plastics. Sulfonic acids are obtained by oxidation of thiols or by sulfonation. Sulfonamides, prepared from the chlorides, were the

mainstay therapeutic agents in infections until the



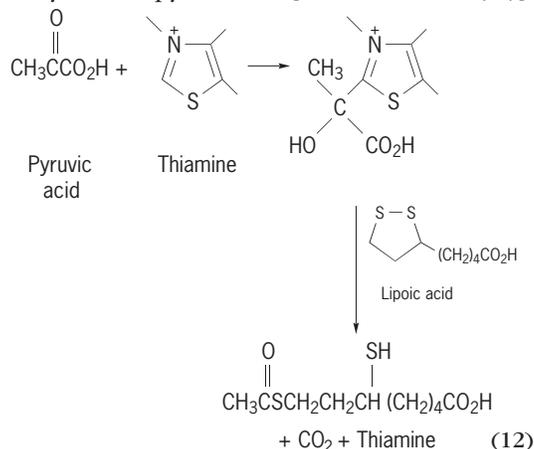
advent of antibiotics; they are still used for some conditions. See POLYSULFONE RESINS; SULFONAMIDE; SULFONIC ACID.

**Biochemical compounds.** A number of proteins and metabolic pathways in systems of living organisms depend on the amino acid cysteine and other sulfur compounds. In many proteins, for example, in the enzyme insulin, disulfide bonds formed from the  $-\text{SH}$  groups of cysteine units are an essential part of the structure. Other examples are the keratins (the tough fibrous proteins of hair, wool, and animal horns), which contain large amounts of cysteine. The  $-\text{SH}$  groups of cysteine, whose structure is shown below, also play a role in the metal-sulfur proteins



that mediate electron-transport reactions in respiration and photosynthesis. See INSULIN; PHOTOSYNTHESIS.

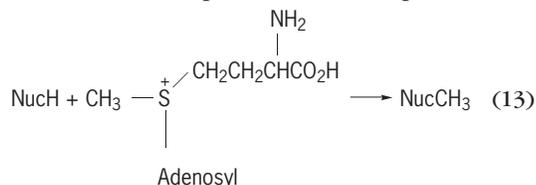
The coenzyme lipoic acid is a cyclic disulfide that functions together with the coenzyme thiamine diphosphate to accept electrons and undergo reduction of the  $-\text{S}-\text{S}-$  bond in the oxidative decarboxylation of pyruvic acid [reaction scheme (12)].



See BACTERIAL PHYSIOLOGY AND METABOLISM; COENZYME; THIAMINE.

Two other major pathways in metabolism, the transfer of acetyl groups ( $\text{CH}_3\text{CO}-$ ) and of methyl ( $\text{CH}_3-$ ) groups, are mediated by organosulfur

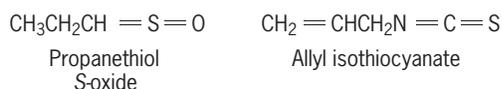
compounds. Acetyl transfer, a key step in lipid and carbohydrate metabolism, occurs by way of thioesters. An acetyl group is activated by formation of the thioester with the —SH of coenzyme A. Biological methylations take place by the formation of S-adenosylmethionine and attack of a nucleophile at the methyl group of this sulfonium ion, as in reaction (13), where Nuc represents the nucleophile.



See CARBOHYDRATE METABOLISM; LIPID METABOLISM.

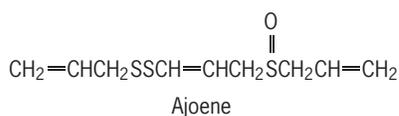
**Naturally occurring compounds.** In addition to cysteine, methionine, and several coenzymes, sulfur is present in numerous other compounds found in natural sources. Petroleum contains variable amounts of sulfur, both as simple thiols and sulfides, and also heterocyclic compounds such as benzothiophene. Removal of these is an important step in petroleum refining. See PETROLEUM; PETROLEUM PROCESSING AND REFINING.

The presence of sulfur is often signaled by characteristic odors. In onions and garlic, the odors released when the bulb is cut are due to several unsaturated organosulfur compounds. The unique flavor of garlic arises from the disulfide S-oxide allicin. The pungent, lacrimatory properties of onions are due to propanethiol S-oxide; in horseradish these properties are due mainly to allyl isothiocyanate (see below). Shiitake mushrooms owe their distinctive



aroma to the cyclic polysulfide lenthionine. A mixture of four carbon thiols is responsible for the odor of skunk secretion.

Several sulfur-containing compounds from natural sources have important pharmacological properties. Examples are the  $\beta$ -lactam antibiotics penicillin, cephalosporin, and thienamycin, and the platelet anticoagulating factor ajoene from garlic, produced by a series of complex enzymatic reactions from al-



licin. See ANTIBIOTIC; HETEROCYCLIC COMPOUNDS; ORGANIC CHEMISTRY; SULFUR. James A. Moore

Bibliography. E. Block (ed.), *Advances in Sulfur Chemistry*, vol. 1, 1988; E. Block, *Reactions of Organosulfur Compounds*, 1978; D. N. Jones (ed.), *Organosulfur Compounds*, vol. 3 of *Barton-Ollis Comprehensive Organic Chemistry*, 1979; A. Senning (ed.), *Sulfur in Organic and Inorganic Chemistry*, vols. 1-4, 1971-1982; T. W. G. Solomons, *Organic Chemistry*, 7th ed., 1999; B. Zwanenburg and A. J. H. Klunder (eds.), *Perspectives in the Organic Chemistry of Sulfur*, 1987.

## Oriental vegetables

Oriental vegetables are very important in Asian countries, but are considered as minor crops in the United States and Europe. However, there has been an increased interest in these crops because of their unusual flavors and textures and in some cases their high nutritional values (see **table**). Some of the more common ones are described below.

**Chinese cabbage.** Chinese cabbage, celery cabbage, napa, or pe-tsai (*Brassica campestris*, pekinensis group; *B. rapa*, pekinensis group; *B. pekinensis*) belongs to the mustard (Cruciferae) family, and is a biennial leafy plant but is grown as an annual. The origin of this crop is obscure; it is thought to have first developed in China, and spread to southeastern Asia and then to Korea and Japan. The harvested part is a head which is composed of broad crinkled leaves with a very wide, indistinct, white midrib. The outer leaves are pale green, and the inner leaves of the head are blanched. Chinese cabbage is usually grown in the cool season.

In the north temperate region, if the summers are relatively cool, seeds are planted in July and August; but if summers are hot, seeding is usually in late August or during September, when the maximum day temperatures have started to decrease. If planted too early during hot weather, plants may bolt (produce seed stalks) before heads can form. The production of Chinese cabbage occurs in the fall, winter, and into early spring. There are many varieties of Chinese cabbage adapted for various climates; the hybrids are very popular because they are more uniform and of better quality than the open-pollinated ones.

The seeds are planted 0.25-0.50 in. (0.6-1.2 mm) in depth on well-prepared beds about 3 ft (90 cm) apart. The seeds may be planted first in a seedling bed for transplanting to the field when plants are at the four- to six-true-leaf stage. Before transplanting to the field, the plants should be hardened by withholding water until they show signs of wilting in the afternoons. Spacing of Chinese cabbage should be 10-12 in. (25-30 cm) between plants. In Asia the thinned plants in direct-seeded fields are often bundled and sold in markets as greens. In 45-75 days from seeding, Chinese cabbage forms a cylindrical compact head. Harvested heads may be stored for several weeks in the refrigerator under high humidity without much loss of quality. It is a good salad vegetable because of the mild flavor and crisp texture.

**Pak choy.** Most varieties of pak choy, bok choy, Chinese mustard, or celery mustard (*Brassica campestris*, chinensis group; *B. rapa*, chinensis group; *B. chinensis*) are biennials, but some are annuals which flower during the first year without vernalization. The crop is grown as an annual. The origin of pak choy was probably in northern China; the plant subsequently spread to southeastern Asia and to other parts of the world in post-Columbian times. Pak choy does not form a head like most varieties of Chinese cabbage, but forms a celerylike stalk of tall, dark green leaves, with prominent white veins and long white petioles (**Fig. 1**). The base of the

Nutrient composition of some oriental vegetables in amounts per 100-g fresh edible portion<sup>a</sup>

Vegetable	Refuse as purchased, %	Approximate household equivalent <sup>b</sup>	Average food energy, cal <sup>c</sup>	Water	Protein	Fat	Total sugar	Other carbohydrates
Chinese cabbage	10	1½ cups	11	91	1.2	0.2	1.3	0.1
Pak choy	5	1½ cups	13	95	1.5	0.2	1.0	0.2
Chinese winter radish	50 <sup>d</sup>	1 cup	13	94	0.6	0.1	2.5	0.2
Edible podded peas	5	¾ cup	35	88	2.8	0.2	4.0	1.8
Yard-long bean	3	¾ cup	30	89	2.8	0.4	3.1	0.7
Bean sprouts								
Mung bean	0	1⅛ cups	25	92	2.7	0.1	2.1	1.4
Jicama <sup>e</sup>	16	1 cup	55	85	1.4	0.2		12.8 <sup>f</sup>
Chinese winter melon <sup>e</sup>	25	Slice, 1.5 in.	9	96	0.2	0.1	1.9	0.2
Balsam pear	25	¾ cup	10	94	1.1	0.2	0.8	0.4
Chinese okra	10	2 med. fruits	20	93	1.2	0.2	3.2	0.8
Water spinach	0	1½ cups	25	92	2.6	0.2	0.3	3.1

SOURCE: F. D. Howard, J. H. MacGillivray, and M. Yamaguchi, *Nutrient Composition of Fresh California-Grown Vegetables*. Calif. Agr. Exp. Sta. Bull. 788, 1962.

<sup>a</sup>For vitamins and minerals, amounts are given in milligrams, except for vitamin A, which is given in International Units; elsewhere amounts are in grams, unless otherwise specified.

<sup>b</sup>1 cup = 237 cm<sup>3</sup>. 1 in. = 25 cm.

petiole may be expanded and spoon-shaped; the blade of the leaf is smooth, and not crinkled like that of Chinese cabbage. When grown in cool regions and in moist soils, some varieties may develop enlarged turniplike roots late in the season. Varieties such as Choy Sum are grown especially for the seed stalks, with flower buds and with a few open yellow flowers. The growing of pak choy is very similar to that for Chinese cabbage; however, pak choy is more tolerant of warmer temperatures.

**Oriental winter radish.** The large, long, white radish (*Rapbanus sativus*, longipinnatus group) is often called oriental winter radish, daikon, Chinese winter radish, lobok, and lob paak. Radish is a dicotyledonous herbaceous plant grown for its long, enlarged roots. The origin of radish is thought to be

in the eastern Mediterranean region. Radish was an important crop in Egypt about 4500 years ago; it spread to China via India about 2500 years ago, and then to Japan about A.D. 700, where it ranks first in production among the vegetables produced there. Radish roots are generally long and cylindrical, and measure about 2–4 in. (5–10 cm) in diameter and up to 18 in. (45 cm) in length. The turnip-shaped Sakurajima variety can attain diameters as large as 30 cm (12 in.) and weigh 15–22 lb (7–9 kg) each. Rat-tailed radish (*R. sativus*, caudatum group) is grown for the young slender seed pods 12 in. (30 cm) in length, which are eaten raw, cooked, or pickled in India.

Winter radish seeds are planted in rows 12–18 in. (30–45 cm) apart in mid-August to mid-November in deep loose soil. Seedlings are thinned to about 8 in. (20 cm). As the hypocotyl and the tap root enlarge, the hypocotyl region pushes out of the ground 3–6 in. (8–15 cm). The roots are ready for harvest 8–10 weeks after planting, but can remain in the ground during the winter months if the temperature remains above freezing; growth continues to take place. The roots should be harvested before the plant bolts (produces a seed stalk) in the spring. Cultivars such as the Tokinashi (all seasons) can be planted in the spring for harvest in the early summer; if the temperature remains cool, it can be planted later for harvest in the summer and into the fall. Spring crops can be harvested 40–50 days after planting.

**Edible podded peas.** Edible podded peas, China peas, sugar peas, or snow peas (*Pisum sativum*, macrocarpon group) belong to the Leguminosae family. Evidence of peas used for food dates back to about 7000 B.C. in the Near East. The primitive forms of pea are slightly bitter, which prevents their being eaten by animals, and have a tough seed coat, which allows for long dormant periods. In the garden pea the pod swells first as the seeds enlarge, but in the edible podded pea the immature seeds bulge the pod, which demarks the developing seeds. Unlike the regular garden peas which have tough fibery seed pods,



Fig. 1. Pak choy or Chinese mustard (*Brassica campestris*). (University of California Agricultural Experiment Station)

	Vitamins				Minerals					
	A	Thiamine	Riboflavin	Niacin	C	Ca	Fe	Mg	P	K
1200	0.04	0.05	0.4	27	92	0.5	14	31	230	70
3000	0.04	0.07	0.5	45	105	0.8	27	37	180	100
0	0.02	0.02	0.2	22	27	0.4	22	24	190	30
580	0.15	0.08	0.6	60	43	0.9	22	53	170	6
1400	0.13	0.11	1.0	32	50	1.0	51	59	210	4
25	0.11	0.03	0.6	12	20	0.6	16	35	130	2
Trace	0.04	0.03	0.3	20	15	0.6	—	18	—	—
Trace	0.02	0.03	0.5	14	14	0.4	16	7	200	2
380	0.04	0.04	0.4	84	19	0.5	—	28	—	—
410	0.05	0.06	0.4	12	20	0.4	—	32	—	—
3500	0.03	0.10	0.9	55	95	2.2	49	40	370	6

<sup>c</sup>1 cal = 4.18 joules.      <sup>d</sup>With top as refuse.  
<sup>e</sup>Data from W. W. Leung, R. I. Pecot, and B. K. Watt, *Composition of Foods Used in Far Eastern Countries*, USDA Agr. Handb. 34, 1952.  
<sup>f</sup>Total carbohydrates.

the edible podded peas were selected for the tender pods and not for the seeds. A variety called Sugar Snap Peas, which has low-fiber-content pods and enlarged seeds, has become popular again. In Taiwan the tender shoots of garden peas are harvested and cooked as greens; the shoots have a very delicate flavor.

Peas are a cool-season crop which can be grown at mean temperatures of 55–65°F (13–18°C). Prior to flowering, pea vines can stand temperatures down to freezing, but the flowers and immature pods are injured by such temperatures. Varieties of edible podded peas have been developed for the tropics and subtropics. Peas are sensitive to acid soils; a pH range of 5.5–7.0 is considered optimal. For early production a light soil is recommended. A good, airy soil is preferred for optimal nitrogen fixation by the symbiotic *Rhizobium* bacteria found in the nodules of the roots. In cool soils some nitrogen fertilizer, together with phosphorus and potassium, is recommended. Late in the fall in mild climates or early in the spring in cold climates, seeds are planted 1–1.5 in. (2.5–4 cm) in depth, 2–3 in. (5–8 cm) apart, and in rows 3 ft (90 cm) apart. For edible podded pea varieties which are usually climbers, poles or trellises should be put up. From seeding to harvest, 60–110 days are required, depending upon the variety and season. For best quality, the pods are harvested just before the seeds begin to swell. The harvested pods can be put in a refrigerator under high humidity for several days without much loss in quality.

**Yard-long bean.** Yard-long bean or asparagus bean (*Vigna sinensis*, sesquipedalis group) belongs to the Leguminosae family. It is an annual climbing plant which can reach heights of 6–12 ft (2–4 m) and produce pods 1–3 ft (30–90 cm) long. It is a relative of the cow pea (*V. unguiculata*), its primary center of origin was in Ethiopia, and its secondary center of development was in India about 3500 years ago. During the first millennium B.C. the cultivation of yard-long beans spread into southeastern Asia and China.

Yard-long beans grow best in loam soil, but can be grown in heavy clay soils as well. It is a warm-season crop, and cannot stand any freezing temperature. Seeds are planted 1–1.5 in. (2.5–4 cm) in depth in fertile, well-drained soil after all danger of frost has passed, 6–8 in. (15–20 cm) apart in rows of two or three seeds per hill 12–24 in. (30–60 cm) apart; rows or hills should be 3–3.5 ft (90–105 cm) apart to allow for poling or trellising. In some regions of the tropics, yard-long beans are often planted with corn. The beans are planted after the corn is well established and tasseling. The growing bean plant can use the corn stalk for support after the ears have been harvested. Pods can be harvested about 50 days from seeding. They are picked before the pods reach full size and are not fibery or tough, and while the seeds are immature and not starchy. There are two varieties of yard-long beans; the green-podded type, with pods 1.5–3 ft (45–90 cm) long, and the white-podded (pale-green) type, with pods 1–1.5 ft (30–45 cm) long.

**Bean sprouts.** Mung bean or green gram (*Phaseolus aureus*) sprouts have been used by the Chinese for their remarkable healing qualities for thousands of years, and were described in their earliest medical treatises. Only recently has the Western world recognized the value of sprouted leguminous seeds (mung, soy, and alfalfa sprouts). The unsprouted mung bean seeds contain negligible amounts of vitamin C (ascorbic acid), whereas the sprouted seeds contain 20 mg per 100 g of sprouts, which is as high as tomato juice.

**Jicama or yam bean.** Jicama (*Pachyrrhizus erosus*), also called the yam bean, is indigenous to Mexico and Central America. It belongs to the pea (Leguminosae) family. The name jicama was derived from the Aztec word *xicama* by the Spaniards who first found the crop growing in Mexico. The crop is grown for its enlarged turnip-shaped root, which is eaten raw or cooked, has a crisp texture, and is sweetish in taste. It is an important ingredient in Chinese cookery. Jicama is an herbaceous vine growing to lengths of more than 10 ft (3 m) and under short-day conditions

produces small white to deep violet flowers shaped typically like those of the common bean or pea. It is a tropical plant, and requires a long, warm, frost-free season for the formation of the enlarged root. In the tropics it is grown from sea level to about 5000 ft (1600 m) in elevation.

The seeds are planted 1-2 in. (2.5-5 cm) deep in rows 3-4 ft (1-1.3 m) apart and in rich, sandy or loose soil. The young plants, when about 4 in. (10 cm) in height, are thinned to about 1 ft (30 cm) between plants. The following fertilizers are recommended: nitrogen (35 lb/acre, about 39 kg/ha), phosphorus (30 lb/acre, 34 kg/ha), and potassium (10 lb/acre, 11 kg/ha). Additional nitrogen (about 30 lb/acre, 34 kg/ha) is supplied about the time the plants begin to bloom. For best-quality roots and high yields, the flower buds are pruned off so that seed pods are not allowed to develop. The roots are harvested when they are about 4 in. (10 cm) in diameter; larger roots, and plants left in the ground too long, tend to become fibery and of poor quality. The roots can be stored in a refrigerator or cool place under relatively high humidity for several weeks. The plants, including the enlarged roots, contain rotenone, a natural insecticide, so that no pesticide is needed to grow the crop.

**Chinese winter melon.** Chinese winter melon, wax gourd, winter gourd, white gourd, ash gourd, Chinese preserving melon, ash pumpkin, or tung kwa (*Benincasa hispida*; *B. cerifera*) is a monoecious (having male and female flowers separate on the same plant) viny annual cucurbit, probably of Indian or Chinese origin. It has been cultivated for more than 2000 years, and is reported to grow wild in Java. The immature fruits of this species have bristlelike hairs on the surface. The mature fruits are watermelon-shaped (Fig. 2); some varieties are somewhat spherical, and are 15-18 in. (38-46 cm) in length, 12-15 in. (30-38 cm) in diameter, and 15-30 lb (7-14 kg) in weight. With increasing maturity, most varieties develop a white waxy bloom on the fruit surface which thickens with time, even after harvest. The flesh of mature fruits is used in making Chinese soups. It can be eaten raw or made into sweet preserves similar to citron or watermelon rind preserves. For special occasions the skin is scraped off the fruit, the pulp and seeds are removed, the



Fig. 2. Mature Chinese winter melon (*Benincasa hispida*). (University of California Agricultural Experiment Station)

fruit cavity is stuffed with previously cooked meat and vegetables, and the entire fruit is steamed for a couple of hours before serving. The fruit is said to have medicinal qualities. A variety of this species is grown especially for the immature fruits, which are 4-6 in. (10-15 cm) long and 2-3 in. (5-7.5 cm) in diameter. These, called mo kwa by the Chinese, are used like summer squash.

The crop is planted like winter squash or melons, and the vines are allowed to creep on the ground. The seeds are planted in hills about 6-10 ft (2-3 m) apart and in rich, loamy soil 1-2 in. (2.5-5 cm) deep. Fertilizer containing nitrogen, phosphorus, and potassium is applied at planting. Additional nitrogen is applied at first bloom. In southeastern Asia the vines and fruit are sometimes supported on trellises. Chinese winter melon fruits, when mature, store for more than 6 months, and often a year when stored in a cool, dry place such as a cellar. The temperature during storage should not be below 45°F (7.5°C) or above 65°F (18°C). The ideal storage temperature is 50-60°F (10-15°C). If kept at warmer temperatures, the storage time is reduced.

**Balsam pear.** Balsam pear, alligator pear, bitter gourd, bitter melon, bitter cucumber, or fu kwa (*Momordica charantia*) is a monoecious annual herbaceous vine of southeastern Asia origin; it is now widespread through the world tropics. The fruits are heart-shaped to cylindrical, tapered at the blossom end, and characterized by longitudinal rounded ridges; the entire surface of the fruit is rugose. The mature fruit size varies from 5 to 10 in. (12 to 25 cm) in length and from 2 to 3 in. (5 to 7.5 cm) in diameter. The fruit is extremely bitter; some of the bitterness is removed before cooking by peeling and steeping in salt water. Immature fruits are less bitter than mature ones.

The seeds of balsam pear are sown in rich loam soil about 1-2 in. (2.5-5 cm) deep in hills or beds 3-4 ft (90-120 cm) apart. When the plants are 3-4 in. (7.5-10 cm) in height, they are thinned to 15-18 in. (18-45 cm) between plants and staked or trellised. A complete fertilizer containing nitrogen, phosphorus, and potassium is applied at the planting of seeds, and additional nitrogen is side-dressed shortly after the vines have commenced to climb. As a vegetable, the fruits are harvested just prior to attainment of full size and while the seeds are still immature. The harvested fruits can be stored in a refrigerator for about a week. As with immature cucurbit fruits, prolonged storage in the refrigerator should be avoided, since physiological breakdown, called chilling injury, occurs at temperatures below 50°F (10°C), with subsequent rotting by decay organisms. The mature fruits, with seeds removed, are parboiled in salt water, and some of the bitterness is removed by draining and pressing or squeezing of the pulp, which is then added as an ingredient to other foods being cooked. The red aril of mature fruits is sometimes used as a condiment, and the red pigment is used as food coloring. The tender shoots and immature leaves are used as greens.

**Chinese okra.** Chinese okra or angled loofa (*Luffa acutangula*) is called zit kwa by the Chinese; it is a



Fig. 3. Chinese okra (*Luffa acutangula*). (University of California Agricultural Experiment Station)

close relative of *L. cylindrica*, known as loffa, sponge gourd, or dishcloth gourd. Chinese okra is a monoecious annual climbing vine grown for the immature fruits, which are harvested when the fibers have not yet developed and when the fruits are 4–6 in. (10–15 cm) in length. The origin of *Luffa* is thought to be in India, since wild forms of this genus have been found there. The club-shaped fruits of Chinese okra are oblong and pointed at the blossom end (Fig. 3); when mature they are 12–15 in. (30–38 cm) long and 3–3.5 in. (7.5–9 cm) in diameter at the thickest part. The 10 prominent sharp longitudinal ridges that extend from the peduncle to the blossom end characterize the fruit.

The seeds, which are poisonous, are planted in rich loam soil about 1–2 in. (2.5–5 cm) in depth in hills or in rows 6–6.5 ft (2–2.5 m) apart. The plants are spaced at 2–2.5 ft (60–75 cm) in the row. A complete inorganic fertilizer or well-decomposed manure is recommended. The crop is trellised or staked to allow the vines to climb for better quality and increased yields. Though not as large a fruit as *L. cylindrica*, the thoroughly mature fruits of Chinese okra can be made into vegetable sponge by retting until the outer walls are rotted, and then the skin, seeds, and pulp are washed away from the network of fibers. The fibers can be whitened with a dilute solution of bleach, followed by rinsing and drying.

**Water spinach.** Water spinach, water convulvulus, swamp cabbage, kang kong, or shui ung tsoi (*Ipomoea aquatica*) is a perennial semiaquatic tropical plant grown for its long, tender shoots. It is a relative of the sweet potato (*I. batatas*), but does not produce an enlarged root. Possibly of East Indian origin, it is now widespread throughout the tropics and semitropics, where it has become a very popular green vegetable. There are two types of water spinach: a narrow, pointed leaf type adapted for dryland culture and a broadleaf type adapted for aquatic culture. In the dryland culture the crop is seeded in beds and grown with abundant water. Irrigation is practiced in regions of low rainfall. It is very fast-growing, requiring only about 2 months under ideal conditions from seeding to first harvest. For aquatic culture the crop is propagated by stem cuttings 12–15 in. (30–40 cm) long. The cuttings are planted in puddled soil, similar to rice paddies, to a depth of 6–8 in. (15–20 cm), and the spacing between plants is about 12 in. (30 cm) apart in blocks. For rapid growth

and high yields, 40–50 lb/acre (45–56 kg/ha) of ammoniacal nitrogen are recommended after the plants are established.

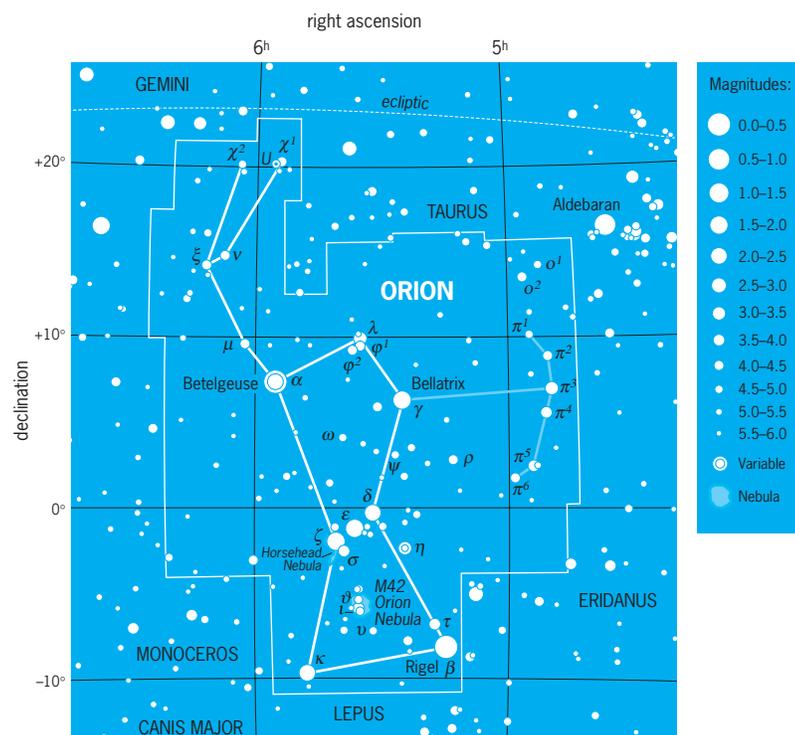
The first harvest for crops grown in water can be made as soon as the new shoots are more than 15 in. (40 cm) long. About 10–12 in. (25–30 cm) of the shoot is cut, and 8–10 shoots are bundled for marketing. New shoots continue to develop, so that harvest can be made every 7–10 days. For crops grown on dry soil, shorter shoots of 8 in. (20 cm) are harvested, since the stem sections further from the growing point are not as tender. Water spinach is used like common spinach in oriental cooking. It can be stored in a refrigerator under high humidity for a few days without much loss of quality. In southeastern Asia, parts not used for food are harvested and fed to livestock. Tender shoots from the sweet potato vine are often used instead of water spinach for human consumption.

M. Yamaguchi

**Bibliography.** M. Yamaguchi, Production of oriental vegetables in the United States, *Hort. Sci.*, 85:362–370, October 1973; M. Yamaguchi, *Vegetable Crops of the World*, 1978.

## Orion

The Hunter, a prominent constellation in the evening winter sky (see illustration). Orion is perhaps the easiest constellation to identify in the sky, because of the three bright stars in a line that form the belt of the mythical figure. The cool, red supergiant star



Modern boundaries of the constellation Orion, the Hunter. The celestial equator is 0° of declination, which corresponds to celestial latitude. Right ascension corresponds to celestial longitude, with each hour of right ascension representing 15° of arc. Apparent brightness of stars is shown with dot sizes to illustrate the magnitude scale, where the brightest stars in the sky are 0th magnitude or brighter and the faintest stars that can be seen with the unaided eye at a dark site are 6th magnitude. (Wil Tirion)

Betelgeuse glows brightly as Orion's shoulder, above which is an upraised club, and the hot, blue star Rigel marks Orion's heel. From the leftmost star of the belt dangles several faint stars and the Horsehead Nebula. The faint stars and the Orion Nebula, M42 (the 42nd object in Messier's eighteenth-century catalogue), one of the most prominent and beautiful emission nebulae and a nursery for star formation, make up Orion's sword. Four bright stars close together (within about 1 light-year) make the Trapezium, which provides the energy to make the Orion Nebula glow. *See* BETELGEUSE; ORION NEBULA; RIGEL.

Orion is pictured holding up his shield to ward off the charging Taurus, the bull. In Greek mythology, Orion met Artemis, goddess of the hunt and of the Moon. To protect her virginity, her brother Apollo sent Scorpius, the Scorpion, to attack Orion. He then tricked Artemis into shooting Orion. When Orion could not be revived, Artemis placed him in the heavens, with the scorpion eternally pursuing him. *See* SCORPIUS; TAURUS.

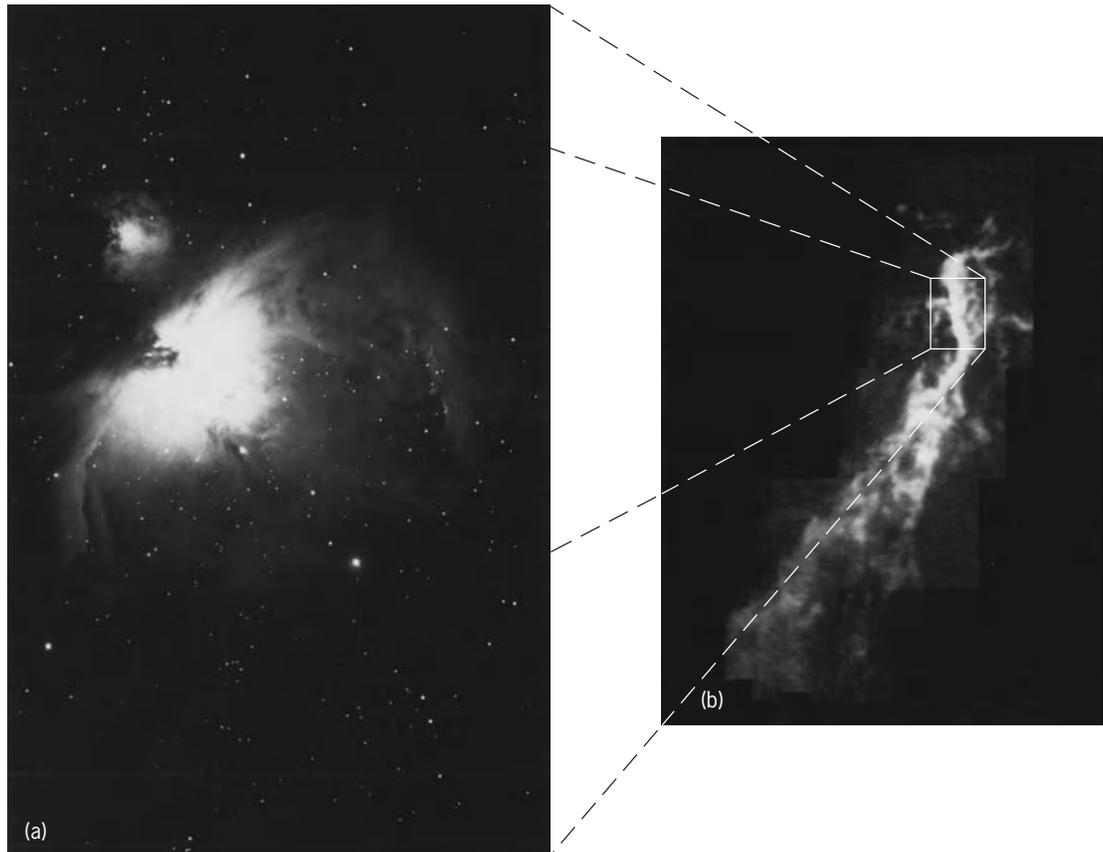
The modern boundaries of the 88 constellations, including this one, were defined by the International Astronomical Union in 1928. *See* CONSTELLATION.

Jay M. Pasachoff

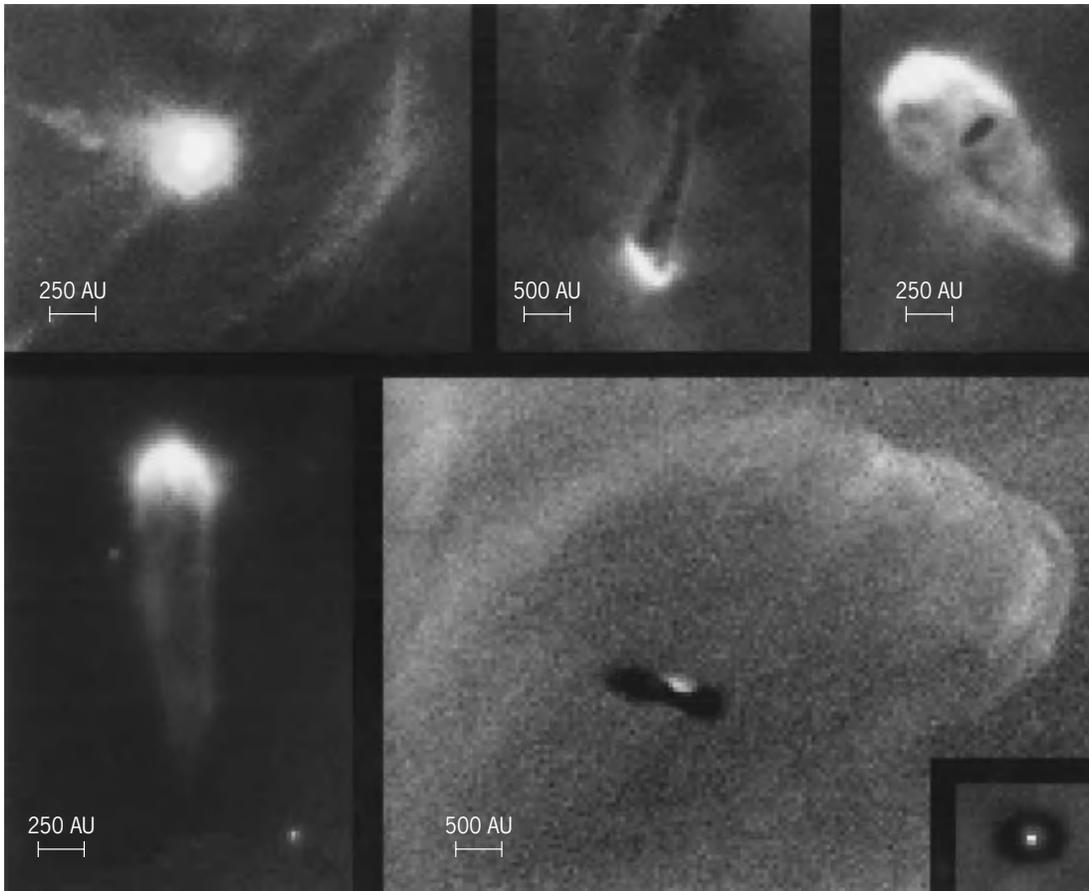
## Orion Nebula

The brightest emission nebula in the sky, designated M42 in Messier's catalog. The Great Nebula in Orion consists of ionized hydrogen and other trace elements (**Fig. 1; Colorplate 1**). The nebula belongs to a category of objects known as H II regions (the Roman numeral II indicates that hydrogen is in the ionized state), which mark sites of recent massive star formation. Located in Orion's Sword at a distance of 460 parsecs or 1500 light-years ( $8.8 \times 10^{15}$  mi or  $1.4 \times 10^{16}$  km), the Orion Nebula consists of dense plasma, ionized by the ultraviolet radiation of a group of hot stars less than 100,000 years old known as the Trapezium cluster (**Colorplate 2**). The nebula covers an area slightly smaller than the full moon and is visible with the aid of binoculars or a small telescope. *See* ORION.

In addition to the Trapezium cluster of high-mass stars, the Orion Nebula contains about 700 low-mass stars packed into an unusually small volume of space. Near the cluster center the average separation between stars is only about 5000 astronomical units ( $4.6 \times 10^{11}$  mi or  $7.5 \times 10^{11}$  km), about 50 times less than the separation between stars in the Sun's



**Fig. 1.** Orion Nebula and its neighborhood. (a) Visual-wavelength view of the Orion Nebula obtained by taking a time exposure on a photographic plate with the 4-m-diameter (158-in.) telescope on Kitt Peak, Arizona (*National Optical Astronomy Observatories*). (b) Orion A giant molecular cloud as seen in the 2.6-mm-wavelength emission line of  $^{13}\text{CO}$ . Data used to construct this image were obtained with a 7-m-diameter (23-ft) millimeter-wave telescope at Crawford Hill, New Jersey. The image shows a region which extends  $5^\circ$  north-south and  $2^\circ$  east-west. The Orion Nebula, located in front of the northern part of the molecular cloud, is not visible in this image. The approximate dimensions of the optical view are shown by the inset box. (*AT&T Bell Laboratories*)



**Fig. 2.** Hubble Space Telescope image showing closeups of several disks and ionized structures surrounding low-mass stars embedded within the Orion Nebula. The stars located near the Trapezium stars exhibit bright comet-shaped regions of gas expanding away from hidden disks. However, these circumstellar disks can be seen directly in silhouette against background nebular light for stars located sufficiently far from the Trapezium stars. The scale of each image is shown in astronomical units. (John Bally, Dave Devine, and Ralph Sutherland)

vicinity. However, as the Orion Nebula evolves, the cluster is expected to expand and possibly dissolve into the smooth background of stars in the Milky Way.

Observations with the Hubble Space Telescope have shown that many low-mass stars in the Orion Nebula are surrounded by disks of dense gas and dust (**Fig. 2**). These so-called *proplyds* (derived from the term “proto planetary disks”) appear to be rapidly losing their mass as they are irradiated by intense ultraviolet radiation produced by high-mass stars in the nebula. They may survive for only  $10^4$ – $10^5$  years in this harsh environment. However, the *proplyds* may eventually evolve into planetary systems if most of the solids in these disks have already coagulated into centimeter-sized objects which are resistant to ablation by the radiation field.

Stars form by the gravitational collapse and fragmentation of dense interstellar molecular clouds. A star-forming molecular cloud core (known as OMC1 for Orion Molecular Cloud 1) is hidden behind the Orion Nebula by a shroud of dust. A group of infrared sources, believed to be stars too young to have emerged from their dusty birth sites, is buried within OMC1. Although invisible at optical wavelengths, OMC1 has been investigated in the infrared and

radio portions of the spectrum. Observations of the massive young stars in OMC1 have shown that stellar birth is associated with powerful outflows of gas, naturally occurring maser emission, and shock waves that disrupt the remaining cloud core.

The cloud core behind the Orion Nebula is only a small part of a giant molecular cloud (the Orion A cloud), 100,000 times more massive than the Sun (**Fig. 1b**). Although mostly made of molecular hydrogen ( $H_2$ ), the cloud is best traced in the 2.6-mm-wavelength emission line  $^{13}CO$ , the rarer isotopic variant of carbon monoxide (CO). Over the past  $1 \times 10^7$  years, the Orion A cloud, together with a second giant molecular cloud in the northern portion of Orion (the Orion B cloud), have given birth to the Orion OB association, a loose, gravitationally unbound grouping of hot massive stars of spectral types A, B, and O. The Orion OB association also contains tens of thousands of low-mass young stars which formed from the Orion molecular clouds. The collective effect of the ionizing radiation and winds produced by these stars, and the supernova explosions occurring at their death, has generated an expanding bubble 70 by 200 pc ( $1.3$  by  $3.8 \times 10^{15}$  mi or  $2.2$  by  $6.2 \times 10^{15}$  km) in extent in the interstellar medium. The eastern portion of this bubble is

running into dense gas near the plane of the Milky Way Galaxy, where faint optical emission is seen as the 8° radius arc of nebulosity called Barnard's Loop. See INTERSTELLAR MATTER; NEBULA; RADIO ASTRONOMY; STAR; STELLAR EVOLUTION; SUPERNOVA.

John Bally

Bibliography. J. Bally, C. R. O'Dell, and M. J. McCaughrean, Disks, microjets, windblown bubbles, and outflows in the Orion Nebula, *Astron. J.*, 119:2919-2959, 2000; A. E. Glassgold, P. J. Huggins, and E. L. Schucking (eds.), Symposium on the Orion Nebula to Honor Henry Draper, *Ann. N. Y. Acad. Sci.*, 395:1-338, 1982; J. M. Pasachoff and A. Filippenko, *The Cosmos: Astronomy in the New Millennium*, 3d ed., Brooks Cole, 2007.

## Ornamental plants

The use of ornamental plants developed early in history for it is known that primitive peoples planted them near their dwellings. There was much use of flowers in the Greek and Roman worlds and a small plant nursery was found in the ruins of Pompeii. Gardening in various forms has been a part of the cultures of all known great civilizations. Although increased urbanization has often reduced the areas which can be devoted to ornamentals, nevertheless there is a growing interest in them and the sums spent for plants and horticultural supplies are quite large. New plants from various parts of the world are constantly being introduced by individuals and by botanic gardens and arboreta.

**Types and uses.** Ornamentals have been selected from virtually the whole kingdom of plants (Fig. 1). Ground covers include turfgrasses, herbaceous annuals, biennials or perennials, and some scandent shrubs or vines. The woody structural landscape plants include vines (deciduous and evergreen), shrubs (deciduous and evergreen), deciduous trees, broad-leaved evergreen trees, gymnosperms and cyads, bamboos, and palms.

Plants for special habitats or uses include lower plants such as mosses, selaginellas or horsetails, ferns and tree ferns, sword-leaved monocots, bromeliads and epiphytic plants, aquatic plants, cacti and succulents, and bulbous and cormous plants. Interior decorative plants in a large measure are shade plants from the tropics.

The best choice of an ornamental plant for a given situation depends on the design capabilities of the plant, the ecological requirements, and various horticultural factors. The plant must fulfill the purposes of the designer regarding form, mass, scale, texture, and color. It must also have the ability to adjust well to the ecology of the site in such factors as temperature range, humidity, light or shade, water, wind, and various soil factors, including drainage, aeration, fertility, texture, and salinity or alkalinity. Some important horticultural factors are growth rate, ultimate size and form, predictability of performance, and susceptibility to insects, diseases, deer, gophers, snails, and other pests. The hardiness of the plant is often

important, and pruning and litter problems may influence choice.

Large decorative foliage plants are used extensively in homes and offices. In many buildings the architects have included large planter areas where plants may be grown both indoors and outdoors. Where the light level is too low, banks of lamps have been placed where they can assist plants to keep in condition for long periods of time. Because some attention is required to keep plants attractive, there have been some failures in their upkeep. This has encouraged the development of artificial plants and flowers which are increasingly accurate and pleasing representations, especially if they are not inspected closely. Sometimes artificial plants are used in the background, with living plants in the foreground. Artificial flowers, on the other hand, do not seem to have lessened the demand for cut flowers. One way of solving the problem of regular care needed by living plants has been to make available special equipment, professional care, and replacement of plants after their installation. Several automatic watering systems have been developed which aid in maintenance.

**Propagation and culture.** Scientific advances have been made in plant propagation and culture. The use of small amounts of auxins, such as naphthalene acetic acid or indole butyric acid, promotes root formation on cuttings. Various types of intermittent mist and fog systems prevent cuttings from drying out and often improve rooting. See AUXIN.

An improvement in propagation is the introduction of tissue (meristem) culture of orchids into nutrient solutions. This permits the rapid large-scale multiplication of superior cultivars. However, the details of procedure are not yet established for all types of orchids. Meristem culture sometimes permits the elimination of a virus or a fungus from the rapidly expanding growing point of a plant. This system has been used with carnations and chrysanthemums to avoid diseases of several types and to produce clean propagating stock. See APICAL MERISTEM; TISSUE CULTURE.

The use of photoperiod to control the blooming time of plants has been important economically. Day length can be augmented by the use of electric lights, or shading with a black cloth may be used to increase the length of the dark periods. Thus it is now possible to buy blooming chrysanthemums and certain other flowers every day of the year. See PHOTOPERIODISM.

Automatic watering systems have been devised for plant growing. These can also be connected to tensiometers or several other devices to measure the soil moisture. See PLANT-WATER RELATIONS.

Plant growth retardants such as B-9, Phosphon, and Ammo 1618, when applied to certain plants such as chrysanthemum, poinsettias, and azaleas, may produce compact specimens which are superior in appearance. In some instances, the lasting qualities of the plant are greatly improved and flowering may be enhanced.

The composition of soil mixes for the growth of

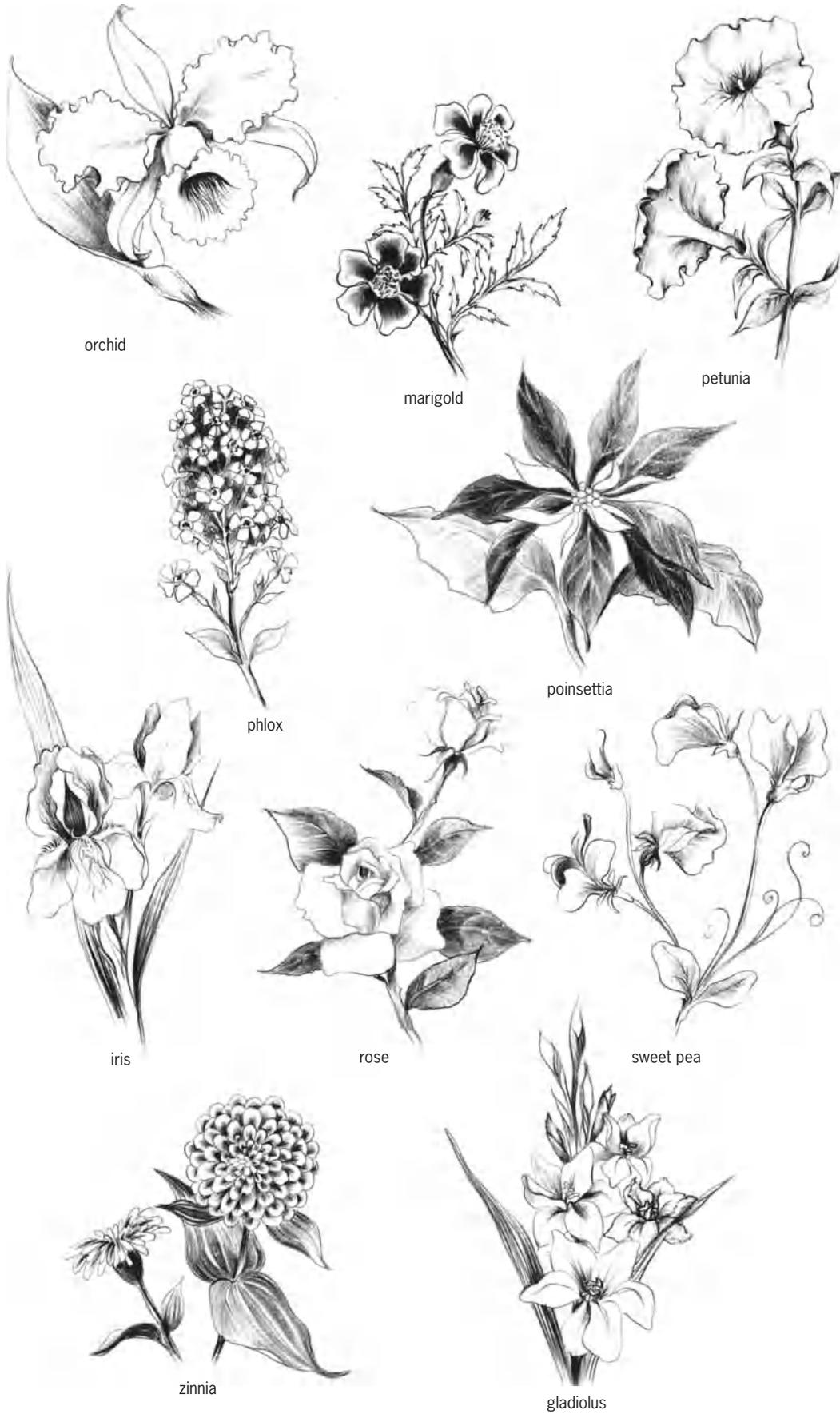


Fig. 1. Representative ornamental flowers used by amateurs and professionals.

container plants either in a commercial nursery or in a garden is important. The outcome of considerable experimentation has been the development of soil mixes for container growing which have a large proportion of relatively inert materials, such as sand, rice hulls, sawdust, forest tree bark, perlite, or vermiculite. Some soil mixes are based mainly on a fixture of fine sand and sphagnum peat moss. These mixtures have an advantage in that they may be used immediately after steam sterilization. Peat and perlite are used in other soil mixes which are quite close in characteristics to those above. Both mixes require a regular application of fertilizers unless slow-release types are incorporated. *See* FERTILIZER.

Many nursery plants are produced in salvage metal containers and several machines are available which assist in shifting small nursery plants. Container growing is more common in warmer areas than in colder parts of the United States.

**Horticultural maintenance.** Many traditional gardening tools have been improved or mechanized, which reduces labor costs. Electric clippers permit much greater use of clipped hedges; pneumatic machinery is available for tree pruning; and power grass mowers and edgers are widely used.

Herbicides meet the needs of practically any situation and save hand weeding. The new slow-release fertilizers furnish plant nutrients over a long period of time and are particularly helpful in container plant growing. The appropriate chelates furnish trace elements such as iron or zinc to plants and thus largely avoid deficiency symptoms. The use of liquid fertilizers in watering has been simplified by the use of automatic proportioning pumps. Often liquid fertilizer is applied with each watering of plants in the nursery. *See* HERBICIDE.

New and improved fungicides and insecticides are constantly being introduced. The systemic insecticides in which the chemical is applied to the soil and taken up by the plant are especially useful since residue problems do not exist with ornamentals. They are being used by both amateurs and professionals. Both groups are also beginning to use soil fumigation prior to planting. Horticultural maintenance on a contract basis is common now in some parts of the United States and tree surgery is done by specialists. *See* FUNGISTAT AND FUNGICIDE; INSECTICIDE.

**Breeding.** Outstanding plant breeding has been done by amateurs working with popular garden flowers of minor economic value. Notable examples include such plants as the daylily, iris, and dahlia. These nonprofessionals have frequently formed societies for the exchange of information and breeding stocks. Some of the outstanding developments in breeding ornamental plants are listed in the **table**.

Professional plant breeders work largely with important commercial flowers such as the rose, lily, orchid, gladiolus, and chrysanthemum. The United States plant patent law has given protection to hybridizers and all vegetatively propagated plants are covered. This law has given a great impetus to plant breeding, particularly that of roses. Test gardens for

#### Notable developments in ornamentals

Flower	Development
African violet	Polyploids; many new foliage and flower characters
Aster	Wilt resistance; cut-flower types
Camellia	The Williamsi hybrids ( <i>Camellia japonica</i> and <i>C. salensis</i> ); spectacular new varieties of <i>C. reticulata</i> from Yunnan, China; intensive breeding of many types
Carnation	The Sims varieties (U.S.A.), outstanding for commercial cut flowers; spray-type carnations
Fuchsia	Heat-resistant types (California); pure white flowers (California)
Gladiolus	New miniature types; breeding for disease resistance and vigor
Hemerocallis	Purer pinks and reds; recurrent blooming habit; polyploid types
Iris	Superior new varieties of Dutch iris; tall bearded iris, with purer pinks and recurrent blooming habit
Marigold	F <sub>1</sub> hybrids; closer approach to white; scentless foliage types
Orchid	Complex species hybrids; polyploid forms with improved flower quality; miniature flowered cymbidiums; breeding for specific blooming season
Petunia	F <sub>1</sub> hybrids; pure scarlet crimson color; giant ruffled and all-double flower types
Phlox	Tetraploid annual varieties
Poinsettia	New flower and foliage types with longer-lasting qualities
Rose	Large-flowered polyantha types ("floribundas" and "grandifloras"); lavender flower color; introduction of varieties with flowers of superior keeping quality; new miniature types; intensive breeding for vigor, disease resistance, fragrance, and new colors
Shasta daisy	New single and double types for florist use
Snapdragon	F <sub>1</sub> hybrids; giant tetraploids; reduction of flower shattering; ruffled double-flowered types; rust resistance
Stocks	Mosaic resistance; high double strains
Sweet pea	Heat-resistant strains; many-flowered (multiflora) strains
Zinnia	F <sub>1</sub> hybrids; new twisted-petal types; giant size

new varieties (cultivars) are located in various parts of the country to assist in evaluating the new varieties. *See* BREEDING (PLANT).

Occasionally, ionizing radiation has been applied to ornamentals to produce new forms. The use of colchicine to double chromosomes has been a valuable technique both for producing direct plant changes and for making new crosses possible. Spontaneous mutations often occur when large plant populations are grown. Many new types of annual flowers introduced by seed growers arose in this manner. *See* MUTATION.

The producers of plants grown from seed have no legal protection for their new productions. However, some protection is being obtained by the production of hybrid seed from parents, which are kept by the originator. In this instance seeds from the hybrids may not come true, and thus the seed producer is protected from competition. The flower seed industry does much plant breeding which is concentrated in the cool coastal valleys of central and southern California.

Air transport has revolutionized the growing of floricultural crops and has brought many

competitive advantages to California and to southern Florida. The production is in glasshouses, in cloth houses, and in many instances in open fields. Various types of plastic coverings are used for some crops as a cheaper substitute for glass; however, many of these last only one season. Plastics laminated with fiberglass have a much longer life and are used more. See ARBORETUM; BOTANICAL GARDENS; FLORICULTURE.

Vernon T. Stoutemyer

**Diseases of woody ornamentals.** Woody ornamentals are subject to a number of diseases. Some diseases are inconspicuous, causing little or no permanent injury, while others, by severely injuring leaves, twigs, branches, or roots, cause plants to become stunted or to die.

*Nonparasitic disease.* Environmental stresses commonly affect landscape plants, which are often not genetically adapted to the sites where they are placed. Nutrient stress may be a problem in disturbed or altered soils. In alkaline soils, essential nutrients may not be available to the plants, and chlorosis or yellowing of leaves is a common symptom. Sensitive plants should be grown in a somewhat acid soil.

Drought or flooding are the climatic factors most likely to cause stress damage. Water or air deficiency in the soil causes similar symptoms in the aerial portion of the plant: wilt, leaf necrosis, and, when persisting, death. Both air and water are also important for healthy roots.

Low-temperature injury is more common than high-temperature injury. Rapid drops in temperature to below freezing, following mild weather, often injure plants that are not fully acclimated to low temperatures. Low temperatures dehydrate stem tissues, and vertical slits in the stems (frost cracks) may occur. Below-freezing temperatures coming after shoot emergence in the spring often kill succulent tissues. See COLD HARDINESS (PLANT).

Pollutants can be present in the soil or in the air and may result in chronic or acute plant damage. The pollutants that kill plants usually originate in stationary manufacturing plants: sulfur dioxide and fluorides are examples. Ozone and peroxyacetyl nitrate, pollutants associated with automobile emissions, injure leaf surfaces but rarely cause defoliation. Herbicides are considered pollutants when misapplied and may cause leaf and stem distortion, necrosis, and plant death.

*Pathogens.* Fungi are the infectious agents most frequently reported to cause diseases of landscape plants, but bacteria, viruses, nematodes, and mycoplasma-like organisms also can be serious pathogens.

Pathogens that cause leaf spots or blights do not greatly weaken the host plant unless premature defoliation occurs. Evergreens may be seriously injured after one defoliation, but two or more consecutive years of defoliation usually are required to kill deciduous plants or to cause extensive dieback in them. Mildews, rusts, blights, and leaf spots can be controlled with applications of protective fungicides.

Cankers and dieback result when fungal pathogens kill portions of the vascular cambium of twigs and stems. The extent of injury by these organisms

is modified by environmental conditions. Vigorous, well-maintained plants are resistant to many of these pathogens. Pruning out the affected limbs is drastic but is often the only control for susceptible plants.

Root and crown rot pathogens cause plants to have small leaves, thin foliage, dieback, and slow decline. These diseases are most common when plants are in moist or improperly drained soil. Fungicide drenches and dips can reduce the likelihood of their occurrence.

Fungal or bacterial pathogens that become systemic in the vascular tissues of plants bring about the most rapid plant mortality. Those in the wood prevent sap flow: those in the bark prevent carbohydrate translocation. Satisfactory control measures are not now available for vascular wilts. Dan Neely

**Diseases of herbaceous ornamentals.** Ornamental plants include a large number of herbaceous genera that are susceptible to one or more diseases that affect growth, flowering, or reproduction of the plant. This section of the article describes some of the most common diseases and the kinds of pathogenic organisms that cause them.

*Classification.* Diseases may be classified as parasitic and nonparasitic. Parasitic diseases are those caused by microorganisms such as fungi, bacteria, viruses, and nematodes. To this group have been added viroids, mycoplasma, and spiroplasma. Nonparasitic diseases are those induced by environmental stress that adversely affect plant growth. Only parasitic diseases will be described below.

Diseases are classified according to the types of symptoms they produce. Microorganisms that cause parasitic diseases may affect roots, stems, leaves, and flowers of fruits. Symptoms of these diseases may be described as leaf spots, blights, cankers, galls, and scorches if they occur on the aboveground portion of the plant where the organism has invaded host tissue. The causal organism can often be isolated from leaves, stems, flowers, and fruits. Fungi and bacteria of different genera can cause one or more of these symptoms.

*Fungal and bacterial diseases.* Leaf spots are commonly observed on many plants. Fungus leaf spot diseases often spread rapidly by spores produced in the leaf spots. Infected leaves drop prematurely, and susceptible varieties may defoliate rapidly. Many fungi that cause leaf spots of ornamentals overwinter in leaf debris and diseased stems; such stems should be removed and diseased leaves should be destroyed to reduce subsequent infections. Fungicides applied after the diseased leaves are removed will protect the newly formed leaves from infection.

Bacteria also produce leaf spots. Black leaf spot on *Delphinium* is caused by *Pseudomonas delphinii*. Diseased leaves are removed and plants sprayed with streptomycin to control the disease.

Powdery mildew is caused by a group of many closely related fungi that commonly affect the leaves but may also affect the flowers and fruit. A white coating that appears like powder is formed in the spring and summer. The fungus survives the winter as strands or hyphae on shoots that again produce

spores in the spring. The fungus can be controlled by interrupting its life cycle by removing and destroying all diseased leaves to eliminate the overwintering stage. Destruction of infected leaves that drop from the plant will reduce the capability of the fungus to reproduce further and spread during the following spring. This sanitation approach to disease control is an important principle in controlling many diseases caused by foliar pathogens of ornamental plants. During the growing season, chemicals may also be applied to control spread of the disease. Powdery mildew commonly causes diseases on rose, chrysanthemum, lilac, and many other ornamentals.

Blight is a nonspecific term that is usually applied to a disease that affects young growing tissue, including developing buds, flowers, and shoot tips. A common garden blight of ornamentals is caused by the fungus *Botrytis cinerea* and is referred to as gray mold. The disease is common on flowers of rose, lily, geranium, chrysanthemum, carnation, snapdragon, gladiolus, and many other garden flowers. The disease often occurs in the garden during wet weather and may cause rapid rotting of the flowers. The disease may be controlled by removing and destroying affected flowers and by frequent fungicidal applications during periods of wet weather. The fungus forms small black structures (sclerotia) on or in infected tissue, and these structures may survive and overwinter on debris in the soil.

Canker refers to disease of the stem and branches that usually occur on woody ornamentals. The canker appears as a localized eruption that may encircle the stem and cause collapse. Several different fungi produce cankers on rose, and they are controlled by removing and destroying affected canes. Fungicidal sprays are also applied to control these diseases.

Galls are swollen tissues. They may appear as a proliferation of cell growth in flowers, leaves, or roots. Crown gall is a bacterial disease caused by *Agrobacterium tumefaciens*. The bacterium has a wide host range and produces galls on the roots and on stems aboveground. The disease is controlled by cutting off the galls and destroying them and dipping the roots and stem in a solution of the antibiotic streptomycin. Another approach to control is to dip the roots in a suspension of an avirulent strain of the bacterium. See CROWN GALL.

Serious root diseases are caused by fungi and nematodes. Disease agents that attack the roots and the bases of the stems are often not recognized until serious damage has been done. Symptoms may appear as wilting and dwarfing. Controls must often be applied as preventive rather than therapeutic measures because the plant may be killed if infection is not prevented. An example of a fungus disease of seedling stems and roots is damping-off caused by *Rhizoctonia solani* and *Pythium* and *Fusarium* species. These fungi often kill germinating seedlings, and are controlled by treating the seed or soil with an appropriate fungicide. Several species of *Fusarium* also cause wilt of plants or individual branches. The conductive tissues in the stem may show a darkening and

discoloration. Infected plants show the most prominent symptoms in hot weather, and the plants die. Another root pathogen is *Verticillium*. This fungus thrives in cool weather and will also kill susceptible species. *Fusarium* species are usually specific to one kind of plant, but *Verticillium* attacks many different plants, including annuals and woody species. Both of these fungi survive for long periods in the soil. Infected plants should be removed and burned and resistant varieties planted in soil where the fungus is known to occur.

Rusts are complex fungi that produce as many as five different forms of spores during the life cycle. Rusts are obligate parasites; that is, they must develop in living host tissue. Symptoms appear as reddish-brown spots on the leaves. Some common rusts occur on carnation, hollyhock, snapdragon, aster, and chrysanthemum. Rust diseases may be difficult to control; whereas some rusts complete their life cycle in one host, others must infect two unrelated species to complete the life cycle. Control measures include the application of fungicides and removing and burning diseased leaves to destroy overwintering stages.

*Viral diseases.* Viral diseases can be very destructive. Viruses move systemically throughout the roots, stems, and leaves. Many viruses are transmitted by rubbing juice from an infected plant onto healthy plants, and many are also transmitted by insects. Diseased plants may be stunted with yellow mottling and ring patterns on the leaves and light- or dark-colored streaks on the flowers. Cucumber mosaic virus infects many species, including gladiolus, lilies, dahlias, and delphiniums. This virus is transmitted by aphids, and the virus spreads rapidly with an increase in aphid populations. Viral disease symptoms may appear only within a particular range of temperature. Disease symptoms may be suppressed in plants grown at unusually high or low temperatures. Infected plants should be removed and destroyed, especially perennial and bulb crops that serve as a source of continuing infection. There are no known chemicals that can be applied to control virus diseases.

*Mycoplasmas.* Yellows diseases are caused by mycoplasmas that are transmitted by leafhoppers. Aster yellows produces stunting and yellowing of the leaves and growth proliferation producing a "witches'-broom" type of growth. Tetracycline antibiotic sprayed on infected plants has induced a remission of symptoms and resumption of normal growth. However, symptoms reappeared when the antibiotic treatment was terminated.

*Viroids.* Viroids are a newly described class of disease agents that are much smaller than viruses. Chrysanthemum stunt and chrysanthemum chlorotic mottle are important viroid diseases. Diseased plants should be destroyed.

*Control.* Parasitic disease control depends on the accurate identification of the pathogen. The kind of pathogen will determine the specific control procedures. Sanitary procedures that reduce the source of infection; healthy seeds, bulbs, and cuttings; disinfection of tools; pasteurization of soil; and chemicals

applied to seed, to soil, or directly to the diseased plant may be used separately or in combination to produce healthy, productive plants. Roger H. Lawson

**Diseases of foliage ornamentals.** Diseases of foliage ornamental plants, as with any plants in fact, can occur only when the proper combination of the pathogen, host plant, and environmental conditions exists. With the surge of interest in foliage plants for utilization in the home, office, shopping plazas, public agencies, and private institutions, disease problems have become more important. The growth and well-being of such plants, when not grown in the natural environment where they have evolved successfully against adversity, inevitably result in myriads of problems. Of these, diseases often become manifest, often resulting in death of the plant.

Foliage ornamental diseases are usually the result of improper care or lack of knowledge of the requirements for healthy plants, that is, light, moisture, temperature, and nutrition, as well as sanitation. When grown in unnatural situations, they are placed under adverse conditions resulting in debilitation, thus subjecting them to plant-invading organisms.

Plant pathogenic bacteria (*Xanthomonas*, *Pseudomonas*, *Erwinia*, *Agrobacterium*, *Corynebacterium*) can be soil-borne or occur in infected plants. The most common symptoms of bacterial infections are leaf spots having a water-soaked appearance, wilting of the foliage, and soft-rotting of underground plant parts. Control of bacterial diseases lies in the use of healthy, disease-free plants, "clean" soil, and prevention of contamination of healthy plants by infected plants or contaminated equipment and splashing water. Chemical control is not considered practical for most bacterial diseases of foliage ornamentals.

Pathogenic fungi (*Rhizoctonia*, *Sclerotium*, *Sclerotinia*, *Cercospora*, *Fusarium*, *Pythium*, *Phytophthora*, *Verticillium*, *Cylindrocladium*, and so on) are capable of attacking virtually all parts of plants, and are the most common and generally most readily visible of the disease-causing organisms. They cause rotting of the roots, plugging of the vascular system, stem cankers and galls, damping-off (death of the plant just before or soon after emerging from the surface of the soil), and leaf spots (Fig. 2) and distortion. Control of fungal diseases can often be achieved with proper chemicals applied at the early stages of disease symptoms, provided that a proper diagnosis is made as to the causal agent.

Viruses or viruslike agents are the most difficult to control. Residing in a plant in almost complete harmony, these organisms are often referred to as the perfect parasites since they rarely, except in a few well-known instances, cause the obvious and readily observed destruction inflicted by bacteria and fungi. Most symptoms of viral diseases are subtle but nevertheless injurious. The range of symptoms include mosaics, flecking and mottling of leaves and sometimes stems, vein necrosis or discoloration, leaf and flower distortion and discoloration, and stunting and general unthriftiness of the plant. Virus-infected plants are best discarded, since there are no practical cures for these diseases. Insects are known to be vectors of



Fig. 2. Leaf spot of *Syngonium podophyllum* caused by the fungus *Cephalosporium cinnamomeum*.

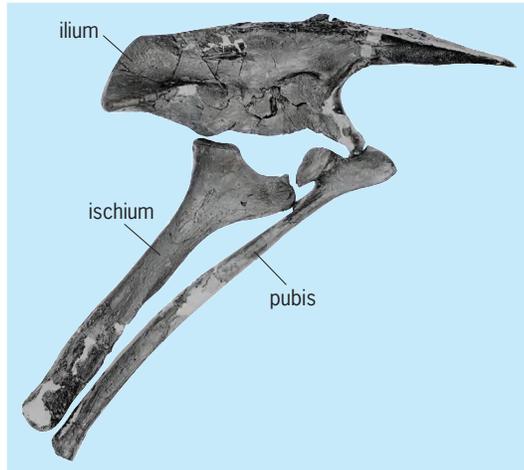
some viruses or viruslike agents; hence, they should be controlled with proper insecticides. See PLANT PATHOLOGY.

S. A. Alfieri, Jr.

## Ornithischia

One of two constituent clades of Dinosauria (the other being Saurischia). Ornithischian dinosaurs are characterized by the possession of numerous anatomical features, notably the configuration of the pelvis. In ornithischian hips, the pubis bone (which points downward and forward in the majority of other reptile groups) has rotated posteriorly to lie parallel to the ischium (Fig. 1). This configuration is superficially similar to that seen in living birds, but actually represents an example of evolutionary convergence (birds are, in fact, descended from saurischian dinosaurs). Other features that characterize ornithischians include the presence of ossified tendons that form a complex lattice arrangement supporting the vertebral column, a single midline bone called the predentary that lies at the front end of and connects the two lower jaws, and a palpebral bone that traverses the orbit (Fig. 2).

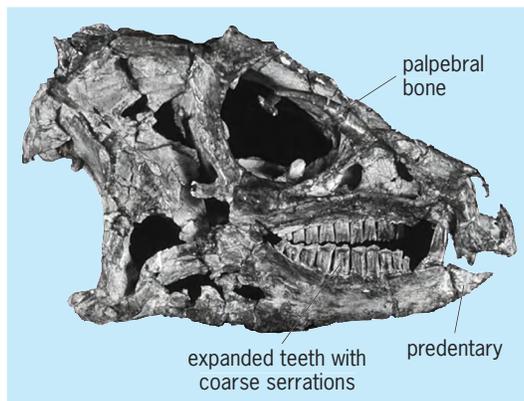
**Distribution.** The earliest known ornithischian, *Pisanosaurus*, was discovered in deposits of Late Triassic age (approximately 225 million years ago) in Argentina. During the Late Triassic–Early Jurassic interval, ornithischians were relatively rare components of dinosaur faunas; however, the group began to radiate spectacularly in terms of species richness, abundance, and morphological diversity during the Middle Jurassic. From the Middle Jurassic onward, ornithischians attained a global distribution and became the dominant terrestrial vertebrates of the Cretaceous. Almost all ornithischians were herbivorous,



**Fig. 1.** Pelvis of the Lower Jurassic ornithischian dinosaur *Scelidosaurus harrisonii* from England (Natural History Museum, London specimen BMNH R1111). The front of the animal is to the right of the photograph. Note that the pubis bone has rotated backward to lie alongside the ischium, as in all ornithischian dinosaurs, in contrast to the usual reptilian condition in which the pubis extends forward, away from the ischium.

though a few primitive forms may have been omnivores.

**Phylogeny and evolution.** Phylogenetic analyses of ornithischian interrelationships indicate that the clade can be subdivided into three major lineages (Thyreophora, Marginocephalia, and Ornithopoda), along with a small number of primitive forms that do not fall within any of these groupings. Primitive ornithischians, such as *Lesothosaurus* from the Early Jurassic (approximately 205 million years ago) of southern Africa, were small [1–2 m (3–6.5 ft) in body length], bipedal animals with leaf-shaped teeth for shearing plant material, long counterbalancing tails, and grasping hands; this basic design was modified and elaborated in later ornithischian groups. Ornithischian evolution was dominated by several major themes: experimentation with stance (quadrupedality vs. bipedality), the appearance of ever more complex feeding mechanisms (such as al-



**Fig. 2.** Skull of the Lower Jurassic ornithischian dinosaur *Heterodontosaurus tucki* from South Africa (Iziko: South African Museum specimen SAM-PK-K1332). Note the predentary bone at the tip of the lower jaw, the palpebral bone crossing the eye socket, and the expanded, coarsely serrated teeth. These features are characteristic of all ornithischian dinosaurs.

terations to teeth and skull morphology), and the development of sociality (gregarious living, parental care, and intra- and interspecific communication).

**Thyreophora.** Thyreophora (the “armored” dinosaurs) consists of two clades: Stegosauria and Ankylosauria. Both of these groups were quadrupedal and characterized by the possession of numerous osteoderms (bony plates) that were embedded within the skin.

*Stegosauria.* Stegosaurs are known from the Middle Jurassic–Early Cretaceous of North America, Africa, Europe, and Asia. The osteoderms were either flat, subtriangular plates (as in *Stegosaurus*: Late Jurassic, North America) or spinelike (as in *Kentrosaurus*: Late Jurassic, Tanzania). The plates and/or spines were arranged in two parallel rows that extended along the back and tail; a large spine was also sometimes present in the shoulder region. The arrangement of plates and spines differs between stegosaur species, perhaps indicating a role in species recognition or intraspecific display. In addition, spines were likely to have had a defensive function; the plates were incised with many grooves for small blood vessels and may have had a thermoregulatory function.

*Ankylosauria.* Ankylosaurs appeared in the Early Jurassic (*Scelidosaurus* from England; Fig. 1) and persisted until the end of the Cretaceous Period. They were relatively rare during the Jurassic but increased in diversity during the Cretaceous, at which time they had a worldwide distribution. Ankylosaurs possessed an extensive covering of armor, consisting of multiple rows of low armor plates and spines covering almost all surfaces of the body except the underside, giving them an almost impenetrable tanklike exterior. Some genera (for example, *Ankylosaurus*: Late Cretaceous, North America) possessed tail-clubs made of solid bone, which were presumably used as defensive weapons. One genus (*Euoplocephalus*: Late Cretaceous, North America) even possessed a bony eyelid.

**Marginocephalia.** Marginocephalia are united by the possession of a distinct shelf of bone at the rear of the skull and consist of two groups: Pachycephalosauria (“bone-headed” dinosaurs) and Ceratopsia (“horned” dinosaurs).

*Pachycephalosauria.* Pachycephalosaurs were bipedal animals whose skulls were thickened, leading to the development of a dome-like structure in some cases. This dome [which could reach up to 25 cm (10 in.) thickness in *Pachycephalosaurus*: Late Cretaceous, North America] was probably used for butting rivals during intraspecific combat. Pachycephalosaurs were relatively rare animals that were restricted to the Cretaceous of North America, Asia, and possibly Europe.

*Ceratopsia.* Early ceratopsians, such as *Psittacosaurus* (Early Cretaceous: China and Mongolia) were small bipeds, lacking horns and possessing only short frills that extended from the back of the skull. However, later forms such as *Triceratops* (Late Cretaceous: North America) were large quadrupeds, which often displayed an array of horns (situated on the tip of the snout and/or above the eyes) and an

impressive frill that could be over 1.5 m (5 ft) in length. All ceratopsians had parrot-like beaks for cropping vegetation and guillotine-like dentitions for processing fibrous plant material. Ceratopsians were limited to the Cretaceous of Asia and North America, though they were locally abundant.

**Ornithopoda.** A conservative but very successful group that retained the bipedal pose of primitive ornithischians, ornithopods are first known from the Middle Jurassic and persisted until the end of the Cretaceous. All ornithopods possessed sophisticated grinding jaw mechanisms, allowing them to process vegetation as efficiently as living grazers and browsers (for example, horses and cows). One subgroup of ornithopods, the hadrosaurs (“duck-billed” dinosaurs), had teeth arrayed in dense “batteries,” consisting of multiple columns and rows of functional teeth that provided a large continuous surface area for chewing. Hadrosaurs were also among the most social of dinosaurs: one genus, *Maiasaura* (Late Cretaceous: North America), is known from dozens of eggs, nests (which are found grouped together in colonies) and babies, as well as numerous associated adult remains. Some evidence (indicating, for example, feeding of babies in the nest) suggests that *Maiasaura* had a high degree of parental care. In addition, colonial nest sites, multiple co-occurring trackways, and accumulations of “bone-beds” indicate gregarious habits for these animals (and for many other ornithopods, such as *Iguanodon* from the Early Cretaceous of Europe, Asia, and North America). Many hadrosaurs (for example, *Parasaurolophus* and *Corythosaurus*: both from the Late Cretaceous of North America) had bizarre cranial crests that were probably used for display and species recognition. See ARCHOSAURIA; CRETACEOUS; JURASSIC; DINOSAURIA; SAURISCHIA.

Paul M. Barrett

**Bibliography.** P. J. Currie and K. Padian (eds.), *The Encyclopedia of Dinosaurs*, Academic Press, San Diego, 1997; J. O. Farlow and M. K. Brett-Surman (eds.) *The Complete Dinosaur*, Indiana University Press, Bloomington, 1997; D. B. Weishampel, P. Dodson, and H. Osmólska (eds.) *The Dinosauria*, 2d ed., University of California Press, Berkeley, 2004.

## Orogeny

The process of mountain building. As traditionally used, the term orogeny refers to the development of long, mountainous belts on the continents called orogenic belts or orogens. These include the Appalachian and Cordilleran orogens of North America, the Andean orogen of western South America, the Caledonian orogen of northern Europe and eastern Greenland, and the Alpine-Himalayan orogen that stretches from western Europe to eastern China (Fig. 1). It is important to recognize that these systems represent only the most recent orogenic belts that retain the high-relief characteristic of mountainous regions. In fact, the continents can be viewed as a collage of ancient orogenic belts, most so deeply eroded that no trace of their original mountainous

topography remains (Fig. 2). By comparing characteristic rock assemblages from more recent orogens with their deeply eroded counterparts, geologists surmise that the processes responsible for mountain building today extended back through most (if not all) of geologic time and played a major role in the growth of the continents. See CONTINENTS, EVOLUTION OF.

The construction of mountain belts is best understood in the context of plate tectonics theory. Earth's lithosphere is currently fragmented into at least a dozen, more or less rigid plates that are separated by three kinds of boundaries: convergent, divergent, and transform. Plates move away from each other at divergent boundaries. On the continents, these boundaries are marked by rift systems such as those in East Africa; in the ocean basins, they correspond to spreading centers, submarine mountain chains (such as the Mid-Atlantic Ridge) where new oceanic crust is produced to fill the gap left behind by plate divergence. Transform boundaries are zones where plates slide past one another; a familiar example on land is the San Andreas fault system. Orogenic belts form at convergent boundaries, where lithosphere plates collide. See FAULT AND FAULT STRUCTURES; MID-OCEANIC RIDGE; PLATE TECTONICS; RIFT VALLEY; TRANSFORM FAULT.

**Convergent plate boundaries.** There are two basic kinds of convergent plate boundaries, leading to the development of two end-member classes of orogenic belts. Oceanic subduction boundaries are those at which oceanic lithosphere is thrust (subducted) beneath either continental or oceanic lithosphere. The process of subduction leads to partial melting near the plate boundary at depth, which is manifested by volcanic and intrusive igneous activity in the overriding plate. Where the overriding plate consists of oceanic lithosphere, the result is an intraoceanic island arc, such as the Japanese islands. Where the overriding plate is continental, a continental arc is formed. The Andes of western South America is an example. See MARINE GEOLOGY; OCEANIC ISLANDS; SUBDUCTION ZONES.

The second kind of convergent plate boundary forms when an ocean basin between two continental masses has been completely consumed at an oceanic subduction boundary and the continents collide. Continent collisional orogeny has resulted in some of the most dramatic mountain ranges on Earth. A good example is the Himalayan orogen, which began forming roughly 50 million years ago when India collided with the Asian continent. Because the destruction of oceanic lithosphere at subduction boundaries is a prerequisite for continental collision, continent collisional orogens contain deformational features and rock associations developed during arc formation as well as those produced by continental collision, and are thus characterized by complex internal structure.

Compressive forces at convergent plate boundaries shorten and thicken the crust, and one important characteristic of orogens is the presence of unusually thick (up to 40–60 mi or 70–80 km) continental crust. Because continental crust is more buoyant than the underlying mantle, regions of thick

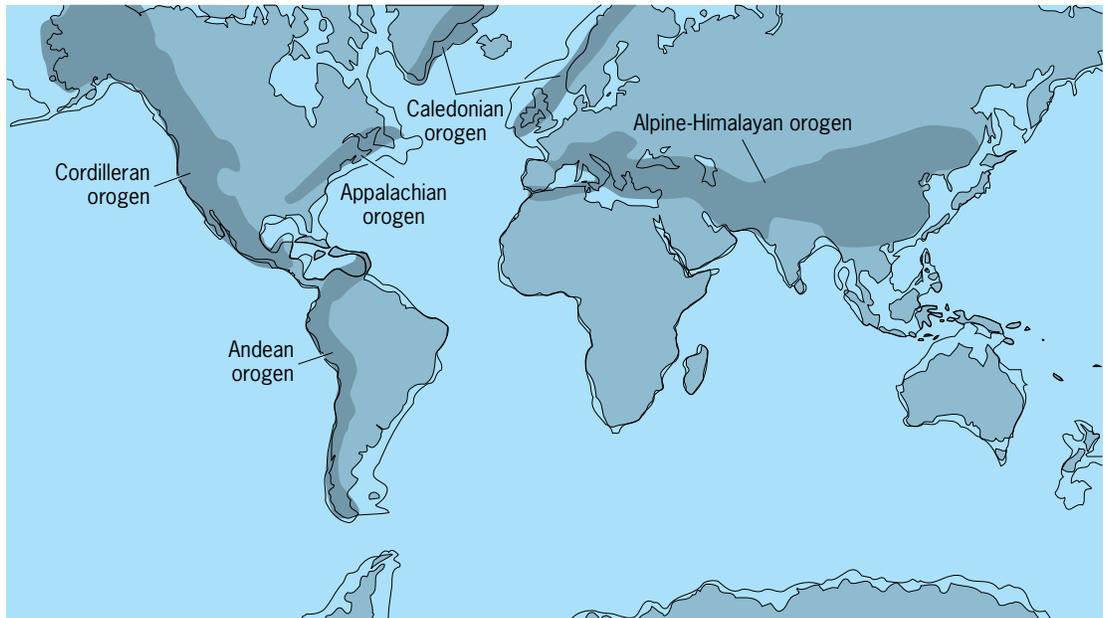


Fig. 1. Distribution of major orogenic belts. (After E. M. Moores and R. J. Twiss, *Tectonics*, W. H. Freeman, 1995)

crust are marked by high elevations of the Earth's surface and, hence, mountainous terrain. Once a mountain range has developed, its high elevation can be maintained by two mechanisms in addition to its own buoyancy. First, compressional forces related to continuing convergence across the plate boundary serve to buttress the orogen against collapse. Second, some mountain ranges are developed on thick, cold, and therefore strong continental lithosphere, and the lithosphere serves as a rigid support. However, such processes work against the gravitational forces that are constantly acting to destroy mountain ranges through erosion and normal fault-

ing. All mountain ranges ultimately succumb to gravity; the resulting denudation exposes deep levels of orogenic belts and provides geologists with important insights regarding the internal architecture. See ASTHENOSPHERE; EARTH CRUST; LITHOSPHERE.

*Continent collisional orogen.* The tectonics of the Himalayas can serve as an example of the basic anatomy of a continent collisional orogen (Fig. 3). Prior to collision between India and Asia, the two continental masses were separated by an ocean referred to as Tethys. Most Himalayan geologists are convinced that the southern margin of Asia in late Mesozoic time was an oceanic subduction boundary where

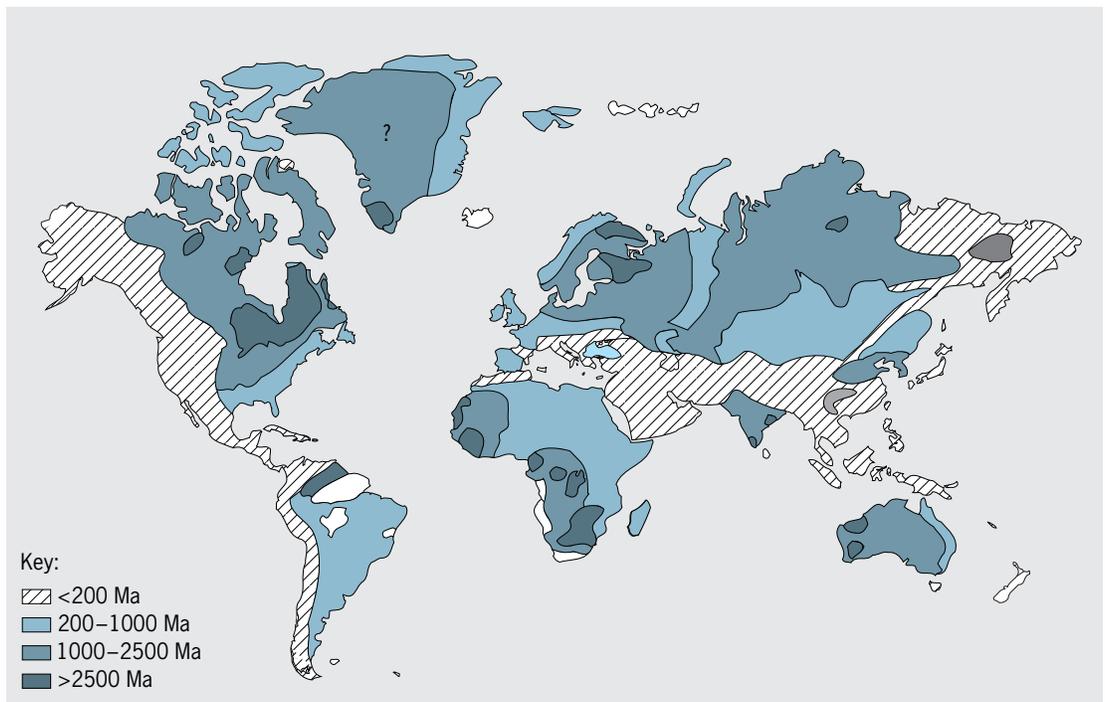


Fig. 2. Distribution of orogenic belts, by time of major orogenic distribution, in the continents. (After B. C. Burchfiel, *The continental crust*, in E. M. Moores, ed., *Shaping the Earth: Tectonics of Continents and Oceans*, W. H. Freeman, 1990)

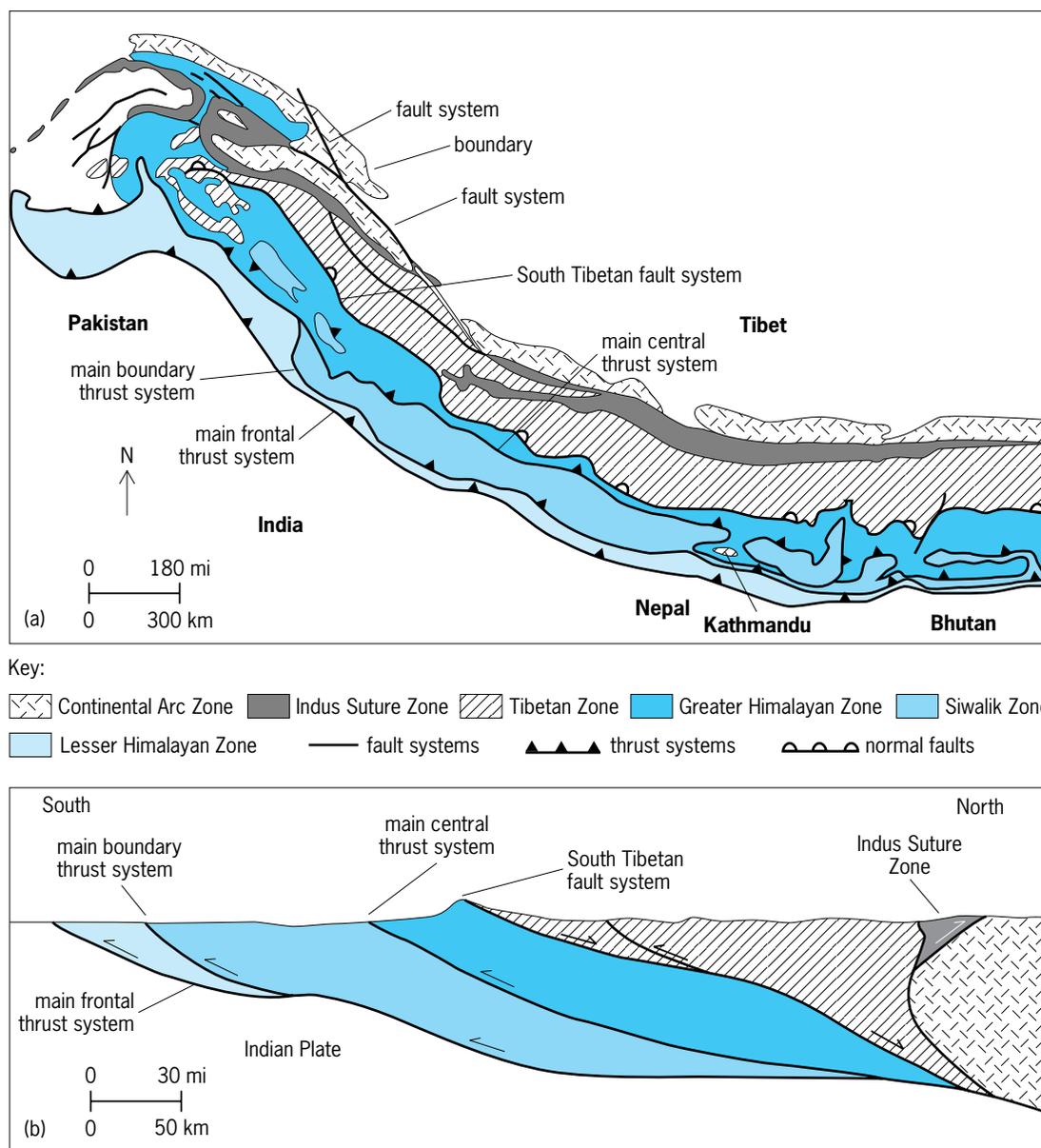


Fig. 3. Himalayan orogen. (a) Simplified tectonic map. (b) Generalized cross section. Fault systems are labeled at their surface exposure. Half-arrows indicate directions of movement on the fault systems.

Tethyan oceanic lithosphere was thrusting beneath Asia. The principal evidence for this interpretation comes from the existence of a continental arc of appropriate age that was built on Asian lithosphere just north of the Himalayan mountain range (Continental Arc Zone). India and Asia collided when all the intervening oceanic lithosphere had been destroyed at the subduction boundary. The collision zone at the surface of the Earth is marked by remnants of Tethyan oceanic crust (ophiolites) and related rocks exposed in the Indus Suture Zone. All tectonic zones south of the suture are part of the Indian continental lithosphere that was folded and imbricated (overlapped at the margins) during collision. The Tibetan Zone includes Paleozoic to early Tertiary sedimentary rocks originally deposited along the northern margin of the Indian continent. The Greater Himalayan Zone consists of crustal rocks that were deformed, metamorphosed, and partially melted at high temperatures

deep within the crust as a consequence of Himalayan orogeny; these rocks form the principal substrate for the high mountains of the Himalayas, including Mount Everest. Farther south, the Lesser Himalayan Zone contains Precambrian to Paleozoic sedimentary and metamorphic rocks that have been deformed at lower temperatures and pressures. As collisional orogeny progressed through Tertiary time, material eroded from the rising Himalayan ranges was transported southward by a variety of ancient river systems and deposited at the southern margin of the orogen to form the Siwalik Zone. These rocks were themselves deformed during the most recent stages of Himalayan orogeny as the locus of surface deformation propagated southward toward the Indian subcontinent. See GEOLOGIC TIME SCALE.

A simplified geologic cross section through the orogen reveals that all of these tectonic zones are separated by major fault systems (Fig. 3b). Most of

these are dominated by thrust faults that accommodated shortening of the continental crust during collision. Geologists have discovered that the boundary between the Tibetan Zone and the Greater Himalayan Zone is marked by normal faults of the South Tibetan detachment system. Unlike thrust faults, normal faults thin and extend the crust, and identification of them in a continent-continent collisional setting like the Himalayan orogen was surprising at first. Many Himalayan geologists have come to believe that the South Tibetan detachment system developed when the range grew high enough so that the buttressing effects of plate convergence were no longer capable of supporting the weight and the orogen began to collapse, spreading out the load laterally over a larger region.

*Other effects.* Although mountain belts are the most conspicuous products of orogeny, the effects of plate convergence—especially continent-continent collision—can be found hundreds or even thousands of miles away from the orogenic belt. For example, the Tibetan Plateau, which lies just north of the Himalayan orogen, has an average elevation of about 15,000 ft (5000 m) over an area about the size of France. Development of the plateau was undoubtedly related to India-Asia collision, but the exact age and mechanism of plateau uplift remain controversial. Large-displacement strike-slip faults (where one block of continental crust slides laterally past another) are common in central and southeast Asia, and some geologists hypothesize that they extend from the surface down to at least the base of the crust, segmenting this part of Asia into a series of rigid blocks that have extruded eastward as a consequence of the collision. Other geologists, while recognizing the significance of these faults for upper crustal deformation, suggest that the upper and lower crust of Tibet are detached from one another and that the eastward ductile flow of Tibetan lower crust has been an equally important factor in plateau growth.

**Topographic expression.** Earthquake activity in Bhutan, northern India, Nepal, northern Pakistan, and southern Tibet demonstrates that the Himalayan orogen continues to evolve, and this is principally why the Himalayas include such high mountains. As convergence across an orogen ends, erosion and structural denudation begin to smooth out the topographic gradient produced by orogeny. Rugged and high ranges such as the Himalayas are young and still active; lower ranges with subdued topography, such as the Appalachians, are old and inactive. Some of the oldest orogenic belts on Earth, such as those exposed in northern Canada, have no remaining topographic expression and are identified solely on the basis of geologic mapping of characteristic orogenic zones such as those discussed above for the Himalayas. Mountain ranges such as the Himalayas were formed and eventually eroded away throughout much of Earth's history; it is the preservation of ancient orogenic belts that provides some of the most reliable evidence that plate tectonic processes have been operative on Earth for at least 2 billion years.

**Ancient geological processes.** One of the most important axioms in the earth sciences is that the present is the key to the past (law of uniformitarianism). Earth is a dynamic planet; all oceanic lithosphere that existed prior to about 200 million years ago has been recycled into the mantle. The ocean basins thus contain no record of the first 96% of Earth history, and geologists must rely on the continents for this information. The careful study of modern orogenic belts provides a guide to understanding orogenic processes in ancient orogens, and permits use of the continents as a record of the interactions between lithospheric plates that have long since been destroyed completely at convergent plate boundaries.

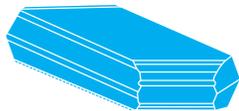
Many of the basic geologic features of modern mountain systems such as the Himalayas can be found in ancient orogens, and geologists have established generic names for the different parts of orogens to facilitate comparisons. The Foredeep contains sediments transported from the rising mountain chain; in the Himalayan orogen, it corresponds to the more southerly portions of the Siwalik Zone (Fig. 3). The Foreland Fold and Thrust Belt (northern Siwalik Zone and Lesser Himalayan Zone) is a complexly deformed sequence of sedimentary and crystalline rocks that were scraped off the upper portion of the down-going plate during collision. The Orogenic Core comprises the central part of an orogen; it may include large tracts of crust that experienced deformation at deep structural levels (Greater Himalayan Zone), rock sequences that were deformed at higher structural levels but are now juxtaposed with metamorphic rocks by movement on normal fault systems such as the South Tibetan detachment (Tibetan Zone), and a discontinuous belt of ophiolites and related materials marking the zone of plate convergence *sensu stricto* (Indus Suture Zone). Ophiolite belts can be viewed as the fossil remnants of oceans that no longer exist. The presence of ophiolite belts of various ages within the continents means that a map of the modern distribution of lithospheric plates provides only a recent snapshot of an ever-changing configuration of lithospheric plates on Earth: mountain ranges such as the Himalayas were formed and eventually eroded away throughout much of Earth's history. Exactly how far back in geologic time plate tectonic processes and their orogenic consequences can be extrapolated is a matter of debate among geologists, but the bulk of evidence suggests that mountain ranges such as the Himalayas provide useful analogs for even the oldest orogens on Earth. The preservation of ophiolite belts as old as 2 billion years in Precambrian orogenic belts provides direct evidence that plate tectonic processes have been operative on Earth for at least the past 50% of Earth history. See CORDILLERAN BELT; MOUNTAIN SYSTEMS; OPHIOLITE.

Kip Hodges

**Bibliography.** K. C. Condie, *Plate Tectonics*, 4th ed., 1997; E. M. Moores (ed.), *Shaping the Earth: Tectonics of Continents and Oceans*, 1990; E. M. Moores and R. J. Twiss, *Tectonics*, 1995; J. T. Wilson (ed.), *Continents Adrift and Continents Aground*, 1996.

## Orpiment

A mineral having composition  $As_2S_3$  and crystallizing in the monoclinic system. Crystals are small, tabular, and rarely distinct (see **illus.**); the mineral



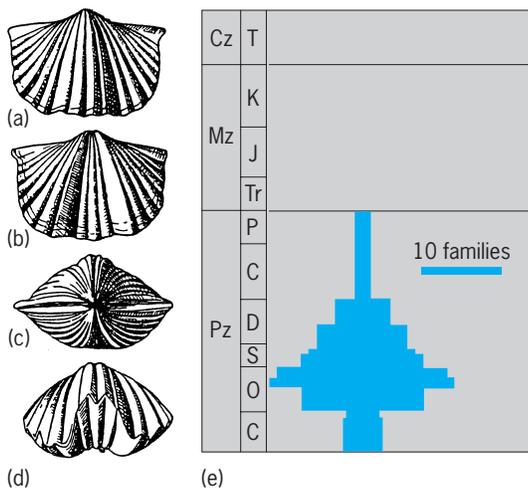
Orpiment crystal habit. (After C. S. Hurlbut, Jr., *Dana's Manual of Mineralogy*, 17th ed., John Wiley and Sons, 1959)

occurs more commonly in foliated or columnar masses. There is one perfect cleavage yielding flexible folia which distinguishes it from other minerals similar in appearance. The hardness is 1.5–2 (Mohs scale) and the specific gravity is 3.49. The luster is resinous and pearly on the cleavage surface; the color is lemon yellow. Orpiment is associated with realgar and stibnite in veins of lead, silver, and gold ores. It is found in Romania, Peru, Japan, and Russia. In the United States it occurs at Mercer, Utah; Manhattan, Nevada; and in deposits from geyser waters in Yellowstone National Park. See ARSENIC; REALGAR; STIBNITE. Cornelius S. Hurlbut, Jr.

## Orthida

An extinct order of brachiopods that lived in shallow-shelf seas during the Paleozoic. It contains the oldest members of the Rhynchonelliformea and is the stem group for all other brachiopods in this subphylum.

Orthids are characterized by unequally biconvex valves containing distinct ribs and a strophic (straight) hinge line with associated well-developed interareas (**illus.** a–d). Openings in the interareas



Orthida: (a) dorsal, (b) ventral, (c) posterior, and (d) anterior views of *Platystrophia*, Ordovician; (e) diversity history of orthid families. (a–d after R. C. Moore, C. G. Lalicker, and A. G. Fischer, *Invertebrate Fossils*, McGraw-Hill, New York, 1952; e after J. J. Sepkoski, Jr., *A Compendium of Fossil Marine Animal Families*, 2d ed., Milwaukee Public Mus. Contrib. Biol. Geol., no. 83, 1992)

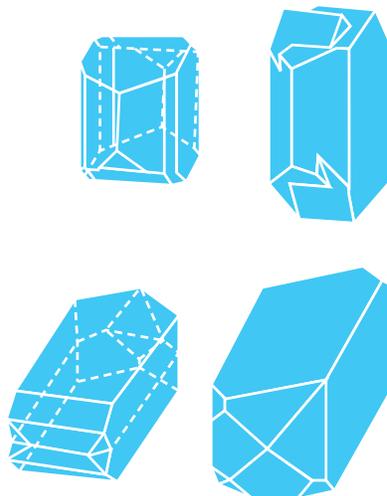
provided space for a fleshy protrusion (pedicle) that was used for attachment to the substrate. One group, the billingsellids, possesses a calcareous structure (pseudodeltidium) that partially covers the pedicle opening; it is a key character suggesting that this orthid group was ancestral to the Strophomenida. All other orthids lack a pseudodeltidium and are the group from which all the rest of the rhynchonelliform brachiopods arose, either directly or indirectly.

Orthids were sessile, attached, epifaunal suspension feeders and were important members of marine epifaunal communities in the Paleozoic. They arose in the Early Cambrian but remained low in diversity and were only minor constituents of Cambrian marine epifaunal communities (**illus.** e). Beginning in the Early Ordovician, orthid brachiopods began to increase dramatically in diversity, taking part in the great Ordovician radiation of marine animal life. During this taxonomic diversification, orthids also became the dominant member of marine epifaunal communities and helped to establish rhynchonelliform brachiopods as one of the main components of the Paleozoic evolutionary fauna. Orthids suffered high extinction in the Late Ordovician mass extinction and never returned to their earlier Ordovician diversity levels. Their importance in marine epifaunal communities also changed as a result of the Late Ordovician mass extinction. Their overall importance decreased mostly at the expense of other brachiopod orders that diversified, such as the strophomenids, pentamerids, and spiriferids. Orthid diversity decreased again as a result of the Late Devonian mass extinction, and orthids were finally extinguished in the late Permian mass extinction. See BRACHIOPODA; ARTICULATA (ECHINODERMATA). Mark E. Patzkowsky

Bibliography. D. A. T. Harper and R. Moran, Brachiopod life styles, *Geol. Today*, 13:235–238, 1997; J. R. Richardson, Brachiopods, *Sci. Amer.*, 255:100–106, 1986; M. J. S. Rudwick, *Living and Fossil Brachiopods*, Hutchinson, London, 1970; A. Williams et al., A supra-ordinal classification of the Brachiopoda, *Phil. Trans. Roy. Soc. London B*, 351:1171–1193, 1996.

## Orthoclase

Potassium feldspar ( $Or = KAlSi_3O_8$ ) that usually contains up to 30 mole % albite ( $Ab = NaAlSi_3O_8$ ) in solid solution. Its hardness is 6; specific gravity, 2.57–2.5, depending on Ab content; mean refractive index, 1.52; color, white to dull pink or orange-brown; good (001) and (010) cleavages at  $90^\circ$  to one another. Some orthoclases may be intergrown with relatively pure albite which exsolved during cooling from a high temperature in pegmatites, granites, or granodiorites. This usually is ordered low albite, but in rare cases it may show some degree of Al,Si disorder, requiring it to be classified as analbite or high albite. If exsolution is detectable by eye, the Or-Ab composite mineral is called perthite; if microscopic examination is required to distinguish the phases, it is called microperthite; and if exsolution is detectable



Orthoclase crystal habits. (After C. Klein and C. S. Hurlbut, Jr., *Manual of Mineralogy*, 21st ed., John Wiley and Sons, 1993)

only by x-ray diffraction or electron optical methods, it is called cryptoperthite.

Orthoclase is optically monoclinic, ostensibly with  $C2/m$  space group and partially ordered Al,Si distribution in the two tetrahedral sites of the feldspar structure. In fact, electron microscopic examinations show most orthoclases to have complex, strained structures that might be considered to represent an intermediate stage in the sluggish Al,Si ordering process between monoclinic sanidine (with its highly disordered Al,Si distribution) and triclinic low microcline (with its nearly completely ordered Al,Si distribution). It has even been described as having a unit-cell-scale, pseudotwinned microstructure. Thus its structure averaged over hundreds of nanometers may be monoclinic, but its true symmetry is triclinic (see *illus.*). See ALBITE; CRYSTAL STRUCTURE; FELDSPAR; IGNEOUS ROCKS; MICROCLINE; PERTHITE; SOLID SOLUTION. Paul H. Ribbe

## Orthogonal polynomials

A special case of orthogonal functions that arise in many physical problems (often as the solutions of differential equations), in the study of distribution functions, and in certain other situations where one approximates fairly general functions by polynomials. See PROBABILITY.

Each set of orthogonal polynomials is defined with respect to a particular averaging procedure. The average value of a suitable function  $f$  is denoted by  $E\{f\}$ . Examples are shown in Eqs. (1)-(4).

$$E\{f\} = \frac{1}{2} \int_{-1}^1 f(x) dx \quad (1)$$

$$E\{f\} = \frac{\int_{-1}^1 f(x) (1-x)^\alpha (1+x)^\beta dx}{\int_{-1}^1 (1-x)^\alpha (1+x)^\beta dx} \quad (2)$$

$$E\{f\} = \int_0^\infty f(x) e^{-x} dx \quad (3)$$

$$E\{f\} = (2\pi)^{-1/2} \int_{-\infty}^\infty f(x) e^{-x^2} dx \quad (4)$$

In general an averaging procedure shown in Eq. (5), a Stieltjes integral, where  $\sigma$  is a distribution

$$E\{f\} = \int_{-\infty}^\infty f(x) d\sigma(x) \quad (5)$$

function, that is, an increasing function with  $\sigma(-\infty) = 0$  and  $\sigma(+\infty) = 1$ . In the above examples  $\sigma$  has the form

$$\sigma(x) = b^{-1} \int_{-\infty}^x \omega(y) dy$$

where  $\omega$  is a nonnegative weight function and

$$b = \int_{-\infty}^\infty \omega(y) dy$$

Consideration will be given only to averaging procedures for which all the moments

$$\mu_n = E\{x^n\} = \int_{-\infty}^\infty x^n d\sigma(x)$$

exist and for which  $E\{|P|\} > 0$  for every polynomial  $P$ .

**Orthogonal functions.** Two functions  $f$  and  $g$  are said to be orthogonal with respect to a given averaging procedure if  $E\{f\bar{g}\} = 0$  where the bar denotes complex conjugation. By the system of orthogonal polynomials associated with the averaging procedure is meant a sequence  $P_0, P_1, P_2, \dots$  of polynomials  $P_n$  having exact degree  $n$ , which are mutually orthogonal, that is,  $E\{P_m \bar{P}_n\} = 0$  for  $m \neq n$ . This last condition is equivalent to the statement that each  $P_n$  is orthogonal to all polynomials of degree less than  $n$ . Thus  $P_n$  has the form  $P_n(x) = a_0 + a_1x + a_2x^2 + \dots + a_nx^n$  where  $a_n \neq 0$  and is subject to the  $n$  conditions  $E\{x^k P_n\} = 0$  for  $k = 0, 1, \dots, n-1$ . This gives  $n$  linear equations in the  $n+1$  coefficients of  $P_n$ , leaving one more condition, called a normalization, to be imposed. The method of normalization differs in different references. Orthogonal polynomials arising from the average of Eq. (1), Legendre polynomials, satisfy Legendre's differential equation. With the normalization  $P_n(1) = 1$  the first few Legendre polynomials are  $P_0(x) = 1, P_1(x) = x, P_2(x) = 3/2x^2 - 1/2, P_3(x) = 5/2x^3 - 3/2x$ . The average in Eq. (1) is the special case of Eq. (2) with  $\alpha = \beta = 0$ ; the orthogonal polynomials corresponding to averages of Eq. (2) are called Jacobi polynomials; those associated with Eq. (3), Laguerre polynomials; with Eq. (4), Hermite polynomials.

The proper setting for the study of expansions in terms of orthogonal polynomials is the Hilbert space  $H$  of functions  $f$  such that  $E\{|f|^2\}$  exists and is finite. The inner product is  $(f, \bar{g}) = E\{f\bar{g}\}$ . In analogy with the procedure for Fourier series one can write down a formal expansion, relation (6), where the

$$f(x) \sim \sum_{n=0}^{\infty} c_n P_n(x) \quad (6)$$

coefficients are given by Eq. (7). The  $N$ th partial sum

$$c_n = \frac{E\{f\bar{P}_n\}}{E\{|P_n|^2\}} \quad (7)$$

of the series shown in Eq. (8) has the property that

$$s_N(x) = \sum_0^N c_n P_n(x) \quad (8)$$

among all polynomials  $p$  of degree not exceeding  $N$ , the minimum of the quadratic deviation  $E\{|f - p|^2\}$  is achieved uniquely by  $p = s_N$ . If the only function  $f$  in  $H$  with the property that  $E\{x^k f\} = 0$  for every  $k$  is the zero function, one says that the polynomials are "complete" in  $H$ . In this case the coefficients in Eq. (7) uniquely determine the function  $f$ , and the properties of the series in Eq. (6) are quite analogous to the properties of Fourier series of functions in  $L^2$ . The polynomials are always complete when the average is taken over a finite interval, but in general some extra assumption is required. The divergence of the series  $\sum \mu_{2n}^{-1/2n}$  is a sufficient condition for the completeness of the polynomials. (It is fulfilled in each of the examples cited.) See FOURIER SERIES AND TRANSFORMS.

The orthogonality property entails certain algebraic properties for the polynomials. For example, the zeros of  $P_n$  are all distinct, they lie in the interior of the interval over which the average is taken, and they separate the zeroes of  $P_{n-1}$ . Let  $X_1^{(n)}, \dots, X_n^{(n)}$  be the zeros of  $P_n$ . One can find constants  $b_1^{(n)}, \dots, b_n^{(n)}$  such that  $Q_n\{1\} = 1, Q_n\{P_k\} = 0$  for  $0 < k < n$ , where

$$Q_n\{f\} = \sum_{j=1}^n b_j^{(n)} f(X_j^{(n)})$$

In the case of an average over a finite interval,

$$\lim_{n \rightarrow \infty} Q_n\{f\} = E\{f\}$$

for every continuous  $f$ . This is of interest in approximate integration, because the integral  $E\{f\}$  is approximated by an expression  $Q_n\{f\}$  which depends only on the values of  $f$  at  $n$  points and, what is remarkable,  $Q_n\{f\} = E\{f\}$  whenever  $f$  is a polynomial of degree  $\geq 2n - 1$ , whereas one would ordinarily expect an  $n$ -point approximation to be exact only for polynomials of degree  $\leq n$ .

**Ultraspherical polynomials.** There exists a theory of orthogonal polynomials in several variables. The most important applications involve averages over spheres in  $m$  dimensions. Complete sets of orthogonal polynomials may be chosen among the homogeneous, harmonic polynomials. A polynomial  $P(x_1, \dots, x_m)$  is homogeneous of degree  $n$  if  $P(\lambda x_1, \dots, \lambda x_m) = \lambda^n P(x_1, \dots, x_m)$  for each  $\lambda$ ; it is harmonic if it satisfies Laplace's differential equation. Let  $P$  be such a polynomial with the property that  $P(1, 0, \dots, 0) \neq 0$ . Consider  $P(x, y_1, \dots, y_{m-1})$  as a polynomial in the  $m - 1$  variables  $y_1, \dots, y_{m-1}$  and take the average over a sphere centered at the origin in  $m - 1$  dimensions. The result is a polynomial  $P_n(x)$  of degree  $n$ . The orthogonality

over the sphere in  $m$  dimensions translates itself into orthogonality on the interval  $[-1, 1]$  with the weight function  $\omega(x) = (1 - x^2)^{(n-3)/2}$ . For fixed  $m$ , the polynomials obtained this way are the ultraspherical polynomials, special cases of the Jacobi polynomials with  $\alpha = \beta = (m - 3)/2$ . Where  $m = 3$  corresponds to three-dimensional space, these are Legendre polynomials. See DIFFERENTIAL EQUATION; LAPLACE'S DIFFERENTIAL EQUATION; POLYNOMIAL SYSTEMS OF EQUATIONS; RIEMANNIAN GEOMETRY.

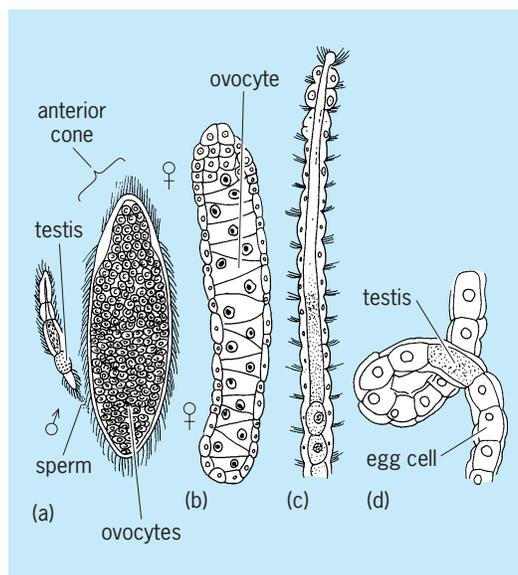
Carl S. Herz

**Bibliography.** R. Askey, *Orthogonal Polynomials and Special Functions*, 1975; T. S. Chihara, *An Introduction to Orthogonal Polynomials*, 1978; H. F. Davis, *Fourier Series and Orthogonal Functions*, 1963, reprint 1989; P. G. Nevai, *Orthogonal Polynomials*, reprint 1991; J. A. Shohat and J. D. Tamarkin, *The Problem of Moments*, Math. Surv. no. 1, 1943, reprint 1983; G. Szegő, *Orthogonal Polynomials*, 1978, reprint 1991.

### Orthonectida

An order of Mesozoa. The orthonectids parasitize various marine invertebrates as multinucleate plasmodia, sometimes causing considerable damage to the host. The plasmodia multiply by fragmentation. Eventually they give rise asexually, by polyembryony, to sexual males and females. Commonly only one sex arises from a given plasmodium.

These sexually mature forms escape as minute ciliated organisms. Structurally they are composed of a single layer of ciliated epithelial cells surrounding an inner mass of sex cells (see **illus.**). The ciliated cells are disposed in rings around the body. Those at the anterior end form the anterior cone, on which the cilia are directed forward, while on the rest of the



**Orthonectids.** (a) *Rhopalura ophiocomae*, male discharging sperm near genital pore of female. (b) *R. metschnikovi* germ cells oriented in a double row. (c) *Stochartum giardia*, anterior end, and (d) part of trunk showing one egg cell to each apparent segment.

body they point backward. Males are smaller than the corresponding females. A few species, however, are hermaphroditic.

After insemination the eggs develop in the female and form ciliated larvae. When they are liberated, these larvae invade new individuals of their host and then disaggregate, liberating germinal cells which give rise to new plasmodia. *See* MESOZOA; REPRODUCTION (ANIMAL).  
Bayard H. McConnaughey

## Orthoptera

An order that includes over 20,000 species of terrestrial insects, including most of the "singing" insects, some of the world's largest insects, and some well-known pests. Most species of Orthoptera (from "orthos," meaning straight, and "pteron," meaning wing) have enlarged hind legs adapted for jumping. These include grasshoppers and locusts (in the suborder Caelifera, a mainly diurnal group); and the crickets, katydids (bush-crickets), New Zealand weta, and allied families (suborder Ensifera, which comprises mainly nocturnal species). Orthoptera share with other orthopteroid insects, such as mantids and stick insects (now in separate orders Mantodea and Phasmatodea), gradual metamorphosis, chewing mouthparts, and two pairs of wings, the anterior pair of which is usually thickened and leathery and covers the fanwise folded second pair. Wings are reduced or absent in many species. Characters that define the Orthoptera as a natural group (the inclusive set of all species stemming from a common ancestor) are the jumping hind legs, small and well-separated hind coxae (basal leg segments), a pronotum with large lateral lobes, and important molecular (genetic) characters. Food habits range from omnivorous to strictly carnivorous or herbivorous. Habitats are nearly all terrestrial, including arctic-alpine tundra and tropical areas with aquatic floating plants. Female orthopterans usually lay their eggs into soil or plant material. There are no parasitic species, but a few crickets live as cleptoparasitic "guests" in ant nests.

**Stridulation and mating.** Males of many species are outstandingly noisy or musical. Their songs typically function in obtaining mates. In some species females move to the singing male, and in others a female answering song is involved in pair formation. Males may interact with bouts of singing until one or the other moves away. Song frequencies can range from the audible to the very high ultrasonic. Song patterns typically consist of a complex species-specific series of clicks, buzzes, or musical chirps produced by stridulation, the rubbing of a movable bowlike structure over a precisely arranged series of pegs or ridges. Grasshoppers (Acridoidea) stridulate by rubbing the hind femora against the outer wings, whereas crickets (Gryllidae), katydids (Tettigoniidae), and humped-winged crickets (Haglidae) rapidly rub the forewings together. Some wingless species stridulate with the overlapping edges of abdominal segments or the mandibles. There are at least 20 different mod-

ifications of surfaces for acoustical communication, including fanlike snapping of brightly colored hindwings in flight (some grasshoppers).

Hearing organs are similarly diverse in form and location. In the grasshoppers, the first abdominal spiracle is modified as a tympanum, whereas the crickets, katydids, and haglids have small tympanic membranes and acoustic receptor cells set into the front tibiae which are connected via the tracheal system to a specialized thoracic spiracle. There are many species of grasshoppers and crickets, however, that lack specialized sound-producing and -receiving organs. Signal communication in these species may be tactile, visual, or olfactory. *See* ANIMAL COMMUNICATION; PHONORECEPTION.

Sound production and later stages of mating behavior have been the focus of much behavioral research. In particular, male orthopterans are known for their nuptial gifts, which are consumed by females. Examples are specialized expendable wings and leg spines, edible substances attached to sperm packages (spermatophores), and dorsal gland secretions of many katydids and crickets, and male glandular secretions absorbed in the female reproductive tract in some grasshoppers. Such donations by males can be costly and can mediate sex-role reversals when females compete to mate and acquire important nuptial offerings from males. *See* REPRODUCTIVE BEHAVIOR.

**Suborder Ensifera.** The first ensiferans appear in the fossil record in the Permian. Three superfamilies are commonly recognized in this natural group: Grylloidea (true crickets and allies), Tettigonioidae (katydids, haglids, and allies), and Gryllacridoidea (camel crickets and allies). Based on analyses of morphological and anatomical characters showing only two natural groups of ensiferans, some authors recognize only the first two superfamilies. Ensiferan antennae are usually longer than the body. Most species are nocturnal, and the night-time stridulation of katydids, crickets, and weta can be very loud, especially when males chorus. The ovipositor is long and sword-shaped, and often bores holes in soft wood, bark, living plant stems, or hard-packed soil to lay eggs singly.

Tettigoniidae (katydids and bush-crickets) and Gryllidae (true crickets) are the most widespread and diverse families. Many species in both families share with acridoidea the completion of the life cycle (egg to adult) in one year, and the use of microhabitats in vegetation (many other ensiferan species have life cycles longer than a year and use burrows and crevices as microhabitats). Many katydids are cryptically shaped and colored, some so similar to leaves that decay spots and insect-chewed margins are mimicked. A major western North American pest is the flightless Mormon cricket (see **illustration**), dense groups of which can reach several kilometers in length and cover many miles, damaging crops along the way.

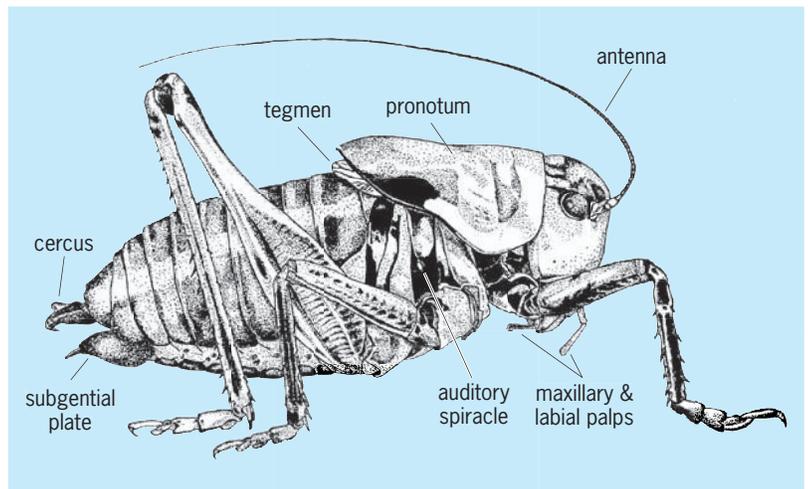
True crickets (Gryllidae) are ground-, tree-, or bush-dwelling, relatively chunky insects with a needle-shaped ovipositor. Members of the genera *Acheta*, *Teleogryllus*, and *Gryllus* are easily reared,

important laboratory animals, and a standard bait and food item for vertebrates. Although they can occur in large populations, they are not usually serious pests. However, some mole crickets (Gryllotalpidae) are serious pests of tropical crops.

Other ensiferan families include many fossil forms and some bizarre living forms, most of which are flightless. The family Anostostomatidae includes the giant weta of New Zealand (large specimens can be 30–40 grams, or 1.1–1.4 oz, in weight) and related, sexually dimorphic weta and king crickets of New Zealand, Australia, and South Africa, males of which fight using tusks or long mandibles. The North American Jerusalem crickets (Stenopelmatidae) are a diverse group of large-headed desert-inhabiting species. The camel and cave crickets, including cave weta of New Zealand (Rhaphidophoridae), are common in caves and damp forests of most continents. Most of these ensiferans are wingless. In contrast are the hump-winged crickets (Haglidae), which comprise a few species inhabiting the forests and high sage of the North American Rockies and the mountains of temperate Asia. One species, the only nocturnal singing insect at high altitudes, stridulates at the lowest ambient temperature known for any insect,  $-2^{\circ}\text{C}$  ( $28^{\circ}\text{F}$ ).

**Suborder Caelifera.** This is a natural group that shares a number of characters, including an ovipositor with only four functional valves at the end of the abdomen. In many species, egg laying is accomplished by working the abdomen slowly into the soil and laying eggs in groups. In some grasshoppers the eggs are surrounded by a foamy matrix. Other characteristics that define Caelifera as a natural group are antennae composed of less than 30 segments (being typically less than half the length of the body) and certain genetic characters. There are eight superfamilies: Tridactyloidea (false and pygmy mole crickets and sand gropers); Tetrigoidea (pygmy grasshoppers and grouse locusts); Eumastacoidea (monkey grasshoppers and false stick insects); Pneumoroidea (flying gooseberries, desert longhorned grasshoppers, and razor-back bushhoppers); Pamphagoidea (rugged earth hoppers and true bushhoppers); Acridoidea (grasshoppers and locusts); and Trigonopterygoidea.

Most caeliferans are diurnal. The most familiar and diverse are short-horned grasshoppers (Acridoidea) which tend to be active in bright sunshine. Some species are cryptic, mimicking stones, debris, and grass stems. Although most species of grasshoppers are never abundant enough to become agricultural pests, some have little-understood cycles of abundance that sometimes phase together to become major crop-destroying events, especially in drier regions of North America, Africa, and Australia. In some years, plague locusts such as *Locusta migratoria*, *Schistocerca gregaria*, and others migrate hundreds of miles from an outbreak center to devastate thousands of acres of crops in the tropics and subtropics. Remarkable structural and behavioral changes occur from the solitary generation to the migratory generation. Several North American species



Male Mormon cricket, *Anabrus simplex*, showing some key morphological features of Orthoptera. (Reprinted from D. T. Gwynne, *Katydid and Bush-Crickets: Reproductive Behavior and Evolution of the Tettigoniidae*. Copyright © 2001 by Cornell University. Used by permission of the publisher, Cornell University Press.)

of *Melanoplus* have a latent ability to shift to a migratory phase, though natural outbreaks have occurred only historically, in a species which is now thought to be extinct. The plague locusts are large, hardy insects, extensively used for neurophysiological and general physiological studies.

Pygmy mole crickets and sand gropers (Tridactyloidea) have evolved specialized front legs for digging that resemble the front legs of true mole crickets. Some burrowing sand gropers can be pests of crop fields in Australia. Pneumoroids include the remarkable bladder grasshoppers of Africa, which lack a tympanum in the abdominal ear yet the very loud male stridulation can be detected by females at a distance of almost 2 km (1.2 mi). See INSECT PHYSIOLOGY; INSECTA. Darryl T. Gwynne; Robert B. Willey

**Bibliography.** J. L. Capinera, R. D. Scott, et al., *Field Guide to Grasshoppers, Katydid, and Crickets of the United States*, Cornell University Press, Ithaca, NY, 2004; R. F. Chapman and A. Joern (eds.), *The Biology of Grasshoppers*, Wiley, 1990; L. H. Field (ed.), *The Biology of Wet, King Crickets and their Allies*, CABI International, Wallingford, 2001; S. K. Gangwere, M. C. Mulalirangen, and M. Mulalirangen (eds.), *The Bionomics of Grasshoppers, Katydid and Their Kin*, CAB International, Oxford, 1997; D. T. Gwynne, *Katydid and Bush-crickets: Reproductive Behavior and Evolution of the Tettigoniidae*, Cornell University Press, Ithaca, NY, 2001; J. A. Marshall and E. C. M. Haes, *Grasshoppers and Allied Insects of Great Britain and Ireland*, Harley Books, Colchester, 1988; D. C. Rentz, *Grasshopper Country: The Abundant Orthopteroid Insects of Australia*, University of New South Wales Press, Sydney, 1996.

### Orthorhombic pyroxene

A group of minerals having the general chemical formula  $\text{XYSi}_2\text{O}_6$ , in which the Y site contains iron (Fe) or magnesium (Mg) and the X site contains Fe,

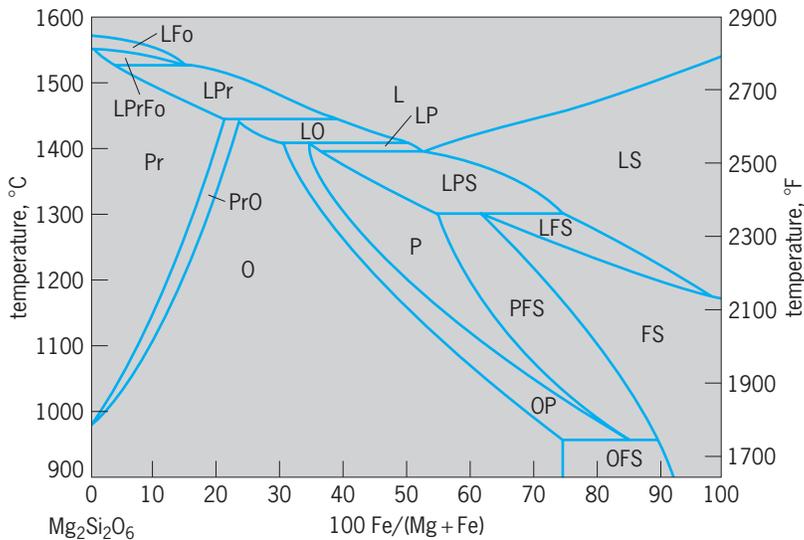


Fig. 1. Temperature-composition relations at low pressure for the system  $Mg_2Si_2O_6$ - $Fe_2Si_2O_6$ . O = orthorhombic phase; Pr = protoenstatite phase; P = monoclinic phase; Fo = magnesian olivine; F = iron-rich olivine; S = silica; L = melt. (After C. T. Prewitt, ed., *Pyroxenes, Reviews in Mineralogy*, vol. 7, Mineralogical Society of America, 1980)

Mg, manganese (Mn), or a small amount of calcium (Ca; up to about 3%). They are characterized by orthorhombic crystal symmetry, which makes it impossible for significant amounts of larger cations such as Ca or sodium (Na) to enter the X position. The most important compositional variation in the orthopyroxene series is variation in the Fe-Mg ratio, leading to a solid solution series bounded by the end members enstatite ( $Mg_2Si_2O_6$ ; abbreviated En) and

ferrosilite ( $Fe_2Si_2O_6$ ; Fs). Names that are used for intermediate members of the series are enstatite ( $Fs_{0-50}$ ) and ferrosilite ( $Fs_{50-100}$ ). More Fs-rich enstatite may be called ferroan enstatite, and more enstatite-rich ferrosilite may be called magnesian ferrosilite. In addition to the major elements noted above, minor components in orthopyroxene may include aluminum (Al), titanium (Ti), chromium (Cr), nickel (Ni), and ferric iron ( $Fe^{3+}$ ), although only Al occurs in substantial amounts. Conditions of formation may be important in determining orthopyroxene compositions: for example, Al content is greater at higher temperatures, and the most Fe-rich compositions are stable only at very high pressures. Further, there is a limited solid solution of orthopyroxene toward calcic clinopyroxene, noted above, which increases with increasing temperature. See ENSTATITE; SOLID SOLUTION.

**Crystal structure.** Orthopyroxenes crystallize in a primitive orthorhombic cell, space group Pbc1 (a high-temperature Mg-rich form called protoenstatite has slightly different symmetry and is in space group Pbcn). The *b* and *c* unit-cell dimensions are about the same as for monoclinic pyroxene (such as diopside), but the *a* dimension is approximately doubled. This suggests that the absence of large cations in orthopyroxene causes a slight structural shift so that the smallest repeated unit contains twice the number of atoms as in a monoclinic pyroxene unit cell. The crystal chemistry is otherwise similar to that of diopside, with single chains of  $(SiO_4)$  tetrahedra alternating in the structure with strips of octahedral sites that accommodate larger cations. As in monoclinic pyroxene, Al is small enough to enter both octahedral and tetrahedral sites. See CRYSTAL STRUCTURE; DIOPSIDE.

**Properties.** Many of the physical and optical properties of orthopyroxene are strongly dependent upon composition, and especially upon the Fe-Mg ratio. Magnesium-rich enstatite ( $Fs_{0-20}$ ) is commonly pale brown in color; the color becomes much darker for more Fe-rich members of the series. The specific gravity and refractive index both increase markedly with increasing Fe content. Enstatite in the range of  $Fs_{10-30}$  (formerly called bronzite) commonly contains a bronze iridescence that is due either to very fine exsolved lamellae of calcic monoclinic pyroxene that diffract light rays (similar to blue-green iridescence of labradorite plagioclase feldspar) or to very thin, translucent, oriented plates of ilmenite, an iron-titanium oxide. In hand specimens, orthopyroxene can be distinguished from amphibole by its characteristic  $88^\circ$  cleavage angles, and from augite by color—augite is typically green to black, while orthopyroxene is more commonly brown, especially on slightly weathered surfaces. In rock thin sections, orthopyroxene is usually recognized by its parallel extinction and blue-green to pink pleochroism. See PETROGRAPHY.

**Experimental phase relations.** A considerable body of experimental data exists on the pressure-temperature (*P-T*) relations of orthorhombic pyroxenes, both for crystallization from silicate

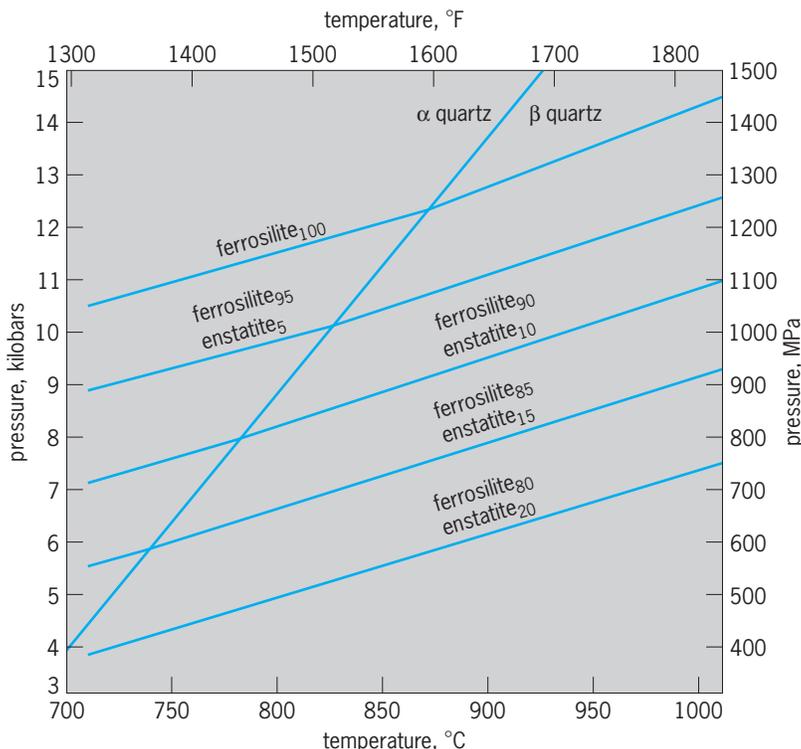


Fig. 2. Pressure-temperature relations of iron-rich orthopyroxenes. Contours show stability of ferrosilite with 0, 5, 10, 15, and 20 mol% of enstatite ( $MgSiO_3$ ). (After S. R. Bohlen and A. L. Boettcher, *Experimental investigations and geological applications of orthopyroxene geobarometry*, *Amer. Mineralog.*, 66:951-964, 1981)

melts and in the subsolidus regions. The low-pressure thermal behavior of phases in the system  $\text{Mg}_2\text{Si}_2\text{O}_6\text{-Fe}_2\text{Si}_2\text{O}_6$  is notably pseudobinary, as reflected by the wide variability of the silica content of the liquids with different Fe-Mg ratios, by the incongruent melting of enstatite to Mg-olivine plus siliceous melt, and by the subsolidus breakdown of ferrosilite to Fe-olivine plus silica (**Fig. 1**).

Although orthopyroxene more Fe-rich than  $\text{Fs}_{75}$  is not stable at low pressure (**Fig. 1**), higher pressures stabilize these pyroxenes. Pure  $\text{Fe}_2\text{Si}_2\text{O}_6$  is stable above about 10 kilobars (100 megapascals), and the stability of orthopyroxene increases markedly with solid solution of small amounts of  $\text{Mg}_2\text{Si}_2\text{O}_6$  (**Fig. 2**). In fact, accurate experimental calibration makes Fe-rich orthopyroxene very useful for geobarometry if the composition is known and the temperature can be independently determined.

Aluminum is essentially the only minor element whose effect on pyroxene phase relations has been studied. Investigation of the  $\text{Al}_2\text{O}_3$  content of enstatite coexisting with forsterite and spinel has demonstrated that pyroxene alumina content increases with increased temperature, and is almost independent of pressure. Aluminous enstatites in spinel peridotites may therefore be very useful as geothermometers. *See* GEOLOGIC THERMOMETRY.

**Occurrence.** Orthopyroxene is a widespread constituent of both igneous and metamorphic rocks. It is common in peridotites of the Earth's upper mantle (found as nodules in kimberlites or alkali basalts), where it coexists with olivine, augite, and spinel or garnet. Orthopyroxene of intermediate Fe/Mg occurs in many basalts and gabbros and in meteorites, but is notably absent from most alkaline igneous rocks. Ferroan enstatite or magnesian ferrosilite is relatively common in intermediate volcanic and plutonic rocks such as andesite, dacite, and diorite, but rarer in more silicic rocks such as rhyolite and granite. Enstatite or ferroan enstatite is an essential constituent of the type of gabbro called norite, in which augite is subordinate to orthopyroxene in abundance. *See* LITHOSPHERE.

Orthopyroxene occurs in the highest concentration in ultramafic rocks, especially the peridotites and harzburgites of the large layered intrusions such as the Bushveld Complex of southern Africa or the Stillwater Complex of Montana. In these complexes enstatite crystallized from the magma early, along with olivine, and settled through the less dense magma to form cumulate rocks at the bottom of the intrusion. Layered complexes may contain hundreds to thousands of meters of these orthopyroxene-rich cumulate rocks. *See* IGNEOUS ROCKS; MAGMA.

Orthopyroxene of intermediate to high Fs content is a characteristic mineral in granulite-facies metamorphic rocks; granulite-facies terranes are the typical locales for application of geobarometry using compositions of Fs-rich orthopyroxene. In feldspathic rocks such as charnockite, orthopyroxene occurs with garnet, augite, hornblende, alkaic feldspars, and quartz; in mafic gneisses it coexists with augite, hornblende, garnet, and cal-

ic plagioclase. In very high-grade metamorphic rocks, especially in low-pressure contact metamorphism, the dehydration of biotite commonly results in a coexistence of orthopyroxene and potassic feldspar. Magnesian orthopyroxene is a typical constituent of medium- to high-grade metamorphosed ultramafic rocks, in which it coexists with olivine, spinel, chlorite, and amphiboles. *See* FACIES (GEOLOGY); METAMORPHIC ROCKS; MINERALOGY; PYROXENE.

Robert J. Tracy

**Bibliography.** W. A. Deer, R. A. Howie, and J. Zussman, *Rock-Forming Minerals*, Vol. 2B, Double-Chain Silicates, 2d ed., 1997; W. A. Deer, R. A. Howie, and J. Zussman, *Rock-Forming Minerals*, vol. 2A: *Single Chain Silicates*, 1978; C. T. Prewitt (ed.), *Pyroxenes*, *Reviews in Mineralogy*, vol. 7, 1980.

## Orthotrichales

An order of the true mosses (subclass Bryidae) consisting of five families and 23 genera. The plants grow in mats or tufts in relatively exposed places, on trunks of trees and on rock. They are rather freely branched and prostrate, or may be sparsely forked and erect-ascending; the habit is more or less pleurocarpous, but sporophytes may be produced at the ends of leading shoots or branches. The leaves are generally oblong and broadly pointed, with a strong midrib. The upper cells are short and often papillose, while the basal cells are longer and often pellucid. The capsules are generally immersed and often ribbed, and the peristome, if present, has a poor development of endostome. The calyptrae are nearly always mitrate and often hairy. The Grimmiiales, likewise somewhat pleurocarpous, differ in having single peristomes and a lesser development of neck tissue below the spore-bearing portion of the capsule. *See* BRYIDAE; BRYOPHYTA; BRYOPSIDA; GRIMMIALES.

Howard Crum

## Osage orange

The genus *Maclura* of the mulberry family, with one species, *M. pomifera*. This tree may attain a height of 60 ft (18 m) and has yellowish bark, milky sap, simple entire leaves, strong axillary thorns (see **illus.**), and



Branch and leaf of *Maclura pomifera*.

aggregate green fruit about the size and shape of an orange. It is planted for hedges and as an ornament, especially in the eastern United States where it is naturalized. The wood is used for fence posts and fuel and as a source of a yellow dye. It has also been used for archery bows, hence one of its common names, bowwood. *See* FOREST AND FORESTRY; TREE.

Arthur H. Graves; Kenneth P. Davis

## Oscillation

Any effect that varies in a back-and-forth or reciprocating manner. Examples of oscillation include the variations of pressure in a sound wave and the fluctuations in a mathematical function whose value repeatedly alternates above and below some mean value.

The term oscillation is for most purposes synonymous with vibration, although the latter sometimes implies primarily a mechanical motion. A device designed to reduce a person's weight by shaking part of the body is likely to be called a vibrator, whereas an electronic device that produces an electric current which reverses its direction periodically is usually called an oscillator. The alternating current and the associated electric and magnetic fields are referred to as electric (or electromagnetic) oscillations.

If a system is set into oscillation by some initial disturbance and then left alone, the effect is called a free oscillation. A forced oscillation is one in which the oscillation is in response to a steadily applied periodic disturbance.

Any oscillation that continually decreases in amplitude, usually because the oscillating system is sending out energy, is spoken of as a damped oscillation. An oscillation that maintains a steady amplitude, usually because of an outside source of energy, is undamped. *See* ANHARMONIC OSCILLATOR; DAMPING; FORCED OSCILLATION; HARMONIC OSCILLATOR; MECHANICAL VIBRATION; OSCILLATOR; VIBRATION.

Joseph M. Keller

## Oscillator

An electronic circuit that generates a periodic output, often a sinusoid or a square wave. Oscillators have a wide range of applications in electronic circuits: they are used, for example, to produce the so-called clock signals that synchronize the internal operations of all computers; they produce and decode radio signals; they produce the scanning signals for television tubes; they keep time in electronic wristwatches; and they can be used to convert signals from transducers into a readily transmitted form.

Oscillators may be constructed in many ways, but they always contain certain types of elements. They need a power supply, a frequency-determining element or circuit, a positive-feedback circuit or device (to prevent a zero output), and a nonlinearity (to define the output-signal amplitude). Different choices for these elements give different oscillator circuits

with different properties and applications.

Oscillators are broadly divided into relaxation and quasilinear classes. Relaxation oscillators use strong nonlinearities, such as switching elements, and their internal signals tend to have sharp edges and sudden changes in slope; often these signals are square waves, trapezoids, or triangle waves. The quasilinear oscillators, on the other hand, tend to contain smooth sinusoidal signals because they regulate amplitude with weak nonlinearities. The type of signal appearing internally does not always determine the application, since it is possible to convert between sine and square waves. Relaxation oscillators are often simpler to design and more flexible, while the nearly linear types dominate when precise control of frequency is important.

### Relaxation Oscillators

Figure 1a shows a simple operational-amplifier based relaxation oscillator. This circuit can be understood in a number of ways (for example, as a negative-resistance circuit), but its operation can be followed by studying the signals at its nodes (Fig. 1b). The two resistors, labeled  $r$ , provide a positive-feedback path that forces the amplifier output to saturate at the largest possible (either positive or negative) output voltage. If  $v_+$ , for example, is initially slightly greater than  $v_-$ , then the amplifier action increases  $v_o$ , which in turn further increases  $v_+$  through the two resistors labeled  $r$ . This loop continues to operate, increasing  $v_o$  until the operational amplifier saturates at some value  $V_{\max}$ . [An operational amplifier ideally follows Eq. (1), where  $A_v$  is very large,

$$v_o = A_v(v_+ - v_-) \quad (1)$$

but is restricted to output levels  $|v_o| \leq V_{\max}$ .] For the purposes of analyzing the circuit, the waveforms in Fig. 1b have been drawn with the assumption that this mechanism has already operated at time 0 and

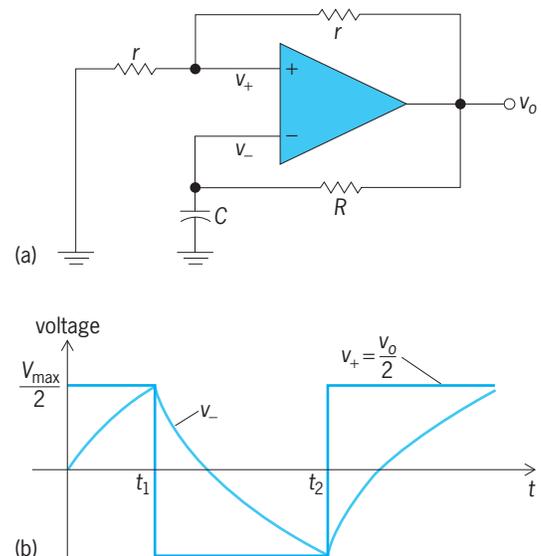


Fig. 1. Simple operational-amplifier relaxation oscillator. (a) Circuit diagram. (b) Waveforms.

that the initial charge on the capacitor is zero. See AMPLIFIER; OPERATIONAL AMPLIFIER.

Capacitor  $C$  will now slowly change from  $v_o$  through resistor  $R$ , toward  $V_{\max}$ , according to Eq. (2). Up until time  $t_1$ , this process continues without any

$$v_- = V_{\max}(1 - e^{-t/RC}) \quad (2)$$

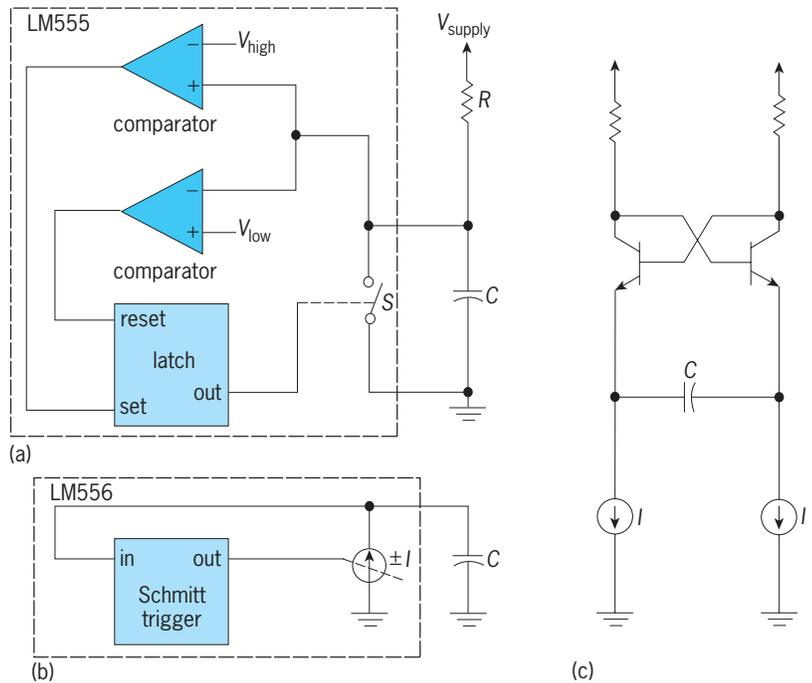
change in the amplifier's output because  $v_+ > v_-$ , and so  $v_o = V_{\max}$ . At  $t_1$ , however,  $v_+ = v_-$  and  $v_o$  will start to decrease. This causes  $v_+$  to drop, and the positive-feedback action now drives the amplifier output negative until  $v_o = -V_{\max}$ . Capacitor  $C$  now discharges exponentially toward the new output voltage until once again, at time  $t_2$ ,  $v_+ = v_-$ , and the process starts again. The period of oscillation for this circuit is  $2RC \ln 3$ .

The basic elements of an oscillator that were mentioned above are all clearly visible in this circuit. Two direct-current power supplies are implicit in the diagram (the operational amplifier will not work without them), the  $RC$  circuit sets frequency, there is a resistive positive-feedback path that makes the mathematical possibility  $v_o(t) = 0$  unstable, and the saturation behavior of the amplifier sets the amplitude of oscillation at the output to  $\pm V_{\max}$ .

**Practical oscillators.** While the circuit of Fig. 1 is easy to understand and to experiment with, it does have some limitations because practical operational amplifiers produce a finite and variable slope in the portions of the waveform shown in Fig. 1 as vertical. The finite slope limits the useful frequencies available to the audio range, and the variability of the slope makes it impossible to obtain precise frequency control. **Figure 2a** and **b** show block diagrams of two widely used commercial oscillators, with generic integrated circuit type numbers 555 and 566, that work on similar principles to the one described above.

The 555 (Fig. 2a) allows capacitor  $C$  to charge through resistor  $R$  until it reaches a threshold that causes the upper comparator to set a latch (memory circuit) that turns on the electronic switch  $S$  (actually a transistor), thus discharging  $C$ . When  $C$  has discharged below the threshold voltage of the lower comparator, the latch is reset (turned off) and the switch opened. The cycle then restarts. This circuit is often used by hobbyists and for applications not requiring high frequencies or great accuracy. See ELECTRONIC SWITCH; TRANSISTOR.

The 566 (Fig. 2b) uses a transistor current source supplying a current  $I$  to charge its capacitor, and then discharges it by switching the sign of the current to  $-I$ . The Schmitt trigger used to determine the appropriate times for turning the second source on and off uses positive feedback in much the same way as did the operational-amplifier circuit of Fig. 1. The most interesting feature of this circuit is that it can be used to make a voltage-controlled frequency, since the current  $I$  can easily be set by an external voltage. In comparison, the circuit of Fig. 1 can be tuned only by varying  $R$  or  $C$ . Voltage-controlled oscillators are important components of phase-locked loops,

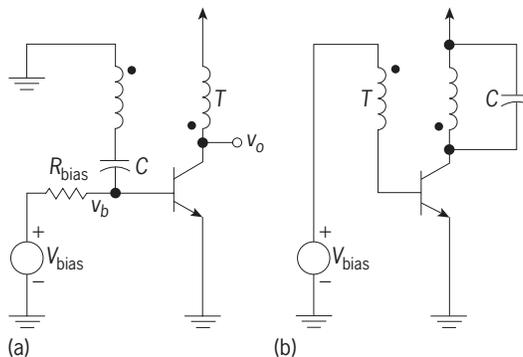


**Fig. 2. Commercial relaxation oscillators. (a) LM555 timer. (b) LM566 voltage-controlled oscillator. (c) Emitter-coupled multivibrator.**

which in turn are widely used in radio and data communications. See CURRENT SOURCES AND MIRRORS; PHASE-LOCKED LOOPS.

Figure 2c shows a simplified circuit for an emitter-coupled multivibrator, which is used as a variable-frequency oscillator in high-speed digital circuits. The cross coupling of each transistor's collector to the other's base is a positive-feedback path, while the charging of  $C$  by the adjustable current sources sets the oscillation frequency. See MULTIVIBRATOR.

**Blocking oscillators.** Relaxation oscillators that have a low duty cycle—that is, produce output pulses whose durations are a small fraction of the overall period—are sometimes called blocking oscillators because their operation is characterized by an “on” transient that “blocks” itself, followed by a recovery period. The circuit of Fig. 3a, for example, can be seen to have a positive-feedback loop through the transistor and transformer: if  $v_o$  starts to fall, the base voltage  $v_b$  will rise (because of the po-



**Fig. 3. Transformer-coupled oscillators. (a) Blocking oscillator. (b) Hartley oscillator.**

larity of windings on the transformer), causing the transistor to turn on and so further decreasing  $v_o$ . Thus the transistor will turn on hard, until it saturates and starts to draw a large amount of base current. Transformers cannot provide direct current, because they have imperfect coupling that manifests itself as inductance, and so eventually base current will begin to fall. This starts to turn the transistor off, and the process quickly regenerates through the positive-feedback path. The output can now be said to be blocked.

All of the base current consumed when the transistor was on went through capacitor  $C$ , which is therefore charged with a negative voltage on the transistor side. Thus, when the transformer voltage drops to zero, this entire negative capacitor voltage appears at the transistor's base and keeps it turned off. The capacitor must recharge slowly from  $V_{bias}$  through  $R_{bias}$  before conduction can begin again.

Figure 3*b*, discussed below, uses similar components but a smaller loop gain to make a sinusoidal oscillator.

**Oscillators with digital gates.** Inverters (digital circuits that invert a logic signal, so that a 0 at the input produces a 1 at the output, and vice versa) are essentially voltage amplifiers and can be used to make relaxation oscillators in a number of ways. A circuit related to that of Fig. 1 uses a loop of two inverters and a capacitor  $C$  to provide positive feedback, with a resistor  $R$  in parallel with one of the inverters to provide an  $RC$  charging time to set frequency. This circuit is commonly given as a simple example, but there are a number of problems with using it, such as that the input voltage to the first gate sometimes exceeds the specified limits for practical gates. See LOGIC CIRCUITS.

A more practical digital relaxation oscillator, called a ring oscillator, consists simply of a ring containing an odd number  $N$  (greater than 1) of inverters. It has no stable logical state (as can be seen by calling the state on any node  $x$ , then following the ring around through the inverters until coming back to the same node, which would have to be the inverse of  $x$ ,  $\bar{x}$ , so that  $x = \bar{x}$ , a contradiction) and in practice (when the number of inverters is three or greater) has no stable state with the outputs at levels intermediate between the voltages representing 0 and 1 either. However, a practical inverter produces an output signal that is a delayed version of the inverse of the input and is given by Eq. (3), where  $x$  is the

$$x_o(t) = \overline{x_{in}(t - \tau)} \quad (3)$$

delay. Because of this delay the circuit solves the problem of the lack of stable states by oscillating. The output of each gate varies periodically, spending time  $T/2$  in the "0" state and  $T/2$  in the "1" state, where  $T$  is the period. A ring oscillator with a ring of  $N$  gates oscillates with a period  $T = 2N\tau$ , where  $\tau$  is the delay of a single gate. The total time displacement of the output waveform of each gate with respect to that of the previous gate, due to the combination of inversion and delay, is given by

Eq. (4). In theory there is another possible mode of

$$\frac{T}{2} + \tau = (N + 1)\tau = \frac{(N + 1)T}{2N} \quad (4)$$

oscillation, in which all states switch back and forth simultaneously between 0 and 1, but this mode is not stable and in practice quickly decays. A ring oscillator is commonly used as a way of measuring gate delays, and also when very fast oscillators with several outputs at finely spaced phases are required.

### Sine-Wave Oscillators

Oscillators in the second major class have their oscillation frequency set by a linear circuit, and their amplitudes set by a weak nonlinearity.

**Linear circuit.** A simple example of a suitable linear circuit is a two-component loop consisting of an ideal inductor [whose voltage is given by Eq. (5), where  $i$

$$v = L \frac{di}{dt} \quad (5)$$

is its current] and a capacitor [whose current is given by Eq. (6)], connected in parallel. These are said to

$$i = C \frac{dv}{dt} \quad (6)$$

be linear elements because, in a sense, output is directly proportional to input, for example, doubling the voltage  $v$  across a capacitor also doubles  $dv/dt$  and therefore doubles  $i$ . The overall differential equation for a capacitor-inductor loop can be written as Eq. (7). Mathematically this has solutions of the form of Eq. (8), where  $\omega = 1/\sqrt{LC}$  [which means that the

$$i + LC \frac{d^2 i}{dt^2} = 0 \quad (7)$$

$$i = A \sin(\omega t + \phi) \quad (8)$$

circuit oscillates at a frequency  $1/(2\pi\sqrt{LC})$  and  $A$  and  $\phi$  are undefined. They are undefined precisely because the elements in the circuit are linear and do not vary with time: any solution (possible behavior) to the equation can be scaled arbitrarily or time-shifted arbitrarily to give another. Practically,  $A$  and  $\phi$  are determined by weak nonlinearities in a circuit. See DIFFERENTIAL EQUATION; LINEARITY.

Equation (7) is a good first approximation to the equation describing a pendulum, and so has a long history as an accurate timekeeper. Its value as an oscillator comes from Galileo's original observation that the frequency of oscillation ( $\omega/2\pi$ ) is independent of the amplitude  $A$ . This contrasts sharply with the case of the relaxation oscillator, where any drift in the amplitude (resulting from a threshold shift in a comparator, for instance) can translate directly into a change of frequency. Equation (7) also fundamentally describes the operation of the quartz crystal that has replaced the pendulum as a timekeeper; the physical resonance of the crystal occurs at a time constant defined by its spring constant and its mass. See HARMONIC MOTION; HARMONIC OSCILLATOR; PENDULUM.

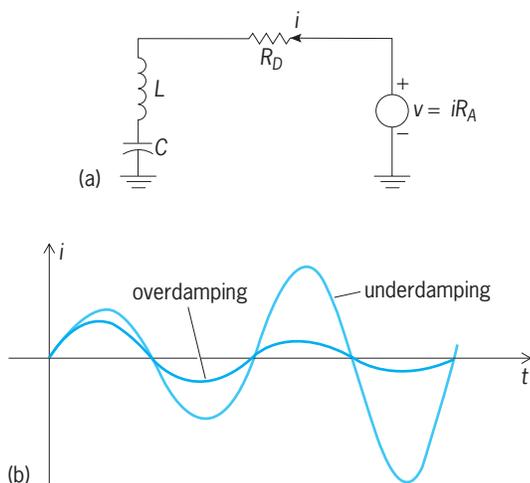


Fig. 4. Negative-resistance oscillator based on a tuned circuit. (a)  $LC$  tuned circuit and controlled-source negative-resistance simulation. (b) Output waveforms with overdamping and underdamping.

**Amplifier element.** Just as the swinging of a practical pendulum gradually dies away, so the oscillations in an  $LC$  circuit and a crystal die away. Clocks solve this problem with an escapement mechanism that gives the pendulum a small kick of additional energy on each cycle, and electronic oscillators use amplifiers for the same purpose. This mechanism is simultaneously used to set amplitude.

Figure 4a shows one implementation of this idea. The heart of the circuit is the resonator formed by the inductor  $L$  and capacitor  $C$ . This resonator also contains a damping resistance  $R_D$ , which might come, for example, from the finite resistance of the wires used to wind the inductor coil. The effect of  $R_D$  is to change the differential equation for a practical  $LC$  loop to Eq. (9). This in turn changes the response

$$i + RC \frac{di}{dt} + LC \frac{d^2i}{dt^2} = 0 \quad (9)$$

of the resonator to an exponentially decaying sinusoid given mathematically by Eq. (10) [where  $\tau$  is

$$i = Ae^{-t/\tau} \sin(\omega t + \phi) \quad (10)$$

called the decay time and  $\omega$  is shifted from its undamped value] and graphically by the curve labeled overdamping in Fig. 4b. See DAMPING.

The transresistance amplifier [current-in, voltage-out, related by Eq. (11), with a gain  $R_A$  whose units

$$v_{\text{out}} = R_A i_{\text{in}} \quad (11)$$

are resistance] simulates a negative resistance, which tends to cancel the effect of the physical damping resistor. If  $R_A = R_D$  exactly, the output will be given exactly by Eq. (8), but if it is smaller the oscillation will decay (overdamping) and if it is larger the output will grow exponentially (underdamping; Fig. 4b). See NEGATIVE-RESISTANCE CIRCUITS.

In practice it is not possible to make  $R_A = R_D$  exactly, and even if the equality did hold, nothing would define oscillation amplitude. Instead a slightly nonlin-

ear amplifier is used, with higher gain at low voltages than at high voltages. If the initial amplitude is small, the amplifier dominates and the oscillations grow; if large, damping dominates and the oscillations decay. An equilibrium is reached with some amplitude of sine wave that is determined by the nonlinearity.

A rigorous treatment of the behavior of this system is relatively difficult, because it now has a nonlinear differential equation. If the amplifier nonlinearity is cubic, the Van der Pol equation (12) applies. This

$$\frac{d^2x}{dt^2} + \epsilon(x^2 - 1) \frac{dx}{dt} + x = 0 \quad (12)$$

equation has been studied extensively, and it displays most of the interesting properties of real oscillators. In particular, as the “strength” of the nonlinear term increases (that is, as  $\epsilon$  increases), the sine wave becomes increasingly distorted and eventually its frequency starts to shift. Therefore these oscillators are at their most accurate when the nonlinearity is weakest.

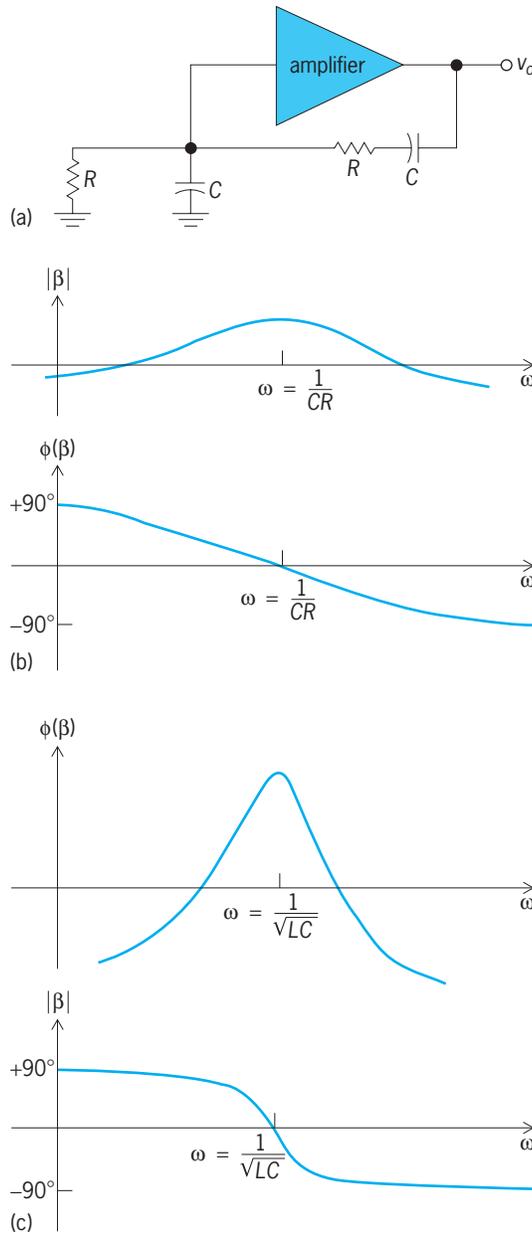
There are practical difficulties with making the nonlinearity vanishingly small. In particular, if the damping resistance is not accurately known or if it may vary, it is necessary to have some safety margin of excess gain and therefore a stronger nonlinearity in order to guarantee oscillation. In some precision circuits a second type of nonlinearity, an automatic gain control, is used to estimate the appropriate gain so that a very weak nonlinearity can be used. See AUTOMATIC GAIN CONTROL (AGC).

**Barkhausen criterion.** The negative-resistance point of view above is one approach to design, but it is often hard to identify the negative resistor in a given circuit. Instead it is common to use a very general amplifier-based analysis and determine stability with the so-called Barkhausen criterion. In this approach the circuit is seen as a loop of an amplifier with gain  $A$  and a linear circuit with frequency-dependent gain  $\beta(j\omega)$ . The loop will oscillate with a perfect sine wave at some frequency  $\omega_0$  if at that frequency  $A\beta(j\omega_0) = 1$  exactly. This means that at this frequency there is positive feedback with a gain of exactly 1, and any oscillation will sustain itself. The equation is usually expressed as a pair of conditions, one on the magnitude of the gain,  $|A\beta(j\omega_0)|$ , required for oscillation, and a second that requires the phase,  $\phi[A\beta(j\omega_0)]$ , to equal  $0^\circ$  or  $360^\circ$ . These conditions define the frequency of oscillation.

Figure 5a shows a circuit called a Wien bridge, commonly given as a simple example, that illustrates the use of the Barkhausen criterion. It consists of a voltage amplifier with some gain  $A$ , and an  $RC$  circuit. The gain  $\beta(j\omega)$  of the  $RC$  circuit is given by expression (13). This complex quantity is plotted

$$\frac{j\omega CR}{1 + 3j\omega CR - (\omega CR)^2} \quad (13)$$

in Fig. 5b, where the variation of the magnitude and phase of  $\beta(j\omega)$  with frequency are shown. When



**Fig. 5.** Design of oscillators based on plotting variation of loop gain with frequency. (a) Wien bridge circuit. (b) Bode plots for the Wien bridge RC-CR network. (c) Bode plots for the LC resonator of Fig. 4a, shown for comparison.

magnitude and frequency are on logarithmic scales, as here, this is known as a Bode plot. See GAIN.

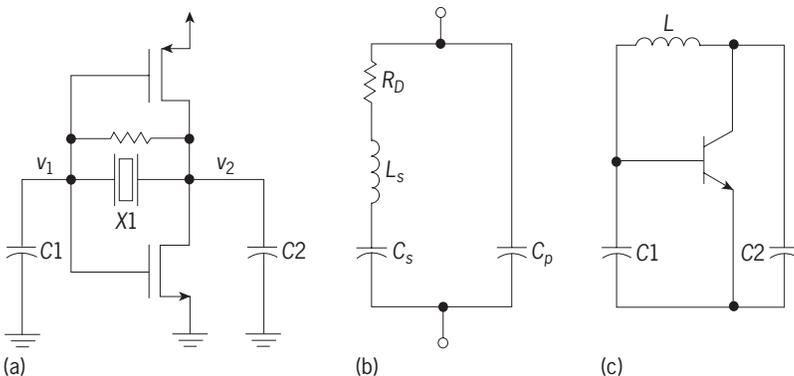
At frequency  $\omega = 1/(CR)$ , the phase of  $\beta$  is  $0^\circ$ . If the amplifier also has zero phase (ideal frequency-independent amplifiers have a phase of  $0^\circ$  or  $180^\circ$ ), then this frequency is a candidate for oscillation if the gain  $A$  is appropriately chosen. The phase portion of Barkhausen's criterion therefore gives the oscillator frequency; the magnitude condition then defines the amplifier gain that is needed. At  $\omega = 1/(CR)$ , the loop gain is  $A/3$ . Thus the amplifier gain must be at least 3. In practice,  $A$  would be chosen somewhat larger than this minimum in order to guarantee oscillation even in the presence of small variations in component values. Oscillation amplitude would be defined by amplifier nonlinearities, such as clipping or by the intentional addition of other nonlinear elements such as diodes to the circuit.

*Choice of circuits.* There are many linear circuits to choose from, and they have different strengths and weaknesses. The Wien bridge circuit of Fig. 5, for example, is less costly to implement at audio frequencies than the LC circuit of Fig. 4, but is less accurate in frequency and produces more distorted waveforms. Bode plots for the Wien bridge (Fig. 5b) and the LC resonator (Fig. 5c) show why.

The phase curve for the LC oscillator passes through the critical frequency where phase is  $0^\circ$  with a much steeper slope than does the Wien bridge curve. This makes the LC oscillator frequency much less sensitive to the amplifier: suppose in each case that the amplifier has  $-1^\circ$  of phase rather than the ideal zero (because a practical amplifier has frequency-dependent gain); then the new oscillation frequency will be that at which  $\phi(\beta) = 1^\circ$ . This is a much smaller error in frequency for the LC circuit than for the Wien bridge.

As to distortion, the LC circuit has a magnitude curve with a much sharper peak. Thus distortion, which introduces harmonic frequencies well above the peak into the loop, can be much more effectively filtered in the LC case than in the Wien bridge case. The Barkhausen criterion and the use of Bode plots provide a straightforward method of oscillator design as well as insight into important design choices.

*Example.* The most common oscillator circuit is probably that of Fig. 6a, which generates the clock signals in most digital circuits, and of which a low-voltage variant keeps time in quartz watches. The two transistors are just a digital inverter in complementary metal-oxide-semiconductor (CMOS) technology, and the resistor biases the circuit to operate with  $v_1 = v_2$ . In these conditions the inverter acts as a transconductance amplifier—the amplifier part of an oscillator. The element X1 is a small slice of quartz crystal with electrodes attached to it, and has a mechanical resonance together with a piezoelectric effect that converts between mechanical motion and electric current. This combination gives it the electrical equivalent circuit shown in Fig. 6b, in which  $L_s$ ,  $C_s$ , and  $R_D$  come from the mechanical resonance as translated by the piezoelectric effect. Capacitor



**Fig. 6.** Clock oscillator, a crystal oscillator. (a) Circuit diagram. (b) Electrical equivalent of piezoelectric crystal resonator. (c) General structure of a Colpitts oscillator.

$C_p$  comes from the ordinary physical capacitance between the electrodes and makes the design somewhat more complicated than for a simple  $LC$  loop. A crystal is an excellent choice for an oscillator, because it has an exceptionally sharp phase curve and the frequency is defined by very stable physical properties. See PIEZOELECTRICITY; QUARTZ CLOCK.

This oscillator is a variant of the classic Colpitts oscillator, of which an idealized equivalent circuit is shown in Fig. 6c (without biasing circuitry). Barkhausen analysis of this circuit shows that it oscillates at a frequency given by Eq. (14). The amplifier

$$\omega = 1 \left/ \sqrt{L \frac{C_1 C_2}{C_1 + C_2}} \right. \quad (14)$$

gain (transconductance) required for oscillation depends on the amount of damping in the circuit. The circuit of Fig. 6a is obtained from the general form by replacing the inductor with a crystal and adding biasing details.

**Practical LC oscillators.** The Colpitts structure of Fig. 6c is used in many forms to make practical oscillators used at radio frequencies. A second widely used structure is the Hartley oscillator, which generally uses an  $LC$  tuned circuit and a transformer in the form shown in Fig. 3b. Transformer  $T$  has some inductance  $L$ , which resonates with  $C$ , and simultaneously couples the transistor's collector to its base to form a feedback loop. Bias voltage  $V_{\text{bias}}$  is chosen to keep the transistor in the so-called on condition at an appropriate current. (In practical circuits a more complicated  $RC$  network is used for bias generation, because operation of the circuit shown is overly sensitive to small changes in  $V_{\text{bias}}$ . At a frequency  $1/(2\pi\sqrt{LC})$ , the phase shift around the loop is zero, and the circuit has the potential to oscillate if the magnitude of loop gain exceeds unity. If loop gain is excessive, the nonlinearity of the transistors will become more pronounced, and the circuit behavior will begin to resemble that of the blocking oscillator in Fig. 3a.

### Frequency Locking

If an external signal is injected into an oscillator, the natural frequency of oscillation may be affected. If the external signal is periodic, oscillation may lock to the external frequency, a multiple of it, or a submultiple of it, or exhibit an irregular behavior known as chaos. For example, in the 555 circuit of Fig. 2a a capacitor voltage charges up until it meets a reference voltage, and then rapidly discharges. Because the reference voltage is now varied with an externally generated signal, the two voltages meet earlier than they would have for a constant reference, and so the oscillation frequency is changed. Furthermore, the signals are more likely to meet when the reference voltage is low than when it is high, which tends to synchronize them. See CHAOS.

This locking behavior occurs in all oscillators, sometimes corrupting intended behavior (as when an oscillator locks unintentionally to a harmonic of the power-line frequency) and sometimes by design.

Whether locking occurs or not depends on the extent of nonlinearity in the oscillator loop, manner of injection, amplitude of the external signal, and the ratio of free-running to external frequencies. The same mechanism can cause collections of oscillators with some coupling mechanism to synchronize to one another. The mathematics of the phenomenon is still being explored.

An important example of an oscillator that exploits this locking principle is the human heart. Small portions of heart muscle act as relaxation oscillators. They contract, incidentally producing an output voltage that is coupled to their neighbors. For a short time the muscle then recovers from the contraction. As it recovers, it begins to become sensitive to externally applied voltages that can trigger it to contract again (although it will eventually contract anyway). Each small section of heart muscle is thus an independent oscillator, electrically coupled to its neighbors, but the whole heart is synchronized by the frequency-locking mechanism. See CARDIAC ELECTROPHYSIOLOGY.

Martin Snelgrove

**Bibliography.** K. K. Clarke and D. T. Hess, *Communication Circuits: Analysis and Design*, 1971, reprint 1994; National Semiconductor Ltd., *Linear Databook 3*, 1988; A. Sedra and K. C. Smith, *Microelectronic Circuits*, 4th ed., 1997.

## Oscillatory reaction

A chemical reaction in which some composition variable of a chemical system exhibits regular periodic variations in time or space. It is a basic tenet of chemistry that a closed system moves inexorably toward an unchanging state called chemical equilibrium. That motion can be described by the monotonic increase of entropy if the system is isolated, and by the monotonic decrease of Gibbs free energy

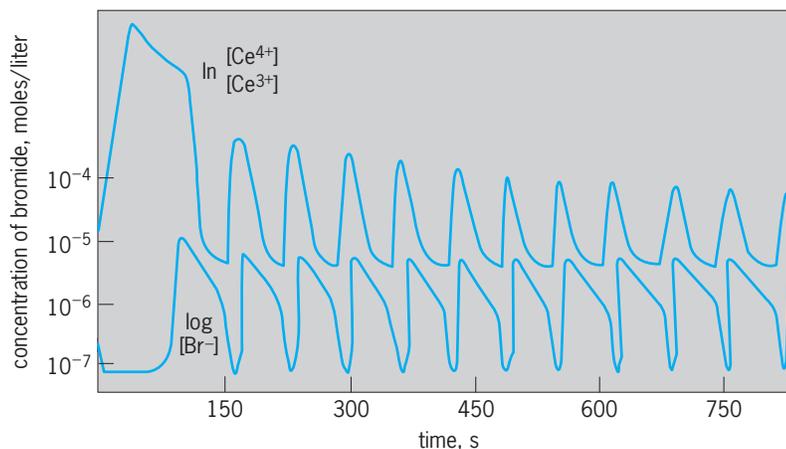


Fig. 1. Oscillatory behavior in a sulfuric acid medium containing potassium bromate, malonic acid, and cerous nitrate. The upper curve is the potential (in unspecified units) of a tungsten electrode and is related to the concentration ratio of cerium(IV)/cerium(III). The lower curve is the potential of an electrode sensitive to bromide ion, and the logarithmic scale at the left relates that potential to the absolute concentration of bromide. (After R. J. Field, E. Körös, and R. M. Noyes, *Oscillations in chemical systems, II. Thorough analysis of temporal oscillations in the bromate-cerium-malonic acid system*, *J. Amer. Chem. Soc.*, 94:8649-8664, 1972)



Fig. 2. Rotating spiral bands of oxidation in a thin layer of solution containing the same reagents as in Fig. 1 except that the redox indicator ferrous phenanthroline has been substituted for cerous ion. (Courtesy of Prof. A. T. Winfree, University of Arizona; from R. J. Field and R. M. Noyes, *Mechanisms of chemical oscillators: Conceptual bases*, *Acc. Chem. Res.*, 10:214–221, 1977).

if the system is constrained to constant temperature and pressure. Because of this universal restriction on what is possible in chemistry, it may appear bizarre when electrodes in a solution generate the oscillating potentials shown in Fig. 1.

The species taking part in a chemical reaction can be classified as reactants, products, or intermediates. In a closed system the concentrations of reactants decrease and those of products increase. Intermediates are formed by some steps and are destroyed by others. If there is only one intermediate, and if its concentration is always much less than the initial concentrations of reactants, this intermediate attains a stable steady state in which the rates of formation

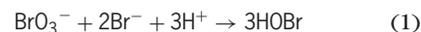
and destruction are virtually equal. The kind of oscillation reflected in Fig. 1 requires at least two intermediates which interact in such a way that the steady state of the total system is unstable to the minor fluctuations present in any collection of molecules. The concentrations of the intermediates may then oscillate regularly, although the oscillations must disappear before the inevitable monotonic approach to equilibrium.

Periodic chemical behavior may be temporal in a uniform system as illustrated in Fig. 1; it may involve simultaneous temporal and spatial variations as in Fig. 2; or it may involve spatial periods in a static system as in Fig. 3.

Well-authenticated examples of periodic chemical behavior have been known for almost a century, but until the 1970s most chemists either did not know about them or deliberately ignored them. In 1976 I. Prigogine of Brussels received the Nobel prize for his demonstration a few years before that periodic behavior was indeed consistent with the accepted principles of chemistry. Since that time, interest has developed rapidly, but most examples are still poorly understood. The phenomena are classified here according to types of chemical processes involved. Very different classification schemes may become more appropriate in the future. See CHEMICAL EQUILIBRIUM; ELECTRODE POTENTIAL; ENTROPY.

**Redox oscillators.** The systems whose chemistries are best understood all involve an element that can exist in several different oxidation states. Figure 1 illustrates the so-called Belousov-Zhabotinsky reaction, which was discovered in the Soviet Union in 1951. A strong oxidizing agent (bromate) attacks an organic substrate (such as malonic acid), and the reaction is catalyzed by a metal ion (such as cerium) that can exist in two different oxidation states.

As long as bromide ion ( $\text{Br}^-$ ) is present, it is oxidized by bromate ( $\text{BrO}_3^-$ ), as in reaction (1).

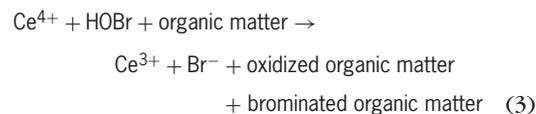


When bromide ion is almost entirely consumed, the cerous ion ( $\text{Ce}^{3+}$ ) is oxidized, as in reaction (2).



Reaction (2) is inhibited by  $\text{Br}^-$ , but when the concentration of bromide has been reduced to a critical level, reaction (2) accelerates autocatalytically until bromate is being reduced by  $\text{Ce}^{3+}$  many times as rapidly as it was by  $\text{Br}^-$  when reaction (1) was dominant.

The hypobromous acid ( $\text{HOBr}$ ) and ceric ion ( $\text{Ce}^{4+}$ ) formed in reaction (2) then react with organic matter by a complicated sequence that can be summarized in reaction (3). Reaction (3) creates the



bromide ion necessary to shut off the fast reaction



Fig. 3. Layers of crystallization from an initially uniform magma in the Skaergaard Intrusion in Greenland. (From A. R. McBirney and R. M. Noyes, *Crystallization and layering of the Skaergaard Intrusion*, *J. Petrol.*, 20:487–554, 1979)

(2) and throw the system back to dominance by the slow reaction (1).

As other redox oscillators become understood, they fit the same pattern of a slow reaction destroying a species that inhibits a fast reaction that can be switched on autocatalytically; after a delay the fast reaction generates conditions to produce the inhibitor again. Most of the known examples involve positive oxidation states of chlorine (Cl), bromine (Br), or iodine (I). Elements like nitrogen (N), sulfur (S), chromium (Cr), and manganese (Mn) can also exist in several positive oxidation states, and it may be found that they also drive chemical oscillations.

The traces in Fig. 1 were obtained in a closed system such that no matter crossed its boundaries; it must eventually have decayed to a stable chemical equilibrium. Even more dramatic oscillations can be made to persist indefinitely if fresh chemicals are continuously added while material from the reactor is simultaneously removed. Such a system is called a continuously stirred tank reactor.

If the solution in Fig. 1 were unstirred but had a gradient in composition, oscillations in different regions would get out of phase, and an apparent wave would traverse the medium much the way flashing lights cause a message to move across a theater marquee.

Figure 2 illustrates a still more complex situation. Each light curve is a region of dominance by reaction (2). A combination of reaction and diffusion triggers an advance outward perpendicular to the wavefront into the dark region dominated by reaction (1). These trigger waves annihilate each other when they meet, and Fig. 2 shows two spirals spinning in opposite directions. This kind of behavior has been suggested to explain the fibrillations when a human heart loses its rhythm and degenerates to uncoordinated local contractions that result in death if the condition is not rapidly reversed. *See* OXIDATION-REDUCTION.

**Nucleation oscillators.** If molecules of a gas are produced at a steady rate in a uniform solution, the evolution of that gas may take place in pulses during which the solution foams up and then subsides. Several examples of this general phenomenon are known. As the chemical reaction proceeds homogeneously, the solution becomes more and more supersaturated, until a threshold is reached at which many microscopic bubbles are nucleated almost simultaneously. These bubbles grow ever faster as they become bigger, until they deplete the supersaturation and escape because of the Earth's gravitational field. The chemical reaction must then raise the supersaturation to the critical level before another burst of nucleation can occur. The result is that the solution foams up repeatedly like a shocked glass of beer and then subsides. Examples of such gas-evolution oscillators are the dehydration of formic acid ( $\text{HCO}_2\text{H}$ ) by concentrated sulfuric acid and the decomposition of aqueous ammonium nitrite ( $\text{NH}_4\text{NO}_2$ ) to produce elementary nitrogen. *See* SUPERSATURATION.

If a solution is not uniform but has gradients in composition or in temperature, then nucleation and

growth of crystals may occur in bands. Figure 3 illustrates a geologic formation on the east coast of Greenland. A large magma chamber cooled very slowly, and the initially uniform material crystallized in a pattern of regular layers.

**Oscillations coupled to mass transport.** If a reaction in a solution is autocatalytically accelerated by a reactant coming from another phase, the rate may increase until transport of the critical reactant can no longer keep up. For certain situations, the system does not go to the anticipated steady state. Instead, some intermediate is depleted, the concentration of the critical reactant is renewed, and the rate begins to increase again. The only known examples involve reactions with dissolved oxygen in a solution in contact with the air. Detailed mechanisms are not yet understood.

**Thermokinetic oscillators.** Many chemical reactions give out heat, and rates are strongly dependent on temperature. If heat is not removed too rapidly from the reactor, reaction rate and temperature may couple to generate oscillations.

The known examples involve highly exothermic reactions like combustion or chlorination of organic compounds. No chemical mechanisms are yet understood in detail. In at least some examples, gradients of temperature and of composition are important in addition to changes in local rates of reaction.

**Reactions on surfaces.** Many important industrial reactions take place on the surfaces of catalysts. The occurrence of such a reaction may temporarily alter that surface or its temperature. Such effects sometimes couple to generate oscillations in the rate of reaction. These oscillations may or may not be of value for the reaction being carried out. Specific examples are being studied actively by chemical engineers.

The surfaces of electrodes may also be influenced by processes taking place on them, and periodic changes in voltage or current are well preceded.

**Biological chemistry.** Living organisms take in nutrients of high free energy and eliminate waste products of lower free energy. They therefore resemble the continuously stirred flow reactors. The degradation of those nutrients drives the essential vital processes. The pathways involve intricate couplings so that a decrease in free energy of one species contributes in forming species like adenosine triphosphate (ATP) that have increased free energy. Many important intermediates follow repeated cyclic paths while the nutrients are degraded. *See* ADENOSINE TRIPHOSPHATE (ATP).

If all of the processes of metabolism took place at the same rate in the same place, the only net effect would be an increase in entropy, and life would not exist. Processes involving some intermediates must be separated in space or in time from processes involving other intermediates.

Separation in space can be accomplished by membranes permeable to some species but not to others. Such membranes are ubiquitous in biological organisms. *See* CELL MEMBRANES.

Separation in time can be accomplished by oscillatory reactions that turn component processes on and

off much as happens with the Belousov-Zhabotinsky reaction described above. It can hardly be accidental that periodicities are observed in many biological activities.

One of the best chemical examples involves oscillations during oxidation of glucose catalyzed by cell-free yeast extracts. The enzyme phosphofructokinase (PFK) is strongly implicated, and its activity is influenced by the oxidation products.

Aggregation of the slime mold *Dictostelium discoideum* takes place because individual cells move in the direction of a received pulse of cyclic adenylic acid (cyclic AMP) and simultaneously emit a pulse in the opposite direction. Aggregating cells may create spiral patterns resembling those in Fig. 2 except that motion is inward instead of outward.

Circadian rhythms with periods of about 24 h are common in biology. Anybody who has taken a jet halfway around the world becomes well aware of their effects.

Undoubtedly many periodic biological processes of chemical origin are not even recognized. None can yet be described in mechanistic detail. Development of the quantitative explanation characteristic of a mature science is anticipated. Richard M. Noyes

Bibliography. I. R. Epstein et al., Oscillating chemical reactions, *Sci. Amer.*, 248(3):112-123, 1983; R. J. Field and M. Burger (eds.), *Oscillations and Traveling Waves in Chemical Systems*, 1985; H. Haken, *Advanced Synergetics*, 1993; G. Nicolis and I. Prigogine, *Self-Organization in Nonequilibrium Systems*, 1977; S. K. Scott, *Oscillations, Waves, and Chaos in Chemical Kinetics*, 1994; A. T. Winfree, *The Geometry of Biological Time*, 2d ed., 2001.

## Oscilloscope

An electronic measuring instrument which produces a display showing the relationship of two or more variables. In most cases it is an orthogonal ( $x,y$ ) plot with the horizontal axis being a linear function of time. The vertical axis is normally a linear function of voltage at the signal input terminal of the instrument. Because transducers of many types are available to convert almost any physical phenomenon into a corresponding voltage, the oscilloscope is a very versatile tool that is useful for many forms of physical investigation. See TRANSDUCER.

The oscillograph is an instrument that performs a similar function but provide a permanent record. The light-beam oscillograph used a beam of light reflected from a mirror galvanometer which was focused onto a moving light-sensitive paper. These instruments are obsolete. The mechanical version, in which the galvanometer drives a pen which writes on a moving paper chart, is still in use, particularly for process control. See GALVANOMETER; GRAPHIC RECORDING INSTRUMENTS.

Oscilloscopes are one of the most widely used electronic instruments because they provide easily understood displays of electrical waveforms and are capable of making measurements over an extremely

wide range of voltage and time. Although a very large number of analog oscilloscopes are in use, digitizing oscilloscopes (also known as digital oscilloscopes or digital storage oscilloscopes) are preferred, and analog instruments are likely to be superseded. See ELECTRONIC DISPLAY.

### Analog Oscilloscopes

An analog oscilloscope, in its simplest form, uses a linear vertical amplifier and a time base to display a replica of the input signal waveform on the screen of a cathode-ray tube (CRT) [Table 1]. The screen is typically divided into 8 vertical divisions and 10 horizontal divisions. Analog oscilloscopes may be classified into nonstorage oscilloscopes, storage oscilloscopes, and sampling oscilloscopes.

**Analog nonstorage oscilloscopes.** These oscilloscopes are the oldest and most widely used type. Except for the cathode-ray tube, the circuit descriptions also apply to analog storage oscilloscopes.

A typical oscilloscope (Fig. 1) might have a bandwidth of 150 MHz, two main vertical channels plus two auxiliary channels, two time bases (one usable for delay), and a  $3.2 \times 4$  in. ( $8 \times 10$  cm) cathode-ray-tube display area; and it might include on-screen readout of some control settings and measurement results. A typical oscilloscope is composed of five basic elements: (1) the cathode-ray tube and associated controls; (2) the vertical or signal amplifier system with input terminal and controls; (3) the time base, which includes sweep generator, triggering circuit, horizontal or  $x$ -amplifier, and unblanking circuit; (4) auxiliary facilities such as a calibrator and on-screen readout; and (5) power supplies.

The first four elements will be described in detail; power supplies are used in all electronic equipment and require no special description. See ELECTRONIC POWER SUPPLY.

**Cathode-ray tube.** The central component in a cathode-ray oscilloscope is the cathode-ray tube, which in its simplest form consists of an evacuated glass container with a fluorescent screen at one end and a focused electron gun and deflection system at the other.

Either magnetic or electric fields may be used for focusing and deflection. Electrostatic deflection is almost universally used for oscilloscopes because it is capable of superior high-frequency response.

The cathode-ray-tube designer is faced with four primary technical objectives: (1) high deflection sensitivity up to the desired frequency; (2) a bright image for ease of observation and photography; (3) small spot size relative to the image area; and (4) accurate

**TABLE 1. Vertical measurement capabilities of analog oscilloscopes**

Maximum bandwidth	Rise time	Deflection sensitivity	Measurement accuracy
1 MHz	350 ns	10 $\mu$ V/division	5%
40 MHz	8.75 ns	5 mV/division	3%
200 MHz	1.8 ns	2 mV/division	2.5%
350 MHz	1 ns	2 mV/division	2%

deflection geometry. All of these are interdependent so that each tube design is a compromise chosen to best meet the needs of the intended users. See CATHODE-RAY TUBE.

**Vertical amplifier.** The signal to be observed is generally applied to the vertical or  $y$  axis of the oscilloscope. The vertical amplifier is required to provide sufficient gain that small signals may be displayed with suitable amplitude on the cathode-ray tube. All modern oscilloscope vertical amplifiers are decoupled so as to allow accurate observation of slowly changing signals. The vertical amplifier must amplify the signal so as to minimize distortion in its wave shape. It is thus required to be highly linear and have a suitable frequency, phase, and transient response. In addition, because it is desirable to view the signal that triggers the time base, a signal delay is incorporated in the vertical amplifier to allow the time base to start before the vertical signal reaches the cathode-ray-tube vertical deflection plates. A delay line, either with coaxial cables or of a specially wound type, is generally used for this purpose and is located in the vertical amplifier after the trigger pick-off point. See AMPLIFIER; DELAY LINE; DIRECT-COUPLED AMPLIFIER; DISTORTION (ELECTRONIC CIRCUITS).

To obtain a suitable image size on the cathode-ray tube, a convenient means of varying the amplifier gain is needed. The method used must vary the gain without introducing frequency or amplitude distortion. Generally, a high-impedance frequency-compensated attenuator switched in an accurate 1-2-5 sequence is used before the amplifier input, and a continuously variable control covering a range of at least 2.5:1 is used within the amplifier to bridge between the steps of the input attenuator. See ATTENUATION (ELECTRICITY).

All oscilloscope vertical amplifiers provide for variable positioning on the cathode-ray-tube screen and a choice of ac or dc coupling for the input signal. The positioning control inserts a variable current controllable by the user at a suitable point in the vertical amplifier, and the input coupling control inserts a capacitor in series with the input terminal when ac coupling is chosen. AC coupling is typically used when a small ac signal is present together with a large dc signal which would otherwise deflect the display off-screen.

The input connector of most oscilloscope signal amplifiers has one terminal connected to the chassis of the instrument. For safety reasons, the chassis is usually connected to ground through the mains lead. Most waveforms also have a ground reference, so the signal and the instrument are compatible. However, there is a danger that the resulting ground loop will introduce errors, especially with low signal levels. Some instruments have isolated inputs to deal with this problem. Add-on input isolators are also available, but these restrict the working bandwidth. Sometimes it is necessary to observe the signal between two points, both of which are at significant potentials relative to ground. Differential amplifiers having two high-impedance input terminals are available for this purpose. Many oscilloscopes allow the differ-

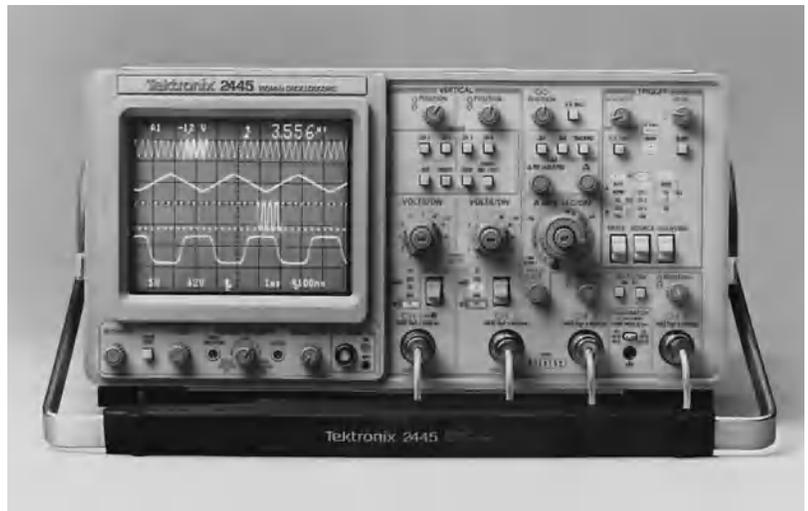


Fig. 1. Four-channel 150-MHz analog oscilloscope. (Tektronix)

ence between two inputs to be displayed, achieving a similar result; however, the two channel gains must be carefully matched at the time of use. See DIFFERENTIAL AMPLIFIER.

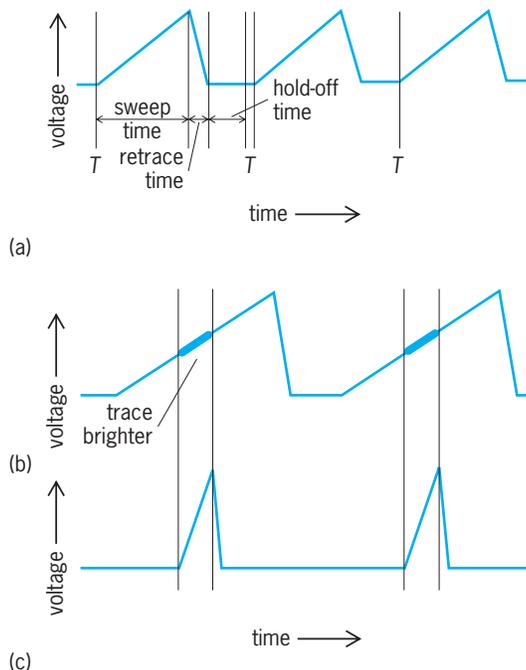
Most oscilloscopes incorporate a multitrace vertical system which provides for the simultaneous display of two or more different signals for comparison purposes. The most common method used to achieve a multitrace display is to incorporate a vertical input amplifier complete with all controls for each channel and then electronically switch between them at a rate faster than the human persistence of vision. In this case the display will appear flicker-free. To achieve this result, two alternative switching methods are employed with a choice available to the user. One method is to switch during the sweep return time (alternate mode) so that succeeding sweeps display first one channel, then another, and then repeat. This gives a flicker-free display provided the sweep repetition rate is high enough. If this is not the case, then the chopped mode must be used. Here the switching between channels is done continuously at a high rate with cathode-ray-tube beam blanking used to avoid displaying the switching waveform itself.

Some applications require that two simultaneous high-speed signals occurring only once (single shot) or at a very low repetition rate be observed or photographed. Neither of the switching systems above is suited for this measurement, and a dual-beam oscilloscope must be employed. A dual-beam oscilloscope comprises two completely independent vertical amplifiers feeding two separate sets of cathode-ray-tube vertical deflection plates. Either one or two time bases and cathode-ray-tube horizontal deflection systems may be employed. Since the vertical systems are independent, there are no compromises imposed by having to choose between alternate or chopped modes as in a time-shared multitrace oscilloscope. Dual-beam oscilloscopes are more expensive than single-beam types and in general have poorer matching between the transfer response of the vertical channels.

**Time bases.** In the standard oscilloscope waveform presentation the horizontal axis on the cathode-ray tube represents a linear function of time. This is achieved by the use of a linear sweep generator feeding the cathode-ray-tube horizontal deflection plates via the horizontal amplifier. For a satisfactory display the sweep generator must be correctly triggered and the cathode-ray-tube beam blanked except when a sweep is present.

A cathode-ray-tube presentation that is a linear function of time requires that a voltage changing linearly with time be applied to the cathode-ray-tube plates, generally in push-pull to avoid geometrical distortion in the cathode-ray tube itself. The sweep generator is started by a fast edge related to the vertical input signal. This edge may be applied from an external trigger input or derived within the vertical amplifier. The sweep generator produces a linearly increasing voltage (Fig. 2a), which is amplified by the horizontal amplifier and applied to the horizontal deflection plates. At the end of the sweep the generator rapidly returns to its initial voltage, the time taken being the retrace time. This is followed by the hold-off time, during which the circuits return to a standard initial condition. The cycle is repeated on the receipt of the first trigger signal after the end of the hold-off period. Sometimes the hold-off time may be adjusted by the user, making it possible to choose the overall cycle time and so repetitively display the same part of a complex signal.

Dual time bases are used to obtain a detailed view of part of a complex waveform. One time base (A) is used to establish a delay; the second, faster time base (B) is initiated either by the end of the delay or by the first subsequent trigger pulse. To set up the



**Fig. 2.** Sweep generator waveforms. (a) Single waveform, showing instant of trigger (T). (b) Dual time base in set-up mode: sweep by A, intensified by B. (c) Dual time base in delayed display mode: scan delayed by A, sweep by B.

two time bases, the delaying sweep is first used as a simple time base and adjusted so that the whole complex waveform is presented on the screen. The delayed sweep B appears as a bright region on the trace (Fig. 2b). Its position and duration are adjusted in order to cover the region of interest. When the delayed display mode is selected, the complete screen horizontal scan is provided by the delayed sweep (Fig. 2c), and the details of a relatively brief part of the waveform are revealed. In better-quality instruments the trace brightness is automatically compensated to suit the change in sweep rate.

**Trigger generator.** To ensure a stable display of the incoming signal, each sweep must start at a time corresponding to precisely the same point on the signal waveform. The information needed by the sweep generator to accomplish this may come either from the signal itself or from another signal which has a consistent time relationship to the signal to be observed. The input of the trigger generator is thus capable of being switched between several sources, the most commonly used being a portion of the signal at a suitable point before the delay line in the vertical amplifier. Whatever its source, the signal is applied to a circuit which generates a very fast output pulse every time the incoming signal reaches a certain voltage. This pulse is used to start the sweep generator, providing the hold-off time has ended. Front panel controls allow selection of the slope and voltage of the incoming signal that results in a trigger signal being generated. Most oscilloscopes have an automatic mode of triggering which arranges the trigger circuit to sense the input signal and provide a trigger pulse output at a point on the waveform which will generally result in a satisfactory display. When no signal is present, the time base is made to run, resulting in a bright horizontal line (baseline) being displayed which assists the user in selecting the kind of display that is needed. See TRIGGER CIRCUIT.

**Calibrators.** Most oscilloscopes incorporate an accurate square-wave calibrator which is used to check the vertical-gain calibration and adjust the probe, if one is used, for optimal response. In addition, some calibrators may be sufficiently accurate in time to allow the oscilloscope time-base calibration to be checked.

**Sampling oscilloscopes.** The limit of speed of vertical response of analog nonstorage oscilloscopes is determined by the design of the vertical amplifier and by the cathode-ray-tube deflection plate bandwidth. The fastest general-purpose oscilloscope available has a bandwidth from 0 (direct current) to 1 GHz; wider bandwidths are possible, but cost and power consumption rapidly increase. Equivalent-time or sampling oscilloscopes provide much higher bandwidths at reasonable cost and power consumption, but the technique requires that the signal to be observed be repetitive at a suitable minimum rate. It makes use of extremely short pulses which are used to sample the incoming signal. Samples are taken at a slightly later time at each recurrence of the trigger pulse. The samples are amplified, lengthened in time, and displayed at a relatively slow sweep speed. Some

important limitations of this type of oscilloscope are that single transients cannot be observed and that it may take a considerable time to build up a complete picture of the incoming signal if its repetition rate is low. A variation on the sampling technique described above is to take samples in a random fashion and then reconstitute the samples in the correct equivalent time and again display at a relatively slow sweep speed. Both sampling methods are used in commercially available oscilloscopes with bandwidths from 0 (dc) to over 10 GHz. See VOLTMEETER.

**Analog storage oscilloscopes.** It is often desirable to store a waveform after the originating signal has ceased to exist, particularly in the case of transient signals. The simplest method is to photograph the transient waveform. Photographic recording is, however, inconvenient, time-consuming, and relatively expensive and it requires some skill. Analog storage oscilloscopes, were developed to solve this problem. Digitizing oscilloscopes also store waveforms and, because of their numerous advantages, are rapidly replacing analog storage oscilloscopes.

The analog storage oscilloscope substitutes a storage cathode-ray tube and control circuits for the standard type. There are several types of storage cathode-ray tube, but all have the property of being able to display a waveform for a limited length of time after the originating signal has disappeared. The storage time is at least sufficient for convenient photography and for relatively leisurely visual observation. Most storage oscilloscopes provide several modes of operation such as (1) nonstorage operation similar to a nonstorage oscilloscope, (2) medium-speed storage with long storage time, and (3) high-speed storage with relatively short storage time.

Storage cathode-ray tubes differ from nonstorage types in that they incorporate a target, in some cases the actual phosphor, which changes its electrical state after being scanned by the electron beam and results in a lingering charge storage pattern representing the original waveform. After the original signal has ended, the cathode-ray-tube operating voltages are changed so as to display a replica of the charge storage pattern; simultaneously, all further vertical input signals are stopped from affecting the cathode-ray tube until the user reactivates the oscilloscope. Cathode-ray-tube storage oscilloscopes are available which are capable of storing waveforms at up to 500 MHz.

**Photography of waveforms.** Many oscilloscopes are provided with fittings that enable matching cameras

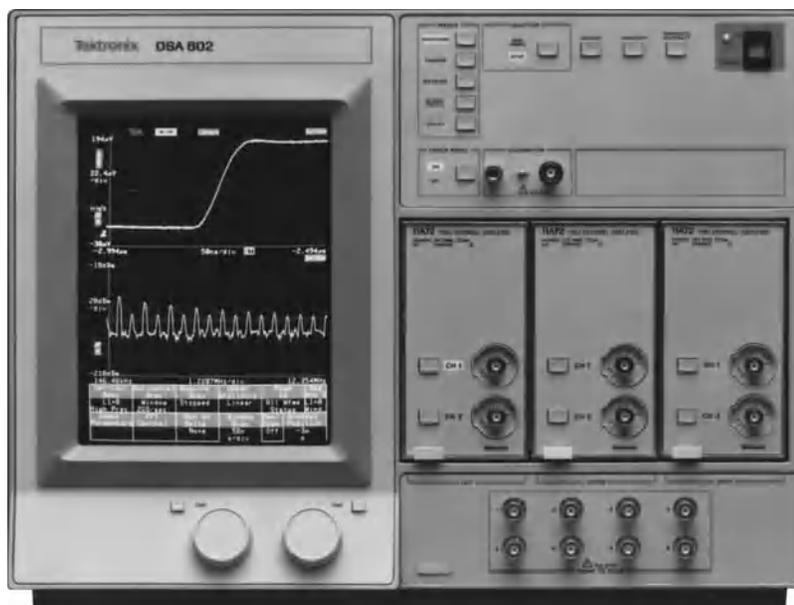


Fig. 3. High-speed general-purpose digitizing oscilloscope. (Tektronix)

to be attached. By using these, permanent records of normal or stored waveforms can be obtained. To photograph single fast transients, special cathode-ray-tube phosphors, wide-aperture lenses, and high-speed films can be used. See CAMERA; LENS (OPTICS); PHOTOGRAPHY.

### Digitizing Oscilloscopes

Digital techniques are applied to both timing and voltage measurement in digitizing oscilloscopes (Fig. 3; Table 2). A digital clock determines sampling instants at which analog-to-digital converters obtain digital values for the input signals. The resulting data can be stored indefinitely or transferred to other equipment for analysis or plotting. See VOLTAGE MEASUREMENT; WAVEFORM DETERMINATION.

**Components.** In its simplest form a digitizing oscilloscope comprises six basic elements: (1) analog vertical input amplifier; (2) high-speed analog-to-digital converter and digital waveform memory; (3) time base, including triggering and clock drive for the analog-to-digital converter and waveform memory; (4) waveform reconstruction and display circuits; (5) display, generally, but not restricted to, a cathode-ray tube; (6) power supplies and ancillary functions. In addition, most digitizing oscilloscopes provide facilities for further manipulation of waveforms prior to

TABLE 2. Capabilities of digitizing oscilloscopes

Maximum bandwidth	Rise time	Maximum digitizing frequency, samples per second	Record length, samples*	Deflection sensitivity	Measurement accuracy	Resolution
50 MHz	7 ns	$2 \times 10^8$	1,000	2 mV/division	2%	8 bit
350 MHz	1 ns	$1 \times 10^8$	60,000	1 mV/division	1.5%	8–12 bit
500 MHz	0.7 ns	$2 \times 10^9$	50,000	1 mV/division	1.5%	8–12 bit
8 GHz	44 ps	$5 \times 10^4$	15,000	1 mV/division	0.7%	8 bit

\* The maximum record length may require the fitting of a nonstandard option.

display, for direct measurements of waveform parameters, and for connection to external devices such as computers and hard-copy units.

*Analog vertical input amplifier.* This device performs functions identical to those of a vertical amplifier in a real-time oscilloscope; however, its output feeds an analog-to-digital converter instead of the cathode-ray tube.

*High-speed analog-to-digital converter.* The analog-to-digital converter samples the incoming vertical signal and provides a digital value for each sample. An 8-bit converter is commonly used, providing a resolution of 1 in 256. Higher resolution can be obtained by using a longer-scale converter or by building up a set of data from many cycles of signal. The latter method can provide the effect of an additional 4 bits of resolution but cannot be applied to single-shot measurements. See ANALOG-TO-DIGITAL CONVERTER.

*Time base.* The analog-to-digital converter is regularly triggered by the time-base clock at a rate selected by the user. During data acquisition the resulting values are stored in adjacent positions of the waveform memory. A trigger circuit similar to that used in an analog oscilloscope identifies a particular point on the waveform. This is used to identify the sample which corresponds to the trigger point. A selected number of further samples are taken after the trigger point. It is therefore possible to arrange that the memory contains a number of samples before, and a different number after, the trigger point. A total record of 1000 samples is typical. When these data are displayed, the effect is as if there was a delay in the signal channel, as information from a period preceding the trigger is included. The effective length of the delay can be adjusted by varying the number of samples which are accepted after the trigger point. No conventional delay line is required in the signal circuit of a digitizing oscilloscope.

*Waveform reconstruction and display.* The digitized input signal, now in memory, is used along with the knowledge of the original digitizing rate to reconstruct an oscilloscope-type display of amplitude versus time. The display device used is generally a cathode-ray tube, but since the reconstructed display rate can be chosen independently of the original signal speed, other types of display devices may be selected if their characteristics are more desirable. The reconstructed display may display one dot for each sample or may use straight or curved lines to join the dots to make the display easier to interpret.

**Real-time and equivalent-time operation.** All digitizing oscilloscopes operate in the real-time mode just described, where the incoming signal is digitized continuously and the trigger, in effect, freezes the digitized data for subsequent processing and display. Because the input signal is sampled at discrete points in time, the resulting digitized data can contain no information at a frequency greater than half the digitizing frequency. For example, an oscilloscope with a 100-MHz maximum sampling frequency can provide no information above 50 MHz. Even at this frequency, only two points per cycle are available, and the display is not very useful. By comparison, an ana-

log oscilloscope shows some information above its  $-3$ -dB bandwidth frequency. Thus, no direct comparison can be made between the bandwidth of an analog oscilloscope and the maximum digitizing frequency of a digitizing oscilloscope when used in its real-time mode.

Most measured signals are repetitive. For such signals, the effective bandwidth of a digitizing oscilloscope may be considerably extended by using the equivalent-time mode of operation. Here, sampling is done as in an analog sampling oscilloscope except that the sample values are subsequently digitized and stored in digital memory. The limits on effective bandwidth are the same for both. Because a very large number of samples may be taken in this mode, the oscilloscope bandwidth may be specified. Most digitizing oscilloscopes provide both modes of operation.

**Advantages and limitations.** In almost all respects, digitizing oscilloscopes can match the capabilities of analog types. The existence of the data in digital form gives the technique many advantages, including:

1. The data can be processed within the instrument or sent elsewhere by using standard interfaces.
2. Mathematical processes, from addition and subtraction to Fourier transformations can be carried out. See FOURIER SERIES AND TRANSFORMS.
3. Signal parameters such as root-mean-square values and rise times can be calculated and displayed.
4. Signals from repeated sweeps may be averaged, giving the effect of up to 3 bits of additional resolution are reducing random noise.
5. Hard-copy representations of waveforms and their parameters can be created easily.
6. Considerable improvements in accuracy are possible, as the resolution is not limited by optical limitations.
7. The display is not limited to a cathode-ray tube designed to suit the requirements of the signal. Color can be used to enhance legibility. See COMPUTER GRAPHICS.
8. Intelligent software can be used to identify sharp transitions and cause a higher sampling rate to be used in such regions. This is called windowing.
9. Completely automatic measuring systems can be constructed.

Digitizing oscilloscopes of modest performance are available in which the techniques are used in conjunction with a liquid-crystal display to provide a versatile hand-held instrument which combines all the functions of an oscilloscope, digital multimeter, and frequency meter. See LIQUID CRYSTALS.

Digitizing oscilloscopes suffer from the phenomenon of aliasing, in which a false waveform is recorded which bears little or no similarity to the actual signal. One possible cause is a sampling rate that is too low compared with the highest signal frequency. In perceptual aliasing, the eye of the observer is deceived and tends to link the dots which are nearest on the screen. These may well not be the samples of the waveform which are closest in time. Perceptual aliasing does not occur with displays where the dots are linked by calculated lines.

### Plug-in Oscilloscopes

So-called plug-in units are often used in high-performance oscilloscopes. These units are interchangeable hardware modules capable of altering certain major characteristics of an oscilloscope without causing the user to have to buy a completely new instrument. Their most common application is in giving a wider choice of vertical input performance characteristics than would be economically possible to build in. They are also used to alter horizontal characteristics such as triggering or time-base type, the latter only in analog oscilloscopes. Certain performance characteristics or combinations thereof are available only with plug-in oscilloscopes.

### Oscilloscope Selection

Higher measurement accuracy is available from digitizing oscilloscopes (Tables 1 and 2). The first decision to be made in choosing an oscilloscope is whether this or any of the other properties exclusive to the digitizing type are essential. If not, the option of an analog design remains. The selected instrument must be appropriate for the signal under examination. It must have enough sensitivity to give an adequate deflection from the applied signal, sufficient bandwidth, adequately short rise time, and time-base facilities capable of providing a steady display of the waveform. An analog oscilloscope needs to be able to produce a visible trace at the sweep speed and repetition rate likely. A digitizing oscilloscope must have an adequate maximum digitizing rate and a sufficiently long waveform memory.

**Signal amplitude.** The lowest expected signal amplitude should provide over two divisions of deflection at the oscilloscope's minimum deflection factor, and the highest amplitude should provide no more than full screen. Probe attenuation must be taken into account.

**Rise time.** The fastest rise time of the signal and the degree of acceptable measurement error determine the needed rise time of the oscilloscope. If  $n$  is the maximum percentage error acceptable when a signal having a rise time of  $t_s$  is viewed, then the oscilloscope's rise time must be no more than that approximately given by expression (1). For instance,

$$\left[ \left( 1 + \frac{n}{100} \right)^2 - 1 \right]^{1/2} \times t_s \quad (1)$$

if  $n = 10\%$  and  $t_s = 10$  ns, then the oscilloscope's rise time must be no more than 4.6 nanoseconds.

Conversely, the true rise time of a signal may be approximately estimated from the measured rise time,  $t_m$ , and the oscilloscope's known rise time,  $t_0$ , using Eq. (2).

$$\text{True rise time} = (t_m^2 - t_0^2)^{1/2} \quad (2)$$

**Oscilloscope bandwidth.** The bandwidth of an oscilloscope is the frequency at which the vertical amplitude response has fallen by 3 dB (about 29%) from its low-frequency value. The way in which the ampli-

tude response deteriorates with frequency needs to be carefully controlled if the best possible response to pulses is to be obtained. All modern instruments achieve this when correctly adjusted. Under these conditions the rise time and bandwidth are related approximately by Eq. (3).

$$\text{Bandwidth (in MHz)} = \frac{350}{\text{rise time (in ns)}} \quad (3)$$

### Signal Acquisition

Oscilloscopes of up to 300-MHz bandwidth generally have an input impedance of their signal input equivalent to 1 megohm with 15–30 picofarads capacitance in parallel. Connection to the circuit under test is generally done by using probes. In order to increase the impedance connected to the test circuit, probes with 10 times attenuation are most commonly used. These have a typical input impedance at their tip equivalent to 10 megohms with 10 pF in parallel. The probes are carefully designed to introduce minimum distortion in the signal they transmit. Above 200–300 MHz, the 10-pF-probe input capacitance has an undesirably low reactance (10 pF has a reactance of 53 ohms at 300 MHz), and other methods are needed. Most common is the provision of an accurate 50-ohm input to the oscilloscope. The user can either terminate the signal at the oscilloscope input, or a low-impedance probe can be used which again divides by 10 times, but has an impedance represented by 500 ohms and about 2 pF; at 300 MHz this gives an impedance of 236 ohms and is thus higher than in the previous case. Active probes are also available; they contain a miniaturized input circuit in the probe itself and transmit the signal over a 50-ohm line to the oscilloscope's input. Active probes provide the highest input impedance in the range of 100–1000 MHz but are expensive and relatively fragile. See ALTERNATING-CURRENT CIRCUIT THEORY; ELECTRICAL IMPEDANCE.

**Calibration.** High-impedance voltage divider probes have a capacitor adjustment in the probe that must be correctly set. This is normally achieved by connecting the probe tip to the oscilloscope's square-wave calibrator and adjusting the capacitor until the observed waveform shows a correct step response. In addition, the voltage measurement accuracy of the oscilloscope-probe combination should be optimized by adjusting the oscilloscope's vertical gain for the correct deflection with the probe tip connected to the calibrator. These two steps ensure that the best amplitude and frequency response are obtained. In addition, connections to the signal should be kept as short as possible, since long connections from either the ground or probe tip can lead to the pick-up of unwanted signals and to deterioration in the system's transient and frequency responses.

**Current probes.** Oscilloscopes indicate and measure voltages. In some circumstances, the current, rather than the voltage, is of interest, and it may not be possible to insert a small resistor in the current path to produce a voltage signal. Systems having probes which clip around a conductor and sense the

magnetic field are available to satisfy this requirement. A variety of products offer the capability of operating down to dc, sensitivities of 1 mA per division, peak current measurements up to kiloamperes, and maximum frequency limits of 15 MHz to 1 GHz.

R. B. D. Knight

**Bibliography.** R. L. Goodman, *Using the Triggered Sweep Oscilloscope*, 1987; I. Hickman, *Oscilloscopes: How to Use Them, How They Work*, 5th ed., 2000; J. D. Lenk, *Handbook of Oscilloscopes*, 1982; R. A. Penfold, *How to Use Oscilloscopes and Other Test Equipment*, 1989; S. Prentiss, *The Complete Book of Oscilloscopes*, 2d ed., 1992.

## Osmium

A chemical element, Os, atomic number 76, atomic weight 190.2. The element is a hard white metal of rare natural occurrence, usually found in nature alloyed with other platinum metals. See METAL; PERIODIC TABLE; PLATINUM.

1																	18				
H																	He				
3	4															5	6	7	8	9	10
Li	Be															B	C	N	O	F	Ne
11	12	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18				
Na	Mg											Al	Si	P	S	Cl	Ar				
19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36				
K	Ca	Sc	Ti	V	Cr	Mn	Fe	Co	Ni	Cu	Zn	Ga	Ge	As	Se	Br	Kr				
37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54				
Rb	Sr	Y	Zr	Nb	Mo	Tc	Ru	Rh	Pd	Ag	Cd	In	Sn	Sb	Te	I	Xe				
55	56	71	72	73	74	75	76	77	78	79	80	81	82	83	84	85	86				
Cs	Ba	Lu	Hf	Ta	W	Re	Os	Ir	Pt	Au	Hg	Tl	Pb	Bi	Po	At	Rn				
87	88	103	104	105	106	107	108	109	110	111	112	113									
Rf	Ra	Lr	Db	Sg	Bh	Hs	Mt	Ds	Rg												

lanthanide series	57	58	59	60	61	62	63	64	65	66	67	68	69	70
	La	Ce	Pr	Nd	Pm	Sm	Eu	Gd	Tb	Dy	Ho	Er	Tm	Yb

actinide series	89	90	91	92	93	94	95	96	97	98	99	100	101	102
	Ac	Th	Pa	U	Np	Pu	Am	Cm	Bk	Cf	Es	Fm	Md	No

Physical properties of the element, which is found as seven naturally occurring isotopes, are given in the **table**. The metal is exceeded in density only by iridium. Osmium is a very hard metal and unworkable, and so it must be used in cast form or fabricated by powder metallurgy. Osmium is a third-row transition element and has the electronic configuration  $[\text{Xe}](4f)^{14}(5d)^6(6s)^2$ ; in the periodic table it lies below iron (Fe) and ruthenium (Ru). In powder form the metal may be attacked by the oxygen in air at room temperature, and finely divided osmium has a faint odor of the tetraoxide. In bulk form it does not oxidize in air below 500°C (750°F), but at higher temperatures it yields OsO<sub>4</sub>. It is attacked by fluorine or chlorine at 100°C (212°F). It dissolves in alkaline oxidizing fluxes to give osmates (OsO<sub>4</sub><sup>2-</sup>). See ELECTRON CONFIGURATION; IRIDIUM; IRON; RUTHENIUM.

The chemistry of osmium more closely resembles that of ruthenium than that of iron. The high oxidation states VI and VIII (OsO<sub>4</sub><sup>2-</sup> and OsO<sub>4</sub>) are much more accessible than for iron. See OXIDATION-REDUCTION.

Osmium forms many complexes. In water, osmium complexes with oxidation states ranging from II to VIII may be obtained. Oxo compounds, which

### Principal properties of osmium

Property	Value
Density, g/cm <sup>3</sup>	22.6
Naturally occurring isotopes (% abundance)	184 (0.018)
	186 (1.59)
	187 (1.64)
	188 (13.3)
	189 (16.1)
	190 (26.4)
	192 (41.0)
Ionization enthalpy, kJ/mol: 1st 2d	840
	1640
Oxidation states	-1 to VIII
Most common	IV, VI, VIII
Ionic radius, Os <sup>4+</sup> , nm	0.078
Melting point, °C (°F)	3050 (5522)
Boiling point, °C (°F)	5500 (9932)
Specific heat, cal/g °C	0.032
Crystal structure	Hexagonal close-packed
Lattice constant a at 25°C, nm c/a at 25°C	0.27341
	0.15799
Thermal neutron capture cross section, barns	15.3
Thermal conductivity, 0-100°C, (cal · cm)/(cm <sup>2</sup> · s · °C)	0.21
Linear coefficient of thermal expansion at 20-100°C, (μin./in./°C)	6.1
Electrical resistivity at 0°C, μΩ-cm	8.12
Temperature coefficient of electrical resistance, 0-100°C/°C	0.0042
Young's modulus at 20°C, lb/in. <sup>2</sup> , static	81 × 10 <sup>6</sup>

contain Os=O, are very common and occur for oxidation states IV to VIII. Although OsO<sub>4</sub> is tetrahedral in the gas phase and in noncomplexing solvents such as dichloromethane (CH<sub>2</sub>Cl<sub>2</sub>), it tends to be six-coordinate when appropriate ligands are available; thus, in sodium hydroxide (NaOH) solution, dark purple OsO<sub>4</sub>(OH)<sub>2</sub><sup>2-</sup> is formed from OsO<sub>4</sub>. Similarly, OsO<sub>2</sub>(OH)<sub>4</sub><sup>2-</sup> is formed by addition of hydroxide to osmate anion. Analogous reactions with ligands such as halides, cyanide, and amines give osmyl derivatives like the cyanide-duct OsO<sub>2</sub>(CN)<sub>2</sub><sup>2-</sup>, in which the trans dioxo group (O=Os=O) is retained.

Osmium tetraoxide, a commercially available yellow solid (melting point 40°C or 104°F), is used commercially in the important *cis*-hydroxylation of alkenes and as a stain for tissue in microscopy. It is poisonous and attacks the eyes. Osmium metal is catalytically active, but it is not commonly used for this purpose because of its high price. Osmium and its alloys are hard and resistant to corrosion and wear (particularly to rubbing wear). Alloyed with other platinum metals, osmium has been used in needles for record players, fountain-pen tips, and mechanical parts. See ALLOY; STAIN (MICROBIOLOGY); TRANSITION ELEMENTS.

Carol Creutz

**Bibliography.** F. A. Cotton et al., *Advanced Inorganic Chemistry*, 6th ed., Wiley-Interscience, 1999; J. R. Davis (ed.), *Metals Handbook: Desk Edition*, 2nd ed., ASM International, 1998.

## Osmoregulatory mechanisms

Physiological mechanisms for the maintenance of an optimal and constant level of osmotic activity of the fluid within and around the cells, considered to be most favorable for the initiation and maintenance of vital reactions in the cell and for maximal survival and efficient functioning of the entire organism.

**Evolution.** Practically all living cells function in a fluid environment. This includes isolated unicellular forms, such as paramecia and amebas, as well as cells that make up tissues in air-breathing terrestrial animals. Thus, the ionic composition and osmotic activity of extracellular fluids have biological significance with respect to survival of the living cells. The fluid environment of simple unicellular forms of life consists of the oceans, lakes, or streams in which the forms are found, while that of the complex animal forms consists of the fluid media enclosed by the various compartments of the body.

Presumably, life originated in the primitive ocean, the salinity of which was lower in prehistoric times than in modern times. The first unicellular forms of life remained in a state of osmotic equilibrium with their dilute seawater surroundings; that is, they had no osmoregulatory devices. In the course of evolution, unicellular and multicellular animals migrated from the sea to freshwater streams and eventually to dry land. Survival of such forms was associated with the maintenance of constant osmotic activity of their body fluids through evolution of osmoregulatory devices. See PREBIOTIC ORGANIC SYNTHESIS.

Changing osmotic conditions occasioned the evolution of special cells and tissues which permitted retention or exclusion of the proper amount of water and solute for the animal. Examples of these special tissues are the skins of amphibians, the gills of fishes, and the kidney tubules of mammals. In cells of these

tissues electrolytes apparently diffuse across one cell membrane and are pumped out across the opposite membrane of the cell. See EXCRETION.

**Biological mechanisms.** The actions of osmoregulatory mechanisms are, first, to impose constraints upon the passage of water and solute between the organisms and its surroundings and, second, to accelerate passage of water and solute between organism and surroundings. The first effect requires a change of architecture of membranes in that they become selectively permeable and achieve their purpose without expenditure of energy. The accelerating effect, apart from requiring a change of architecture of cell membranes, requires expenditure of energy and performance of useful osmotic work. Thus substances may be moved from a region of low to a region of higher chemical activity. Such movement can occur in opposition to the forces of diffusion of an electric field and of a pressure gradient, all of which may act across the cell membrane. It follows that there must be an energy source derived from the chemical reactions of cellular metabolism and that part of the free energy so generated must be stored in molecules driven across the membrane barrier. Active transport is the modern term for such processes. The manner whereby chemical energy can be transferred into osmotic work has not yet been determined. Some examples that illustrate osmoregulation follow.

Fresh-water fish and frogs can pump water out of their bodies via the urine. This process enables these animals to survive in dilute fluids. The salt-water fish pumps salt from its body to the sea across the gills. This removes the salts of ingested ocean water, thus permitting survival with brine as the source of water. Of interest is the absence of glomeruli in the kidneys of some marine forms like the toadfish and the goosfish. Such aglomerular kidneys excrete smaller volumes of urine than do glomerular kidneys and

Summary of osmotic performance in various animals\*

Osmotic characteristics	Principal mechanisms	Examples
Osmotic adjustment	No volume regulation	Marine invertebrate eggs; <i>Phascolosoma</i>
	Volume regulation	Marine mollusks; <i>Maja</i> ; <i>Nereis pelagica</i> ; <i>N. cultrifera</i>
Limited osmoregulation	Low permeability; salt reabsorption in nephridia (?) Water storage	<i>Nereis diversicolor</i> <i>Gunda</i>
Fair osmoregulation in hypotonic media	Selective absorption of salts from medium; kidney reabsorption or secretion; low permeability	<i>Carcinus</i>
Regulation in hyper- and hypotonic media except at extremes	Unknown	<i>Uca</i>
Unlimited regulation in hypotonic media	Hypotonic copious urine; salt reabsorption or water secretion; low surface permeability Water impermeability	Crayfish; fresh-water teleosts, Amphibia Fresh-water embryos
Maintenance of hypertonicity in all media	Urea retention	Elasmobranchs
Regulation in hypertonic media	Extrarenal salt excretion; low water intake Unknown	Marine teleosts <i>Artemia</i>
Regulation in moist air	Low skin permeability; salt absorption from medium; salt reabsorption in kidney	Earthworm; frog
Regulation in dry air	Impermeable cuticle; hypertonic urine Hypertonic urine, water reabsorption in kidney	Insects Birds and mammals

\*After C. L. Prosser et al. (eds.), *Comparative Animal Physiology*, 2d ed., Saunders, 1956.

possess, in their tubular cells, transport systems for the excretion of salt. The albatross and penguin, by secreting an extremely hypertonic nasal fluid, can survive with seawater as the sole source of water, a unique biological advantage for a terrestrial animal. See KIDNEY; SALT GLAND.

The **table** presents a zoological classification of various modes of osmotic defense in several phyla of animals. Efficient osmoregulation is accomplished largely by way of the inherent properties of the cell membranes: (1) The hydrophobic nature of the bilamellar phospholipid-protein array of most membranes constitutes an effective barrier to the free diffusion of polar substances (water and water-soluble solutes); (2) simultaneously, however, most membranes actually facilitate the passage of certain substances into and out of the cell, thus providing selective permeability; and (3) many membranes can actually actively transport, or "pump," certain solutes (for example,  $\text{Na}^+$  or glucose) into or out of the cell, which is an important facet of the osmoregulatory process.

In multicellular organisms such mechanisms are integrated into osmoregulatory organ systems under the control of chemical, nervous, and hormonal stimuli. See URINARY SYSTEM.

### Transport Processes

A stringent definition of active transport of ions requires that a net amount of material be moved unidirectionally across a biological membrane against the forces of diffusion, against the forces of an electrical field, and even against the force of a hydrostatic pressure gradient. For individual ions such a movement is said to be against the electrochemical potential gradient. An equally stringent definition applied to water transport would require movement of water against gradients of hydrostatic pressure and of its chemical potential. In the most general sense active transport of a substance means its movement against its free-energy gradient; therefore, the process requires a source of free energy from cellular metabolism. Exactly how metabolic energy is funneled into a cellular mechanism capable of doing osmotic work is unknown.

The term active transport is generally reserved for the pumping of solutes across membranes. Active transport involves a chemical reaction between the transported solute and some constituent of the membrane. In contrast, the osmotic flow of water through cell membranes is a passive process and does not involve strong chemical interactions. However, during osmoregulation by organs such as the kidney or intestine, water can be moved across the epithelial cell membranes from a lower to a higher chemical concentration of water. This implies that water transfer across osmoregulating epithelial membranes is active—as would be required for an osmoregulating rather than for an osmoaccommodating process. However, the metabolic energy source of osmoregulatory transport of water is derived from the active transport of solute across the membrane. It is generally accepted that the actively transported solute,

and not the water, reacts with the membrane constituents. This implies that all transmembrane movement of water is primarily passive and secondary to local osmotic gradients and hydrostatic pressures induced by solute pumping.

**Carrier systems.** In the majority of cases the actively transported substances involved in osmoregulation are univalent inorganic ions, which are supposed to react with membrane constituents called carriers. Apart from attempts to identify and isolate such carriers, there are good reasons for believing that they exist. To account for active transport in the absence of physical forces, it has been necessary to invoke chemical reactions with carrier substances in the membrane. However, chemical reactions are by their nature scalar, and the rate of reaction depends only upon the concentration of the reactants and products and not upon the spatial location of the reactants and products. In contrast, transport processes are vectorial in nature and result in the translocation of substrate from one place to another. Thus the vectorial property of the transport process must be ascribed to a physical property over and above the chemical process in the membrane. There are two kinds of physical processes which could account for the directionality of transport. These can be classified as moving site models and as mobile carrier models.

*Moving site model.* This can be represented as an aqueous filled pore, the walls of which are lined with sites having a high affinity for the transported ion. The ion to be transported is first adsorbed to the site adjacent to the edge of the membrane. The occupied site then undergoes a decrease in affinity for the transported substance, and the ion is transferred to the next site of high affinity deeper in the membrane. These sequential changes in the affinity of the adsorption sites for the transported ion continue from one side of the cell membrane to the other, in a wavelike manner, thus imparting directionality to the transport process. While the moving site model is adequate, there is no experimental evidence for its existence.

*Mobile carrier model.* This requires that the transported ions combine with a limited number of mobile carrier molecules and that they then be carried across the cell membrane in the form of carrier-ion complexes. The main tenets of this theory are that (1) the carrier molecule is free to move about in the cell membrane; (2) the range of motion of the carrier molecule is limited to the interior of the cell membrane; (3) the carrier molecule undergoes a change in affinity for the transported ion at one edge of the cell membrane; and (4) the cell membrane is relatively impermeable to the free ion being transported.

The precise mode of motion of the carrier molecule is not known. It could undergo free diffusion; swing across the cell membrane in a pendular motion; revolve about an axis in the plane of the cell membrane; or invert along an axis of symmetry normal to the plane of the membrane.

The experimental evidence for the existence of carrier molecules rests on kinetic studies of the transport of isotopes. Specifically, the kinetic phenomena

of transport are those of saturation kinetics cis inhibition, trans stimulation, and counterflow.

The term cis inhibition describes the following sequence of events. A biological membrane is transporting a mixture of  $^{22}\text{Na}$  and  $^{23}\text{Na}$  having a given specific activity from left to right at a constant rate. Additional  $\text{Na}^{23}$  is added to the bathing fluid on the left side of the membrane, reducing the specific activity of the mixture and resulting in a decrease in the absolute rate of transport of  $^{22}\text{Na}$  from left to right. Thus the term cis inhibition arises because addition of one isotope of an ion inhibits the transport of another isotope of the same ion when the addition is made to the same side as that from which the ion is being transported.

Trans stimulation is said to occur when addition of one isotope of a transported ion, to the side toward which the transported ion moves, causes an increase in the absolute rate of transport of another isotope of the same ion toward that side.

It has been shown on theoretical grounds that the two kinetic phenomena of cis inhibition and trans stimulation are peculiar to systems where the membrane penetration process is carrier-mediated.

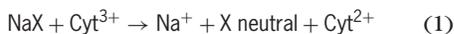
Further important information on carriers has been obtained in certain strains of bacteria which can transport various galactosides, sugars, amino acids, and sulfates. Proteins isolated from the membranes of these bacteria bind specifically to the transportable substrate. Nontransporting mutants of these bacteria do not yield specific binding proteins but can be induced to do so.

In biological systems water flow is usually associated with the transport of ions or sugars. There are four general schemes for the transport of ions: (1) the oxidation-reduction pump; (2) the adenosine triphosphate-driven pump; (3) forced ion exchange; and (4) ion-pair transport. The following is a brief description of these schemes.

**Oxidation-reduction systems.** The characteristics of such a scheme are that there is an oriented electron-transporting mechanism in the membrane; the reaction results in the movement of a given ion, at a single site from the region of low to a region of high electrochemical activity, with generation of an electromotive force; and the electric field so generated transports another ion of opposite charge in the same direction. Alternatively, the field could move a similarly charged ion, at a separate site in the membrane, in a direction opposite to that of the primary ion movement. In either case the laws of electroneutrality are satisfied for the whole system, while osmotic work is done on the transported particles. The active transport of sodium ion ( $\text{Na}^+$ ) with the passive transport of chloride ion ( $\text{Cl}^-$ ) across the frog skin can be described in a stepwise manner by using the assumptions of an oxidation-reduction mechanism.

**Figure 1** illustrates the operation of an oxidation-reduction mechanism designed for the active transport of  $\text{Na}^+$  ions. The reaction in the membrane is oriented by means of unspecified constraints at the interface across which the actively transported ion is ejected. Thus  $\text{Na}^+$  ion enters the cell membrane by

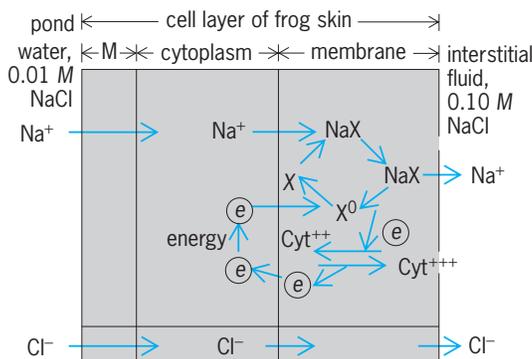
combining electrostatically with a carrier substance  $\text{X}^-$  to yield undissociated  $\text{NaX}$ . The undissociated complex diffuses across the membrane, wherein the oriented reaction is between cytochrome-oxidized and cytochrome-reduced. This results in the reaction shown in reaction (1), and the free  $\text{Na}^+$  ion formed



is ejected from the cell as the electron moves in the opposite direction and reduces the oxidized cytochrome. In order to satisfy electroneutrality, a negative ion ( $\text{Cl}^-$ ) must be transported at a site spatially separated from that of cation transport. (There is evidence in some systems that  $\text{Cl}^-$  is also actively transported, but active transport of the companion ion is not inconsistent with an electrogenic model.) A separate electron donor substance must be present to convert X neutral to  $\text{X}^-$  ion, a process requiring metabolic energy. Although never proven rigorously, the assumptions of this scheme may be used to account for many osmoregulatory processes in cells across which a measurable electric field exists. While the scheme depicted in Fig. 1 is adequate for an explanation of sodium transport, it is not a unique oxidation-reduction system. Redox pairs other than cytochrome could be employed to drive the transport. See BIOLOGICAL OXIDATION.

**ATP systems.** An enzyme called sodium-potassium-ATPase ( $\text{Na-K-ATPase}$ ) has been isolated in the microsomal, or membrane fraction, of many tissues that transport sodium, for example, frog skin, toad and turtle bladders, red cells, and brain. The enzyme hydrolyzes the high-energy phosphate bond of adenosine triphosphate (ATP), producing adenosine diphosphate (ADP), inorganic phosphorus, and energy. See ADENOSINE DIPHOSPHATE (ADP); ADENOSINE TRIPHOSPHATE (ATP).

The cell stores much of its metabolic energy in the ATP, and the hydrolysis of ATP supplies sufficient free energy to drive an oriented transport mechanism in the membrane as well as many chemical reactions in the cell. The remarkable features of  $\text{Na-K-ATPase}$  are that it is found only in microsomes; it is the only known ATPase requiring sodium and potassium ions for activation; and it is the only



**Fig. 1.** Transport of  $\text{Na}^+$  and  $\text{Cl}^-$  from pond water across the frog skin to the interstitial fluid. (After R. F. Pitts, *Physiology of the Kidney and Body Fluids*, 2d ed., Year Book Medical Publishers, 1968)

ATPase whose hydrolytic activity is inhibited by the cardiac glycoside, ouabain—which appears to be a specific inhibitor of sodium transport. Whether Na-K-ATPase acts as a carrier and combines directly with sodium during the transport process or whether Na-K-ATPase is involved only in the transfer of energy to the transport machinery is not clear. However, it is obvious that Na-K-ATPase plays some necessary role in sodium transport. In addition, Na-K-ATPase is the first enzyme identified that appears to play a direct role in the transport of an inorganic ion.

**Forced ionic exchange.** Some investigators believe that electrical potentials across cell membranes are merely diffusion potentials secondary to the ionic gradients produced by a carrier-type active transport. They postulate the existence in the membrane of a charged substance, say  $X^-$ , in the case of cation transport. At one interface the carrier of the mechanism ejects  $Na^+$  ion from the cell and simultaneously injects  $K^+$  ion into the cells. This is called a forced ionic exchange process and produces no electrical potential. At another site of the membrane there are water-filled pores, the walls of which are lined with fixed negative charges. Hence, electropositive ions can enter such pores more readily than can electronegative ions. Therefore the potential difference across the membrane interface (at equilibrium) ought to be predicted by the Nernst equation, shown as Eq. (2), where  $E$  is the potential difference,  $R$  and

$$E = \frac{RT}{F} \ln \frac{a_2}{a_1} \quad (2)$$

$T$  are the gas constant and absolute temperature,  $F$  the Faraday and  $a_1$  and  $a_2$  are the chemical activities of the passively transported ion on each side of the membrane. This means that the transmembrane distribution of ions produced by the forced exchange mechanism results in the production of diffusion potentials across the negatively charged pores. According to this scheme most of the observed potential difference is due to the high intracellular concentration of potassium. The absence of a diffusion potential from the high extracellular concentration of sodium has been attributed to a high specific resistance for this ion across the membrane. A Nernst-type relationship has been observed within a limited range of concentrations of  $K^+$  and  $Na^+$  in the external medium of systems like frog skin and squid axon.

The forced ion-exchange model has most frequently been applied to, and investigated in, systems which transport sodium and potassium. This model has certain difficulties. First, derivations of the Nernst equation tacitly assume equilibrium conditions rather than the steady-state condition that prevails in most living systems. Second, the negatively charged pores must select not only between negative and positive ions but also between sodium and potassium. Third, recent isotopic measurements indicate that the ratio of sodium to potassium transport is often not 1 to 1, as would be expected in such a system.

**Ion-pair transport.** Another mechanism for doing osmotic work is the transport of salts as ion pairs. That is, a carrier could transport neutral sodium chloride molecules. Such a system would do osmotic work, would be nonelectrogenic, and would not necessarily give rise to electrical potentials. Such a transport system has been demonstrated in the proximal renal tubule and in the gallbladder of fish and mammals. See ION TRANSPORT.

### Membrane Permeability

Implicit in all theories of transport of water or solutes are assumptions concerning the permeability properties of the membrane. For example, in a  $Na^+$  transporting system the carrier concept implies that  $NaX$  will diffuse rapidly across the membrane, while free  $Na^+$  ions diffuse slowly or not at all. This is a device for potentiating diffusion of a sluggishly diffusible substance and is economical from a thermodynamic viewpoint. If the membrane were freely permeable to a transported solute, the rate of back diffusion of that solute would increase as its concentration on the transported side increased. Since concentration gradients across cell membranes are high, back diffusion rates would be high. Thus the mechanism would require a large amount of energy to accomplish the net transport of a solute. These and other considerations led to much experimental and theoretical work from which emerged concepts of membrane permeability. See CELL MEMBRANES; CELL PERMEABILITY.

**Molecular architecture.** The cell membrane is viewed as a phospholipid-protein complex. Pairs of phospholipid molecules (such as lecithin) in the form of a bilamellar leaflet are oriented so that the ionizable phosphatidyl head groups of the lipid face toward each aqueous boundary fluid, and the fatty acyl chains of each phospholipid pair face toward each other at the hydrophobic interior of the membrane. Protein molecules, mostly in the alpha-helical and random coil conformation and intercalated with the lipids, penetrate through the entire thickness of the membrane.

Despite the advancing knowledge in the molecular structure of cell membranes (based on patterns of x-ray diffraction, electron microscopy, optical rotary dispersion, and circular dichroism), the exact nature of the path through which solute and solvent molecules pass remains uncertain. Two main models of the intramembrane path have been proposed, the lattice and the aqueous-filled pore. See DICHOISM (BIOLOGY); ELECTRON MICROSCOPE; X-RAY DIFFRACTION.

*Lattice model.* Biologists have proposed that the fatty acyl and protein chains of the cell membrane present a latticelike barrier to the passage of molecules. Thus a penetrant molecule could form a hydrogen bond with an atom of the lattice, then shake loose by thermal agitation, move through an interstitial space to the next bonding site, and so on, until it has traversed the membrane. This picture fits the observed rates of penetration of cell membranes by a large number of nonelectrolyte molecules wherein the rate can be

predicted from the molecular size, the number of hydrogen bonding atoms per molecule, and oil-water solubility ratios. However, the model as constituted falls short in explaining penetration of water (osmosis) and of the electrolytes. See OSMOSIS.

**Aqueous pore model.** Assuming the presence of aqueous-filled pores in membranes, the theory of fixed charges on pore walls was formulated. It is known that artificial membranes, such as collodion, silicates, or protein, when interposed between two electrolyte solutions of different concentration, give rise to an electrical potential difference, depending on the sign and the density of fixed charges lining the pores of the membrane. Such membranes have been called perm-selective, that is, cation- or anion-permeable. When a membrane with positively charged pores separates two salt solutions of different concentration, the dilute solution is negative to the concentrated one; for a membrane with negatively charged pores, the reverse orientation holds. The assumption of fixed charges on pore walls has explained a wide variety of data in artificial perm-selective membranes, and useful analogies between artificial perm-selective and biological membranes have been made. However, many biological membranes (nerve and muscle) can distinguish not only cations from anions but one cation from another, for example.  $\text{Na}^+$  from  $\text{K}^+$ , a finding not accounted for by the fixed charges in the aqueous pore model.

**Facilitated diffusion path.** Some biologically important materials, such as ions, sugars, and amino acids, move across cell membranes more rapidly than can be explained by the parameters of a lattice or fixed-charge path. Therefore, the transmembrane motion has been explained by assuming that the penetrant combines transiently with a mobile constituent molecule of the membrane. These processes of combination with the carrier, translocation of the loaded complex, and dissociation of penetrant from the carrier are termed facilitated diffusion.

**Phenomenological equations.** Due to the lack of knowledge of the exact mechanisms whereby materials cross biological membranes, a theoretical approach which circumvented the need for invoking specific mechanisms of motion was applied. The approach of irreversible thermodynamics originated in the work of L. Onsager and was applied to biological systems by C. Kedem and A. Katchalsky. Essentially, a set of simultaneous equations, called phenomenological equations, is used, based upon well-known empirical laws of motion, such as those of A. Fick, J. Fourier, and G. S. Ohm.

It is assumed that all flows can be represented by a system of linear equations in which each flow is set equal to a linear combination of forces multiplied by appropriate coefficients; that is, each flow is taken to be proportional to each of the forces and the system of flows and forces can be represented in matrix form as Eq. (3). Here  $J_i$  are the flows,  $L_{ik}$  are

$$J_i = \sum_{k=1}^n L_{ik} X_k \quad (i = 1, 2, 3, \dots, n) \quad (3)$$

the coefficients,  $X_k$  are the conjugated forces, and the subscript  $i$  denotes the row, and  $k$  the column position of each term ( $J_i$ ,  $L_{ik}$ , and  $X_k$ ) in the matrix of simultaneous equations.

To illustrate, by an oversimplified example, consider a nonphysiological problem. Suppose that two sucrose solutions are separated by a rigid membrane and that an osmotic or a hydrostatic pressure can be applied to either one of the two solutions. Since there are but two mobile components in the system, a minimum of two flows—one of sucrose and one of water—all will be considered, and the aforementioned matrix, in algebraic form is expressed as Eqs. (4a) and (4b), where  $J_w$  is the flow of water,  $J_s$

$$J_w = L_{11} \Delta P + L_{12} RT \Delta C \quad (4a)$$

$$J_s = L_{21} \Delta P + L_{22} RT \Delta C \quad (4b)$$

is the flow of sucrose,  $\Delta P$  is the hydrostatic pressure difference across the membrane,  $\Delta C$  is the difference in the sucrose concentration in the two solutions,  $L_{11}$  is the hydraulic permeability of the membrane to water,  $L_{22}$  is the permeability of the membrane to sucrose, and  $L_{12}$  and  $L_{21}$  are cross coefficients representing the influence of the concentration difference on water flow and the influence of hydrostatic pressure difference on sucrose flow. In this case  $\Delta P$  and  $\Delta C$  are set at known values experimentally,  $J_w$  and  $J_s$  are measured, and the four  $L$ 's which operationally describe the membrane are unknowns.

Onsager has shown on thermodynamic grounds that  $L_{21} = L_{12}$ , which effectively reduces the unknown to three in number, namely,  $L_{11}$ ,  $L_{22}$ , and  $L_{12}$  or  $L_{21}$ . By successively forcing  $\Delta P$ ,  $\Delta C$ , and  $J_w$  to be zero in the experiment, it is possible to obtain three equations in three unknowns and solve the system for the  $L$ 's. In more complicated situations, with more than two flows, more than three experiments are needed to evaluate all of the unknowns, but the method is the same. The advantage of this system of analysis is that it is not necessary to understand the basic mechanisms to evaluate the permeability of the membrane with respect to a given permeant. The disadvantage is that results from one membrane cannot be generalized to another because the method is essentially empirical.

### Mechanisms of Water Transport

The forces for water transport in living and nonliving systems may be listed as follows: hydrostatic forces; osmotic gradients; electroosmosis; and miscellaneous—chemical reactions, thermal gradients, and unknown forces in contractile vacuoles and pinocytosis.

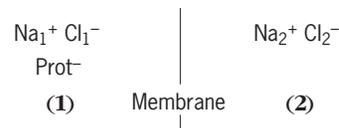
In animals these forces are usually generated by the active transport of solutes and by the circulatory system. When such forces are applied across membranes having perm selective and ion selective properties, such as are possessed by cell membranes, they give rise to the fundamental processes involved in osmoregulation. These processes are filtration, osmosis, diffusion, and maintenance of Gibbs-Donnan

equilibria. In addition to the four simple physical mechanisms, there are three main biological systems which integrate one or more of the simple mechanisms into a device for moving water. They are the three-compartment system, the countercurrent system, and the standing-gradient system. These systems incorporate the processes of filtration, osmosis, diffusion, and the active transport of ions, with multicellular structures of specialized geometric form designed for osmoregulatory functions.

**Hydrostatic forces.** Operationally, flow of fluid through a tube is usually defined in terms of two main parameters: pressure and resistance to flow. It follows that a bulk flow of fluid across any boundary requires a force and that the direction of such flow is down the gradient of pressure. Filtration refers to the passage of water or solution under the influence of a hydrostatic force through the pores of a membrane. A filtration process separates undissolved constituents from a solution, while ultrafiltration separates dissolved constituents of high molecular weight (protein) from a solution.

In filtration a net movement of a finite mass of solute and solvent occurs, while in diffusion there is no net movement of solvent. By definition, the only force operative in a diffusion cell is that of the gradient of chemical potential of the transported solute material. H. Ussing made use of such differences when he measured permeability coefficients of various biological membranes with and without net movement of solution.

**Gibbs-Donnan forces.** Ultrafiltration refers to the passage of solute and water across a membrane between two nonidentical solutions in a system like that conceived by J. Gibbs and F. Donnan. The simplest example of such a system is shown below.



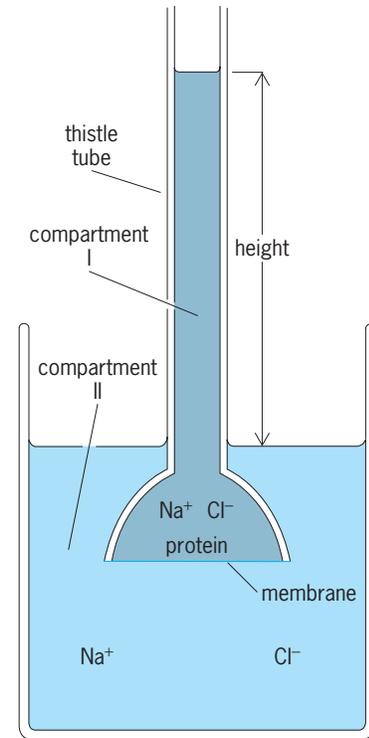
**Figure 2** presents an experimental model of the system at equilibrium. Conditions are such that the membrane permits passage of  $\text{Na}^+$ ,  $\text{Cl}^-$ , and water but not of proteinate. At equilibrium it is possible to derive Eq. (5).

$$[\text{Na}_1^+][\text{Cl}_1^-] = [\text{Na}_2^+][\text{Cl}_2^-] \quad (5)$$

Since  $[\text{Na}_2^+] = [\text{Cl}_2^-]$ , it follows that  $[\text{Na}_1^+] + [\text{Cl}_1^-] > [\text{Na}_2^+] + [\text{Cl}_2^-]$ . This means that a hydrostatic pressure must be applied to side I to prevent the osmotic flow of solvent from II to I. The osmotic conditions at equilibrium (no net flow of solvent or solute) are expressed in Eq. (6), where  $P_1$  and  $P_{II}$  are

$$\begin{aligned}
 & [\text{Na}_1^+] + [\text{Cl}_1^-] + [\text{Prot}^-] \\
 & = [\text{Na}_2^+] + [\text{Cl}_2^-] + \frac{P_{II} - P_1}{RT} \quad (6)
 \end{aligned}$$

the hydrostatic pressures of compartments I and II, respectively,  $R$  is the gas constant, and  $T$  is the absolute temperature. Ultrafiltration will occur when the pressure applied to solution I is sufficient to move



**Fig. 2.** Simple model illustrating a Gibbs-Donnan equilibrium system. (After W. D. Stein, *Theoretical and Experimental Biology*, vol. 6: *The Movement of Molecules Across Cell Membranes*, Academic Press, 1967)

the solution from I to II, that is, in a direction opposite to that of the osmotic gradient. See DONNAN EQUILIBRIUM.

**Filtration across capillaries.** Examples of hydrostatic movement of water or solution in biological systems may be found in capillary beds, including that in glomeruli of kidneys. As blood is pumped by the heart into the arterial side of a capillary system, the hydrostatic pressure head pushes an essentially protein-free filtrate of the blood plasma across the capillary wall into the interstitial fluid. When the blood reaches the venous side of the capillaries, the blood pressure has been dissipated and, as predicted from the balance of Donnan forces and of hydrostatic pressure forces, an osmotic flow of fluid occurs from the interstitial fluid, across the capillary wall, and into the venous capillary. In the steady state the amount of fluid filtered equals the amount of fluid reabsorbed from the capillaries. This is a skeletal description of the classic concept of capillary function patterned mainly after the work of E. Starling and E. Landis.

The aforementioned considerations apply not only to all animals possessing a vascular system but to any cellular forms containing protein-rich cytoplasm within a plasma membrane surrounded by a protein-free fluid environment. This criterion can be applied to unicellular animals and to multicellular animals, as well as to the tissues of practically all invertebrates and vertebrates, aquatic and terrestrial. In animal cells hydrostatic pressures are small (about 0.4–2.0 in. Hg or 10–50 mmHg) but sufficient to balance osmotic differences occasioned by the Donnan

forces operative across the plasma membrane. However, in plant cells the hydrostatic pressure can be relatively tremendous (15–30 atm), owing to the tough cellulose wall encasing the plasma membrane.

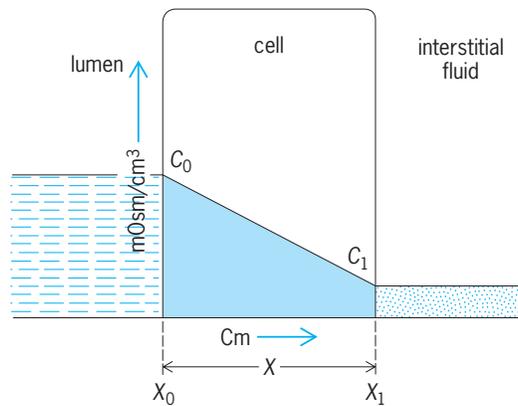
**Osmotic gradients.** The presence of osmotic pressure differences across porous membranes is considered to be one of the most important factors causing a net osmotic flow of water in biological systems. The usual assumption made for the system is that the membrane is permeable to solvent but not to solute. Such a requirement for the membrane is approximated in highly selective cation- or anion-permeable membranes. However, most living membranes, such as amphibian skin, nerve, eggs, erythrocytes, stomach, intestine, bladder, capillaries, and even mitochondria, appear to be permeable to solutes of small molecular weight—that is, of molecular weights up to at least 2000—as well as to water. This means that the water activity of cellular fluid tends to equilibrate rapidly with that of extracellular fluid so that appreciable osmotic differences between the phases rarely exist. Despite wide differences in chemical composition between cellular and extracellular fluid, there exist no measurable differences of water activity. Water activity has been evaluated by the usual measures of colligative properties such as freezing-point depression, melting point, or vapor-tension lowering.

The aforementioned remarks are not so general as they might appear, because large osmotic pressure differences are present, or appear to be present, across membranes in many biological systems. All of the osmoregulating forms present such osmotic pressure differences.

Examples of systems with apparent osmotic gradients are [body fluid—gills—pond water] in freshwater fish, crabs, and worms; [body fluid—renal tubular cell—urine] in kidneys producing urine either hypertonic or hypotonic to the body fluids; and [soil water—cell wall—cytoplasm] in plants.

According to van't Hoff's law, the osmotic pressure developed across an ideal membrane is  $22.4\Delta C = \pi$ , in atmospheres, where  $\Delta C$  is the concentration difference in the nonpenetrating solute on the two sides of the membrane. However, if the membrane is leaky and allows some of the solute to cross from one side to the other, the stopping pressure will be less than  $22.4\Delta C$ . The ratio of the measured stopping pressure to that predicted by van't Hoff's law is called the Staverman reflection coefficient,  $\sigma$ . Hence, for real membranes the stopping osmotic pressure will be predicted by  $\pi = \sigma \times 22.4\Delta C$ , where  $\sigma$  has a value between 1 and 0.

**Franck-Mayer hypothesis.** The steady-state maintenance of osmotic gradients across cells has been a biological problem for years. The magnitude of some gradients—such as the urine of a hydropenic dog, 1500 milliosmoles per kilogram (mOsm/kg), and plasma, 300 mOsm/kg—is too great to be explained by hydrostatic pressure gradients, except in plant cells. This led to the Franck and Mayer hypothesis of osmotic gradients. Apart from maintenance of gradients, their hypothesis included a mechanism



**Fig. 3.** Scheme of an intracellular osmotic gradient in a kidney tubule during production of osmotically concentrated urine. Osmotic activity  $C$  is plotted along the ordinate, and cell thickness  $X$  is plotted along the abscissa. (After C. A. Giese, *Cell Physiology*, Saunders, 1959)

for transporting water (solvent) from a solution of high osmolarity to a solution of low osmolarity; that is, the mechanism could move water up its gradient of activity.

**Figure 3** illustrates the essentials of the Franck-Mayer scheme as applied to the process of formation of hypertonic urine. When the intracellular osmotic activity  $C_0$  at the lumen side of a renal tubular cell is slightly greater than that of the luminal fluid (urine) and the intracellular osmotic activity at the interstitial side of the cell  $C_1$  is equal to that of the interstitial fluid, water could be transported from the lumen to the interstitial fluid. The postulated source of solutes at  $C_0$  was a depolymerizing reaction, enzymatically catalyzed, whence a large molecule was split into numerous small particles at the interface. Thermodynamically, the minimum free energy expenditure for maintenance of the gradient by the mechanism, regardless of its origin, must be at least equal to the heat dissipation or to the decrease of free energy of solute diffusing from  $X_0$  to  $X_1$ . The number of particles  $Q$  diffusing across an area  $A$  of cell in unit time for steady-state conditions is expressed as Eq. (7), where

$$Q = \frac{-DA(C_1 - C_0)}{X_1 - X_0} \quad (7)$$

$D$  is the diffusion coefficient and  $\Delta X = X_1 - X_0$ , the cell thickness of path length for diffusion. The equation is an integrated form of Fick's equation with the implicit assumption of a flat sheet of cells and zero bulk flow. The rate of decrease of free energy for the diffusion is expressed as Eq. (8), where  $\Delta F$  is

$$\frac{\partial \Delta F}{\partial t} = QRT \ln \frac{C_0}{C_1} \quad (8)$$

the change of free energy,  $t$  is the time,  $R$  is the gas constant, and  $T$  is the absolute temperature. Since diffusion is an irreversible process, the free-energy loss cannot be funneled back into the transporting or gradient-creating mechanism. An evaluation of the minimum rate of expenditure of free energy, made by A. Brodsky, B. Rehm, C. Dennis, and D. Miller, in the case of mammalian renal cell yielded a value of

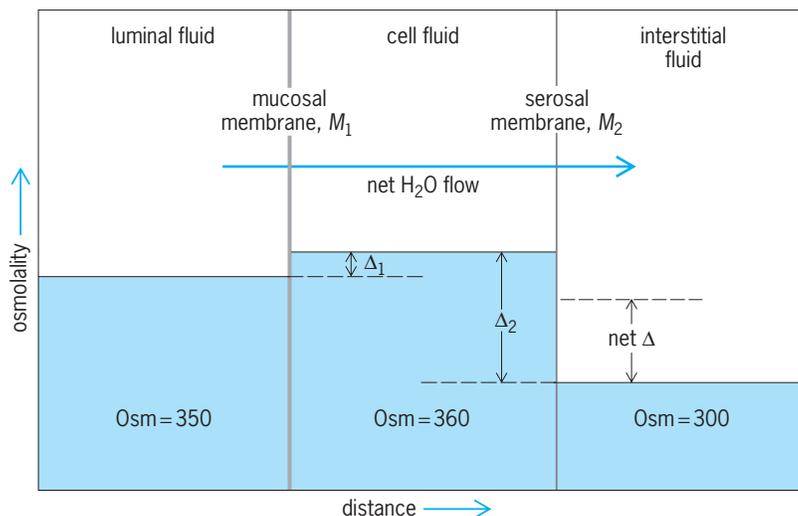
87.8 kilojoules/(kg)(h) or 21,000 kcal/(kg)(h). This is about 1000 times the maximal rate of respiration for living cells and imposes a severe limitation on the theoretical use of intracellular osmotic gradients as devices for water transport in biological systems. A major limitation of the Franck-Mayer scheme is the short diffusion path—one cell width, or about 20 micrometers.

Multicellular organisms are often required to maintain substantial osmotic gradients between various body fluids and secretory fluids. It is sometimes necessary for the organism to move water from a concentrated to a dilute solution against the chemical potential gradient for water. Moreover, these tasks must be accomplished without an inordinate expenditure of energy.

Two biological systems which move water against concentration gradients without too much energy expenditure are the three-compartment, two-membrane system and the countercurrent system.

*Three-compartment, two-membrane system.* This consists of three successive fluid compartments separated by two successive osmotic membranes. One of the two membranes of the system is a tight osmotic membrane, and the other is a leaky osmotic membrane.

**Figure 4** is a schematic representation of a three-compartment, two-membrane system. The figure could represent a mucosal cell separating the luminal fluid in the intestine from the interstitial or blood side of the intestinal membrane. In the figure the luminal fluid is represented as being hypertonic to the plasma by 50 mOsm, and the cell fluid as being hypertonic to the luminal fluid by 10 mOsm and hypertonic to the interstitial fluid by 60 mOsm. The net direction of water movement is from hypertonic luminal fluid to the interstitial fluid, against the overall concentration gradient for water.



**Fig. 4.** Osmolality (Osm) versus distance in a three-compartment, two-membrane system. Cell fluid bounded by membranes separates the hypertonic luminal fluid from the interstitial fluid. Cell fluid is hypertonic to both of the other fluids. Magnitudes of osmotic gradients across membranes  $M_1$  and  $M_2$ , indicated by  $\Delta_1$  and  $\Delta_2$ , are 10 and 60 mOsm, respectively. The magnitude of the overall gradient between luminal and interstitial fluids, indicated by net  $\Delta$ , is 50 mOsm. (After A. S. Crafts, H. B. Currier, and C. R. Stocking, *Water in Physiology of Plants*, Ronald Press, 1949)

The system operates as follows: Water (with practically no solute) is pulled across  $M_1$  into the cell by the transmembrane osmotic gradient. This causes an increased hydrostatic pressure within the cell fluid. The intracellular pressure developed is large enough to force water and solute through  $M_2$  in the direction opposite to the osmotic gradient. Thus the water and solute cross the leaky membrane  $M_2$  by a process of pressure-ultrafiltration with some sieving. In other words, water moves across  $M_1$  by an osmotic process and across  $M_2$  by a pressure-filtration process. Nevertheless, the system can be analyzed in terms of two opposing osmometers, one more efficient than the other in separating water from solute.

Now consider the system as two opposing osmometers, one composed of [luminal fluid—membrane ( $M_1$ )—cell fluid] and the other composed of [cell fluid—membrane ( $M_2$ )—interstitial fluid]. The degree of sieving, or the Staverman reflection coefficient of the tight osmotic membrane  $M_1$ , is larger than that for the leaky osmotic membrane  $M_2$ ; that is,  $\sigma_1 > \sigma_2$ . The osmotic pressure developed across a membrane is  $\Delta\Pi = \sigma 22.4\Delta$  Osm. Thus the net driving force on the water would be expressed as Eq. (9),

$$\Delta\Pi_{\text{net}} = 22.4[\sigma_1(\text{Osm}_{\text{cf}} - \text{Osm}_{\text{if}}) - \sigma_2(\text{Osm}_{\text{cf}} - \text{Osm}_{\text{isf}})] \quad (9)$$

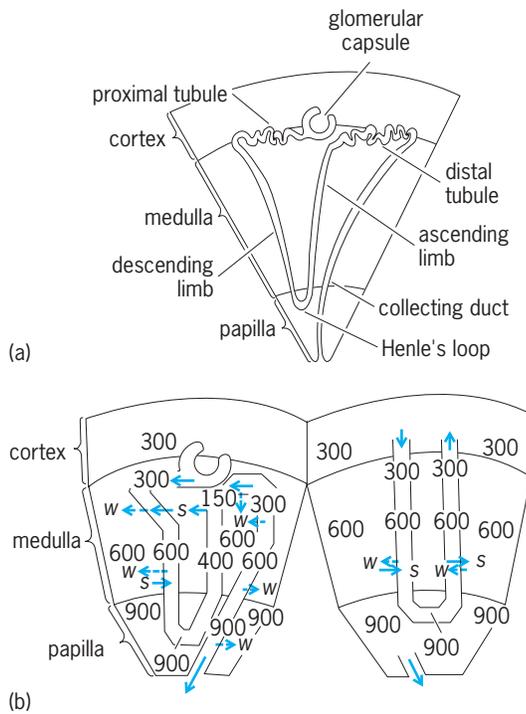
where  $\sigma_1$  or  $\sigma_2$  are the Staverman reflection coefficients of the membranes  $M_1$  and  $M_2$ , respectively, and the subscripts on the osmolarities refer to fluid in the lumen fluid (lf), cell fluid (cf), and interstitial fluid (isf).

In the example of Fig. 4, the values for the sigmas must be such that  $10\sigma_1 > 60\sigma_2$  in order to cause a net movement of water to the right.

The flow of water through the cell will carry solutes through the leaky membrane  $M_2$  and cause the cell fluid to become diluted. In order to keep the process going there must be a pump which transports the solute into the cell from one side or the other. By choosing the proper location for the solute pump and the proper values for  $\sigma_1$  and  $\sigma_2$  in such a model, water or solute, or both, can be made to move in either direction.

Given that the three-compartment system contains two or more solutes (as is the case in biological systems), it can be shown that water will move against the overall osmotic gradient from compartment two to compartment three even in the absence of any intracellular hydrostatic pressure gradient.

*Countercurrent system.* An ingenious model for moving water economically against osmotic gradients in the kidney is the so-called countercurrent multiplier system. This scheme was originated by analogy with the well-known physical laws of heat flow in concentric pipes containing counterflowing fluids. The elements of the scheme applied to the mammalian kidney are illustrated in Fig. 5. The first step in the production of osmotically concentrated urine is glomerular filtration, which produces an ultrafiltrate of 300 mOsm/liter, isoosmotic with plasma. The filtrate in the proximal tubule suffers a reduction of



**Fig. 5. Osmotic activity in the human kidney. (a) An anatomical reference diagram. (b) Schematic presentation of a renal countercurrent system for production of an osmotic gradient in the interstitial fluid and consequently for production of urine of osmotic activity greater than that of plasma. Numbers indicate osmotic activity in milliosmoles per liter; *s* = direction of the diffusion of salts; and *w* = direction of diffusion of water. (After A. Krogh, *Osmotic Regulation in Aquatic Animals*, Peter Smith Magnolia, 1939)**

volume without change of osmotic activity. Next the fluid enters the descending limb of the loop of Henle, where it suffers further reduction of volume but increases in osmotic activity by equilibrating with hypertonic interstitial fluid. This raises the questions of how hypertonic interstitial fluid is produced and how the osmotic gradient from cortex to papilla is maintained.

It has been shown that the ascending limb of the loop of Henle contains a salt pump which can pump salt from the lumen to the interstitial fluid and that this region of tubule is impermeable to water. Such a pump would tend to produce a region of hypertonicity in the interstitial fluid. The maintenance of the gradient along the long axis of the renal pyramids under steady-state conditions is occasioned by the countercurrent design of the vasa recta (postglomerular branches of arterioles from glomeruli in the subcortical regions of the kidney). As the descending part of the vasa recta plunges into the hypertonic papilla, salts diffuse into the blood, concentrating it toward the hypertonic level of papillar interstitial fluid. As the ascending branch of the vasa recta emerges from the papilla, it moves salt-laden blood toward the cortex, a region isotonic with peripheral blood. Therefore, salts diffuse from the blood toward the interstitial fluid and the osmotic gradient is maintained by what is called a countercurrent exchange system.

The source of energy for this process, referred to as the countercurrent multiplier, consists of the  $\text{Na}^+$  pumps along the ascending limb of the loop of Henle. As the hypertonic fluid entering the lumen moves up the water-impermeable ascending limb, it is diluted by the sodium pumps to such an extent that it becomes hypotonic to systemic plasma by the time it reaches the distal tubule. The distal tubular walls are water-permeable; water moves from the lumen to the isotonic interstitial fluid and back of the body via the renal vein. This step is the true water-economy mechanism for the body. A tubular fluid, isotonic with peripheral blood, then enters the cortical portion of the collecting duct. As the collecting duct proceeds toward the papilla, its contents equilibrate osmotically with those of the interstitial fluid. Since osmotic activity of interstitial fluid increases in moving from cortex to papilla, the osmotic equilibration renders the final urine hypertonic to the systemic plasma.

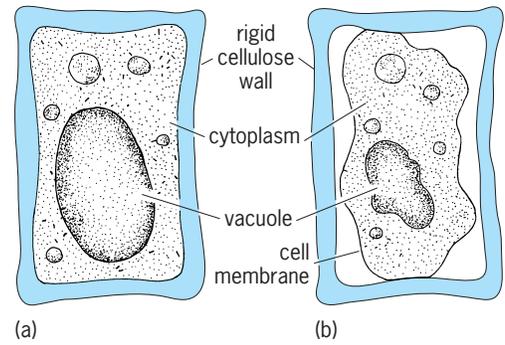
During periods of excess water intake the mammalian kidney elaborates a urine hypotonic to the plasma. Presumably, dilution of body fluids during water ingestion reduces secretion of the antidiuretic hormone of the posterior pituitary gland. Absence of this hormone reduces the water permeability of the distal tubules and the collecting ducts. Thus the dilute luminal fluid in these segments does not equilibrate osmotically with the surrounding interstitial fluid.

The biological applications of the countercurrent system have been fruitful for the explanation of diverse phenomena. It has been invoked to account for heat conservation in the extremities or skin of aquatic warm-blooded animals such as seals, whales, or ducks. It is present in gill systems, in swim bladders, or in any system containing or separating two solutions of greatly different osmotic activity. K. Schmidt-Nielsen has noted the importance of diffusion length in countercurrent systems, showing that the relative length or renal papillae in desert animals is greater than that in other animals. See COUNTERCURRENT EXCHANGE (BIOLOGY).

*Standing gradient system.* The countercurrent system described above is essentially a standing gradient system. Another type of standing gradient system, less complex than that of the kidney, may be involved in the movement of hypertonic or isotonic solutions across simple epithelial membranes. The mucosal cells of some secretory epithelia at the foot, or the interstitial boundary, of the cell and joined together at the top, or the luminal surface, by short tight junctions. The side walls of such cells are not strongly interdigitated and are not joined together, except at the tight junction at the top of the cell. Thus there is a system of one-ended channels between the cells. The channels, sealed at the luminal end, open into the interstitial fluid at the level of the basement membrane. The lateral walls of the cells are relatively permeable to water and impermeable to ions. Sodium pumps located on the lateral wall of the cell near the tight junction transport sodium from cell fluid into the channel, causing the channel to be filled

with hypertonic fluid at its closed end. As the hypertonic fluid moves down the channel toward the foot of the cell, water moves osmotically from the cell into the channel. As water moves into the channel along the length of its walls, the velocity of fluid flow in the channel increases and the salt concentration decreases toward isotonicity at the open end of the channel. In the steady state the osmotic profile of the fluid along the axis of the channel changes with distance but not with time. The steady-state gradient is referred to as a standing gradient. Such a system has been proposed by J. Diamond, who has provided a mathematical analysis of the system. The hypertonicity caused by the salt pumps at the head of the channel provides an osmotic force for the movement of the final isotonic fluid against an osmotic or hydrostatic pressure.

**Figure 6** is a schematic representation of a standing gradient system. The standing gradient resides in the channel between the two epithelial cells. The lower panel of the figure represents the osmotic profile of luminal fluid—the fluid in the channel and the interstitial fluid. The shaded area at the beginning of the channel represents the zona occludens, or the tight junction between the luminal surfaces of the cells. The thin arrows marked  $H_2O$  represent the path of water movement across the epithelial membrane. Sodium chloride diffuses into the cell across the luminal membrane of the cell and is actively trans-



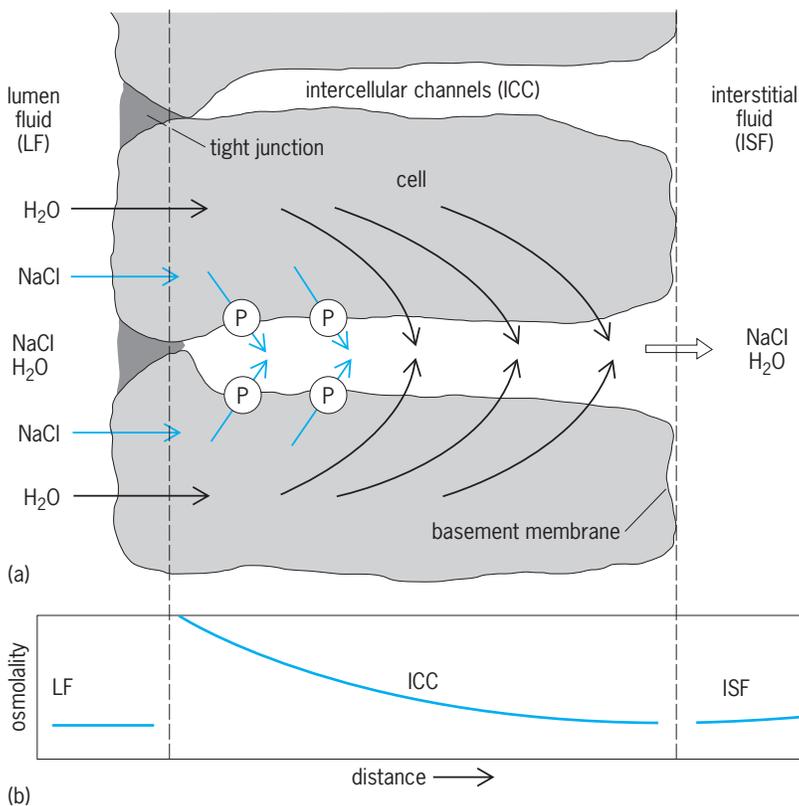
**Fig. 7.** Osmotic behavior of a plant cell during plasmolysis. (a) The plant cell is immersed in an external medium which is hypotonic to the cytoplasm. (b) The cell is immersed in an external medium which is hypertonic to the cytoplasm. (After J. T. Edsall and Wyman, *Biophysical Chemistry*, vol. 1, Academic Press, 1958)

ported into the channel near the site marked, P, the NaCl pump.

**Osmotic gradients in plant cells.** A typical plant in is that found lining the pith, phloem, cortex, or xylem of succulent plant organs. It consists of a cell wall of great tensile strength, cytoplasm, and nuclei. An osmotic gradient between the cell and its surrounding fluid is created by a mechanism transporting salts from the external region of low concentration to the intracellular region of relatively high salt concentration. The mechanism for the transport could be the transmembrane transport of an anion such as nitrate ( $NO_3^-$ ) at a site separate from that of cation movement. Alternatively, it could be by formation of ion pairs ( $NaNO_3$ ) in the membrane. In either case the osmotic pressure difference across the membrane provides a force oriented to drive water from the exterior to the interior of the cell. Water penetrates into the cell until the internal (turgor) pressure is sufficient to stop the flow. Such turgor pressures are of the order of 10–20 atm (100–200 kilopascals). See PLANT CELL; PLANT-WATER RELATIONS.

**Plasmolysis.** The process (**Fig. 7**) of shrinking the cytoplasm and vacuole by immersing an isolated plant cell in a solution of higher osmotic activity than that of a plant cytoplasm is called plasmolysis. Presumably, no osmotic shrinking or swelling would occur if the cells were immersed in a solution of the same osmotic activity as that of plant cytoplasm. Such data of plasmolysis may not give a true measure of osmotic activity of plant cells because of solute penetration. Nevertheless, a good approximation may be made in cells from which solute leaks very slowly and into which external solute does not penetrate. The data are usually checked by cryoscopic determinations on the juice of pulverized cells. Discrepancies in the results have been found when both techniques have been applied to the same plant tissue.

**Electroosmosis.** If identical aqueous solutions are separated by a porous membrane containing fixed charges along the walls of the pores and a voltage is applied across the membrane, water will flow across the membrane, against opposing hydrostatic or osmotic forces. If the walls of the pores are lined with negative charges, then some of the positive ions or



**Fig. 6.** Movement of an isotonic solution through cells. (a) Layer of epithelial cells transporting an isotonic solution of NaCl and  $H_2O$ . P = salt pumps. (b) Osmotic profile of the standing gradient in the intracellular channels. Vertical broken lines show the spatial relation between the plot and the sketch in a. (After C. L. Prosser et al., eds., *Comparative Animal Physiology*, 2d ed., Saunders, 1950)

$H^+$  from the dissociated water will be adsorbed to the walls of the pore, leaving the bulk solution in the pore with a net negative charge. Thus, when a voltage is applied across the membrane, water will move toward the positive side of the membrane.

Numerous biological membranes involved in osmoregulatory functions are known to be ion-selective and to have electrical fields across them. Thus electroosmosis may play a role in the movement of water across biological membranes. Electroosmosis and the converse phenomenon of streaming potentials have been observed in the wall of the gallbladder.

**Miscellaneous forces.** Thermal gradients, if present in living cells, could provide a force for the movement of water. However, the magnitude of temperature gradients is small across most cell membranes and, consequently, the amount of water moved would be small.

Chemical reactions with production or consumption of water would provide a device for water transport. However, the molar quantity of water involved in most known reactions is small compared to the known rates of water transport in animals.

Hydration and dehydration of protein molecules have been invoked as water-transporting processes. It is difficult to tell whether a protein molecule adsorbs pure solvent or bulk solution, and it is even more difficult to see how much a system could drive water in a given direction.

Contractile vacuoles, the structures in the cytoplasm of protozoans such as the paramecia, have been observed extensively while the paramecia were immersed in solutions of varying osmotic activity. The vacuole appears to fill with water, burst, and consequently extrude its contents. The rate of pulsation of the vacuole increases as the external medium is rendered dilute. Vacuolar activity depends on metabolic energy, as shown by the suppressing action of cyanide. The mechanism of water transport of contractile vacuoles could be by a secretion of water into the vacuole or by secretion of solute and diffusion of water into the vacuole.

Endocytosis is a term applied to the engulfing of water or solution by pseudopodlike processes of leukocyte membranes. The droplet of water is engulfed by the outer surface of the membrane, wherein it migrates to the inner surface, at which point it is pushed into the cell. See ENDOCYTOSIS.

**Hormonal control systems.** Given the presence of water and solute-transporting mechanisms, the organism requires controlling machinery to govern such processes so that effective osmotic regulation will be maintained. The system in mammals, though complex, has been studied extensively. Therefore, a presentation of the elements of such a system can provide a model that is applicable by analogy to the osmoregulating system of other animal forms. Anatomical elements of the mammalian system are the neurohypophyseal tract of the hypothalamus and the posterior pituitary and the kidney tubule, probably the distal part plus collecting ducts. Osmoreceptor cells, sensitive to small changes of osmolality

of the perfusing fluids, are present in the suprahypophyseal tract of the hypothalamus. When these cells respond to osmotic stimuli, impulses are sent through the nerve tract to the posterior portion of the pituitary gland. The pituitary gland responds by increasing or by decreasing the rate of secretion of the antidiuretic hormone (ADH), vasopressin. The hormone, after reaching the blood, is carried to the kidney, where it affects the water-transporting cells of the tubule and collecting ducts.

When the plasma becomes hypertonic (as after dehydration or excessive salt intake), secretion of ADH is increased, causing the walls of the collecting ducts of the kidney to become permeable to water. The fluid in these ducts then equates with the hypertonic papilla of the kidney. Hypertonic urine is produced and free water is returned to the circulation.

When the plasma becomes hypotonic (after excessive water intake), secretion of ADH is reduced and the walls of the collecting duct become impermeable to water; the hypotonic fluid in the collecting ducts fails to equilibrate with the renal papilla. Thus hypotonic urine is produced, whence free water is removed from the circulation.

By virtue of its primary action in the retention or elimination of water per se, vasopressin plays a role in volume control of body fluids. Other hormones, notably the adrenal cortical hormone aldosterone, are involved in volume control. However, their primary action is on the retention or elimination of salts which are followed passively by water. See ENDOCRINE MECHANISMS. William A. Brodsky; T. P. Schilb  
Bibliography. T. E. Andreoli et al. (eds.), *Membrane Physiology*, 1980; C. J. Duncan and C. R. Hopkins (eds.), *Secretory Mechanisms*, 1980; J. F. Hoffman (ed.), *Membrane Transport Processes*, vol. 1, 1978.

## Osmosis

The transport of solvent through a semipermeable membrane separating two solutions of different solute concentration. The solvent diffuses from the solution that is dilute in solute to the solution that is concentrated. The phenomenon may be observed by immersing in water a tube partially filled with an aqueous sugar solution and closed at the end with parchment. An increase in the level of the liquid in the solution results from a flow of water through the parchment into the solution. The process occurs as a result of a thermodynamic tendency to equalize the sugar concentrations on both sides of the barrier. The parchment permits the passage of water, but hinders that of the sugar, and is said to be semipermeable. Specially treated collodion and cellophane membranes also exhibit this behavior. These membranes are not perfect, and a gradual diffusion of solute molecules into the more dilute solution will occur. Of all artificial membranes, a deposit of cupric ferrocyanide in the pores of a fine-grained porcelain most nearly approaches complete semipermeability.

The walls of cells in living organisms permit the passage of water and certain solutes, while preventing the passage of other solutes, usually of relatively high molecular weight. These walls act as selectively permeable membranes, and allow osmosis to occur between the interior of the cell and the surrounding media.

The flow of liquid through such a barrier may be stopped by applying pressure to the liquid on the side of higher solute concentration. The applied pressure required to prevent the flow of solvent across a perfectly semipermeable membrane is called the osmotic pressure and is a characteristic of the solution.

**Osmotic pressure.** The driving force for osmosis is the decrease in Gibbs free energy accompanying the dilution of the solution by the transfer of solvent to the solution. With  $G_1^0$  representing the molar free energy of the pure solvent, the molar free energy of solvent in the solution is shown in Eq. (1), where  $R$  is

$$G_1 = G_1^0 + RT \ln a_1 \quad (1)$$

the molar gas law constant,  $8.314 \text{ J mole}^{-1} \text{ K}^{-1}$ , and  $a_1$  is the solvent activity in the solution. This is conventionally taken as unity for the pure solvent, and in ideal solutions, as the mole fraction,  $x_1$ . The pressure that must be applied to the solution phase to increase the solvent free energy to that of the pure solvent is the osmotic pressure,  $\pi$ , of the solution. The rate at which the free energy of a system is changed at a fixed temperature by pressure is given by the basic thermodynamic relationship (2). For an incompress-

$$(\delta G / \delta P)_T = V \quad (2)$$

ible liquid solution, the osmotic pressure is, then, shown in Eq. (3), where  $V_1$  is the molar volume of sol-

$$\pi = -(RT / V_1) \ln a_1 \quad (3)$$

vent. Equation (3) represents the fundamental relationship between the osmotic pressure and solution thermodynamics. In an ideal solution containing  $n_1$  moles of solvent and  $n_2$  moles of solute, the solvent activity is  $a_1 = x_1 = n_1 / (n_1 + n_2)$ . In terms of the solute, the solvent activity is  $(1 - x_2)$ , where  $x_2$  is the solute mole fraction,  $n_2 / (n_1 + n_2)$ . For dilute solutions,  $\ln a_1 = \ln(1 - x_2) \sim -x_2$  since  $n_1 \gg n_2$ . Then Eq. (3) becomes Eq. (4), where  $n_1 V_1$  is replaced by the total

$$\pi = n_2 RT / (n_1 V_1) = n_2 RT / V = cRT \quad (4)$$

volume,  $V$ . Equation (4) is recognized as identical in form to the ideal gas law, in which  $c$  represents the solute concentration in moles  $\text{L}^{-1}$ . As applied to osmotic phenomena, Eq. (4) is called the van't Hoff equation. Measurements of osmotic pressure made early in the twentieth century on dilute sucrose solutions using copper ferrocyanide membranes confirm the applicability of (4). For example, a 0.0291 molar aqueous sucrose solution at  $15.6^\circ\text{C}$  exhibited an osmotic pressure of 0.684 atm compared with the calculated value of 0.689 atm.

Solution properties reflecting changes in solvent free energy caused by a nonvolatile solute include vapor-pressure lowering, boiling-point elevation, freezing-point depression, as well as osmotic pressure. Of these, the last is the most sensitive to solute concentration. Membrane requirements, however, limit its usefulness in studying solution thermodynamics.

An important application of measuring osmotic pressure is determining a polymer's molar mass. The polymer sample is dissolved in an organic solvent or in a mixture of solvents, and the osmotic pressure is measured using a pretreated cellophane film membrane. An osmometer is used in which the pressure is measured in terms of the height to which liquid rises in a solution cell capillary above that in the surrounding solvent. In accounting for nonideality, Eq. (3) is used in a modified form (5), where  $c^*$  is

$$\pi / c^* = RT / M + Bc^* + Cc^{*2} \quad (5)$$

the concentration in g/L, and  $M$ , the molar mass in g/mole.  $B$  and  $C$  are correction terms analogous to gas-phase virial coefficients.  $RT/M$  is obtained as a limit of  $\pi/c^*$  as  $c^*$  approaches zero. The molar mass obtained is a number average of the polymer molar mass spectrum of the sample.

Aside from the inorganic-type membranes, which are formed by deposition of the salt in a porous matrix, membranes suitable for measurements of osmotic pressures in aqueous systems are not available. A number of synthetic polymeric films, which allow the passage of water and small ionic or molecular species, are sufficiently strong to act as selectively permeable membranes in aqueous solutions. These have been successfully used to study, for example, micellar interactions in salt and buffered media. Although the pressures measured are not true osmotic pressures, thermodynamic solute data may be obtained through a redefined  $G_1^0$  and use of standard Gibbs-Duhem activity relationships. See FREE ENERGY.

**Reverse osmosis.** Just as the osmotic pressure is that pressure which when applied to the solution phase will prevent the solvent from passing through a semipermeable membrane into the solution, the application of greater pressure will cause solvent to pass from the solution into the pure solvent, or reverse osmosis. Reverse osmosis has long been considered for water purification. Home units for improving the quality of a water supply are effectively pressure filtration devices capable of removing colloidal impurities but are largely ineffective in reducing salt concentrations. On a larger scale, polymeric films or ceramic porous membranes, for example, of aluminum oxide, are described in terms of the fraction of salt rejection as a measure of efficiency. The prospect of using reverse osmosis for desalinating seawater is energetically attractive but technically difficult to realize. The osmotic pressure of seawater is approximately 23 atm. Membrane design and mechanics would have to allow for use at substantially higher pressures with large volumes of water,

over long periods. With advances in membrane engineering, it is possible that reverse osmosis could be used to purify ocean water. See EDEMA; OSMOREGULATORY MECHANISMS; PLANT-WATER RELATIONS; SOLUTION.

Francis J. Johnston

Bibliography. D. A. Amos et al., Osmotic pressure and interparticle interactions in ionic micellar surfactant solutions, *J. Phys. Chem. B*, pp. 2739–2753, 1998; P. W. Atkins, *Physical Chemistry*, 5th ed., W. H. Freeman, 1994; P. R. Bergethon and E. R. Simons (eds.), *Biophysical Chemistry*, 1984; S. Glasstone, *Textbook of Physical Chemistry*, Van Nostrand, 1946; K. Kotyk, K. Koryta, and K. Janacek, *Biophysical Chemistry of Membrane Functions*, 1988; S. Sourirajan and T. Matsuura, *Reverse Osmosis and Ultrafiltration*, American Chemical Society, 1985.

## Ostariophysi

A very large group of fishes containing 27% of all known fishes of the world and accounting for 64% of the freshwater species. Although the perciforms dominate modern seas, the ostariophysans command preeminent faunal importance in freshwaters. At least one-half of the freshwater fishes of each continent, with the exception of Australia, belong to this group. See PERCIFORMES.

The classification hierarchy is as follows:

- Class Actinopterygii
  - Subclass Neopterygii
    - Division Teleostei
      - Subdivision Euteleostei
        - Superorder Ostariophysi
          - Series Anotophysii
            - Order Gonorhynchiformes
          - Series Otophysi
            - Order Cypriniformes
            - Order Characiformes
            - Order Siluriformes
            - Order Gymnotiformes

Several characteristics alone unite the Ostariophysi: modification of the anterior three to five vertebrae as a protective encasement for a series of bony ossicles that form a chain connecting the swim bladder with the inner ear, the so-called Weberian apparatus; multicellular horn tubercles, in the form of breeding and nuptial tubercles or pearl organs; minute unicellular horny projections called unculi; and an alarm substance (pheromone) that can be released by injured tissue to elicit a fright reaction.

Other identifying characteristics include abdominal pelvic fins, if present, usually with many rays; a pelvic girdle which does not have contact with the cleithra; and a more or less horizontal pectoral fin placed low on the side. Usually there are no fin spines, but one or two spines derived from consolidated soft rays may be found in dorsal, anal, or pectoral fins, and a spinelike structure is present in the adipose fin of certain catfishes. The upper jaw

is often bordered by premaxillae and maxillae, the latter commonly toothless and rudimentary. An orbitosphenoid is present, and a mesocoracoid is usually present. There is a swim bladder, usually with a duct (physostomous). Scales are the cycloid type, which may be modified into bony plates, and branchiostegal rays are variable in number and arrangement. See EAR (VERTEBRATE); SALMONIFORMES; SWIM BLADDER.

Herbert Boschung

## Osteichthyes

A group that is recognized herein as consisting of the classes Actinopterygii (rayfin fishes) and Sarcopterygii (lungfishes and lobefins, or coelacanth, but excluding tetrapods). (Euteleostomi is a monophyletic group consisting of Actinopterygii and Sarcopterygii, including tetrapods.) The Sarcopterygii and Actinopterygii are well-defined phyletic lines which had evolved by the Middle Devonian, and many members of the two taxa persist today.

The osteichthyans (bony fishes) include most of the Recent fishes. They differ from the living Agnatha in having jaws, paired nostrils, true teeth, paired pelvic and pectoral fins supported by girdles (unless lost secondarily), three semicircular canals, and bony scales (unless lost or modified). Many fossil agnathans possess bony scales but differ from the higher fishes in the above-mentioned features. Separation of the osteichthyans from the Paleozoic Placodermi and Acanthodii is more difficult because these three groups agree in most basic vertebrate features. The osteichthyans contrast with the Chondrichthyes in having a bony skeleton (although some Recent bony fishes possess a largely cartilaginous skeleton), a swim bladder (at least primitively), a true gill cover, and mesodermal scales (sharks possess dermal denticles, or placoid scales) which are sometimes modified or lost. Fertilization is usually external, but if it is internal, the intromittent organ is not derived from pelvic-fin claspers. Most often a modified anal fin or a fleshy tube or sheath functions in sperm transfer. See COPULATORY ORGAN; SCALE (ZOOLOGY); SWIM BLADDER.

**Classification.** Some authorities rank the Actinopterygii and Sarcopterygii in three subclasses: Actinopterygii (rayfin fishes), Coelacanthiformes or Crossopterygii (lobefin fishes), and Dipnoi (lungfishes). Others may combine the lobefin fishes and lungfishes in one subclass, Sarcopterygii or Choanichthyes. The Paleozoic Acanthodii have been variably assigned as a subclass of the Osteichthyes, of the Placodermi, or (as in this encyclopedia) as a separate class. The Polypteriformes (bichirs) are ranked by some as a subclass of the Osteichthyes, but are here placed as an order of the chondrosteian Actinopterygii. The osteichthyans as perceived herein had evolved by the Middle Devonian and all persist today; thus each of these subclasses has a longer history than any of the tetrapods, none of which had then appeared.

The lobefin fishes and lung fishes were destined to

experience moderate success until the early Mesozoic, after which they barely persisted as an element in the world's ichthyofauna. Today's fauna of these fishes includes only four families, four genera, and seven species. Of great interest are the rhipidistians (sarcopterygians in the order Osteolepiformes), the paternity of tetrapods. *Latimeria* is a member of the Coelacanthiformes, the other major radiation of crossopterygians. The Dipnoi are especially interesting scientifically because of their air-breathing habit and the capacity of some to estivate for a protracted period. See DIPNOI.

In the late Paleozoic the slowly developing Actinopterygii realized marked success in the palaeonisciform line, a group that was superseded by ascendant development of amiiforms and related groups in the Mesozoic only to be replaced by the enormous and highly successful outburst of salmoniform, cypriniform, siluriform, and perciform descendants in the Upper Cretaceous and early Cenozoic. Most modern fishes belong to these latter groups, which are conveniently termed teleosts (that is, the end of the bony fish line.) The Recent osteichthyan fauna includes 2 classes, 45 orders, 435 families, 4076 genera, and 24,497 species.

A classification of extant and selected fossil osteichthyans follows. Equivalent names are given in parentheses. For more detailed information, see separate articles on each group.

#### Osteichthyes

##### Class Actinopterygii

###### Subclass Cladistia

Order Polypteriformes  
(Brachiopterygii)

###### Subclass Chondrostei

Order Acipenseriformes

###### Subclass Neopterygii

Order Lepisosteiformes (in part,  
Semionotiformes)  
Amiiformes

###### Division Teleostei

###### Subdivision Osteoglossomorpha

Order Hiodontiformes  
Osteoglossiformes

###### Subdivision Elopomorpha

Order Elopiformes  
Albuliformes  
Anguilliformes (Apodes)  
Saccopharyngiformes

###### Subdivision Clupeomorpha

Order Clupeiformes  
Subdivision Euteleostei

###### Superorder Ostariophysi

Series Anotoptysi  
Order Gonorhynchiformes

###### Series Otophysi

Order Cypriniformes  
Characiformes  
Siluriformes  
(Nematognathi)  
Gymnotiformes

###### Superorder Protacanthopterygii

Order Argentiniformes,  
Osmeriformes  
Salmoniformes  
Esociformes (Haplomi,  
Esocae)

###### Superorder Stenopterygii

Order Stomiiformes  
(Stomiatiiformes)  
Ateleopodiformes

###### Superorder Cyclosquamata

Order Aulopiformes

###### Superorder Scopelomorpha

Order Myctophiformes

###### Superorder Lampridiformes

(Lampriformes)  
Order Lampriformes

###### Superorder Polymixiomorpha

Order Polymixiiformes

###### Superorder Paracanthopterygii

Order Percopsiformes  
Gadiformes  
Ophidiiformes  
Batrachoidiformes  
(Haplodoci)  
Lophiiformes

###### Superorder Acanthopterygii

###### Series Mugilomorpha

Order Mugiliformes

###### Series Atherinomorpha

Order Atheriniformes

Beloniformes

Cyprinodontiformes

(Microcyprini)

###### Series Percomorpha

Order Stephanoberyciformes

(Xenoberyces, in part)

Beryciformes

Zeiformes

Order Gasterosteiformes

###### Suborder Gasterosteoidi (Thoracostei)

###### Suborder Syngathoidi (Solenichthys)

Order Synbranchiformes

(Symbranchii)

Scorpaeniformes

Perciformes

###### Suborder Percoidei

Elassomatoidei

Labroidei

Zoarcoidei

Notothenioidei

Trachinoidei

Pholidichthyoidei

Blennioidei

Icosteidei

Gobiesocoidei (Xenopterygii)

Callionymoidei

Gobioidei

Kurtoidei

Acanthuroidei

Scombrolabracoidei

Scombroidei

Stromateoidei

- Suborder Anabantoidei (Labyrinthici, in part)
- Channoidei (Ophiocephaliformes)
- Caproidei
- Order Pleuronectiformes (Heterostomta)
- Tetraodontiformes (Plectognathi)
- Class Sarcopterygii
  - Subclass Coelacanthimorpha
    - Order Coelacanthiformes
  - Subclass Dipnotetrapodomorpha
    - Infraclass Dipnoi (Dipneusti)
      - Order Ceratodontiformes

There is a lack of general agreement as to the nomenclature of higher groups of fishes. Broadly, two systems are employed, that of C. T. Regan and the later one of L. S. Berg. In the Berg system, followed here, ordinal names are formed by a standard suffix “-formes” to the step of a type genus.

Reeve M. Bailey; Herbert Boschung

**Phylogeny.** The Osteichthyes evolved from some group among the primitive, bony gnathostome fishes. Plated agnathans and primitive acanthodians are known from the Silurian, but the oldest osteichthyans do not appear until the Lower Devonian, when lobefins and lungfishes enter the paleontological record. The osteichthyans became well represented in the Middle Devonian, at which time the rayfin fishes, lobefin fishes, and lungfishes were well differentiated. The ancestral stock of the osteichthyans may be sought among the other major groups of gnathostomes, the Placodermi and Acanthodii, but the still-fragmentary character of the early fossil record prevents clear elucidation of the origin and initial history of the group. The Placodermi, abundant fishes of the Devonian, were heavily armored and highly specialized, and they display scant evidence of relationship to early osteichthyans. Advanced acanthodians of the Devonian also were highly modified and are unlikely progenitors of osteichthyans. The occurrence of more generalized types in the Late Silurian, about the time when osteichthyans presumably originated, makes the acanthodians a more plausible ancestral stock. Their overall morphology, bony skeleton, including true bony scales of structure not unlike that of osteichthyans, and the presence of paired pectoral and pelvic fins are all suggestive of kinship with that group. The caudal fin of acanthodians is heterocercal, as in primitive osteichthyans. The other fins had strong spines at their leading edges. Commonly there were paired spines between the pectoral and pelvic fins. Such structures are unknown in primitive osteichthyans. In view of the above, it is a plausible inference that the Osteichthyes and Acanthodii diverged from a common basal stock, presumably in Silurian time. See ACANTHODII; PLACODERMI.

Although the beginnings of the osteichthyans are shrouded in uncertainty, advancements in paleontology have laid to rest the belief that they

evolved from that other great class of modern fishes, the Chondrichthyes. According to this belief, cartilage preceded bone in phylogeny as well as in ontogeny; cartilaginous fishes were seen as ancestral to bony fishes. Support for this assumption was drawn from the occurrence of largely cartilaginous endoskeletons in modern lungfishes and sturgeons, representatives of ancient lineages of osteichthyans. In contradiction, however, it has been shown that early members of these osteichthyan groups are better ossified than are their secondarily chondrified descendants. The adaptive advantage of cartilage in developmental stages appears to explain its prevalence in young fishes. Bone long preceded the oldest known chondrichthyans, and the Osteichthyes themselves precede the Chondrichthyes in the fossil record. See CHONDRICHTHYES.

**Fossils.** Osteichthyes contains many fossil fishes, including the ancestors of tetrapods. Their earliest satisfactory record is Lower Devonian, but remains of fossils of possible osteichthyan lineage have been found in the Upper Silurian.

The upper biting edge of the mouth is formed by dermal premaxillary and maxillary bones to which teeth are fused, rather than by palatoquadrates (as in acanthodians and elasmobranchs). Lateral line canals run through the dermal bones and scales rather than between them. A dermal parasphenoid forms the roof of the mouth, and other dermal bones form an opercular-shoulder girdle complex around the branchial chamber. Primitively the neurocranium is perichondrally ossified in two parts, anterior and posterior, which are partly separated by cranial fissures (lost in higher bony fishes). Also primitively, the fins are supported by dermal lepidotrichia. An air bladder is usually present, sometimes modified into a lung or hydrostatic organ.

John G. Maisey

**Bibliography.** L. S. Berg, *Classification of Fishes, Both Recent and Fossil*, 1947; P. H. Greenwood et al., Phyletic studies of teleostean fishes, with a provisional classification of living forms, *Bull. Amer. Mus. Nat. Hist.*, 131:339-456, 1966; P. H. Greenwood et al. (eds.), *Interrelationship of Fishes*, 1973; E. S. Herald, *Living Fishes of the World*, 1961; J. A. Moy-Thomas and R. S. Miles, *Palaeozoic Fishes*, 2d ed., 1971; J. S. Nelson, *Fishes of the World*, Wiley, 1994; C. Patterson and D. E. Rosen, Review of ichthyodectiform and other Mesozoic teleost fishes and the theory and practice of classifying fossils, *Bull. Amer. Mus. Nat. Hist.*, 158:81-172, 1977; A. S. Romer, *Vertebrate Paleontology*, 3d ed., 1966; D. E. Rosen et al., Lungfishes, tetrapods, paleontology, and pleisiomorphy, *Bull. Amer. Mus. Nat. Hist.*, 167:163-275, 1981; B. Schaeffer and D. E. Rosen, Major adaptive levels in the evolution of the actinopterygian feeding mechanism, *Amer. Zool.*, 1:187-204, 1961.

## Osteoglossiformes

An order of teleost fishes consisting of two monophyletic clades, suborders Osteoglossoidei and Notopteroidei. The Osteoglossoidei consists of one

family, the Osteoglossidae (bonytongues and butterfly fishes), which occur in freshwaters of tropical South America, Africa, and Southeast Asia to northern Australia. The Notopteroidei consist of three extant families, Notopteridae (Old World knifefishes and featherfin knifefishes, or featherbacks), Mormyridae (elephantfishes), and Gymnarchidae (gymnarchid eel). The featherbacks are native to central tropical Africa, India, and the Malay Archipelago. The elephantfish family, having a long proboscislike snout in many species, is the most species-rich osteoglossiform taxon but is limited to tropical Africa and the Nile River. The family Gymnarchidae is represented by a single species, *Gymnarchus niloticus*, of tropical Africa including the upper Nile.

These fishes, formerly assigned to the orders Clupeiformes (Isospondyli) and Mormyriiformes, are accorded a revised classification as a result of modern phylogenetic research on lower teleosts.

**Anatomy and development.** Osteoglossiformes have the primary bite between the well-toothed tongue and the strongly toothed parasphenoid and certain pterygoid bones in the roof of the mouth. The mouth is bordered by a small premaxilla, which is fixed to the skull, and the maxilla. The anterior part of the intestine passes posteriorly to the left of the esophagus and stomach, whereas in all but a very few fishes the intestine pass to the right. A unique feature is the presence of paired bony rods at the base of the second gill arch. Additional characters include the absence of epineural intermuscular bones; a single dorsal fin of soft rays (or absent in some featherbacks) [Figs. 1 and 2]; lack of adipose fin; caudal fin either well developed, reduced, or absent, its supporting skeleton usually compact and consolidated; pelvic fin usually abdominal, not attached to cleithrum; pelvic rays usually 6 or 7, the fin sometimes rudimentary or absent; supraoccipital separated from frontals by parietals; orbitosphenoid usually present; a mesopterygoid arch; no Weberian apparatus; branchiostegals 3-17; and cycloid scales that are commonly large and often have a complex reticulum formed by the radii. Unlike the Clupeiformes, there is no extension of the cephalic canal

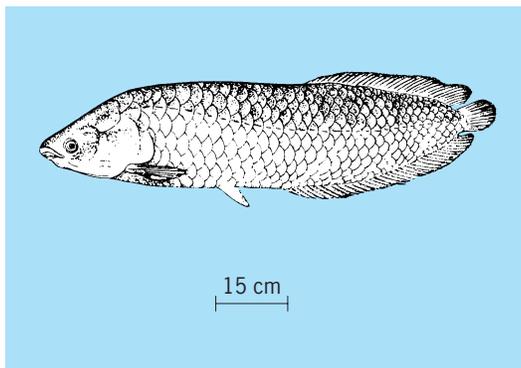


Fig. 1. African bonytongue (*Heterotis niloticus*). (After G. A. Boulenger, *Catalogue of the British Museum, Natural History*, vol. 1, 1909)

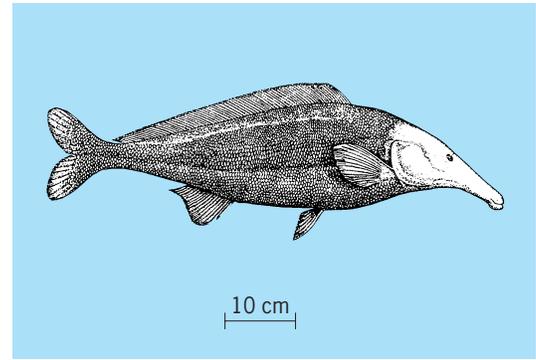


Fig. 2. African elephantnose (*Mormyrus probosciostris*). (After G. A. Boulenger, *Catalogue of Fresh Water Fishes of Africa in the British Museum, Natural History*, vol. 1, 1909)

system onto the opercle. In contrast to the Elopiformes, development is without a leptocephalous (elongate and flattened side to side) larva. See CLUPEIFORMES; ELOPIFORMES; TELEOSTEI.

**Evolution and adaptations.** Osteoglossiforms represent one of the basal stocks among the teleosts. The Ichthyodectidae, including such giants as *Xiphoctis* (or *Portheus*), and two related families are well represented in Cretaceous marine deposits, and *Allobrissops* and *Pachybrissops* of the Jurassic and Cretaceous probably belong here.

Recent members may be classified in 2 suborders, 6 families, 29 genera, and about 217 species. All inhabit freshwater, and all are tropical.

Bonytongues are well known because of their economic importance, large size, and frequent exhibition in public aquariums. They are represented by five Recent species: the microphagous, African *Heterotis* (Fig. 1), the giant *Arapaima gigas* (over 10 ft or 3 m long), and the arawana (*Osteoglossum bicirrhosum*) of the Amazon, and two species of *Scleropages* of southeastern Asia, New Guinea, and Australia. The North American Eocene *Phareodus* is a typical osteoglossid. A related group consists of a single small species known as African butterflyfish (*Pantodon buchholzi*). Butterflyfish leap a meter or more out of the water and glide with the help of their expansive pectoral fins. The featherbacks, Notopteridae, with five species from Africa and southeastern Asia, are compressed fishes with tapering bodies and long anal fins that are continuous with the caudal fin; the dorsal fin, if present, is a tiny tuft of a few rays.

Numerically Osteoglossiformes is dominated by the Mormyridae. Of the roughly 198 species, many are valued as food. The elongate snout of some species accounts for the vernacular name elephantfishes (Fig. 2). The caudal peduncle is slim in all species; in most the dorsal and anal fins are approximately equal, but in some the anal is much the longer and in others the dorsal greatly exceeds the anal (Fig. 2). In the related family Gymnarchidae, consisting of a single species that attains a length of 1.5 m, the pelvic, anal, and caudal fins are absent. Mormyrids and gymnarchids are of especial interest

in that they are electrogenic. Modified muscles in the caudal peduncle generate and emit a continuous electric pulse. The discharge frequency varies, being low at rest and high when the fish is under stress. The mechanism operates like a radar device, since the fish is alerted whenever an electrical conductor enters the electromagnetic field surrounding it. The receptor mechanism is not fully understood, but the brain of mormyrids (especially the cerebellum) is the largest of all fishes. Mormyrids have small eyes and often live in turbid water, so this electric mechanism is an effective evolutionary solution to a need for keen awareness of the environment and is comparable, for example, to acute hearing or vision. Electric discharge has been developed in several groups of fishes; in the African mormyrids the discharge is convergent with a similar but wholly unrelated acquisition in the gymnotids or knifefishes of South America. See ACTINOPTERYGII; CYPRINIFORMES; ELECTRIC ORGAN (BIOLOGY).

Osteoglossiforms vary greatly in size, shape, and biology. Their trophic biology is greatly diverse, the various species being either piscivores, insectivores, filter feeders, or omnivores. The well-vascularized swim bladder can act as a lung, which is especially helpful in oxygen-deficient habitats. *Gymnarchus niloticus* builds a floating nest; and the osteoglossid genera *Arapaima* and *Heterotis* are substrate nest builders, whereas *Scleropages* and *Osteoglossus* are buccal incubators. *Arapaima gigas* of South America is one of the largest freshwater fishes of the world, at least 3 m, perhaps 4.5 m, in length. Noeopterids reach a meter in length; they are egg layers and the male guards the eggs. Most mormyrids are small, up to 50 cm in length; most are much less, but 1.5-m specimens have been reported. Species vary in having a blunt snout with a terminal or inferior mouth, to a long snout and small terminal mouth, hence variable feeding strategies.

Reeve M. Bailey; Herbert Boschung

**Bibliography.** T. M. Berra, *An Atlas of Distribution of the Freshwater Fishes of the World*, University of Nebraska Press, 1986; P. H. Greenwood, Interrelationships of osteoglossomorphs, in P. H. Greenwood, R. S. Miles, and C. Patterson (eds.), *Interrelationships of fishes*, J. Linn. Soc. (Zool.), 53(Suppl. 1):307-332, Academic Press, 1973; V. G. Lauder and K. F. Leim, The evolution and interrelationships of the actinopterygian fishes, *Bull. Mus. Comp. Zool.*, 150:95-197, 1983; J. S. Nelson, *Fishes of the World*, 3d ed., Wiley, 1994.

## Osteolepiformes

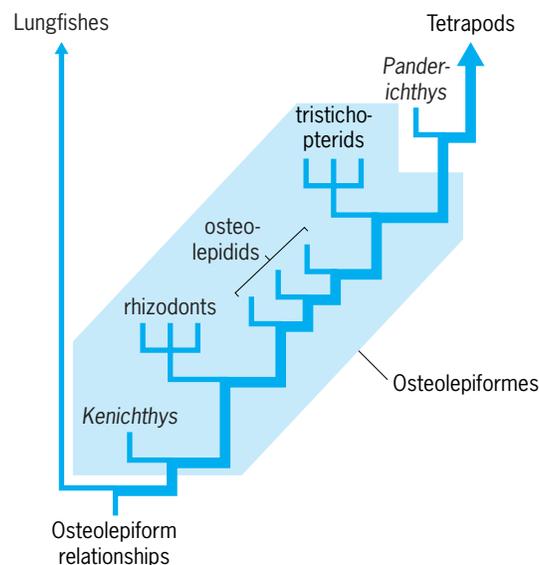
Fossil lobe-finned (sarcopterygian) fishes of Devonian to Permian age (395-270 million years ago), the ancestral group for the land vertebrates, or "tetrapods." See SARCOPTERYGII; TETRAPODA.

**Phylogeny.** The term "osteolepiforms" is applied to a range of fossil fishes, found worldwide in Devonian to Permian strata, that appear to be closely

related to tetrapods. After decades of debate about their precise position in the vertebrate family tree, a consensus has emerged that they occupy the bottom part of the tetrapod stem lineage, that is, the evolutionary line that leads toward land vertebrates from the last common ancestor we share with our nearest living fish relatives, the lungfishes (Fig. 1). Immediately above the osteolepiforms in the stem lineage, just below the most primitive tetrapods, we find the transitional fossil *Panderichthys*. See PANDERICHTHYS.

Rather than forming a single coherent group, the osteolepiforms spread out along the tetrapod stem lineage (Fig. 1, shaded zone): some are very primitive and fall near the base of the lineage, others are close to the fish-tetrapod transition. Such an arrangement is known as a paraphyletic group. The precise relationships of the different osteolepiforms are still debated, but this article follows the arguments presented by P. Ahlberg and Z. Johanson in 1998.

**Morphology.** Despite their paraphyletic arrangement along the tetrapod stem lineage, all osteolepiforms look reasonably similar and share many characteristics. This implies that the characteristics in question are ancestral for tetrapods; osteolepiform fossils thus give a clear picture of the specific fish body plan that gave rise to land vertebrates. General characteristics of the osteolepiforms include an intracranial joint running vertically through the skull behind the eyes, which allowed the front part of the head to move slightly up and down relative to the back part; a pair of internal nostrils, or choanae, connecting the nose to the palate; and paired fin



**Fig. 1. Schematic phylogeny (family tree) of the osteolepiforms, based on arguments presented by P. Ahlberg and Z. Johanson (1998).** The fishes within the grey zone are conventionally designated as Osteolepiformes (the formal group name), but they actually form a paraphyletic segment of the tetrapod stem rather than a well-defined group. The tetrapod stem lineage is indicated in thick line. Arrowheads indicate lineages surviving to the present day.

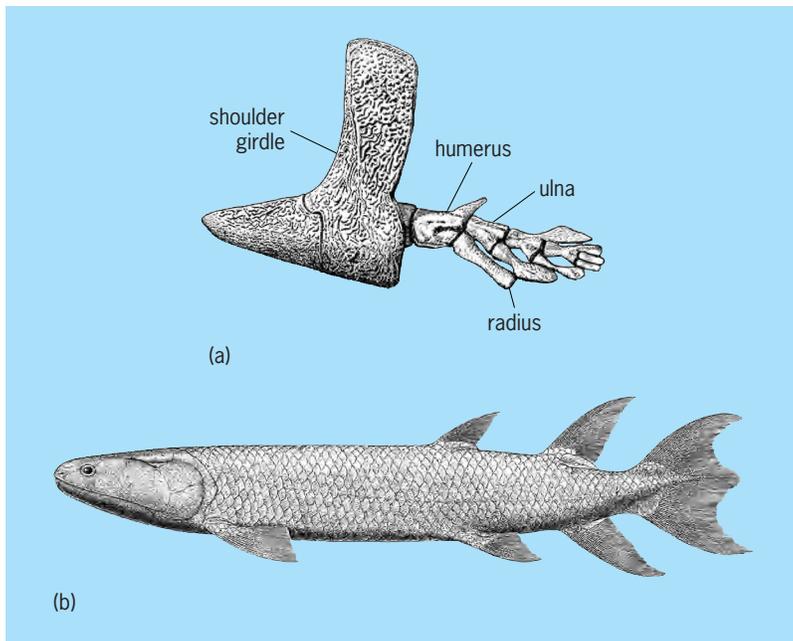


Fig. 2. The tristichopterid *Eusthenopteron foordi*. (a) A close-up of its pectoral fin skeleton. (b) An illustration of the full specimen. *Eusthenopteron* reached approximately 50–70 cm in length. Its pectoral fin skeleton contains structures equivalent to the tetrapod humerus, radius and ulna, but no equivalent of digits. (Reprinted from *Basic Structure and Evolution of Vertebrates*, vol. 1, E. Jarvik, pp. xx, ©1980, with permission from Elsevier)

skeletons in which one can recognize equivalents of the tetrapod limb bones humerus, radius, ulna (forelimb/pectoral fin) and femur, tibia, fibula (hindlimb/pelvic fin) (Fig. 2a). The intracranial joint was lost during the fish-tetrapod transition—it is immobilized in *Panderichthys* and completely obliterated in early tetrapods—but the internal nostrils and limb skeletons persisted and became basic components of the tetrapod body plan. The similarities between osteolepiform fin skeletons and tetrapod limb skeletons are quite specific, extending to the detailed pattern of muscle attachments on the bones. However, despite their tetrapod-like anatomy, osteolepiforms retained the overall body form of “normal” open-water fishes and do not seem to have been adapted for an amphibious life. Osteolepiforms are known from both marine and nonmarine environments.

**Kenichthys.** The earliest and most primitive osteolepiform is *Kenichthys* from the Early Devonian of China (about 395 million years ago). It was a small fish, with a head just over 2 cm (0.8 in) long (complete bodies have not been found). The external bones of the skull are covered in cosmine, a smooth shiny coating of dentine and enamel that is also found in many other lobe-finned fishes. The jaws of *Kenichthys* carry broad fields of bulbous, bullet-shaped crushing teeth, suggesting that it may have fed on shelly invertebrates. Its most interesting feature is the snout, which gives a “snapshot” of the evolution of the internal nostril. Most fishes have two external nostrils on each side, both connected to the nasal sac where the olfactory nerve endings are located, whereas tetrapods and osteolepiforms

have one external and one internal nostril, or choana, on each side. *Kenichthys*, uniquely, has one external nostril and one right on the edge of the mouth, opening within the upper jaw tooth row. This neatly intermediate state shows that one of the external fish nostrils (the posterior one) migrated onto the palate to become the choana, pushing through the tooth row in the process.

**Rhizodonts.** Immediately above *Kenichthys* in the tetrapod stem lineage are the rhizodonts, a group of gigantic predatory osteolepiforms. They appear rather late in the fossil record, at the beginning of the Late Devonian (about 370 million years ago), but their position in the tree suggests that they have an unrecorded “prehistory” of some 15 million years or more. It is clear that the rhizodonts originated in Gondwana (the southern supercontinent that includes Africa, South America, India, Antarctica, and Australia), as this is where all the earliest fossils are found. By the end of the Devonian, rhizodonts had spread to the northern Euramerican supercontinent (North America and Europe), and they remained a prominent feature of freshwater faunas through the earlier part of the Carboniferous Period, until their extinction about 300 million years ago.

All known rhizodonts were large fishes. *Gooloogongia* from the Late Devonian of Australia, one of the earliest and most primitive, was a slightly flattened and probably bottom-dwelling predator just over 1 m (3 ft) in length, whereas *Rhizodus hiberni* from the Carboniferous of Scotland reached at least 7 m (23 ft) and is the largest freshwater fish known. Rhizodonts had formidable dentitions of piercing teeth, sometimes equipped with cutting edges, and must have been the top predators of their ecosystems. Their large, rounded pectoral fins, which have very complex fin skeletons, may have been used as paddles to scull through weed-choked waters; they do not appear to have been active swimmers. Later forms such as *Strepsodus* have almost eel-like bodies with very reduced midline fins and were probably lurking ambush predators. This is also suggested by their extremely elaborate lateral line systems, sensory organs that detect vibrations in the water. Rhizodonts lacked cosmine and had round scales.

**Osteolepidids and tristichopterids.** The remaining “typical” osteolepiforms can be roughly divided into osteolepidids, which had cosmine and rhombic scales, and tristichopterids, which lacked cosmine and had round scales. This similarity between tristichopterids and rhizodonts seems to be convergent (that is, independently evolved in the two groups), as other characteristics indicate that the tristichopterids are much closer to tetrapods than the rhizodonts. The osteolepidids probably form a paraphyletic segment of the tetrapod stem lineage, but the tristichopterids are a coherent group with a single common ancestry (a clade) that forms a well-defined side branch on the tree (Fig. 1).

**Osteolepidids.** Osteolepidids appear in the fossil record early in the Middle Devonian, about 390 million years ago, and have a worldwide

distribution. Early forms such as *Osteolepis* are usually quite small, about 20–40 cm (8–16 in) total length, and broadly resemble *Kenichthys* except that they had fully developed choanae and a more obviously predatory dentition with narrow rows of piercing teeth. Some later osteolepidid groups, such as the megalichthyids, (which become prominent in the Carboniferous and persisted into the early Permian as the last of all osteolepiforms), reached lengths up to 2 m (7 ft). Megalichthyids often occur alongside rhizodonts but never reached the same gigantic proportions. Like *Kenichthys* and rhizodonts, osteolepidids had broad, rounded heads with short snouts. They usually had heterocercal (that is, asymmetrical, “shark-like”) tail fins.

**Tristichopterids.** The tristichopterids (or eusthenopterids, according to some authors) are the most advanced osteolepiforms. They had narrower, more sharply pointed heads than the other groups, and, in particular, longer snouts. These features are shared with *Panderichthys* and tetrapods, and may reflect a change from suction feeding to a “snapping” mode of prey capture. Tristichopterids had more or less symmetrical tail fins, often with a distinctive three-pronged outline, suggesting they were neutrally buoyant and did not use their tails to generate lift.

The best-known tristichopterid, and arguably the most thoroughly studied of all fossil fishes, is *Eusthenopteron foordi* from the early Late Devonian of Quebec, Canada (Fig. 2b). A perfect specimen of this fish was investigated during the 1940s–1960s by Dr. Erik Jarvik of the Swedish Museum of Natural History (Naturhistoriska Riksmuseet), Stockholm, using the painstaking serial grinding technique. This method involved grinding away the head of the fossil a fraction of a millimeter at a time and using the cross sections revealed by the grinding process as templates for building a large wax model; this allowed Jarvik to study not only the external features but also the internal anatomy of the brain cavity and inner ears.

The earliest tristichopterids, such as *Tristichopterus* from the late Middle Devonian of Scotland, were only about 30 cm (12 in) long. Later forms were generally larger, culminating in the 2–3-m (7–10 ft) long *Eusthenodon* and *Hyneria* from the latest Devonian, probably major predators on the early tetrapods that occur in the same faunas. These late, large tristichopterids were in some respects convergent on *Panderichthys* and early tetrapods—they had greatly lengthened snouts and somewhat reduced median fins—but unlike *Panderichthys*, their paired fins do not show any modification toward walking. They died out at the Devonian-Carboniferous boundary and may have been replaced ecologically by the rhizodonts.

**Conclusion.** Overall, the osteolepiforms are rather conservative fishes. Although they share a suite of tetrapodlike characteristics, there is no sign that they were becoming progressively more amphibious. There is, however, a repeated trend toward greatly increased size, coupled with a reduction of the median

fins and the acquisition of enlarged teeth at the front of the jaw, in most groups. Only in the *Panderichthys* lineage was this trend coupled with the emergence of incipient terrestrial adaptations; this became the starting point for the origin of tetrapods. See ANIMAL EVOLUTION; CARBONIFEROUS; DEVONIAN; FOSSIL; PERMIAN.

Per E. Ahlberg

**Bibliography.** J. A. Clack, *Gaining Ground*, Indiana University Press, 2002; J. A. Long, *The Rise of Fishes*, The John Hopkins University Press, 1996; C. Zimmer, *At the Water's Edge*, The Free Press, 1998.

## Osteoporosis

A metabolic bone disease in which the amount of bone tissue is reduced sufficiently to increase the likelihood of fracture. Fractures of the vertebrae, femur (hip), and wrist are the most common osteoporotic fractures, but other bones such as the ribs, upper arm, and pelvis may also fracture.

Although low bone mass is the major factor in osteoporotic fractures, there may also be qualitative and architectural changes in bone with aging that lead to increased fragility. Osteoporosis can be primary or secondary. Primary osteoporosis occurs independently of other causes. The secondary osteoporoses result from identifiable causes, such as exogenous cortisone administration, Cushing's disease, hyperparathyroidism, hyperthyroidism, hypogonadism, multiple myeloma, prolonged immobilization, alcoholism, anorexia nervosa, and various gastrointestinal disorders. Primary osteoporosis occurring in children is called juvenile osteoporosis; that occurring in premenopausal women and middle-aged or young men is known as idiopathic osteoporosis. Osteoporosis, which is found in older persons, can be classified as postmenopausal (type I) or involutional (type II) osteoporosis (see **table**). See ALCOHOLISM; ANOREXIA NERVOSA; GASTROINTESTINAL TRACT DISORDERS; METABOLIC DISORDERS.

**Pathogenesis.** The pathogenesis of osteoporotic fracture is related not only to the presence of low bone mass but also to the mechanical stress placed on bone. The pathogenesis of low bone mass is multifactorial. Indeed, the changes with aging (see **illus.**) may be viewed as occurring in different stages of the life cycle. The development of maximal peak bone mass prior to involutional bone loss is thought to be related to heredity, nutrition, and physical activity. In young adulthood, men have a higher bone mass than women, and blacks have a higher bone mass than whites. The pattern of bone loss from various skeletal sites is primarily a reflection of the proportion of compact bone near the surface of a bone to less dense bone deeper within a bone. There appears to be substantial bone loss from the femur and spine before menopause. Significant bone loss from the radius bone in the forearm appears to begin after menopause. In older age, renal function

Features of postmenopausal (type I) and involuntional (type II) osteoporosis

Factor	Type I	Type II
Age	55–75	Over 70
Sex (female-to-male ratio)	6:1	2:1
Fracture sites	Wrist/vertebrae	Hip/vertebrae, long bones
Hormonal cause	Estrogen deficiency	Calcitriol deficiency
Calcium absorption	Decreased	Decreased
Parathyroid hormone level	Normal	Increased
Importance of dietary calcium	Moderate	High

declines, with diminished ability of the kidney to produce calcitriol—the active form of vitamin D. As a result, calcium malabsorption and secondary hyperparathyroidism occur with further bone loss in the aged. *See* CALCIUM METABOLISM; MENOPAUSE; VITAMIN D.

**Prevention.** Goals should include prevention of both the underlying disorder (the disease) and its effects (osteoporotic fractures).

*Disease prevention.* Secondary osteoporoses are managed by eliminating the underlying disorder. To prevent primary osteoporosis, good health-related behavior during childhood and young adulthood has been suggested as the most important factor. Such behavior includes avoiding cigarette smoking and excess alcohol intake, maintaining a normal body weight, and maintaining optimal dietary intake of calcium; preferably, calcium should be obtained by including dairy products with each meal, rather than by taking calcium supplements. It has been suggested that the recommended daily allowance (RDA) for calcium in the postmenopausal period and the recommended intake of vitamin D in the aged should be increased. At menopause, the possibility of estrogen replacement therapy should be considered for

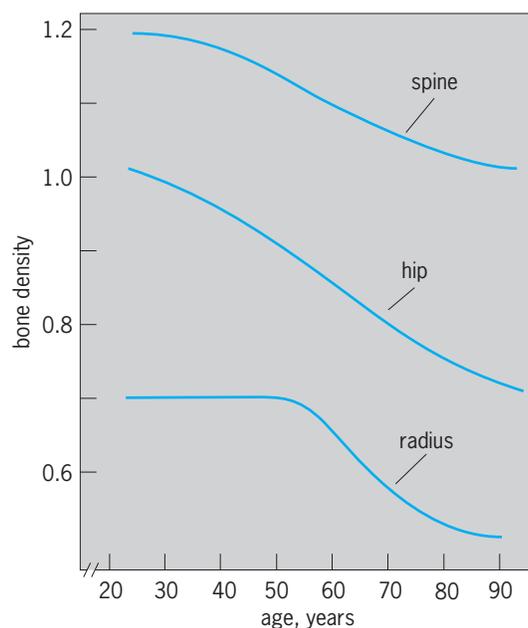
women who are at high risk for osteoporosis; this therapy decreases bone loss and reduces the occurrence of osteoporotic fractures. *See* ESTROGEN; NUTRITION.

*Fracture prevention.* Fracture prevention should be a lifelong effort. During childhood and the premenopausal years, a maximal peak bone mass should be developed through weight-bearing exercise. Exercise to improve coordination and flexibility and to maintain good posture is also useful. As discussed above, there should be an adequate intake of calcium and vitamin D. The types of exercises noted above should be continued during and after menopause. In the elderly, walking and exercises to promote agility should be encouraged, but the risk of spinal injury must be avoided. It is important to deal with health factors that may lead to falls, such as loss of hearing and vision, the use of certain medications, and postural instability. In addition, efforts should be made to minimize the risk of falling. Installation of handrails, nonskid stair treads, and adequate lighting can help prevent injuries. *See* HORMONE; PARATHYROID HORMONE.

**Diagnosis and treatment.** The diagnosis of primary osteoporosis is made in the presence of either low bone mass or a characteristic fracture that cannot be attributed to some other cause.

Standard x-rays can show the shape of bones and indicate the presence of fracture. A number of methods that measure bone density may be used to estimate risk for osteoporosis and to assess responses to therapy. These methods include quantitative computed tomography, single-energy photon absorptiometry, and dual-energy x-ray and photon absorptiometry. Neutron activation analysis with whole-body counting is a research tool that can be used to measure total body calcium, which is a measure of skeletal mass. Secondary osteoporoses are excluded by specific laboratory tests. Bone biopsy for histologic evaluation is ordinarily performed only when another form of metabolic bone disease is suspected.

Bone mineral measurement has been recommended for specific patient groups. In estrogen-deficient women, it can be used to identify bone mass low enough to warrant hormone replacement therapy. In patients with vertebral abnormalities or evidence of reduced bone mass on x-rays, it helps establish a diagnosis of spinal osteoporosis. It can assist in monitoring bone mass in patients receiving long-term glucocorticoid therapy and in evaluating



Decrease in bone density with age in women. Before menopause, bone loss affects primarily the spine and hip; after menopause, it is most pronounced in the radius.

primary asymptomatic hyperparathyroidism. See ACTIVATION ANALYSIS; COMPUTERIZED TOMOGRAPHY; MEDICAL IMAGING.

The goals of treatment are rehabilitation and minimization of the risk of future fractures. Goals for rehabilitation include restoring a positive outlook on life, treating depression if it exists, increase of physical activity, restoring independence, relieving pain, restoring muscle mass, and improving posture. Various medications, including estrogen and calcitonin, can maintain bone mass. Basic research on the local control of bone cells and on growth factors holds promise for therapy that can safely increase bone mass. See AGING; BONE; SKELETAL SYSTEM DISORDERS.

John F. Aloia

Bibliography. J. F. Aloia, *Osteoporosis: A Guide to Prevention and Treatment*, 1989; L. V. Avioli and S. M. Krane, *Metabolic Bone Disease and Clinically Related Disorders*, 3d ed., 1998; H. Broell and M. Dambacher (eds.), *Osteoporosis*, 1994.

## Osteostraci

An order of extinct jawless vertebrate fishes, also called Cephalaspida, known from the Middle Silurian to Upper Devonian of Europe, Asia, and North America. They were mostly small, about 2 in. to 2 ft (5 to 60 cm) in length. The head and part of the body were encased in a solid armor of bone, and the posterior part of the body and the tail were covered with thick scales. Some early forms lacked paired fins, though most possessed flaplike pectoral fins. One or two dorsal fins were present. Their depressed shape and the position of the eyes on the top of the head suggest that Osteostraci were bottom dwellers (Fig. 1). The underside of the throat region was cov-

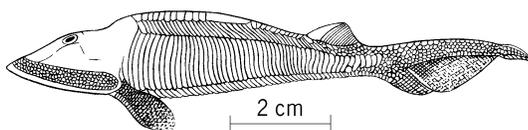


Fig. 1. The ostracoderm *Hemicyclaspis*, a cephalaspid, a Lower Devonian jawless vertebrate. (After E. H. Colbert, *Evolution of the Vertebrates*, Wiley, 1955)

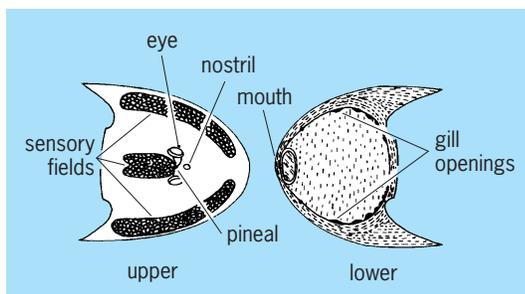


Fig. 2. Upper and lower surfaces of the head in the Devonian ostracoderm *Cephalaspis*. (After E. H. Colbert, *Evolution of the Vertebrates*, Wiley, 1955)

ered by small plates and there was a small mouth in front.

The gill chamber was large and contained as many as 10 pairs of gills, each opening separately to the exterior (Fig. 2). It is believed that they fed by sucking in water and straining out food particles in the gills. Because the internal skeleton of the head was often bony, its structure is well known. Details of this structure, particularly the single, dorsal, and median nostril, and the presence of only two pairs of semicircular canals indicate a relationship to extinct Anaspida and to living Petromyzonida. The Devonian *Cephalaspis* is the best-known genus. See ANASPIDA; JAWLESS VERTEBRATES; OSTRACODERM. Robert H. Denison

## Ostracoda

A major taxon of the Crustacea containing small bivalved animals 0.004–1.4 in. (0.1–33 mm) long, with most between 0.04 and 0.08 in. (1 and 2 mm). They inhabit aquatic environments in nearly all parts of the world. Semiterrestrial species have been described from moss and leaf-litter habitats in Africa, Madagascar, Australia, and New Zealand, and from vegetable debris of marine origin in the Kuril Archipelago. In the oceans, ostracodes live from nearshore to abyssal depths, some swimming and others crawling on the bottom; several are adapted to estuaries of rapidly changing salinity. In fresh water, ostracodes inhabit lakes, ponds, streams, swamps, caves, and ephemeral little standing bodies of water. Of the more than 2000 species extant, none is truly parasitic and most are free-living. However, a few fresh-water and marine forms live commensally on other animals: among the Podocopina, *Entocythere* clings to the gills of crayfish and *Sphaeromicola* and *Paradoxostoma rostratum* to appendages of isopods and amphipods; among the Myodocopina, *Vargula parasitica* and *Sheina orri* are found on the gills of sharks and rays. Ostracodes themselves are parasitized; some marine species are infested with parasitic isopods and copepods, and some fresh-water species serve as intermediate hosts for a cestode (the adult of which parasitizes the black duck) and for an acanthocephalan (the adult of which parasitizes the black bass). Most ostracodes are scavengers, some are herbivorous, and a few are predacious carnivores. Exceptional biological features are known; there are myodocopine ostracodes that produce bioluminescence, and some species of podocopines form a secretion from spinning glands to enable them to climb polished surfaces.

Researchers disagree on the hierarchical classification of the Ostracoda. It is recognized as a distinct class by some and a subclass within the Maxillopoda by others. Consequently, the subdivisions within the Ostracoda may be ranked as subclasses, superorders, or orders. Six major subdivisions are currently recognized: Bradoriida, Phosphatocopida, Leperditicopida, Paleocopa, Myodocopa, and Podocopa. The first four taxa are extinct. The Myodocopa are further subdivided into the

Myodocopida and Halocyprida, and the Podocopa into Platycopida and Podocopida. All fresh-water Ostracoda belong to the Podocopida. Many groups have highly ornamented carapaces, which can be readily classified; a few groups have smooth, bean-shaped carapaces that differ only slightly in shape and present difficulties in taxonomy, particularly in the fossil forms (which have no appendages preserved). See MAXILLOPODA; PODOCOPA.

**Morphology.** Knowledge of the morphology of the Ostracoda is based primarily on extensive study of many species.

**Carapace.** The two valves, sufficient to enclose the rest of the animal, are joined dorsally along a hinge, that may vary from a simple juncture to a complex series of teeth and sockets. They open by a ligament and close by adductor muscles, which pass through the body. The valves provide an attachment for muscles operating the appendages. Each typically consists of an outer lamella and an inward-facing layer, the inner lamella. The triple-layered outer lamella is commonly composed of a layer of calcite between thin films of chitin, but some large pelagic ostracodes lack the calcareous layer having a nearly transparent, flexible carapace. The rim of the inner lamella may also contain a calcareous central layer to form the duplicature. The chitinous coating is highly colored in some living forms, but it disappears in fossilization.

**Body.** From the dorsal part of the carapace, the elongate body is suspended as a pliable sac, with lateral flaps of hypodermis extending between the lamellae of the valves. It is unsegmented, but reinforced by chitinous processes for rigidity in the vicinity of the appendages. There is no true abdominal region.

**Appendages.** Ostracodes of the Podocopida and Myodocopida possess seven pairs of segmented appendages: antennules, antennae, mandibles, maxillules, and three pairs of thoracic legs (Fig. 1a). The body terminates posteroventrally in a pair of furcae or caudal processes. The Platycopida have two pairs of thoracic legs, and the Halocyprida frequently only one pair. Appendages have muscles within them and are connected by musculature to the valves and to a centrally located chitinous structure, the endoskeleton. They are usually specialized for particular functions, such as swimming, walking, food gathering, mastication, and cleaning the interior of the carapace.

**Digestive system.** The alimentary canal is complete (Fig. 1d), consisting of the mouth, esophagus, foregut, midgut, hindgut, and anus. Some have a pair of midgut glands, which discharge into the gut.

**Glands.** Excretory and secreting glands are scattered in the body, showing no distribution which could indicate original segmentation of the animal.

**Respiratory system.** A few species of the subfamily Cyclasteropinae have rudimentary gills. All other ostracodes respire through the thin body wall. Water is circulated by waving of branchial plates on certain appendages.

**Nervous system.** The central system consists of a cerebrum, circumesophageal ganglion, and a ven-

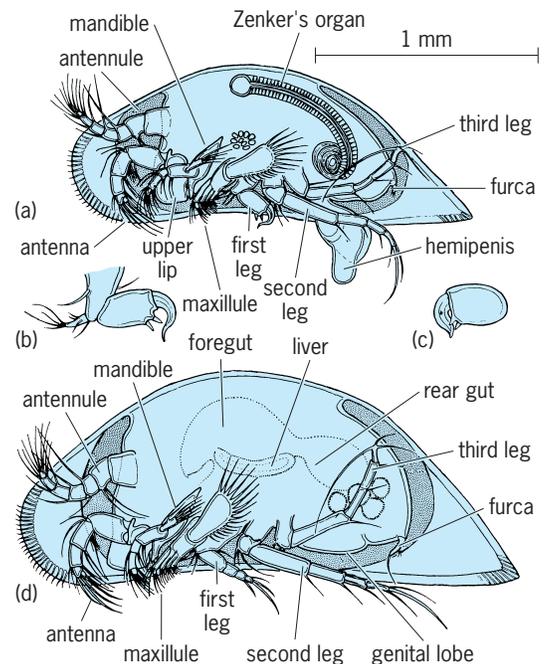


Fig. 1. Ostracode, a typical marine species of the suborder Podocopina, *Macrocypris minna*. (a) Male with left valve removed to show appendages; the internal paired Zenker's organs act as pumps to eject the exceptionally large sperm. (b) Male left first leg, enlarged; the palp is modified into a pincerlike clasp which holds fast to the female during copulation. (c) Male palp of right first leg, enlarged. (d) Female with left valve removed; a part of the digestive system is indicated by dotted lines. (After R. C. Moore, ed., *Treatise on Invertebrate Paleontology*, pt. Q, 1961, and R. V. Kesling, *Contrib. Mus. Paleontol. Univ. Mich.*, 1969)

tral chain of partly fused ganglia. Nerves extend to sensory hairs or setae on the outer surface of the carapace and on the appendages. Modified setae are thought to be olfactory or chemoreceptors. Most Myodocopida have lateral eyes with several lenses, in addition to a median frontal organ (presumably light-sensitive); the Podocopina have lateral eyes (one or two lenses each) fused with a median eye. The suborders Cladocopina and Platycopina lack eyes entirely.

**Reproductive system.** Ostracode reproductive systems usually are duplicated on left and right sides, the halves not connected. Ostracodes have separate sexes, although many species are parthenogenetic and lack males. Whereas the carapace is a protective armor, its presence creates great difficulties in copulation; males have hemipenes of exceptional size and complexity that can be folded for retraction within the closed valves. In most groups, gonads of male and female lie in the rear part of the body (Fig. 1a, d), but in some they extend out into the hypodermis alongside the valve wall. Ostracode sperm are among the largest known in the animal kingdom, some exceeding the length of the carapace. To eject these bulky cells, males of the families Cyprididae and Macrocyprididae (suborder Podocopina) have a pair of large seminal pumps, the Zenker's organs (Fig. 1a), provided with intricate musculature. Various stages of egg development can be traced in the course of the female ovaries and uteri.

**Reproduction and ontogeny.** In some species males are unknown and the eggs develop without fertilization (parthenogenesis), but in other species males are produced and fertilize the eggs (syngamy). Most ostracodes lay their eggs on the substrate or on vegetation, but some transfer them to the posterior space within the carapace, where they hatch and the young brood is retained for a time. The eggs of fresh-water species withstand desiccation for many years and hatch only when again placed in water. This adaptation against drought has enabled many fresh-water species to spread over vast regions, the dried eggs being carried by winds or by mud on the feet of migrating birds.

The egg hatches into a bivalved metanauplius. Growth is by ecdysis; the whole of the covering of carapace and appendages is cast off almost simultaneously with secretion of the new enlarged covering. Six to eight instars (growth stages) plus the adult stage are present, according to the group. During the brief, critical interval of ecdysis, the animal adds new appendages and organs and increases the volume of the carapace to twice its former size. Fresh-water ostracodes reach maturity in about a month, depending on temperature of the water, but marine species may require years for their growth.

**Dimorphism.** In some syngamic species, the valves of male and female are nearly alike. In others, however, the differences are conspicuous. The male may be larger, using the additional volume to house the complicated sex organs, which may constitute one-third of the whole animal; or the female may be larger, using the additional space for egg and brood

care. Certain appendages may also be dimorphic, those in the male being modified into pincerlike claspers, chitinous hooks, or suctorial setae for assistance in copulation (Fig. 1*b, c*). Patsy A. McLaughlin

**Fossils.** The fossil record of the Ostracoda is one of the longest and most complete in the animal kingdom. Three aspects contribute to this excellent fossil record. Being microfossils, most with a length of about 0.5–3 mm, the ostracodes are abundant and likely to be buried rapidly, thus enhancing their likelihood of preservation. Most are benthic organisms, and they have dense, heavily calcified carapaces of calcite that are geochemically quite stable. Finally, they have evolved the capacity to live in a very wide variety of environments that range from shallow-marine to abyssal depths, quiet to turbulent water, and even in the fresh-water environments of lakes, ponds, puddles, and streams.

In spite of the excellent fossil record, several problems remain to be solved before the understanding of early Paleozoic Ostracoda can be regarded as satisfactory. The orders Bradoriida and Phosphatocopida, both restricted primarily to Cambrian rocks, are especially problematic. The bradoriids are widely regarded as polyphyletic, and many taxa assigned to the order are certainly not ostracodes at all. Phosphatocopida is marked by carapaces made of calcium phosphate, a feature that, unlike any in other ostracodes, suggests that they too may be unrelated to the Ostracoda. Moreover, two other orders of lower to middle Paleozoic ostracodes present problems as well. Some specialists regard members of the order Eridostracoda as brachiopods, instead, or perhaps

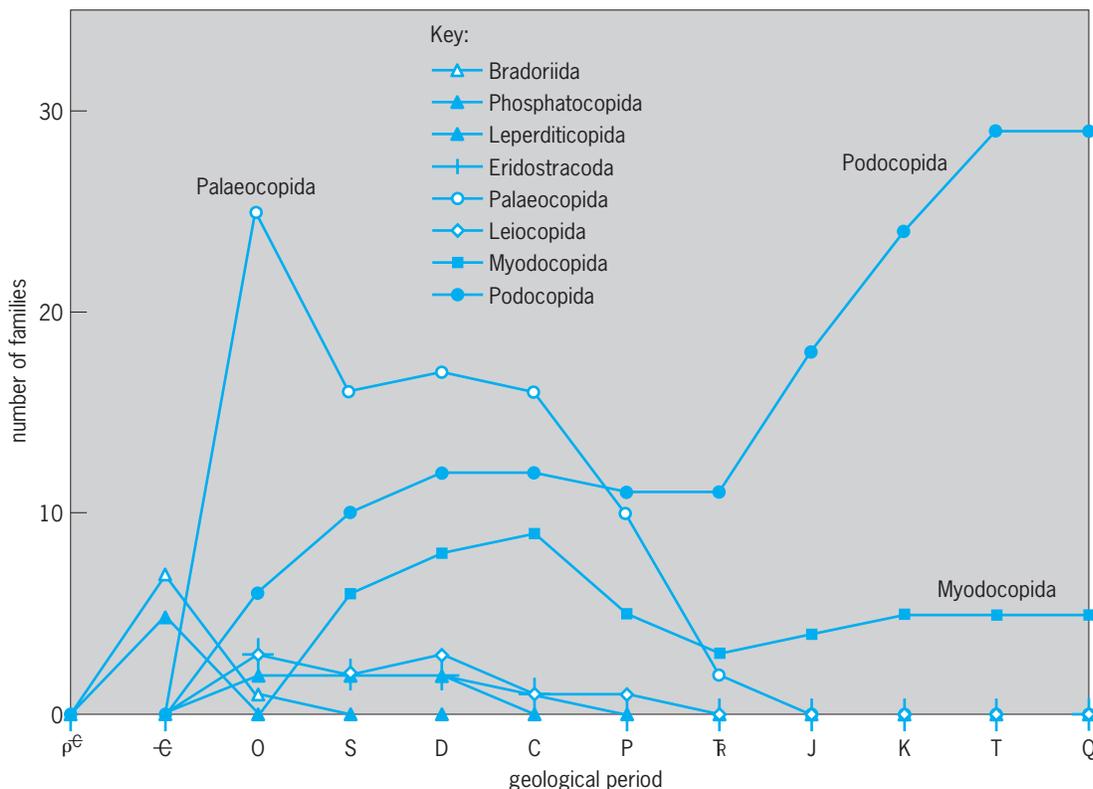


Fig. 2. Biodiversity of the eight orders of Ostracoda showing numbers of families known from the fossil record.

as ostracodes of the order Palaeocopida. Some very large species of the order Leperditicopida are also doubtfully assigned to the class Ostracoda.

Among the most interesting of any ostracodes are the species of the order Palaeocopida, which range from the Ordovician into the Triassic. Many of these ostracodes are marked by pronounced sexual dimorphism. The females develop brood pouches in their carapaces, quite a number of different kinds of which have been described.

The ostracodes that are most common in the modern world, those of the order Podocopida, have been remarkably successful and are found in nearly every marine environment. On at least two separate occasions and perhaps three, the ostracodes have made the difficult transition from marine to fresh-water environments. All these invasions of fresh water have been by podocopids. Their occurrence in fresh-water deposits is especially significant because few invertebrate organisms with mineralized skeletons live in fresh water, and the other kinds of invertebrates that do so are likely to have skeletons of aragonite, which tend to recrystallize with much loss of information.

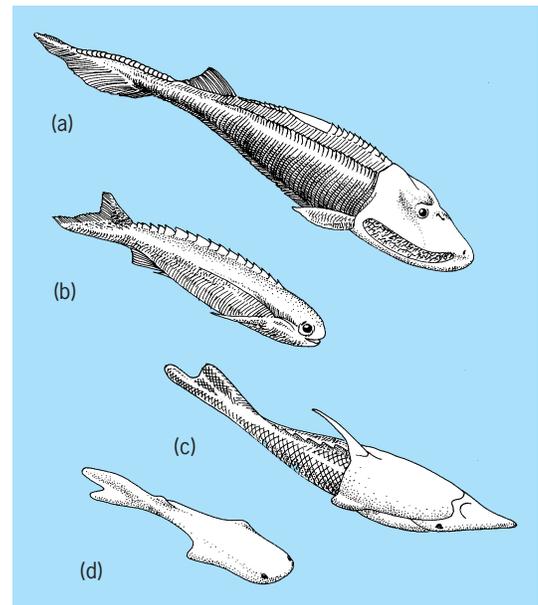
The biodiversities of families known in the fossil record in each of the eight orders of the Ostracoda are shown in Fig. 2. Five aspects of the ostracodes' biodiversity are noteworthy. (1) The ostracodes—or some organisms that were confusingly similar to ostracodes—participated in the Cambrian explosion in diversity, but these early orders had become extinct by the middle of the Ordovician Period. (2) Several less diverse orders of Paleozoic ostracodes persisted from the Ordovician to the Permian. (3) The dominant order of Paleozoic ostracodes, the Palaeocopida, reached a peak in familial diversity in the Ordovician and has been declining almost monotonically since, finally becoming extinct by the end of the Triassic Period. (4) The Ostracoda underwent a fundamental change as a result of the end-Permian extinction event, which some paleontologists think may have eradicated as many as 95% of the Earth's marine species. (5) The familial diversity of the Ostracoda in the modern world is due almost entirely to the order Podocopida, which has been steadily increasing in diversity since the Triassic Period. Important to note is the fact that familial diversity does not necessarily provide information about diversity at other taxonomic levels, especially species diversity. See CRUSTACEA; PALEOCOPA. Roger L. Kaesler

Bibliography. R. H. Bate, E. Robinson, and L. M. Sheppard (eds.), *Fossil and Recent Ostracods*, 1982; H. G. Bronns, *Klassen und Ordnungen des Tierreichs*, Band 5, *Arthropoda*, 1 Abt.: *Crustacea*, 2 Buch (4), 1966, 1975; R. V. Kesling, Anatomy and dimorphism of adult *Candona suburbana* Hoff, *Four Reports of Ostracod Investigations*, 1965; R. V. Kesling, Terminology of ostracod carapaces, *Contrib. Mus. Paleontol. Univ. Mich.*, 9:93–171, 1951; R. C. Moore (ed.), *Treatise on Invertebrate Paleontology*, pt. Q: *Arthropoda* 3, 1961; K. J. Müller, Phosphatocopine ostracods with preserved appendages from the Upper Cambrian of Sweden, *Lethaia*, vol. 1, 1979;

S. P. Parker (ed.), *Synopsis and Classification of Living Organisms*, 2 vols., 1982; F. R. Schram (ed.), *Crustacean Issues*, vol. 1: *Crustacean Phylogeny*, 1983; R. N. Smith, Musculature and muscle scars on *Cblamydotheca arcuata* (Sars) and *Cypridopsis vidua* (O. F. Müller), *Four Reports of Ostracod Investigations*, 1965; R. C. Whatley, D. J. Siveter, and I. D. Boomer, Arthropoda (Crustacea: Ostracoda), in M. J. Benton (ed.), *The Fossil Record 2*, pp. 343–356 Chapman & Hall, London, 1993.

## Ostracoderm

A popular name applied to several groups of extinct jawless vertebrates (fishes). Most of them were covered with an external skeleton or armor of bone, from which is derived their name, meaning “shell-skinned” (see *illus.*). They are known from the



Ostracoderms, drawn to same scale. (a) *Hemicyclaspis*, a cephalaspid or Osteostraci. (b) *Pterygolepis*, an anaspid. (c) *Pteraspis*, a pteraspid or Heterostraci. (d) *Logania*, a coelolepid. (After E. H. Colbert, *Evolution of the Vertebrates*, Wiley, 1955)

Ordovician, Silurian, and Devonian periods, and thus include the earliest known vertebrates. See JAWLESS VERTEBRATES.

Robert H. Denison

## Otter

A member of the family Mustelidae, which also includes weasels, mink, ferrets, martens, sables, fishers, badgers, and wolverines. Otters are found worldwide except in Australia, Madagascar, and on the islands of the South Pacific. Most inhabit freshwater lakes and rivers. The sea otter (*Enhydra lutris*), however, inhabits the waters and shores of the northern Pacific Ocean from southern California to the Kurile Islands.

**General morphology.** All otters are much alike no matter where they live. The North American otter (*Lontra canadensis*) has a lithe, muscular body; a broad, flat head; small ears; and a long, powerful, tapering tail which serves as a rudder (see **illustration**). The ears and nostrils are capable of being closed when the otter is underwater. The limbs are short, but the strong hind feet are large, and the five toes are broadly webbed. The short, dense fur is dark brown and is impervious to water. The densely packed underfur consists of approximately 70,000 hairs/cm<sup>2</sup> (450,000 hairs/in.<sup>2</sup>) and is overlaid by longer guard hairs. Perianal scent glands are present in all species except the sea otter.

**Habits and reproduction.** Otters are skillful, predators. They usually prey on sluggish varieties of fish, but may also feed on trout, frogs, crayfish, crabs, ducks, muskrats, and young beavers. Otters catch fish with their forepaws, then rip them apart with their teeth. An otter's den is usually a hole in the bank of a stream or lake and has an underwater entrance. A second entrance for ventilation is hidden in bushes on the bank. Except during the breeding season, otters are continually on the move and will travel a 20-mi (32-km) circuit of connecting lakes and rivers in 2 or 3 weeks. Baby otters are born in April or May after a gestation of 60 to 70 days, although gestation in some species with delayed uterine implantation may extend to 10 to 12 months. Average litter size is two. The eyes open at about 4 weeks of age. Normal lifespan in the wild is 8 or 9 years.

**Sea otters.** Sea otters are among the largest of all otters. Adults may be 4 to 5ft (1.2 to 1.5 m) in length, including a 12-in. (30.5-cm) tail. They may weigh up to 80 pounds. The dental formula is  $1 \frac{3}{2}$ , C  $1/1$ , Pm  $3/3$ , and M  $1/2$ , for a total of 32 teeth. The heavy, thickset body is much less sleek and graceful in appearance than the river otter. The pelt is very full, soft, and deep. It is brownish-black and more or less finely grizzled. See DENTITION.

Sea otters rarely come to land, preferring to live in great beds of floating kelp, a type of brown seaweed. Individuals are born, eat, sleep, grow old, and die in the water. They prefer to float on their back and propel themselves with their tail. For greater speed, they turn right side up and, by undulating their body and using their webbed hind feet, they may travel at a rate of 10 mi (16 km) an hour. Sea otters feed on fish, sea urchins, crustaceans, cuttlefish, mussels, clams, abalone, and other shellfish. They often dive to depths of 30 m (100 ft), with the record dive being 97 m (318 ft). Food is normally spread out on the belly and chest and eaten as the sea otter floats on its back. A flat stone is often carried to the surface, placed on the chest, and used as an anvil to open shell-encased prey.

Sea otters are believed to mate for life. Following a gestation of 9 months, a single pup is born on a thick bed of floating kelp, usually in a sheltered natural harbor. The eyes are open at birth, unlike those of other species of otters. Pups remain with their mother for 6 to 12 months and are often carried on their mother's chest where they can nurse and be



North American river Otter (*Lontra canadensis*). (Alan and Sandy Carey/Getty Images)

cleaned. They are full-grown at 4 years of age. Wild sea otters may live up to 20 years.

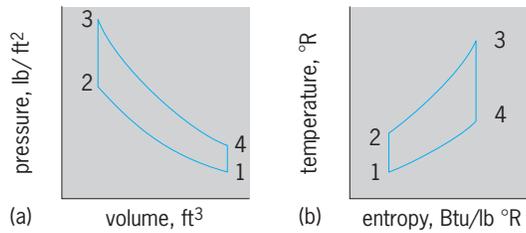
Sea otters have few natural enemies, chief among them being the killer whale. However, this species was almost driven to extinction in the late nineteenth and early twentieth centuries by overhunting for its valuable pelt. It now enjoys complete international protection as the result of a 1911 agreement between the United States, Russia, Japan, and Great Britain. Populations now number approximately 150,000.

**Other otter species.** Other otters include the giant otter (*Pteronura*) that is native to the Amazon River basin of Brazil and reaches a length of about 5 ft (1.5 m), the giant African otter (*Aonyx*) which may be longer than 5 ft and weigh 60 pounds, and whose claws are rudimentary or absent altogether; the clawless otter (*Paraonyx*) of Africa whose forefeet are small with five naked fingers without claws and whose hindfeet bear minute claws on only the third and fourth fingers; the hairy-nosed otter (*Lutra sumatrana*) of Asia and Sumatra, whose nose pad, naked in most species, is covered with fine hair; and the Eurasian otter (*Lutra lutra*), which ranges from England west to southern China and from the Mediterranean region of Africa north to the Arctic coast. See CARNIVORA; MAMMALIA. Donald W. Linzey

Bibliography. H. Kruuk, *Wild Otters: Predation and Populations*, Oxford University Press, 1995; C. F. Mason and S. M. MacDonald, *Otters: Ecology and Conservation*, Cambridge University Press, 1986; R. Nowak, *Walker's Mammals of the World*, Johns Hopkins University Press, 1999; D. Macdonald (ed.), *The Encyclopedia of Mammals*, Andromeda Oxford Limited, 2001.

## Otto cycle

The basic thermodynamic cycle for the prevalent automotive type of internal combustion engine. The engine uses a volatile liquid fuel (gasoline) or a gaseous fuel to carry out the theoretic cycle illustrated in **Fig. 1**. The cycle consists of two isentropic (reversible adiabatic) phases interspersed between two



**Fig. 1.** Diagrams of (a) pressure-volume and (b) temperature-entropy for Otto cycle; phase 1-2, isentropic compression; phase 2-3, constant-volume heat addition; phase 3-4, isentropic expansion; phase 4-1, constant volume heat rejection.

constant-volume phases. The theoretic cycle should not be confused with the actual engine built for such service as automobiles, motor boats, aircraft, lawn mowers, and other small (generally  $<300 \pm \text{hp}$  or  $225 \pm \text{kW}$ ) self-contained power plants.

The thermodynamic working fluid in the cycle is subjected to isentropic compression, phase 1-2; constant-volume heat addition, phase 2-3; isentropic expansion, phase 3-4; and constant-volume heat rejection (cooling), phase 4-1. The ideal performance of this cycle, predicated on the use of a perfect gas, Eqs. (1), is summarized by Eqs. (2) and (3) for thermal efficiency and power output.

$$\frac{V_3}{V_2} = \frac{V_4}{V_1} \quad \frac{T_3}{T_2} = \frac{T_4}{T_1} \quad (1)$$

$$\frac{T_2}{T_1} = \frac{T_3}{T_4} = \left(\frac{V_1}{V_2}\right)^{k-1} = \left(\frac{V_4}{V_3}\right)^{k-1} = \left(\frac{P_2}{P_1}\right)^{(k-1)/k}$$

$$\text{Thermal eff} = \frac{\text{net work of cycle}}{\text{heat added}}$$

$$= \left[1 - \frac{T_1}{T_2}\right] = \left[1 - \left(\frac{1}{r^{k-1}}\right)\right] \quad (2)$$

$$\begin{aligned} \text{Network of cycle} &= \text{heat added} - \text{heat rejected} \\ &= \text{heat added} \times \text{thermal eff} \\ &= \text{heat added} [1 - (T_1/T_2)] \\ &= \text{heat added} [1 - (1/r^{k-1})] \quad (3) \end{aligned}$$

In Eqs. (2) and (3)  $V$  is the volume in cubic feet;  $P$  is the pressure in pounds per square inch;  $T$  is the absolute temperature in degrees Rankine;  $k$  is the ratio of specific heats at constant pressure and constant volume,  $C_p/C_v$ ; and  $r$  is the ratio of compression,  $V_1/V_2$ .

The most convenient application of Eq. (3) to the positive displacement type of reciprocating engine uses the mean effective pressure and the horsepower equation (4), where hp is horsepower; mep is mean

$$\text{hp} = \frac{\text{mep } Lan}{33,000} \quad (4)$$

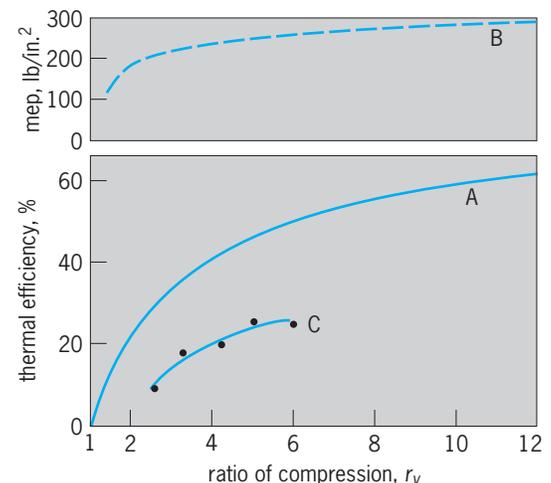
effective pressure in pounds per square foot;  $L$  is stroke in feet;  $a$  is piston area in square inches; and  $n$  is the number of cycles completed per minute. The mep is derived from Eq. (3) by Eq. (5), where 778 is

$$\text{mep} = \frac{\text{net work of cycle} \times 778}{144(V_1 - V_2)} \quad (5)$$

the mechanical equivalent of heat in foot-pounds per Btu; 144 is the number of square inches in 1 ft<sup>2</sup>; and  $(V_1 - V_2)$  is the volume swept out (displacement) by the piston per stroke in cubic feet. See MEAN EFFECTIVE PRESSURE.

**Air standard.** In the evaluation of theoretical and actual performance of internal combustion engines, it is customary to apply the above equations to the idealized conditions of the air-standard cycle. The working fluid is considered to be a perfect gas with such properties of air as volume at 14.7 lb/in.<sup>2</sup> absolute and 492°R equal to 12.4 ft<sup>3</sup>/lb, and the ratio of specific heats  $k$  as equal to 1.4. **Figure 2** shows the thermal efficiency for this air-standard cycle as a function of the ratio of compression  $r$ , and the mep for a heat addition of 1000 Btu/lb of working gases. These curves demonstrate the intrinsic worth of high compression in this thermodynamic cycle.

The **table** gives a comparison of the important gas-power cycles on the ideal air-standard base for the case of compression ratio = 10 and 1000 Btu added per pound of working gases. The Otto, Brayton, and Carnot cycles show the same thermal efficiency of 60%. The mean effective pressures, however, show that the physical dimensions of the engines will be a minimum with the Otto cycle but hopelessly large with the Carnot cycle. The Brayton cycle ideal mep is only about one-fifth that of the Otto cycle, and it is accordingly at a distinct disadvantage when applied to a positive displacement mechanism. This disadvantage can be offset by use of a free-expansion, gas-turbine mechanism for the Brayton cycle. The Diesel cycle offers a lower thermal efficiency than the Otto cycle for the same conditions, for example, 42 versus 60%, and some sacrifice of mep, 200 versus 290 lb/in.<sup>2</sup> As opposed to the Otto engine, the diesel can utilize a much higher compression ratio without preignition troubles and without excessive peak pressures in the cycle and mechanism. Efficiency and



**Fig. 2.** Effect of compression ratio on thermal efficiency and mean effective pressure of Otto cycle. Curve A shows thermal efficiency, air-standard cycle; curve B mean effective pressure, air-standard cycle; and curve C thermal efficiency of an actual engine.

**Thermal efficiency, mean effective pressure, and peak pressure of air-standard gas-power cycles\***

Cycle	Efficiency	mep	Peak pressures lb/in. <sup>2†</sup>
Otto	60	290	2100
Diesel	42	200	370
Brayton	60	61	370
Carnot	60	Impossibly small	Impossibly high

\* Ratio of compression = 10; heat added = 1000 Btu per pound working gases (2.3 megajoules per kilogram).

† 1 lb/in.<sup>2</sup> = 6.895 kilojoules.

engine weight are thus nicely compromised in the Otto and diesel cycles. See GAS TURBINE.

**Actual engine process.** This reasoning demonstrates some of the valid conclusions that can be drawn from analyses utilizing the ideal air-standard cycles. Those ideals, however, require the modifications of reality for the best design of internal combustion engines. The actual processes of an internal combustion engine depart widely from the air-standard cycle. The actual Otto cycle uses a mixture of air and a complex chemical fuel which is either a volatile liquid or a gas. The rate of the combustion process and the intermediate steps through which it proceeds must be established. The combustion process shifts the analysis of the working gases from one set of chemicals, constituting the incoming explosive mixture, to a new set representing the burned products of combustion. Determination of temperatures and pressures at each point of the periodic sequence of phases (Fig. 1) requires information on such factors as variable specific heats, dissociation, chemical equilibrium, and heat transfer to and from the engine parts.

N. A. Otto (1832–1891) built a highly successful engine that used the sequence of engine operations proposed by Beau de Rochas in 1862. Today the Otto cycle is represented in many millions of engines utilizing either the four-stroke principle or the two-stroke principle. See INTERNAL COMBUSTION ENGINE.

The actual Otto engine performance is substantially poorer than the values determined by the theoretic air-standard cycle. An actual engine performance curve *c* is added in Fig. 2, in which the trends are similar and show improved efficiency with higher compression ratios. There is, however, a case of diminishing return if the compression ratio is carried too far. Evidence indicates that actual Otto engines offer peak efficiencies (25±%) at compression ratios of 15±. Above this ratio, efficiency falls. The most probable explanation is that the extreme pressures associated with high compression cause increasing amounts of dissociation of the combustion products. This dissociation, near the beginning of the expansion stroke, exerts a more deleterious effect on efficiency than the corresponding gain from increasing compression ratio. See BRAYTON CYCLE; CARNOT CYCLE; DIESEL CYCLE; THERMODYNAMIC CYCLE.

Theodore Baumeister

Bibliography. E. A. Avallone and T. Baumeister III (eds.), *Marks' Standard Handbook for Mechanical Engineers*, 10th ed., 1996; J. B. Jones and R. E. Dugan, *Engineering Thermodynamics*, 1994; E. F. Obert, *Internal Combustion Engines and Air Pollution*, 1990.

## Ovarian disorders

A variety of neoplastic and nonneoplastic disorders that occur in the ovary. Ovarian neoplasms are of greater diversity in histologic appearance and biologic behavior than for any other organ. The nonneoplastic disorders include physiologic cysts, pregnancy luteomas, and polycystic ovarian disease (Stein-Leventhal syndrome). The ovary can also be a site of metastasis from malignant tumors originating in the genital tract, breast, and gastrointestinal tract. See CANCER (MEDICINE).

**Symptoms.** As ovarian enlargement due to tumors occurs, compression of pelvic and abdominal structures produces vague symptoms such as constipation, pelvic discomfort, a feeling of heaviness, and frequent urination. If the size of the ovary exceeds 5 in. (12 cm), the ovary rises out of the pelvis, and the individual may notice abdominal enlargement, sometimes mistakenly attributed to pregnancy or weight gain. Pain can be an initial symptom of both benign and malignant ovarian disorders. The symptoms of malignant ovarian disorders include abdominal pain and swelling, bloating, heartburn, nausea, and anorexia. Because these symptoms are more suggestive of gastrointestinal problems, these individuals are often first evaluated by internists and family practitioners before the ovary is identified as the cause of symptoms.

**Physical findings.** An examination of the pelvis is the most crucial component in the diagnosis and evaluation of ovarian disorders. Although a mass may be felt during examination of the vagina and abdomen, palpation of the pelvic organs via the rectum is more informative. This often allows the physician to differentiate between disorders of the ovary and other pelvic structures such as the uterus, rectum, and bladder. The location of abnormal ovarian enlargement may offer a clue to the diagnosis as benign cystic teratoma or cysts of endometriosis; most ovarian tumors lie posterior to the uterus. The addition of ultrasound to the pelvic examination may help differentiate benign from malignant disorders. Unilateral, cystic, mobile masses less than 4 in. (10 cm) in size likely indicate benign disease; solid, immobile masses greater than 4 in. (10 cm) likely indicate malignancy.

**Benign disorders.** Physiologic cysts occur in response to cyclic stimulation of the ovary by hormones. Simple cysts in women of reproductive age which are greater than about 1.2 in. (3 cm) can often be treated with oral contraceptives. Persistence of the cyst for 2 months warrants surgical exploration with removal of the cyst, leaving the ovary intact.

**Polycystic ovarian syndrome.** This syndrome is caused by abnormal regulation of the hypothalamic-pituitary-ovarian axis. In this condition, the ovaries are bilaterally enlarged with multiple follicular cysts. Clinically, symptoms include amenorrhea, menstrual abnormalities, infertility, and hirsutism (a condition characterized by the growth of hair in unusual places and in unusual amounts). Therapy consists of administering oral contraceptives, progesterone preparations, or clomiphene. Because this is a condition where there is no ovulation for prolonged periods, individuals with this syndrome are at increased risk for hyperplasia and carcinoma of the endometrium.

**Endometriosis.** This disorder is characterized by ectopic endometrial glandular and stromal tissue that often involves the ovary as well as other pelvic structures. Ovarian endometriosis often produces a cystic mass filled with old blood termed a chocolate cyst. The two most common symptoms of endometriosis are pelvic pain and infertility. This can be treated medically with agents that suppress ovulation, or surgically by removing the ovary.

**Neoplasms.** Neoplasms of the ovary can originate from epithelial, stromal, or germ cells, and produce a variety of symptoms. Epithelial tumors account for 50–65% of benign neoplasms prior to menopause and 80% after menopause. Mature cystic teratomas account for 50% of benign neoplasms in women between the ages of 20 and 50, and 25–40% of all ovarian neoplasms. Treatment is tailored to the individual's age and desire for future fertility. In a reproductive-age woman who has not completed childbearing or desires conservative therapy, removal of the cyst or one ovary is adequate treatment. In the individual who has completed childbearing or is peri- or postmenopausal, removal of both ovaries is acceptable.

**Malignant disorders.** As with benign tumors, malignant ovarian neoplasms can arise from epithelial, stromal, or germ cells. Epithelial tumors account for 85% of ovarian cancer, while germ-cell and stromal tumors account for 10 and 5%, respectively. With a high suspicion of malignancy, useful diagnostic tests include serum tumor markers (CA-125, human chorionic gonadotropin, alpha-fetoprotein), transvaginal color doppler imaging, chest x-ray, and computerized tomography scan. The primary treatment is surgery; procedures include removal of the ovaries, fallopian tubes, uterus, omentum, and other tumor masses within the abdominal cavity. Survival is directly related to the amount of cancer remaining at the end of surgery. Many individuals receive adjuvant therapy with chemotherapy following surgery. See MEDICAL IMAGING.

The ovary is often a site of metastatic cancer from other organs, the most common sources being breast, gastrointestinal tract, and other pelvic organs. Up to 20% of women undergoing removal of an ovary for breast cancer have been found to have metastasis. See ONCOLOGY; OVARY; TUMOR. David L. Tait

Bibliography. C. P. Morrow, J. P. Curtin, and D. E. Townsend (eds.), *Synopsis of Gynecologic Oncology*, 5th ed., 1998.

## Ovary

A part of the reproductive system of all female vertebrates. Although not vital to individual survival, the ovary is vital to perpetuation of the species. The function of the ovary is to produce the female germ cells or ova, and in some species to elaborate hormones that assist in regulating the reproductive cycle.

The ovaries develop as bilateral structures in all vertebrates, but adult asymmetry is found in certain species of all vertebrates from the elasmobranchs to the mammals. This asymmetry may be morphological as a result of fusion or atrophy, or it may be physiological.

The position of the ovaries in the coelom varies within different vertebrate groups. Those animals which produce large numbers of eggs possess ovaries that almost fill the coelomic cavity during the breeding season, but most vertebrates have relatively small ovaries. The ovaries may be anchored to the dorsal body wall anywhere between the transverse septum (lower vertebrates) and the pelvic cavity (higher mammals).

### Histology

The ovary of all vertebrates functions in essentially the same manner. However, ovarian histology of the various groups differs considerably. Even such a fundamental element as the ovum exhibits differences in various groups. The ovum of oviparous forms is large, and it synthesizes and stores large amounts of yolk. The ovum of viviparous forms, especially the mammals, is small and contains little yolk.

**Mammals.** The mammalian ovary is attached to the dorsal body wall by a mesovarium. The free surface of the ovary is covered by a modified peritoneum called the germinal epithelium (Fig. 1a). This layer may consist of a single or a stratified layer of cells, and the cell shape may vary from squamous to columnar, depending upon the species. The activity of this layer, especially during the embryonic development of the gonad whereby cells are proliferated into the interior, is responsible for its name. The potentialities of the germinal epithelium are still a matter of controversy.

Just beneath the germinal epithelium is a layer of fibrous connective tissue, varying in thickness and density in different species and at different phases of the reproductive cycle in the same species. This is the tunica albuginea (Fig. 1a). Most of the rest of the ovary is made up of a more cellular and more loosely arranged connective tissue (stroma) in which are embedded the germinal, endocrine, vascular, and nervous elements (Fig. 1a and b).

The most obvious ovarian structures are the follicles and the corpora lutea. The smallest, or primary, follicle consists of an oocyte surrounded by a layer of follicle (nurse) cells (Fig. 1a and b). Follicular growth results from an increase in oocyte size, multiplication of the follicle cells to form several concentric layers called the granulosa, and the differentiation of the perfollicular stroma to form a fibrocellular

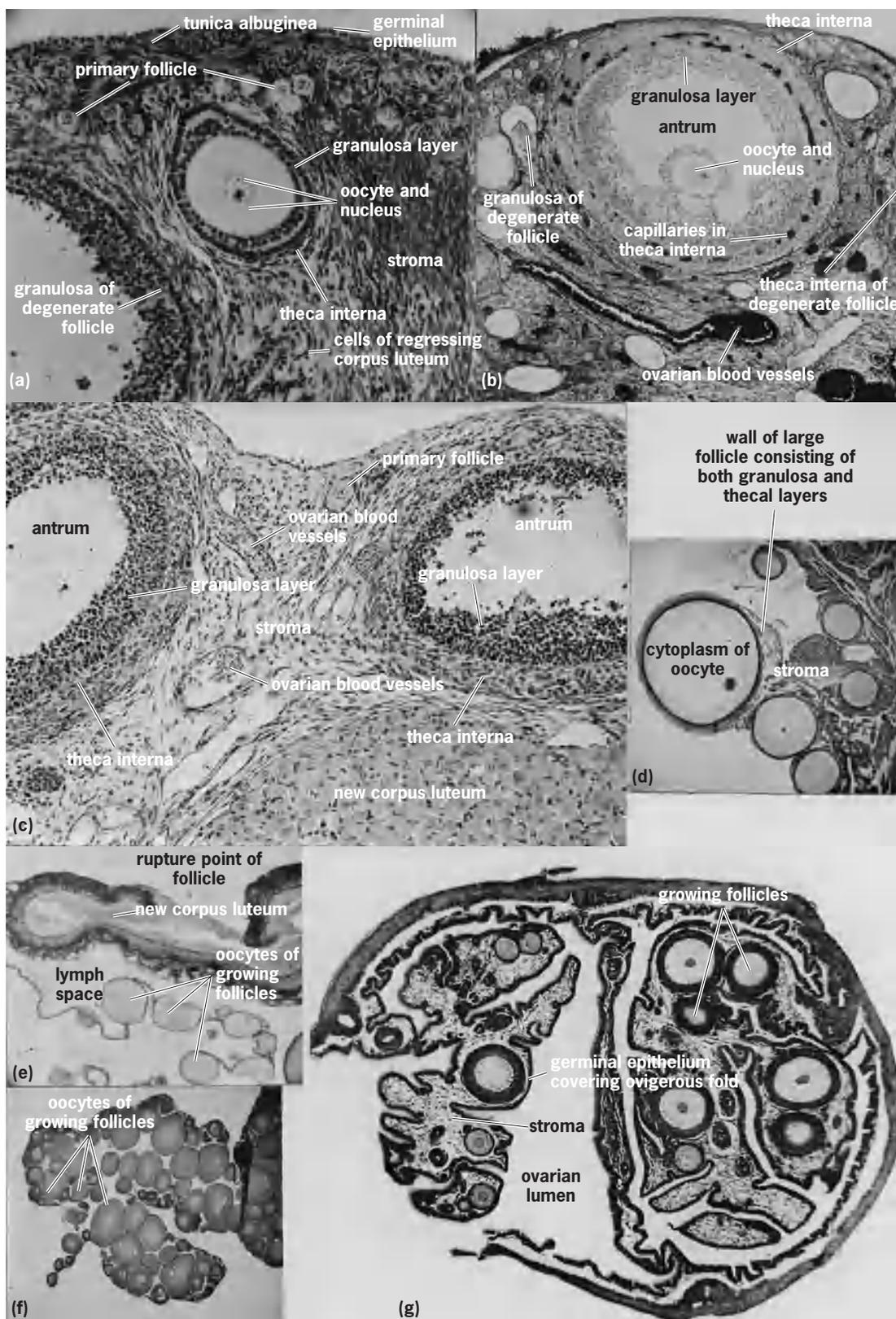


Fig. 1. Ovarian histology. (a) Pika (*Ochotona princeps*); (b) pocket gopher; (c) prairie dog (*Cynomys leucurus*); (d) chicken (*Gallus domesticus*) (from W. Andrew, *Textbook of Comparative Histology*, Oxford, 1959). (e) Lizard (*Xantusia vigilis*) (courtesy of M. Miller). (f) Frog (*Rana* sp.) (from W. Andrew, *Textbook of Comparative Histology*, Oxford, 1959). (g) Viviparous fish (*Neotoca bilineata*) (from G. Mendoza, *Biol. Bull.*, vol. 84, 1943).

envelope called the theca interna (Fig. 1*b* and *c*). Finally, a fluid-filled antrum develops in the granulosa layer, resulting in a vesicular follicle (Fig. 1*b*).

The cells of the theca interna hypertrophy during follicular growth and many capillaries invade the layer (Fig. 1*b*), thus forming the endocrine element that is thought to secrete estrogen. The other known endocrine structure is the corpus luteum (Fig. 1*c*), which is primarily the product of hypertrophy of the granulosa cells remaining after the follicular wall ruptures to release the ovum. Ingrowths of connective tissue from the theca interna deliver capillaries to vascularize the hypertrophied follicle cells of this newly formed corpus luteum; progesterone is secreted here.

Most follicles degenerate before attaining ovulatory size (Fig. 1*a*, *b*, and *c*). In some species the theca interna of such follicles hypertrophies to form the so-called interstitial cells (Fig. 1*b*). The function of this tissue is still uncertain. However, evidence is accumulating that suggests the tissue is involved in steroid secretion.

**Birds.** The avian ovary has much the same general arrangement as that of the mammal, the chief difference being the size of the mature ova. Other differences are a greatly reduced amount of stroma, the thinness of the theca interna, the reduced number of follicle-cell layers, and the absence of follicular antra and corpora lutea (Fig. 1*d*).

**Reptiles and amphibians.** The reptilian ovary is quite similar in appearance to that of the bird. The follicles are similar, and some viviparous forms are reported to have corpora lutea, although the physiology of them is uncertain. Centrally placed, epithelial-lined spaces, said to be lymphatic spaces, are present in reptilian ovaries (Fig. 1*e*).

The amphibian ovary is similar to the reptilian type but contains many more smaller follicles. No corpora lutea are formed (Fig. 1*f*).

**Fishes.** The fishes represent a diverse group as far as ovarian morphology is concerned. Many forms have an arrangement similar to that of the amphibians. Others have the germinal epithelium covering ovigerous folds within a centrally placed ovarian lumen, instead of covering the outer surface (Fig. 1*g*).

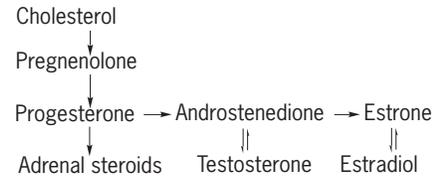
Kenneth L. Duke

### Physiology

Ovarian tissue is responsible for the production of germ cells (ova) and hormones. Lifetime production of ova is variable and may range from about 400 in the human female to millions in some fishes. Ova typically mature within a follicle, and these germ cells are then periodically released from the ovary (ovulation). In many vertebrates ova pass through accessory ducts, where they receive coverings such as jelly capsules or shells. Fertilization of ova can occur within the ducts or externally. In higher mammals, some reptiles, and elasmobranchs the fertilized ovum develops within an enlarged region of the duct system called the uterus. In mammals this development period (gestation) requires the formation and function of a temporary structure termed a placenta.

In some viviparous fish, development of fertilized ova occurs within the ovary.

**Steroid synthesis.** Several endocrine organs (ovary, testis, adrenal gland, and placenta) have the ability to produce various steroid hormones. The main biosynthetic pathway is similar in all steroidogenic tissues. A simplified scheme is shown below. Each step in the



pathway requires the action of one or more enzymes. The presence of the appropriate enzymes and their relative activities can therefore determine which steroid is produced. Ovarian follicular cells possess all the enzymes necessary for estradiol production, whereas the corpus luteum is unable to carry on efficiently the reactions beyond progesterone in the pathway. Tropic hormones from the pituitary, which act on particular steroid-producing tissues, also control the rate and direction of steroid synthesis. See ADENOHYPHOPHYSIS HORMONE; CHOLESTEROL.

**Ovarian hormones.** Hormones produced by the ovary maintain reproductive organs and their accessory structures, and they control cyclic events such as estrus, nidation of the fertilized ovum, and parturition. Ovarian hormones are primarily steroid hormones, but in mammals at least one peptide hormone, relaxin, is produced. Relaxin appears to be involved in enhancing the separation of the symphysis pubis, which brings about an enlargement of the birth canal. Relaxin may also stimulate dilation of the uterine cervix.

Steroids elaborated by the ovary are of two principal types; estrogens which produce sexual receptivity (heat) and growth of the uterus and mammary glands; and progesterone which augments estrogen-stimulated growth of the uterus and mammary gland. See ESTROGEN; PROGESTERONE; STEROID.

The production of estrogen and progesterone are controlled by follicle-stimulating hormone (FSH) and luteinizing hormone (LH) from the pituitary gland. Estradiol is the most potent natural estrogen and is the major hormone synthesized by granulosa cells of the maturing follicle. Estradiol may be converted to estrone (Fig. 2); and estriol, a weak estrogen, is produced as a metabolic end product. Estradiol and estrone have been found in the ovaries of representative species of all vertebrate classes, and estradiol has also been detected in starfish ovaries.

Progesterone (Fig. 3) is produced by the ovarian corpus luteum which develops from the ruptured follicle after ovulation. Progesterone production and secretion are stimulated by pituitary LH. The liver converts large amounts of progesterone to the inactive form, pregnanediol, which is excreted in urine. Progesterone, like estradiol, has been found in the ovaries of all vertebrate classes.

**Ovarian steroid action.** Organs such as the uterus and mammary glands are dependent upon ovarian

estrogens and progesterone for their functional status. These tissues contain receptor molecules (proteins of  $\sim 75,000$  molecular weight) with high affinities for estradiol and progesterone, a property that allows these tissues (targets) to remove the ovarian steroids from circulating blood. When receptors bind estradiol, the receptor-estradiol complex translocates rapidly to target cell nuclei where specific genes may be activated. As a consequence of its action, estradiol causes a general anabolic response in uterine tissue, which includes increases in a number of biochemical events such as glucose metabolism, lipid synthesis, and ribonucleic acid (RNA) and protein synthesis. Further development of uterine tissue requires the action of progesterone, which is again mediated by a progesterone-receptor complex. This steroid is responsible for the decidual reaction, which is related to the phenomenon of implantation of the fertilized ovum in the uterus. Estrogens and progesterone are also responsible for the development of ducts in mammary glands. See LACTATION; MAMMARY GLAND; UTERUS.

**Hypothalamic-pituitary-ovarian relationships.** In mammals an intricate relationship between hypothalamic gonadotropin-releasing hormone (GnRH), pituitary gonadotropins (FSH and LH), and ovarian steroids is essential for normal reproductive function. GnRH promotes release of FSH and LH, and these pituitary hormones stimulate steroid production by the ovary. The ovarian steroids in turn regulate GnRH and gonadotropin levels and effectiveness by positive- or negative-feedback mechanisms (Fig. 4).

Maturation of the follicle requires the action of LH and FSH at precise times. LH binds to receptors on follicular theca cells and stimulates testosterone synthesis. This testosterone is converted to estrogen by follicular granulosa cells under the influence of FSH, which also acts to increase LH receptors on follicular cells. Increasing amounts of LH, including a critically timed LH "surge," bring about ovulation which terminates the follicular phase. The empty follicle becomes transformed into a corpus luteum which responds to FSH and LH by secreting progesterone in addition to estrogens. In rodents, a third gonadotropin, luteotropin (LTH or prolactin) participates in stimulating progesterone secretion.

Circulating estrogen and progesterone act on the pituitary (negative feedback) to inhibit FSH and LH production, and without these hormones the corpus luteum degenerates. As steroid levels fall, FSH and LH secretion resume, and the ovarian cycle may begin again. Although estrogen normally acts to inhibit FSH and LH production, high levels of estrogen at the peak of follicular growth stimulate GnRH secretion and sensitize pituitary cells to GnRH, thus causing the LH surge (positive feedback). In addition to regulation by ovarian steroids, the hypothalamus may also be controlled by other brain centers. Thus, extrinsic environmental stimuli, such as light and temperature, may influence the timing of reproductive cycles by starting a chain of events leading to ovarian activity. See ENDOCRINE MECHANISMS.

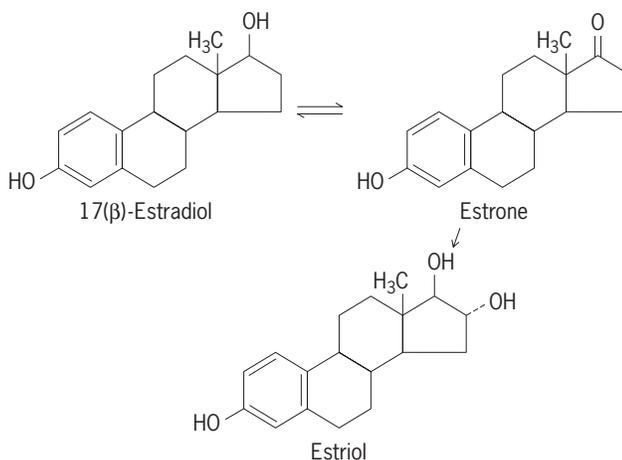


Fig. 2. Structural formulas of the principal natural estrogens.

**Sexual cycles.** Nonmammalian vertebrates generally exhibit seasonal breeding cycles, and their pattern of ovarian function is quite variable. For example, avian and amphibian ovaries do not form true corpora lutea. In most mammals a period of sexual arousal called estrus typically occurs during the reproductive cycle. The recurrence of estrus (estrus cycle) is variable in mammalian species: for example, rat, 4–5 days; dog, 3–4 months; sheep, 16 days; and pig and cow, 21 days. The estrus period corresponds to the time of ovulation and therefore increases the likelihood that mating and fertilization of ova will take place. In some species ovulation only occurs when stimulated by mating. The sexual cycle in human females (menstrual cycle) lasts for about 28 days and is terminated by the degeneration and loss of the uterine lining (menstruation). This occurs because the corpus luteum of the cycle regresses, resulting in the withdrawal of estrogen and progesterone. Ovulation occurs on about the 13th or 14th day of the menstrual cycle. See ESTRUS; MENSTRUATION.

**Pregnancy.** If the ovum is fertilized, a state of pregnancy will begin. Estrogen and progesterone produced during the cycle promote uterine development, which is conducive to the growth of an embryo. In mammals, a fertilized ovum implants in the uterine wall and begins the formation of a placenta. The placenta helps to maintain pregnancy in some species by its role in prolonging the lifetime of the corpus luteum and also through placental production of progesterone. The corpus luteum

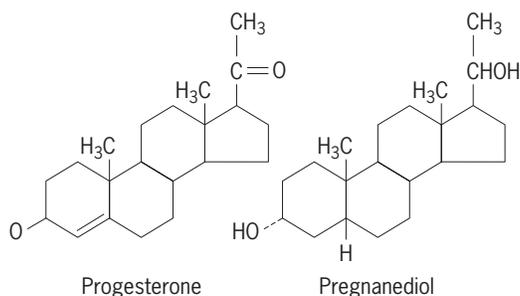
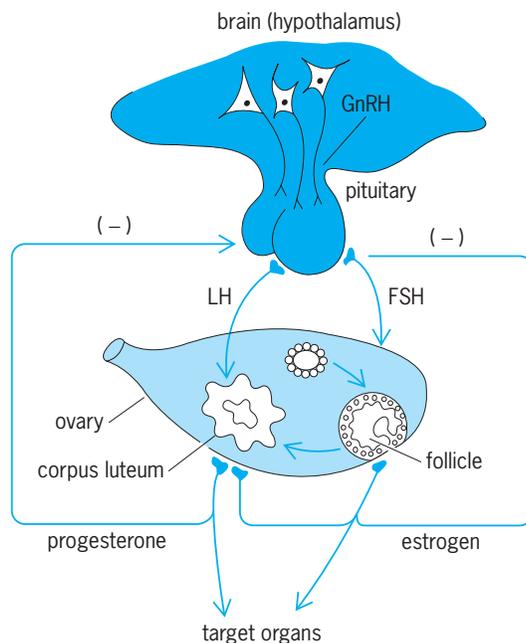


Fig. 3. Active and inactive forms of pregnancy hormone.



**Fig. 4.** Hormonal interactions of the hypothalamus, pituitary, and ovary.

persists during pregnancy, and its secretion of progesterone is important in maintaining the early stages of pregnancy. Later stages are dependent upon increased progesterone secretion from the placenta. Progesterone dominance of the uterus during pregnancy keeps the tissue quiescent and promotes its growth. At the end of the gestation period progesterone dominance wanes, uterine muscle begins to contract rhythmically under the influence of estrogen, and the fetus is expelled (parturition). See PLACENTATION; PREGNANCY.

**Fertility regulation.** In cases where human fertility is low because of anovulatory cycles, it has been possible to induce ovulation by administering pituitary gonadotropins. Because of difficulties in controlling gonadotropin doses, however, these treatments frequently result in multiple births. In domestic animal breeding, the administration of GnRH, FSH, and LH at precisely controlled times shows great promise for increasing both numbers of offspring per pregnancy and numbers of pregnancies per lifetime.

Oral contraceptive pills have proved to be extremely effective in decreasing human fertility. These pills contain synthetic progestagens and estrogens which are effective orally because they are not inactivated by the liver. These "combination" pills are taken daily from days 5–24 of the menstrual cycle, and menstruation will usually occur on day 28. An alternative form is the "minipill," which contains only a small quantity of a synthetic form of progesterone. Although less effective than the combination pill, the minipill avoids side effects caused by estrogens. The synthetic steroids in the pills inhibit or alter gonadotropin secretion at the level of the hypothalamus or pituitary. They thus interfere with maturation of follicles and abolish the LH surge which causes ovulation. It is also possible that these compounds af-

fect implantation, fertilization of ova, or the transport of ova. Although rare instances of abnormal blood clotting have been observed in some users of these pills, this method of contraception is considered to be quite safe for most women after a physician's evaluation. See BIRTH CONTROL; REPRODUCTIVE SYSTEM.

Donald E. Smith

**Bibliography.** E. Y. Adashi and P. C. Leung (eds.), *The Ovary*, 1993; A. Altcheck and L. Deligdisch, *Ovarian Disorders: Pathology, Diagnosis, and Management*, 1995; S. I. Fox, *Human Physiology*, 7th ed., 2001; V. Hayssen and A. Van Tienhoven, *Asdell's Patterns of Mammalian Reproduction: A Compendium of Species—Specific Data*, 1993; E. Knobil et al., *The Physiology of Reproduction*, vols. 1–2, 2d ed., 1994; C. G. Scanes and P. K. Pang (eds.), *The Endocrinology of Growth, Development, and Metabolism in Vertebrates*, 1992; R. L. Stouffer (ed.), *The Primate Ovary*, 1988; J. Tepperman, *Metabolic and Endocrine Physiology*, 1980.

## Overvoltage

The difference between the electrical potential of an electrode or cell under the passage of current and the thermodynamic value of the electrode or cell potential under identical experimental conditions in the absence of electrolysis; it is also known as overpotential. Overvoltage is expressed in volts, often in absolute value; it is a measure of the rates of the different processes associated with an electrode reaction.

An understanding of the factors that contribute to the overvoltage is important in the operation of practical electrochemical systems. In batteries, the overvoltage plays a significant role in the available voltage and power. In large-scale industrial electrolysis, overvoltage is a major factor in determining the energy efficiency of a process, and hence, the cost of electricity. See DECOMPOSITION POTENTIAL; ELECTROCHEMICAL PROCESS; ELECTRODE POTENTIAL; ELECTROLYSIS.

**Basic concepts.** Since the overvoltage is governed by kinetic considerations, all of the experimental conditions that can affect the rate of an electrolytic reaction are of importance. These include concentration of electrolyzed substance, temperature, composition of solvent and electrolyte, nature of the electrode surface, mode of mass transfer, and the current density (current per unit area of electrode). A rapid reaction, such as the reduction of mercurous ion on a mercury electrode, occurs with a small overvoltage (a few millivolts). A slow reaction, for example, the reduction of protons or water on a mercury electrode to produce hydrogen gas, requires a large overvoltage (a few volts).

The rates of electrode reactions are frequently determined from current density–potential curves (Fig. 1). Since the departure of the electrode or cell potential from the thermodynamic value upon passage of current is sometimes termed polarization, such curves are also known as polarization

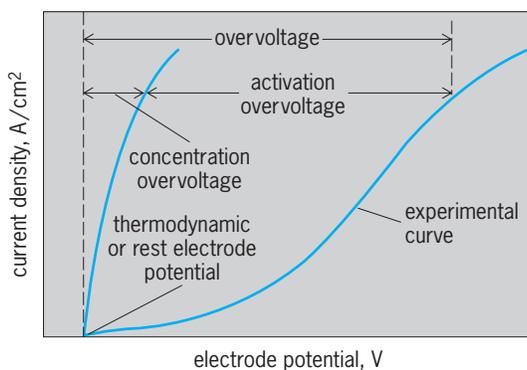


Fig. 1. Current density-potential curve.

curves. These curves are obtained by measurements with three-electrode electrolytic cells. The current density that flows through the electrode of interest (the working electrode) is adjusted with an external direct-current power supply, and the potential of this electrode is measured with respect to a reference electrode whose potential is known and fixed. The measured potential contains a contribution, known as the ohmic drop, that results from the flow of current through the solution resistance between the working electrode and the reference electrode. This drop is minimized by placing these electrodes close together and using various experimental approaches. The thermodynamic potential is obtained from available data for the electrode reaction of interest, corrected for the concentration of the reaction species in the solution under the experimental conditions. It is also sometimes given by the working electrode potential in the electrolysis cell when no current flows (the rest potential). The overvoltage can be read directly from the current-potential curve, as shown in Fig. 1. See ELECTRODE; REFERENCE ELECTRODE.

**Components.** The total overvoltage can be decomposed into different components, which are assigned to different sources of rate limitations, for example, concentration overvoltage, activation overvoltage, reaction overvoltage, and crystallization overvoltage.

**Concentration overvoltage.** This type of overvoltage occurs when the concentration of the reactants or products at the electrode surface are different from those in the bulk solution. These differences arise because the electroactive reactant is consumed, and products are produced by the passage of current. An example is an aqueous solution containing equal concentrations of iron(III) and iron(II) ions. The rest potential of a platinum electrode immersed in this solution will be near the thermodynamic formal potential for these species. When a cathodic current is passed through the platinum electrode, iron(III) is reduced to iron(II), and the concentrations of these species will be changed near the electrode surface; the concentration of iron(III) increases and that of iron(II) decreases. This causes the potential of the electrode to be more negative than the rest potential. This deviation of potential is known as concentration polarization, and its extent is the concentration overvoltage. The magnitude of this overvoltage de-

pends strongly on the mass-transfer rates existing in the electrolysis cell and is smaller under vigorous stirring conditions. Mathematical treatments are sometimes available to account for transport by diffusion and convection in such cells that allow calculation of the concentration overvoltage. Concentration polarization is the only significant contributor to the overvoltage when all other steps in the electrode reaction are rapid and remain essentially at equilibrium. Such electrode reactions are known as reversible or nernstian reactions.

**Activation overvoltage.** This component of the overvoltage arises from slowness in the rate of the electron transfer reaction at the electrode surface. To drive an electrode reaction at a given rate (current density), it is necessary to overcome an energy barrier, the energy of activation for the reaction. The additional energy (that is, beyond the thermodynamic requirements) needed to overcome this barrier is provided by the electrical energy supplied to the cell in the form of an increase in the applied potential. The magnitude of this activation overvoltage often depends upon the nature of the electrode material. For example, the reduction of hydrogen ion to elemental hydrogen is sluggish and occurs with a high activation overvoltage at mercury and lead electrodes, but it is quite rapid and occurs with lower overvoltage at platinum. A substance that decreases the activation overvoltage is known as an electrocatalyst.

The theoretical treatment of the relation between potential and current density for an electrode process with a single rate-determining step is based on a model that considers the overall reaction in terms of a forward (reduction) and backward (oxidation) reaction. The forward-reaction contribution to the overall current density is a cathodic one, and the backward-reaction contribution is an anodic current density (Fig. 2a). The net observed current density at any potential is the algebraic sum of these two components. At the rest potential the net current is zero. The magnitudes of the anodic and cathodic current density components under these conditions are equal in absolute magnitude but opposite in sign. This magnitude of a current density component at the rest potential is known as the exchange current density. The exchange current density is generally expressed in units of amperes per square centimeter ( $A/cm^2$ ) and is directly related to the rate constant for the electrode reaction; a large exchange current density represents a rapid reaction with a large rate constant. The exchange current density cannot be measured directly, but it can be obtained from an analysis of the current-potential behavior, as described below. Both the cathodic and anodic components of the current density vary exponentially with the overvoltage. The cathodic component increases and the anodic one decreases as the overvoltage becomes more negative, while the anodic component increases and the cathodic one decreases as the overvoltage becomes more positive, as shown in Fig. 2a. Reactions whose cathodic and anodic components are both appreciable at potentials near the rest

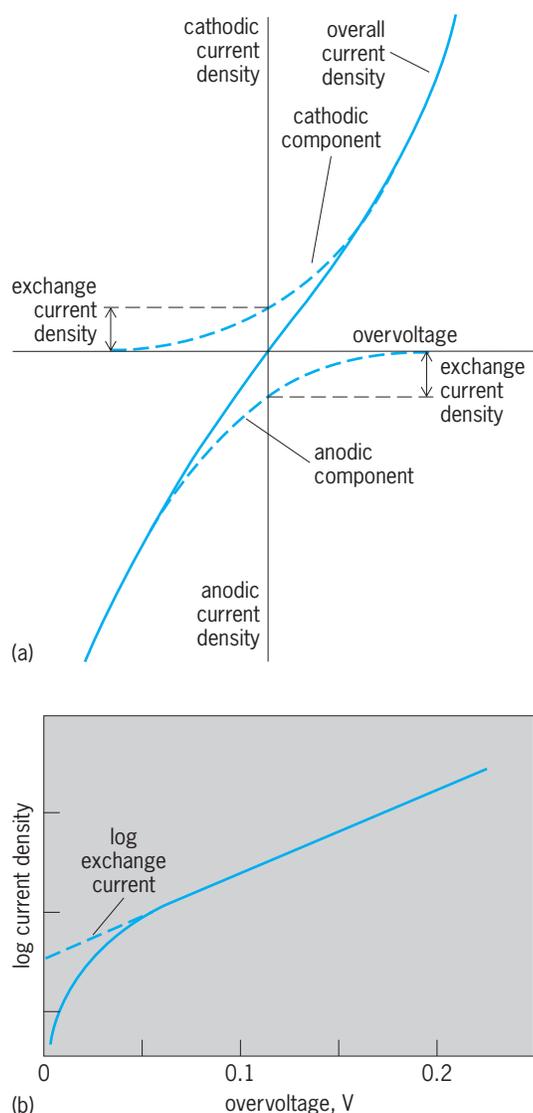


Fig. 2. Current density curves. (a) Experimental current-potential curve showing theoretical cathodic and anodic components. (b) Tafel plot.

potential, that is, for overvoltages between about  $+0.1$  and  $-0.1$  V, are termed quasireversible reactions. See OXIDATION-REDUCTION.

When the overvoltage becomes larger than about  $0.1$  V in absolute value, one of the current components becomes negligible. Under these conditions a plot of the logarithm of the current density against overvoltage is linear (Tafel law). A Tafel plot based on this relationship (Fig. 2b) shows an intercept with the log current density axis (at zero overvoltage) equal to the log exchange current density and a slope that is proportional to a parameter known as the transfer coefficient. An electrode reaction under conditions such that the back reaction makes a negligible contribution to the overall current is known as a totally irreversible reaction.

The exponential dependence of the current density components on overvoltage and the Tafel law were established experimentally. Theoretical treatments of the rates of electrode reactions have led

to a justification of these relations and a more detailed understanding of the factors that contribute to the exchange current density and the transfer coefficient. In general, the exchange current density is a function of temperature, concentration of reactants, nature of the solvent and electrolyte, and electrode surface structure.

*Other overvoltage contributions.* Reaction overvoltage arises when a chemical reaction is associated with the overall electrode reaction. For example, if the electroactive substance is generated from the major reactant by a chemical reaction that precedes the electron transfer, its concentration at the electrode surface will be governed by the rate of this reaction. This preceding reaction will thus affect the potential at which the electrode reaction occurs. Slow steps in the formation of nuclei and the crystal lattice, for example, in the electroplating of a metal, can lead to nucleation and crystallization overvoltages, respectively.

**Electrode reaction kinetics measurements.** While the overvoltage is a manifestation of slow-rate processes in the electrode reaction, the processes themselves are better characterized by rate constants or exchange currents that are independent of current densities or potentials. If the reaction rates are not too rapid, these rate constants can be determined from the steady-state current-potential curves. However, for more rapid reactions, the observed steady-state behavior is often very similar to that of a reversible reaction, and transient methods at short experimental times (microseconds to milliseconds) must be employed. One approach is to obtain the nonsteady-state current-potential curve at a very rapid scan rate (linear scan or cyclic voltammetry). The potentiostatic method involves application of a potential step to the electrode and observation of the resulting current-time curve. Alternatively, a current step can be applied and the potential-time curve recorded (galvanostatic method). Other methods that have been used involve rapid injection of small quantities of charge (coulostatic methods) or studies of the effect of the superposition of a sinusoidal potential variation on the measured current (ac or sine-wave methods). See ELECTROCHEMICAL TECHNIQUES; POLAROGRAPHIC ANALYSIS.

Allen J. Bard

**Bibliography.** A. J. Bard and L. R. Faulkner, *Electrochemical Methods: Fundamentals and Applications*, 2d ed., 2000; J. O'M. Bockris and A. Reddy, *Modern Electrochemistry*, 2d ed., 2000; J. S. Newman, *Electrochemical Systems*, 2d ed., 1991; P. H. Rieger, *Electrochemistry*, 2d ed., 1993; K. J. Vetter, *Electrochemical Kinetics*, 1967.

## Ovum

The egg or female sex cell. Strictly speaking, the term refers to this cell when it is ready for fertilization, but it is often applied to earlier or later stages. Confusion is avoided by using qualifying adjectives such as immature, ripe, mature, fertilized, or developing ova. The mature ova are generally spheroidal

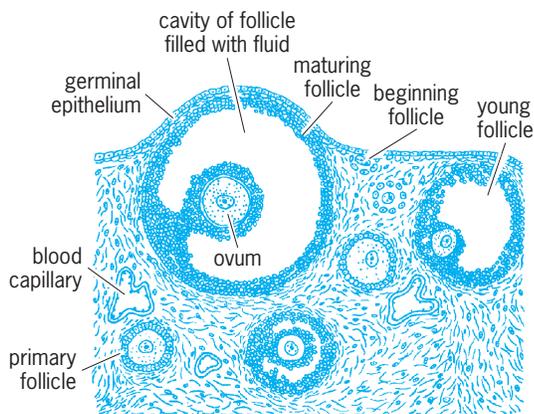


Fig. 1. Section of a mammalian ovary.

and large. In fact, the largest known single cells of a living animal are the mature ova of the ostrich and the shark *Cblamydoselache*, which are about 3 in. (8 cm) in diameter. Among oviparous animals, which spawn eggs at or before the time of fertilization, those that produce larvae capable of feeding at an early stage have small eggs, and their development is generally characterized by radical transitions in appearance, called metamorphosis, before the adult form is attained. The typically viviparous mammals, in which the developing embryo receives nourishment for growth through the uterine tissues of the mother, also characteristically have relatively small eggs which are about 0.1 mm (0.004 in.) in diameter. The number of ova produced at one time varies in different animals, from millions in many marine animals that spawn into the surrounding seawater to about a dozen or less in mammals in which adaptations for internal nourishment of the developing embryo and care of the young are highly developed.

In the ovary the immature ovum is associated with follicle cells through which it receives material for growth. In mammals, as the egg matures, these cells arrange themselves into a structure known as the Graafian, or vesicular, follicle, consisting of a large fluid-filled cavity into which the ovum, surrounded by several layers of cells, projects from the layer of follicle cells that constitutes the inner wall (Fig. 1). The fluid contains estrogenic female sex hormone secreted by cells in an intermediate layer of the follicular wall.

Yolk, or deutoplasm, is essentially a food reserve in the form of small spherules, present to a greater or lesser extent in all eggs. It accounts largely for the differences in size of eggs. Eggs are classified according to the distribution of yolk. In the isolecithal type there is a nearly uniform distribution through the cytoplasm (Fig. 2a), as in most small eggs. The yolk in telolecithal eggs is increasingly concentrated toward one pole (Fig. 2b), as in the large eggs of fish, amphibians, reptiles, and birds. Centrolecithal, or centrally located, yolk (Fig. 2c) occurs in eggs of insects and cephalopod mollusks.

Polarity of organization is manifest in telolecithal eggs by the higher concentration of yolk at the vegetal than at the animal pole, the distribution being

radially symmetrical about the polar axis. This distribution is present, though less marked, in isolecithal eggs. The animal pole of the egg is also the place where the polar bodies are formed. It marks the region where the future head of the embryo develops. See GAMETOGENESIS; OOGENESIS.

Bilateral symmetry is sometimes evident in the shape of the egg, as in insects. The bisecting plane

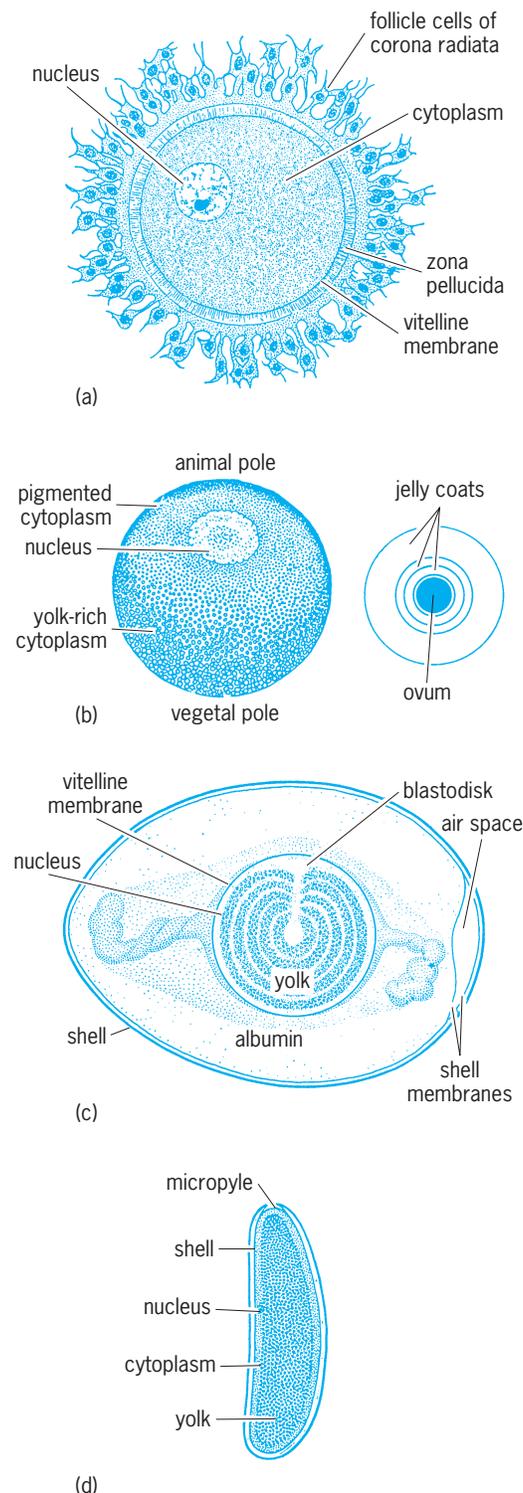


Fig. 2. Types of ova. (a) Isolecithal, human. (b) Telolecithal, frog. (c) Telolecithal, hen. (d) Centrolecithal, fly.

represents the future plane of bilateral symmetry of the embryo. In chickens the plane of bilateral symmetry of the developing embryo is generally perpendicular to the long axis of the shell, and its left side is toward the blunt end of the shell. In frogs and many invertebrates it has been discovered that the path which the fertilizing spermatozoon follows in entering the egg is related to the position of the future plane of bilateral symmetry of the embryo, including its dorsoventral axis. The significance of this may be appreciated when it is realized that, together with the polar axis, the position of the dorsoventral axis specifies completely the future location of all the organs of the developing embryo and adult. The establishment of these axes provides the basic pattern of internal organization of the egg. *See* ANIMAL SYMMETRY.

**Regulative and mosaic eggs.** Despite this patterning at an early stage, it has been possible experimentally in many species of animals to obtain complete individuals from fragments of unfertilized or of developing eggs. This is generally successful when the cuts are made through the polar axis. Eggs in which such fragments develop as a whole are termed regulative and are widely distributed throughout the animal kingdom, including man, as evidenced by the occurrence of identical, multiple births. Eggs in which the fragments develop as structurally defective embryos have been termed mosaic and have been reported principally among the annelids, mollusks, and tunicates. However, the distinction has lost most of its original significance with the demonstration that twins and double monsters were also experimentally obtainable in eggs of the latter group by appropriate procedures. Experiments involving cutting across the polar axis have provided a further clue as to the nature of the internal organization of the egg. When thus cut in half, each fragment develops defectively, but a combination of animal and vegetal quarters develops into a normal individual, as does also the remaining middle fragment. Thus, there are interactions of materials distributed along the polar axis, and a proper balance is essential for normal development. *See* DEVELOPMENTAL BIOLOGY.

**Egg membranes.** The membranes which surround the egg are designated as primary, secondary, and tertiary according to their derivation from the ovum itself, the surrounding follicular cells of the ovary, and the lining of the oviduct, respectively. The primary membrane in practically all animals is called the vitelline membrane. In many animals this becomes elevated from the surface at fertilization and is then called the fertilization membrane. Secondary membranes are represented by the zona pellucida and the surrounding gelatinous, follicle cell-containing material known as the cumulus oophorus in the mammalian egg. Often, however, it is difficult to decide if a particular coat of the egg is produced by the egg itself, by the surrounding follicle cells, or by both. The distinction also loses some of its significance in view of the evidence that most of the materials of the growing oocyte are supplied to it in practically fully

formed state. Tertiary membranes are illustrated by the gelatinous coats of amphibian eggs and by the albumin, shell membranes, and the leathery or calcareous shells of eggs of reptiles and birds. These structures are applied to the egg as it descends the oviduct after ovulation.

In higher plants the egg cell of the embryo sac represents the equivalent of the ovum of animals. In lower plants eggs are formed in an archegonium more nearly analogous to the ovary in animals.

Albert Tyler; Howard L. Hamilton

## Oxalidales

An order of flowering plants (angiosperms) in the eurosid I group of the rosid dicots. The order is previously unrecognized in classifications of the angiosperms but is indicated by numerous studies of DNA sequences. Oxalidales consist of five small families: Cephalotaceae (one species), Connaraceae (300 species of tropical trees and vines), Cunoniaceae (250 species of trees and shrubs mostly from the Southern Hemisphere), Elaeocarpaceae (350 species of trees and shrubs from the Southern Hemisphere and Asian tropics), and Oxalidaceae (350 species, mostly in *Oxalis*, mostly herbs that are found throughout the world). Like many of the other newly defined orders based on studies of DNA sequences, Oxalidales are heterogeneous in their morphological traits; some are bizarre, such as the *Cephalotus* (Cephalotaceae), a carnivorous pitcher plant from Australia. Many species of the order are locally economically important, producing timbers and fruits, including zebra wood (*Connarus*, Connaraceae), star fruit (*Averrhoa*, Oxalidaceae), and lightwood (*Ceratopetalum* and *Eucryphia*, Cunoniaceae). *Oxalis* (Oxalidaceae) has some species that are grown as ornamentals and several that are noxious introduced weeds. *See* MAGNOLIOPHYTA; MAGNOLIOPSIDA; ORNAMENTAL PLANTS; PITCHER PLANT; ROSIDAE; WEEDS.

Mark W. Chase

## Oxidation process

Processes in which oxygen is caused to combine with other molecules. The oxygen may be used as elemental oxygen, as in air, or in the form of an oxygen-containing molecule which is capable of giving up all or part of its oxygen. Oxidation in its broadest sense, that is, an increase in positive valence or removal of electrons, is not considered here if oxygen itself is not involved. *See* OXIDATION-REDUCTION.

Most oxidations occur with the liberation of large amounts of energy in the form of heat, light, or electricity. The stable ultimate products of oxidation are oxides of the elements involved. These oxidations occur in nature as corrosion, decay, and respiration and in the deliberate burning of matter such as wood, petroleum, sulfur, or phosphorus to oxides of the constituent elements. This article deals only with cases where the object of the oxidation process is

the manufacture of a chemical product rather than the production of energy.

The principal variables to be considered and controlled in any partial oxidation are temperature, pressure, reaction time (or contact time), nature of catalyst, if any, mole ratio of oxidizing agent, and whether the substance to be oxidized is to be kept in the liquid or vapor phase. Only a narrow range of conditions unique to each substance being oxidized and each product desired will give satisfactory yields. It is also essential to maintain conditions outside the range of spontaneous ignition, to avoid explosive mixtures or the accidental accumulation of unstable peroxides, and to choose materials which not only can resist the environmental conditions but also which do not have adverse catalytic effects or otherwise interfere with the desired reaction. *See* COMBUSTION; EXPLOSIVE.

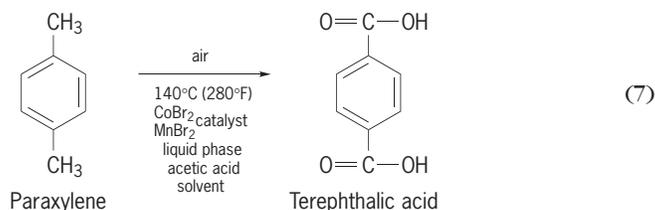
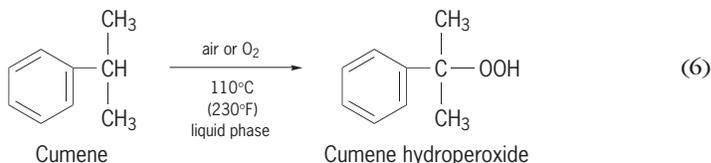
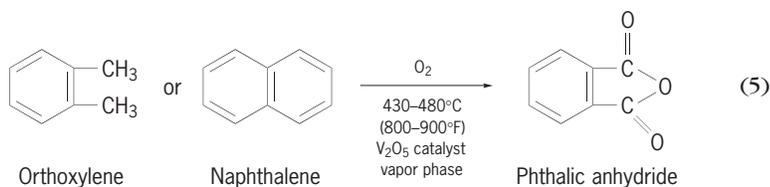
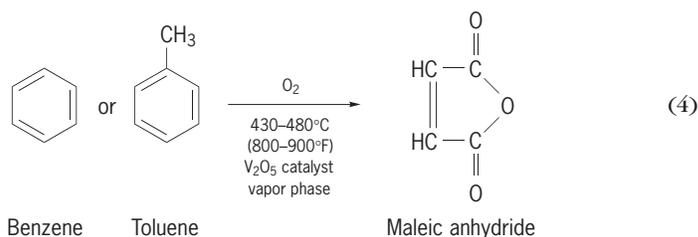
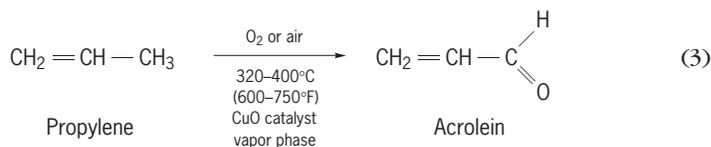
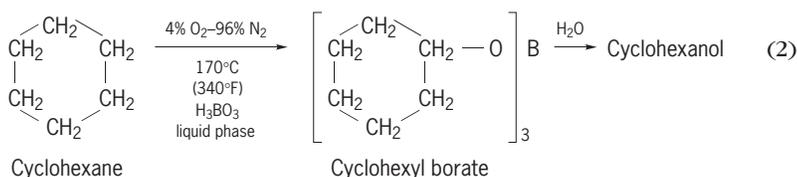
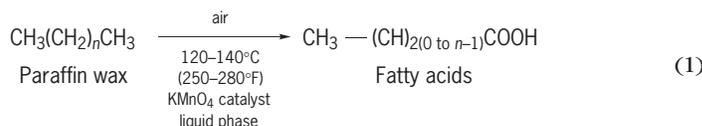
Most oxidations of organic compounds with oxygen appear to proceed through free-radical chain reactions. The specific transient intermediates and sequence of reactions are very complex and are still not completely understood. Many oxidation reactions are autocatalytic. An induction period is therefore often observed during which the concentration of catalyst or intermediate is being built up to the level required for the reaction to proceed at a measurable rate. The reaction rate can continue to increase due to reaction chain branching unless controlled. *See* CHAIN REACTION (CHEMISTRY).

Catalytic effects can be obtained with solid surfaces, generally used in vapor-phase oxidations; with soluble salts, generally used for liquid-phase oxidations; with gases, added in small amounts to the air; or with radiation. Very often catalysts are mixtures in which the action of the major component is modified or maintained by the addition of other agents to the gas, liquid, or solid phase. A limited number of generalizations can be made about oxidation catalysts. The metals, used in the form of their oxides or salts, have a variable valence under the reaction conditions. In vapor-phase reactions the active catalyst is generally deposited on an inert support. Liquid-phase catalysts are usually in the form of a salt, very often containing cobalt or manganese, which is soluble in the organic medium. Inhibitors are also important. Catalyst supports and materials of construction must not have an inhibitory effect and, in liquid-phase oxidations, reactive molecules which terminate free-radical chains, such as phenols, must be avoided. *See* CATALYSIS; INHIBITOR (CHEMISTRY).

In partial oxidations, high selectivity, which may be defined as the equivalents of desired product formed divided by the total moles of feed oxidized, is often made possible by differing degrees of resistance to attack among the various atoms in an organic molecule. Hydrogen atoms attached to aliphatic carbon atoms are more easily oxidized than those attached to carbon atoms in aromatic rings. Among nonaromatics, the ease of oxidation is in the order tertiary > secondary > primary; so that methyl groups are relatively resistant. On the other hand, hydrogen atoms attached to aliphatic carbon atoms are

activated by adjacent methyl groups, double bonds, and aromatic rings in roughly that increasing order. These observations are qualitatively illustrated by the examples shown in reactions (1)–(7). Among the compounds already containing oxygen, a rough order of decreasing stability under oxidizing conditions is carboxylic acid anhydrides > esters > carboxylic acids > ketones > secondary alcohols > aldehydes.

In reactions (4) and (5) the anhydride group is relatively stable under the reaction conditions, whereas the corresponding carboxyl group is not. In cases where the reaction conditions required to initiate oxidation of the starting material are severe, it is generally necessary to employ short reaction times, for



example, 0.001–1 s, by rapidly quenching the reaction mixture to a temperature at which the desired products can survive. In reaction (6) the hydroperoxide is sufficiently stable under the relatively mild conditions required to oxidize cumene for reaction times of an hour or more to be employed. In reaction (2) the unstable secondary alcohol is prevented from decomposing or oxidizing further by formation of the more stable borate ester. Double bonds themselves appear to be relatively resistant to attack by oxygen. Ethylene can be oxidized to ethylene oxide with silver catalysts at high temperatures, but higher olefins require chemical oxidizing agents, such as peroxides, ozone, or hypochlorite, since oxygen attacks other parts of the molecule first.

Plant and equipment design is of utmost importance in achieving efficient control of the oxidation. Vapor-phase reactors with fixed catalyst beds are generally built in the form of tubular heat exchangers if excess air is used. A thermally stable heat exchange medium is circulated outside of the tubes. Molten salts such as sodium nitrate–sodium nitrite mixtures are used for temperatures in the 260–540°C (500–1000°F) region. For temperatures between 100 and 320°C (212 and 600°F), stable organic oils or water under pressure are employed.

If the reaction permits, the catalyst is often used in the form of a fluidized bed; that is, the particle size of the catalyst and support is small enough so that it is suspended in a large vessel by the upward flowing gas. For a given oxidation, reaction time is generally longer than with a fixed bed and the temperature lower. The gas velocity is such that most of the catalyst is not blown out of the reaction zone, and the catalyst which is blown out is removed by cyclone separators or ceramic filters or both and returned. The entire bed is subject to continual mixing, has a high heat capacity, and is easily maintained at a predetermined uniform temperature. Heat is removed by circulating a portion of the hot catalyst through a heat exchanger. Cooling coils can also be put inside the reactor. *See* FLUIDIZATION.

Liquid-phase oxidations are usually carried out in large vessels with internal cooling coils. Inlet air is distributed by spargers or by introducing it at high velocity through a small orifice. Gases leaving the reactor can carry large amounts of the reactants and products, which must be removed by chilling or absorption to minimize losses. In most large oxidation plants, gases are incinerated before release to the atmosphere to avoid pollution.

**Commercial processes using oxygen or air.** Synthesis gas is manufactured by partial oxidation of hydrocarbons with oxygen. The carbon monoxide–hydrogen mixture obtained is used to make primary alcohols from olefins through the Oxo process and for methanol synthesis. Hydrogen is manufactured from synthesis gas by selective oxidation of the carbon monoxide in the mixture to carbon dioxide and by removal of the latter from the hydrogen by combination with an amine. Carbon dioxide is sometimes recovered from the amine as a separate commercial product.

A number of processes for the oxidation of light aliphatic hydrocarbons are in use for the production of mixtures of acetic acid, methanol, formaldehyde, acetaldehyde, and other products in lesser amounts. The processes are reported to use air or oxygen and to operate with and without catalysts over a wide range of temperatures and pressures in both the liquid and gas phases. Processes also have been developed for oxidizing ethylene to acetaldehyde and vinyl acetate. Aqueous systems with palladium–copper salt catalysts are reported. Ethylene is also oxidized to ethylene oxide in the vapor phase with air in tubular reactors with supported silver catalysts. Temperatures are in the range 200–320°C (400–600°F), and pressures are between 100 and 300 lb/in.<sup>2</sup> (700–2100 kilopascals).

Propylene oxide is manufactured by oxidizing propylene with the hydroperoxides of either cumene or isobutane, as in reaction (6). Isobutene or styrene are formed as by-products. Cumene hydroperoxide is cleaved in acid solution to phenol and acetone. Phenol is also manufactured by oxidation of benzoic acid with cupric salts, which are regenerated in place with air.

In the manufacture of phthalic anhydride, reaction (5), both fixed- and fluid-bed vanadium oxide catalysts are used. Excess air is employed at substantially atmospheric pressure. Processes for oxidizing benzene to maleic anhydride and of durene to pyromellitic anhydride are similar to those for phthalic anhydride.

Large amounts of terephthalic acid are manufactured from paraxylene by liquid-phase oxidation with air in the presence of manganese and cobalt bromide catalysts. Acetic acid is used as a solvent. A similar process co-oxidizes acetaldehyde along with the xylene in the presence of cobalt acetate, forming acetic acid as a by-product. The acetaldehyde acts to assist oxidation of the xylene by way of a peroxide intermediate, which is probably the actual oxidizing agent. Methyl paratoluate is oxidized to methyl hydrogen terephthalate in the liquid phase with soluble cobalt salts. Processes for isophthalic acid are analogous to those for terephthalic.

Many attempts have been made to manufacture fatty acids commercially by oxidation of paraffin wax. Synthetic fats are reported to have been manufactured from them in Germany during both world wars. It is difficult to separate the oxyacid and lactone by-products from the desired fatty acid, however, and the process does not appear capable of competing in a free economy with fatty acids and fats from natural sources. Fatty alcohols are manufactured by air oxidation of aluminum alkyls and hydrolyzing the resulting aluminum alcoholate.

Cyclohexane is oxidized in the liquid phase with air to form a mixture of cyclohexanol and cyclohexanone. This mixture is further oxidized with nitric acid to adipic acid. Another development is the use of boric acid in the liquid oxidation medium to form the borate ester of cyclohexanol. The ester is hydrolyzed to recover cyclohexanol, and the boric acid is recycled. The use of boric acid is claimed to give

higher total yields and to minimize the formation of troublesome by-products. This process is also used to make secondary alcohols from paraffins for use as detergent intermediates.

Several oxidation processes to produce acetylene have been developed. In these cases, however, the function of the oxidation itself is to produce temperature of around 1650°C (3000°F). Acetylene is produced from natural gas or naphtha introduced into the hot gases.

Two commercial oxidation processes involve oxidation of hydrogen chloride to chlorine over a manganese catalyst in the presence of ethylene to form vinyl chloride and the reaction of propylene and ammonia with oxygen or air to form acrylonitrile. Methanol and ethanol are oxidatively dehydrogenated to the corresponding aldehydes in the vapor phase over silver, copper, or zinc catalysts. Acetaldehyde is oxidized to acetic acid in the liquid phase in the presence of manganese acetate and to acetic anhydride if mixtures of cobalt and copper acetates are used. Polyunsaturated oils are oxidatively polymerized at room temperature in the presence of soluble salts of cobalt and other metals of variable valence. The physical properties of certain asphalts are improved by oxidation with air at elevated temperature, and phenol can be converted to a high polymer by direct oxidation. Microbiological oxidation to produce animal feed and sodium glutamate, as well as to treat domestic and industrial wastes, is becoming increasingly important.

**Chemical oxidants.** Adipic and terephthalic acids are manufactured from cyclohexanol-cyclohexanone mixtures and paraxylene, respectively, by oxidation with nitric acid of about 30% concentration. Organic and inorganic peroxides are used to manufacture higher olefin oxides, glycols, and hydrogen peroxide. Ozone is used to cleave unsaturated fatty acids to form dibasic acids. Other chemical oxidizing agents for special purposes include sulfur and sulfur compounds, permanganate, perchlorate, hypochlorite, and dichromate. *See* BLEACHING; OZONOLYSIS; PEROXIDE.

**Inorganic processes.** These are generally carried to the highest stable oxidation state and process control is relatively simple.

Vapor-phase oxidations are used to produce major heavy inorganic chemicals, for example, air oxidation of hydrogen sulfide or sulfur dioxide to sulfur trioxide (sulfuric acid); of ammonia to nitric acid; of phosphorus vapor to phosphorus pentoxide (phosphoric acid); of hydrogen chloride to chlorine; and of vaporized zinc to zinc oxide.

Liquid-phase oxidations of inorganic compounds are rare because so few are liquids. Liquid sulfur is burned to sulfur dioxide. At high temperatures mercuric oxide is made from its elements and litharge from molten lead. Air and oxygen, blown through molten iron, make steel by oxidizing such impurities as carbon, sulfur, and phosphorus. *See* STEEL MANUFACTURE.

Solid-phase oxidations are applied most commonly to obtain oxides from the elements. High-

purity carbon dioxide is made from coke in this way. Mixed lead oxides are purified to the monoxide litharge by roasting. Barium peroxide forms from the oxide. Two of the more powerful and costly inorganic oxidizing agents are obtained by processes involving gas-solid phase reactions. Potassium permanganate is produced by roasting a mixture of manganese dioxide and potassium hydroxide with air in a kiln or muffle furnace. In an analogous way, chromite ore and sodium carbonate yield sodium chromate.

I. E. Levine

**Bibliography.** N. Emanuel and D. Gal, *Modelling of Oxidation Processes*, 1986; N. M. Emanuel, G. E. Zaikov, and Z. K. Maizus, *Oxidation of Organic Compounds: Solvent Effects in Radical Reactions*, 1984; M. Murari, *Mechanism of Oxidation of Organic Compounds by Electron Oxidants*, 1985; R. G. Rice, *Advanced Oxidation Processes*, 1995; R. Sheldon and J. Koch, *Metal-Catalyzed Oxidations of Organic Compounds: Mechanistic Principles and Synthetic Methodology Including the Biochemical Process*, 1981.

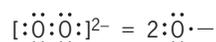
## Oxidation-reduction

An important concept of chemical reactions which is useful in systematizing the chemistry of many substances. Oxidation can be represented as involving a loss of electrons by one molecule and reduction as involving an absorption of electrons by another. Both oxidation and reduction occur simultaneously and in equivalent amounts during any reaction involving either process.

Some important processes which involve oxidation are the rusting of iron or corrosion of metals in general, combustion of hydrocarbons, and the oxidation of carbohydrates (this takes place in a controlled manner in living cells). In each of the foregoing reactions the agent which is reduced is oxygen. Some important reduction processes are the transformation of carbon dioxide to carbohydrates (this takes place in photosynthesis with water being oxidized), the winning of metals from oxides (carbon is often the reducing agent), electrodeposition of metals (this takes place at the cathode, and an equivalent amount of oxidation occurs at the anode), hydrogenation of fats and of coal, and the introduction of electronegative elements such as oxygen, nitrogen, or halogens into hydrocarbons. *See* BIOLOGICAL OXIDATION.

**Oxidation number.** The oxidation state is a concept which describes some important aspects of the state of combination of the elements. An element in a given substance is characterized by a number, the oxidation number, which specifies whether the element in question is combined with elements which are more electropositive or more electronegative than it is. It further specifies the combining capacity which the element exhibits in a particular combination. A scale of oxidation numbers is defined by assigning to an oxygen atom in a molecule such as

$\text{SO}_4^{2-}$  the value of 2-. That for sulfur as 6+ then follows from the requirement that the sum of the oxidation numbers of all the atoms add up to the net charge on the species. The value of 2- for oxygen is not chosen arbitrarily. It recognizes that oxygen is more electronegative than sulfur, and that when it reacts with other elements it seeks to acquire two more electrons, by sharing or outright transfer from the electropositive partner, so as to complete a stable valence shell of eight electrons. For compounds of the halogens an analogous rule is followed, but when a halogen atom is in combination with atoms of a more electropositive element, the oxidation number is taken as 1- because only one electron needs to be added to the valence shell to yield a stable octet. The system amounts to a bookkeeping operation on the electrons, so that for this purpose the more electronegative partner is assigned some agreed upon stable electronic configuration, and after taking into account the total charge on the molecule, the net charge left on the electropositive partner is its particular oxidation number. When the combining capacity of an element toward another one is not completely exhausted in a particular combination, as is the case for oxygen in barium peroxide ( $\text{BaO}_2$ ), the electrons shared between atoms of the same kind are evenly divided between them in carrying out the formal decomposition. Thus in the peroxide unit  $\text{O}_2^{2-}$ , the oxidation number of oxygen is 1-. This is the net charge left on oxygen in the formal decomposition. The ox-

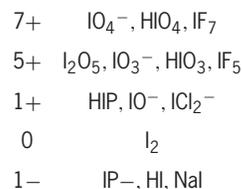


idation number by no means gives a complete description of the state of combination of an atom. Specifically, it is not designed to give the actual charge on an atom in a particular compound. Thus it makes no distinction between fluorine in  $\text{HF}$ ,  $\text{AlF}_3$ , or  $\text{NaF}$ , even though the actual charges residing on the fluorine atoms in these three compounds are different.

The utility of the concept is based in part on just this feature because much of the chemistry of these substances can be understood when it is realized that each of them readily yields  $\text{F}^-$ , as is the case when they dissolve in water. The chemistry of the three substances, in regard to the component fluorine, is concerned with reactions of  $\text{F}^-$ . Although oxidation number is in some respects similar to valence, the two concepts have distinct meanings. In the substance  $\text{H}_2$ , the valence of hydrogen is 1 because each H makes a single bond to another H, but the oxidation number is 0, because the hydrogen is not combined with a different element. See VALENCE.

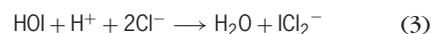
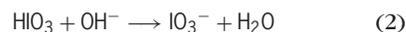
The systematization of chemistry based on the concept of oxidation number can be illustrated with reference to the chemistry of iodine. The usual oxidation states exhibited by iodine are 1-, 0, 1+, 5+, and 7+. Examples of substances corresponding to

each oxidation state are

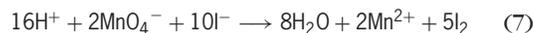
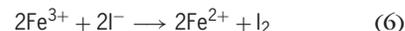


When the oxidation number of an atom in a species is increased, the process is described as oxidation, no matter what reagent produces it; when a decrease in oxidation number takes place, the process is described as reduction, again without regard to the identity of the reducing agent. The term oxidation has been generalized from its historical meaning, implying combination with oxygen, to combination of an element with an element more electronegative than itself.

When classification by oxidation number is adopted, the reactions fall naturally into two classes. In the first class, no change in oxidation number takes place and, in the second, the class of oxidation-reduction reactions, changes in oxidation number do take place. Some examples of the first class are reactions (1)-(4).



Some samples of the second class are reactions (5)-(8). [In reaction (8) it is implied that electrons



are being extracted from  $\text{I}^-$  by an anode in an electrolytic process.]

In reactions of the first class, some center regarded as positive undergoes a change in the nature of the groups associated with it, but provided that the group which replaces the electronegative portion is more electronegative than the center, there is no change in oxidation state. Reaction (3) describes the replacement of  $\text{OH}^-$  on  $\text{I}^+$  by  $\text{Cl}^-$ ; both Cl and OH are more electronegative than I. In reactions of the second class, changes in oxidation number occur which may or may not be accompanied by changes in the state of association of the centers in question.

Reactions (5), (6), (7), and (8) illustrate the utility of the concept of oxidation number. A variety of reagents as different in their properties as  $\text{Cl}_2$ ,  $\text{Fe}^{3+}$ ,  $\text{MnO}_4^-$ , and an anode serve to bring about the change, or oxidation, of  $\text{I}^-$  to  $\text{I}_2$ . However, their chemical individuality does not affect the state of the product iodine, and nongroup characteristic of the

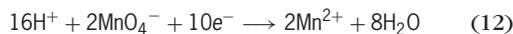
oxidizing agent is necessarily transferred in the net change. This situation obtains only for reactions in a strongly solvating medium such as water, which provides the groups that associate with the atom being oxidized or reduced. Thus, when the reactions take place in the solid, it is necessary to specify what iodide is being used, whether sodium iodide, NaI, or silver iodide, AgI, for example, and the properties of the reaction would be dependent on the choice.

In representing an element in a particular environment, it is often convenient to specify only the oxidation state, without attempting to identify all of the groups which are attached to the atom in question. Thus, the iron content of a solution made up by dissolving, say, ferric chloride in water will be composed, among others, of the species  $\text{Fe}^{3+}$ ,  $\text{FeCl}^{2+}$ ,  $\text{FeOH}^{2+}$ . Collectively, they are correctly, though of course not fully, described by the notation Fe(III). In this kind of usage, the roman numeral represents the oxidation state.

**Oxidation-reduction reactions.** In an oxidation-reduction reaction, some element decreases in oxidation state and some element increases in oxidation state. The substances containing these elements are defined as the oxidizing agents and reducing agents, and they are said to be reduced and oxidized, respectively. The processes in question can always be represented formally as involving electron absorption by the oxidizing agent and electron donation by the reducing agent. For example, reaction (6) can be regarded as the sum of the two partial processes, or half-reactions, (9) and (10).



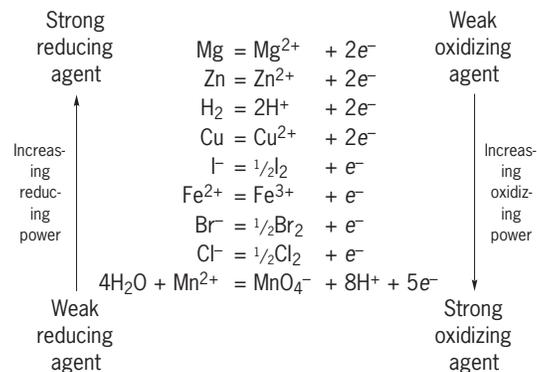
Similarly, reaction (7) consists of the two half-reactions (11) and (12), with half-reaction (11) being



taken five times to balance the electron flow from reducing agent to oxidizing agent.

Each half-reaction consists of an oxidation-reduction couple; thus, in half-reaction (12) the reducing agent and oxidizing agent making up the couple are manganous ion,  $\text{Mn}^{2+}$ , and permanganate ion,  $\text{MnO}_4^-$ , respectively; in half-reaction (11) the reducing agent is  $\text{I}^-$  and the oxidizing agent is  $\text{I}_2$ . The fact that  $\text{MnO}_4^-$  reacts with  $\text{I}^-$  to produce  $\text{I}_2$  means that  $\text{MnO}_4^-$  in acid solution is a stronger oxidizing agent than is  $\text{I}_2$ . Because of the reciprocal relation between the oxidizing agent and reducing agent comprising a couple, this statement is equivalent to saying that  $\text{I}^-$  is a stronger reducing agent than  $\text{Mn}^{2+}$  in acid solution. Reducing agents may be ranked in order of tendency to react, and this ranking immediately implies an opposite order of tendency to react for the oxidizing agents which complete the couples. In the list below some common oxidation-reduction

couples are ranked in this fashion:



A complete list contains the displacement series of the metals. The most powerful reducing agent shown in the list is magnesium, Mg, although this is not the most powerful known. Magnesium is capable of reacting with any oxidizing agent below it in the list to yield  $\text{Mg}^{2+}$  and to produce the reduced product resulting from the oxidizing agent. Similarly, permanganate ion,  $\text{MnO}_4^-$ , in acid, the strongest oxidizing agent shown, is capable of reacting with any reducing agent above it in the list. Conversely, an oxidizing agent at the top of the list will not react appreciably with the reducing agent of a couple below it. The list given, containing nine entries, can be used to predict the results of 72 separate experiments (for example,  $\text{Mg} + \text{Zn}^{2+}$  on the one hand and  $\text{Mg}^{2+} + \text{Zn}$  on the other would be counted as separate experiments in arriving at this figure). See ELECTROCHEMICAL SERIES; ELECTRONEGATIVITY.

Since the driving force for a reaction depends on concentrations, the concentrations of all reactants and products must be specified, as well as other conditions, in compiling a list such as that given. The order shown obtains for water solutions at  $25^\circ\text{C}$  ( $77^\circ\text{F}$ ), approximately 1 M in all soluble reagents and having gases present at approximately 1 atm pressure. A second limitation on the use of this list lies in the fact that it applies only when the expected reaction products are compatible. Although copper is capable in principle of reducing iodine to form  $\text{I}^-$  and  $\text{Cu}^{2+}$  at high concentration, these products are not compatible with each other, but they react to form copper (I) iodide,  $\text{CuI}$ . Allowance for such features can always be made by incorporating the necessary additional half-reactions into the list. Finally, it must be stressed that the list can be used to predict the results of experiments only for systems which reach equilibrium sufficiently rapidly; it does not serve to predict the rate of reaction. To achieve the reduction of  $\text{Fe}^{3+}$  by  $\text{H}_2$  predicted in the list, it would be necessary to use a catalyst in order to realize the reaction in a reasonable time.

The equilibrium information implied by a table of half-reactions can readily be put into quantitative form. Thus, the standard free-energy change for the reaction of 1 equivalent weight of each reducing agent with some common oxidizing agent can be entered opposite each half-reaction. The numerical values of these quantities will be in the same order as

are the half-reactions and can be combined algebraically to yield the standard free energy change, and therefore the equilibrium constant, for any reaction which can be written from the table. *See* CHEMICAL EQUILIBRIUM.

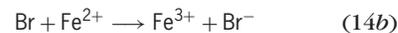
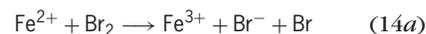
A chemist concerned with reactions of the type under discussion will have a ready vocabulary of facts concerning oxidizing agents and reducing agents, such as their oxidizing or reducing powers, the speed with which they react, and the characteristics which may complicate their application. A typical problem in analytical chemistry is to reduce  $\text{Cr}_2\text{O}_7^{2-}$  to  $\text{Cr}^{3+}$  in acidic (perchloric acid) solution without introducing elements which are not already present. Metallic reducing agents such as zinc and iron or metal ion reducing agents are immediately eliminated from consideration because the products of oxidation may be difficult to remove from the resulting solution. A solution of hydrogen iodide, HI, would be suitable, except that it would be necessary to take special pains to add it in equivalent amount because excess HI would be difficult to remove (the iodine,  $\text{I}_2$ , produced by oxidation of  $\text{I}^-$ , however, is easy to remove by extracting it with a suitable solvent such as carbon tetrachloride). Hydrogen would be ideal (the product of its oxidation,  $\text{H}^+$ , is already present in solution, and excess reducing agent can easily be removed) except that the rate of reaction would be disappointingly slow. A suitable reducing agent would be hydrogen peroxide,  $\text{H}_2\text{O}_2$ ; it reacts rapidly, the product of oxidation is oxygen, which escapes from solution, and excess oxidizing agent is easily destroyed by heating the solution. *See* ELECTRODE POTENTIAL.

**Mechanisms.** The data needed to predict the outcome at equilibrium of the reaction of most common oxidizing and reducing agents are known. A list of the kind shown above can be extended, and when it is elaborated with entries carrying the quantitative information, accurate calculations of the equilibrium state for all the reactions implied by the table can be made. By contrast, though the rates of reaction are also of great importance, they are much less completely understood and less completely described. To understand the rates of reaction, it is necessary to consider how the reactions take place. To illustrate one of the problems of mechanism, a reaction is selected which, though not nearly as complicated as some, suffices for present purposes. When an aqueous solution containing  $\text{Fe}^{2+}$  is added to one containing  $\text{Br}_2$ , reaction (13) takes place. For the final stable



products to be produced in a single step would require that two  $\text{Fe}^{2+}$  and one  $\text{Br}_2$  be brought together in an encounter. This course for the reaction is found to be less probable than one in which a single  $\text{Fe}^{2+}$  encounters one  $\text{Br}_2$ . Since in forming stable products the oxidation state of iron increases by one unit while that of a bromine molecule decreases by two (one for each atom), the reaction resulting from the encounter must leave either iron or bromine in an

unstable state. Reasonable alternatives for the reactive intermediates produced are represented in reactions (14a) and (15b), together in each case with



a sequel reaction which leads to the correct overall stoichiometry. *See* REACTIVE INTERMEDIATES.

In the present system, the evidence points to the mechanism represented by the first alternative [reactions (14a) and (14b)]. But even after a reaction such as (13) is resolved into the different steps, and reactive intermediates are identified, there remain questions about how the changes in oxidation state are brought about in the individual steps. Thus, when  $\text{Fe}^{2+}$  reacts with  $\text{Br}_2$ , do the two reactants make direct contact—this would involve replacement of a water molecule on  $\text{Fe}(\text{H}_2\text{O})_6^{2+}$ , the form which  $\text{Fe}^{2+}$  adopts in water, by  $\text{Br}_2$ —or does reaction occur by electron transfer from intact  $\text{Fe}(\text{H}_2\text{O})_6^{2+}$  to  $\text{Br}_2$ ? If the latter occurs, does electron transfer take place over large distances from  $\text{Fe}(\text{H}_2\text{O})_6^{2+}$  to  $\text{Br}_2$ ?

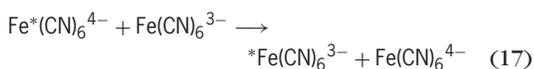
These questions are important not only for inorganic systems of the kind illustrated by the steps making up reaction (13) but also for reactions at electrodes, and for oxidation-reduction reactions catalyzed by enzymes in biological systems. These kinds of questions are under investigation and have been partly answered for certain systems. *See* ENZYME.

Two different kinds of mechanisms are recognized for oxidation-reduction reactions. So-called outer-sphere reactions are easier to understand in a fundamental way and will be described first. Reaction (16) introduces a typical such reaction. Here the changes in oxidation state, 4+ to 3+ for Ir and 2+ to 3+ for Fe, take place without bonds to either atom being broken, and in this particular system there is not even much change in bond distances attending the changes in oxidation state. Electron transfer is explicit in such a system, and the electron transfer act is subject to the Franck-Condon restriction. This imposes a barrier to electron transfer in that it requires that the environments (coordination sphere and solvent) readjust prior to electron transfer so that after readjustment the energy of the system is the same whether the electron is on one metal or the other. The rates of reactions such as (16) can be estimated



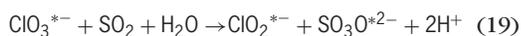
at least approximately by calculating the work to bring the partners together, the work of meeting the Franck-Condon restriction, and assuming that electron delocalization when the partners are in contact is adequate. Greater success has been met in attempts at correlating rates and here the equation developed by R. A. Marcus, relating the rate of reaction such as (16) to the standard free energy change

and to the rates of the so-called self-exchange reactions [(17) and (18)], is proving to be useful not only



in simple systems but also in understanding electron transfer reactions for large and complex biological molecules.

Much more difficult to systematize and understand is the extensive and very important class of oxidation-reduction reactions in which the changes in oxidation state are linked to bond breaking or bond making. A simple example is provided by reaction (19). Isotopic labeling experiments have shown



that in the course of the reaction an oxygen atom originating on  $\text{ClO}_3^-$  is transferred to the reducing agent. Though the process can formally be represented as involving electron loss by  $\text{SO}_2$  and electron gain by  $\text{ClO}_3^-$ , electron transfer as the means by which the changes in oxidation state are brought about is not at all explicit in a reaction of this kind. These so-called inner-sphere mechanisms operate also for reactions involving metal ions. For example, when  $[(\text{NH}_3)_5\text{CoCl}]^{2+}$  reacts with  $\text{Cr}(\text{H}_2\text{O})_6^{2+}$ , the Cr(III) product has the formula  $\text{Cr}(\text{H}_2\text{O})_5\text{Cl}^{2+}$ , and Cl is transferred from Co(III) to Cr(II) when the former oxidizes the latter. This kind of reaction clearly has much in common with that represented by reaction (19). In the latter system atomic oxygen is formally transferred; in the former, atomic chlorine.

A class of inner-sphere reactions of metal-containing molecules which are now recognized as playing an important role in many catalytic processes involves so-called oxidative addition. Reactions of this kind have been known for a long time, but their significance was not appreciated until interest in homogeneous catalytic processes began to develop. A commonplace example of oxidative addition is provided by reaction (20). It will be noted that in reac-



tion (20) both the oxidation number and the coordination number of Sn increase. Oxidative addition for a strong oxidizing agent such as  $\text{Cl}_2$  is not surprising, but the reaction type took on new significance when it was discovered that with a suitable metal complex, oxidative addition can be realized also with hydrogen halides, alkyl halides, and even hydrogen. Among the metal complexes which undergo this kind of reaction are Rh(I), Ir(I), or Pt(O) species with 4 or fewer groups attached. In each case, there is the opportunity for an increase in both oxidation and coordination number. A specific example of a molecule which undergoes oxidative addition with  $\text{H}_2$  is  $[(\text{C}_6\text{H}_5)_3\text{P}]_3\text{RhCl}$ , which is a useful catalyst for the hydrogenation of alkenes and alkynes. A reaction step in the catalytic sequence is the addition of  $\text{H}_2$  to the metal so that the H-H bond is

severed; the adduct then reacts with the alkene (or alkyne) transferring two atoms of hydrogen. A substance which will activate H-R bonds (where R is an alkyl radical) in the same way would be desirable. See HOMOGENEOUS CATALYSIS.

The fundamental aspects of electron transfer processes in oxidation-reduction reactions have much in common with the electron jump processes in semi-conductors. Recognizing this connection is productive both for those interested in chemical effects accompanying electron transfer (that is, in oxidation-reduction processes) and those interested in electron mobility as a subject in its own right.

Henry Taube

Bibliography. J. P. Collman and L. S. Hegedus, *Principles and Applications of Organotransition Metal Chemistry*, 2d ed., 1987; L. E. Ebersson, *Electron Transfer Reactions in Organic Chemistry*, 1987; A. G. Lappin, *Redox Mechanisms in Inorganic Chemistry*, 1993; W. M. Latimer, *Oxidation States of the Elements and Their Potentials in Aqueous Solution*, 2d ed., 1952; H. Taube, *Electron Transfer Reactions of Complex Ions in Solution*, 1970.

## Oxide

A binary compound of oxygen with another element. Oxides have been prepared for essentially all the elements, with the exception of the noble gases. Often, several different oxides of a given element can be prepared; a number exist naturally in the Earth's crust and atmosphere: silicon dioxide ( $\text{SiO}_2$ ) in quartz; aluminum oxide ( $\text{Al}_2\text{O}_3$ ) in corundum; iron oxide ( $\text{Fe}_2\text{O}_3$ ) in hematite; carbon dioxide ( $\text{CO}_2$ ) gas; and water ( $\text{H}_2\text{O}$ ).

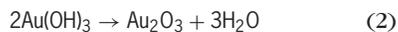
Most elements will react with oxygen at appropriate temperature and oxygen pressure conditions, and many oxides may thus be directly prepared. Phosphorus burns spontaneously in oxygen to form phosphorus pentoxide,  $(\text{P}_2\text{O}_5)_2$ . Sulfur requires ignition and thereafter burns to sulfur dioxide ( $\text{SO}_2$ ) gas if the supply of oxygen is limited. The relative amounts of oxygen and element available often determine which of several oxides will form; in an excess of oxygen, sulfur burns to form some sulfur trioxide ( $\text{SO}_3$ ) gas. Most metals in massive form react with oxygen only slowly at room temperatures because the first thin oxide coat formed protects the metal; magnesium and aluminum remain metallic in appearance for long periods because their oxide coatings are scarcely visible. However, diffusion of the oxygen and metal atoms through the film becomes rapid at high temperatures, and these metals will burn intensely to their oxides if ignited. The oxides of the alkali and alkaline-earth metals, except for beryllium and magnesium, are porous when formed on the metal surface, and they provide only limited protection to the continuation of oxidation, even at room temperatures. Gold is exceptional in its resistance to oxygen, and its oxide ( $\text{Au}_2\text{O}_3$ ) must be prepared by indirect means. The other noble metals, although

ordinarily resistant to oxygen, will react at high temperatures to form gaseous oxides.

Indirect preparation of the oxides may be accomplished by heating hydroxides, nitrates, oxalates, or carbonates, as in the production from the latter of quicklime (CaO) by reaction (1), in which carbon



dioxide is driven off. Gold oxide may be prepared by heating gold hydroxide, as shown by reaction (2). Higher oxides of an element may be reduced



to lower oxides, usually at high temperatures, for example, the reduction of tungsten trioxide by tungsten, as shown by reaction (3). Complete reduction



to the element may be performed by other elements whose oxides are more stable, as in the formation of calcium oxide from titanium dioxide by reaction (4).



Although the solid oxides of a few metals, such as mercury and the noble metals, can be easily decomposed by heating, for example, reaction (5), most



metal oxides are very stable and many constitute important refractory materials. For example, magnesium oxide, calcium oxide, and zirconium dioxide do not melt or vaporize appreciably at temperatures up to 2500°C (4500°F). A great number of refractories consist of compounds of two or more oxides; silicon dioxide and zirconium dioxide form zirconium silicate by reaction (6).



Because so many oxides can be easily formed, studies of them have been most important in establishing relative atomic weights of the elements based on the defined atomic weight for oxygen. Furthermore, these studies were fundamental in forming the basis for the laws of definite proportions and multiple proportions for compounds. It is of special significance that, although any gaseous oxide species must necessarily have a definite oxygen-to-element proportion, a number of solid and liquid oxides can be prepared with proportions which may vary continuously over a considerable range. This is particularly true for oxides prepared under equilibrium conditions at high temperatures. Thus, titanium exposed to oxygen until reaction equilibrium is reached at a number of selected conditions of temperature and oxygen pressure will form the solid oxide TiO. It has the same crystal structure as rock salt; that is, every other site along the three coordinate directions of the crystal will be occupied by titanium atoms and the alternate sites by oxygen atoms (each in

its ion form  $\text{Ti}^{2+}$  and  $\text{O}^{2-}$ ) to give the simple Ti/O ratio of 1:1. However, with other selected pressure-temperature conditions, oxides of same structure at every Ti/O ratio from 1:0.7 to 1:1.25 may be prepared. The variable proportions show that variable numbers of oxygen or titanium sites can simply remain vacant in a homogeneous way. The range is referred to as the TiO/O 1:07–1.25 phase or, more loosely, the TiO solid-solution phase.

Most of the nonmetal oxides commonly encountered as gases, such as  $\text{SO}_2$  and  $\text{CO}_2$ , form solids and liquids in which the molecular units of the gas are retained so that the simple definite proportions are clearly maintained. Such oxides melt and boil at low temperatures, because the molecular units are weakly bonded to adjoining molecular units.

Oxides may be classified as acidic or basic according to the character of the solution resulting from their reactions with water. The nonmetal oxides generally form acid solutions, for example, reaction (7),



the formation of sulfuric acid. The metal oxides generally form alkaline solutions, for example, reaction (8), for the formation of calcium hydroxide or slaked



lime. However, given metals of the groups IV and higher of the periodic table will often have basic, intermediate, and acidic oxides. Here the acid character increases with increasing oxygen-metal ratio. See ACID AND BASE; EQUIVALENT WEIGHT; OXYGEN; REFRACTORY. Russell K. Edwards

Bibliography. F. A. Cotton and G. Wilkinson, *Advanced Inorganic Chemistry: A Comprehensive Text*, 6th ed., 1999; G. V. Samsonov, *The Oxide Handbook*, 2d rev. ed., 1982.

## Oxide and hydroxide minerals

Mineral phases containing only oxide or hydroxide anions in their structures. By volume, oxide and hydroxide minerals comprise only a small fraction of the Earth's crust. However, their geochemical and petrologic importance cannot be overstated. Oxide and hydroxide minerals are important ores of metals such as iron, aluminum, titanium, uranium, and manganese. Oxide and hydroxide minerals occur in all geological environments. Some form as primary minerals in igneous rocks, while others form as secondary phases during the weathering and alteration of silicate and sulfide minerals. Some oxide and hydroxide minerals are biogenic; for example, iron(III) and manganese(IV) hydroxides and oxides often result from bacterial oxidation of dissolved  $\text{Fe}^{2+}$  and  $\text{Mn}^{2+}$  in low-temperature aqueous solutions. See HYDROXIDE; MINERAL; ORE AND MINERAL DEPOSITS; RID="622600">SILICATE MINERALS; WEATHERING PROCESSES.

Iron and manganese hydroxide minerals often occur as nanocrystalline or colloidal phases with

high, reactive surface areas. Adsorption of dissolved aqueous ions onto colloidal iron and manganese oxides plays a major role in the fate of micronutrients and heavy metals in soil and ground water and the trace-element chemistry of the oceans. Much current research is focused on measuring the thermodynamics and kinetics of metal adsorption by colloidal Fe-Mn hydroxides and oxides in the laboratory. In anoxic sedimentary environments, bacteria may use iron(III) and manganese(IV) hydroxide minerals as electron acceptors. Consequently, these minerals may facilitate the biodegradation of organic pollutants in soil and ground water. See ADSORPTION; COLLOID.

**Bonding and crystal chemistry.** To a first approximation, the bonding in oxide minerals can be viewed in the ionic (electrostatic) model. According to Pauling's rules, the coordination number of a metal cation (such as  $Mg^{2+}$ ,  $Al^{3+}$ , and  $Ti^{4+}$ ) is determined by the radius of the cation relative to that of the oxide anion ( $O^{2-}$ ). This allows one to predict the structures of some of the simpler oxide minerals (Table 1). Cations with similar ionic radii (such as  $Mg^{2+}$  and  $Fe^{2+}$ ) are able to substitute for each other and form a solid solution. Sulfide minerals (which are more covalent) show little solid solution. The simple ionic radii arguments will fail when electronic configurations preclude the spherical symmetry of the cations. The ions  $Cu^{2+}$  and  $Mn^{3+}$ , with nine and four *d* electrons, tend to adopt distorted coordination environments because of the Jahn-Teller effect. Also, *d*-electron configurations can give rise to large octahedral site-preference energies that cause small cations,

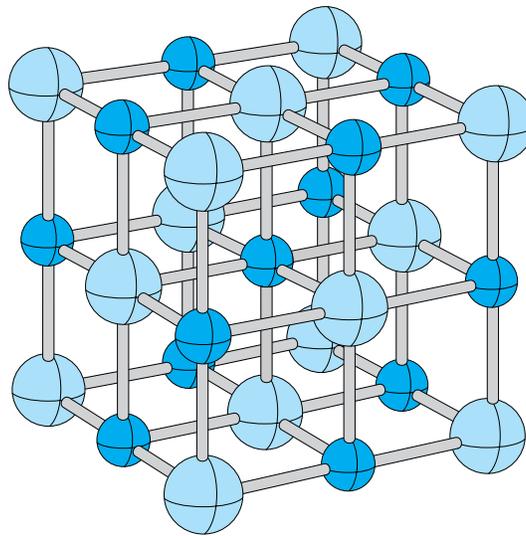


Fig. 1. NaCl structure adopted by XO oxides such as MgO. Small spheres are divalent cations such as  $Mg^{2+}$ . Large spheres are  $O^{2-}$  anions.

such as  $Mn^{4+}$ , to always adopt octahedral coordination. The magnetic and semiconducting properties of transition-metal oxide minerals, such as magnetite ( $Fe_3O_4$ ), give useful geophysical signatures for subsurface exploration. See IONIC CRYSTALS; JAHNTELLER EFFECT; PROSPECTING; SOLID-STATE CHEMISTRY; STRUCTURAL CHEMISTRY; VALENCE.

**Survey of minerals.** The following is a summary of the important oxide and hydroxide minerals by structural classification.

*X<sub>2</sub>O oxides.* The mineral cuprite ( $Cu_2O$ ) forms during the low-temperature oxidation of primary copper sulfide minerals. It is sometimes an important ore of copper. In the structure of cuprite,  $Cu^+$  ions are in twofold coordination with oxygen. No other monovalent cations form stable  $X_2O$  oxide minerals. See COPPER.

*XO oxides.* Oxides of divalent metals, such as  $Fe^{2+}$  and  $Mg^{2+}$ , with the formula XO will adopt the NaCl structure (Fig. 1) in which the metal cation is in sixfold coordination.  $Fe^{2+}$  and  $Mg^{2+}$  have nearly identical ionic radii, and a phase made up of the solid solution (Mg,Fe)O, called ferropericlase, is probably the second most abundant mineral in the Earth's lower mantle. Experimental evidence indicates that it maintains the NaCl structure over the pressure range of the Earth's interior. The end member MgO (periclase) is quite rare as a crustal mineral; it occurs in some metamorphosed limestones. Most other XO oxide minerals (such as manganosite MnO and bunsunite NiO) are also very rare. Not all XO oxides adopt the NaCl structure. Because the ionic radius of  $Zn^{2+}$  is somewhat smaller than that of  $Mg^{2+}$ , the mineral zincite ZnO has a structure based on the tetrahedral coordination of Zn. The mineral tenorite ( $CuO$ ) is a secondary alteration phase of copper sulfides and has a structure based on the square planar coordination of  $Cu^{2+}$ . This structure results from the Jahn-Teller distortion of the  $CuO_6$  coordination polyhedron. Larger divalent cations, such as Sr and Ba,

TABLE 1. Summary of important simple oxide minerals; classified by structure

Mineral	Formula
<b>X<sub>2</sub>O Oxides</b>	
Cuprite	$Cu_2O$
Argentite	$Ag_2O$
<b>NaCl Structure</b>	
Periclase	MgO
Manganosite	MnO
Wustite	FeO
<b>Other MO Oxides</b>	
Tenorite	$CuO$
Zincite	$ZnO$
<b>Corundum Structure</b>	
Corundum	$Al_2O_3$
Eskolaite	$Cr_2O_3$
Hematite	$Fe_2O_3$
Ilmenite	$Fe^{2+}Ti^{4+}O_3$
<b>Rutile Structure</b>	
Rutile	$TiO_2$
Cassiterite	$SnO_2$
Pyrolusite	$MnO_2$
<b>Spinel Structure</b>	
Spinel	$MgAl_2O_4$
Magnetite	$Fe_3O_4$
Chromite	$FeCr_2O_4$
Franklinite	$ZnFe_2O_4$
Ulvo-spinel	$Fe^{2+}(Fe^{2+}Ti^{4+})O_4$
Hausmannite	$Mn_3O_4$ (distorted)
<b>Fluorite Structure</b>	
Uraninite	$UO_2$
Thorianite	$ThO_2$

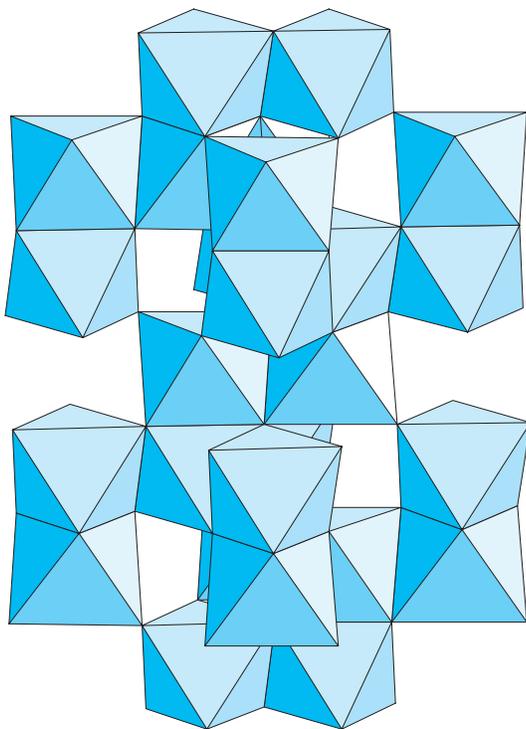


Fig. 2. Polyhedral representation of  $\text{Al}_2\text{O}_3$  structure adopted by  $\text{X}_2\text{O}_3$  oxides. Each octahedron represents an  $\text{M}^{3+}$  cation surrounded by six  $\text{O}^{2-}$  anions, which define the vertices of the octahedron.

form oxide structures based on eightfold coordination, but these are not stable minerals. See CRYSTAL STRUCTURE; SOLID SOLUTION.

*X<sub>2</sub>O<sub>3</sub> oxides and ilmenite.* Trivalent cations, such as  $\text{Fe}^{3+}$  and  $\text{Al}^{3+}$ , having radii appropriate for sixfold coordination with oxygen, will adopt the corundum ( $\alpha\text{-Al}_2\text{O}_3$ ) structure (Fig. 2). Hematite ( $\alpha\text{-Fe}_2\text{O}_3$ ) is a common phase in soils and sediments and forms by the oxidation of  $\text{Fe}^{2+}$  in primary silicates. (Rust is mostly hematite.) Hematite is the most important ore mineral of iron and is the dominant mineral found in Precambrian banded iron formations. These vast deposits formed because of the oxidation of dissolved  $\text{Fe}^{2+}$  in the oceans when the Earth's atmosphere accumulated oxygen from photosynthetic bacteria. Corundum is a minor accessory mineral in metamorphic rocks and occurs in peraluminous igneous rocks. Partial solid solution is found between corundum and hematite. The gemstone ruby is  $\text{Al}_2\text{O}_3$  with minor  $\text{Cr}^{3+}$ , while sapphire is  $\text{Al}_2\text{O}_3$  with other chromophores. A modification of the corundum structure is found in the mineral ilmenite ( $\text{FeTiO}_3$ ). This is an important accessory mineral in felsic igneous rocks. Above  $950^\circ\text{C}$  ( $1740^\circ\text{F}$ ), there is complete solid solution between hematite and ilmenite. Bixbyite ( $\text{Mn}_2\text{O}_3$ ) is a distorted corundum structure resulting from the Jahn-Teller effect in  $\text{Mn}^{3+}$ . See BANDED IRON FORMATION; IGNEOUS ROCKS; IRON; METAMORPHIC ROCKS; RUBY; SAPPHIRE.

*XO<sub>2</sub> oxides.* Tetravalent cations, such as  $\text{Ti}^{4+}$ ,  $\text{Sn}^{4+}$ , and  $\text{Mn}^{4+}$ , whose ionic radii favor sixfold coordination, adopt the rutile structure (Fig. 3). Rutile ( $\text{TiO}_2$ ) is a common accessory mineral in felsic igneous

rocks, gneisses, and schists. It also has two low-temperature polymorphs, anatase and brookite, but these are less common. Cassiterite ( $\text{SnO}_2$ ) is the only significant ore mineral of tin. Cassiterite occurs mostly in granite-hosted hydrothermal deposits such as those in Cornwall, England. Pyrolusite ( $\beta\text{-MnO}_2$ ) is found in low-temperature hydrothermal deposits. It is less common than previously supposed.  $\text{MnO}_2$  also forms another polymorph (ramsdellite,  $\alpha\text{-MnO}_2$ ) based on double chains of  $\text{MnO}_6$  polyhedra. This has a structure similar to that of goethite ( $\alpha\text{-FeOOH}$ ). Like pyrolusite, ramsdellite forms in low-temperature hydrothermal deposits. To complicate matters, a phase called nsutite ( $\gamma\text{-MnO}_2$ ) is a disordered intergrowth of pyrolusite and ramsdellite that forms by the oxidation of manganese carbonate minerals. Synthetic nsutite is used in dry-cell batteries. See GNEISS; GRANITE; MANGANESE; SCHIST; TIN.

Large tetravalent cations, such as  $\text{U}^{4+}$  and  $\text{Th}^{4+}$ , prefer to be in eightfold coordination with oxygen and form oxides with the fluorite structure (Fig. 4). Uraninite ( $\text{UO}_2$ ) is the most important ore of uranium and is a primary mineral in granites. See RADIOACTIVE MINERALS; THORIUM; URANIUM.

*Spinel oxides.* The spinel structure (Fig. 5) is adopted by oxides with the formula  $\text{X}^{2+}\text{Y}^{3+}_2\text{O}_4$ . The spinel structure has one tetrahedral cation site and two octahedral cation sites per four oxygens. In a normal spinel, the tetrahedral site is occupied by a divalent cation such as  $\text{Mg}^{2+}$ ,  $\text{Fe}^{4+}$ , while the octahedral sites are occupied by trivalent cations such as  $\text{Fe}^{3+}$ ,  $\text{Cr}^{3+}$ , or  $\text{Al}^{3+}$ . The inverse spinel structure is a variation where the tetrahedral sites are occupied by trivalent cations and the octahedral sites are occupied by a mixture of divalent and trivalent cations. A variety of solid solutions are possible within the spinel structure oxides.

The most important spinel structure oxide is magnetite ( $\text{Fe}_3\text{O}_4$ ). Magnetite is an inverse spinel, so half of the  $\text{Fe}^{3+}$  cations are in the tetrahedral sites and the remaining  $\text{Fe}^{3+}$  cations, along with the  $\text{Fe}^{2+}$

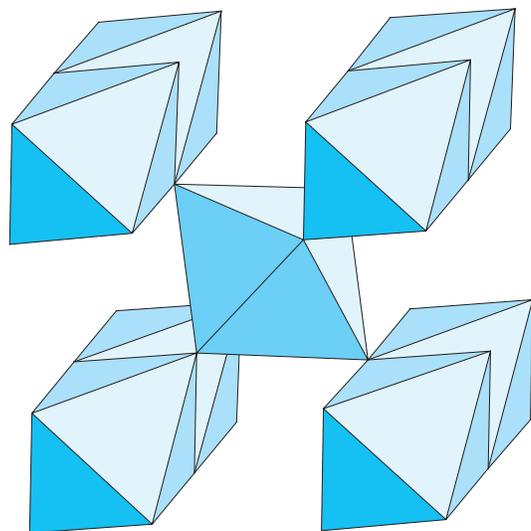


Fig. 3. Rutile structure adopted by  $\text{XO}_2$  oxides.

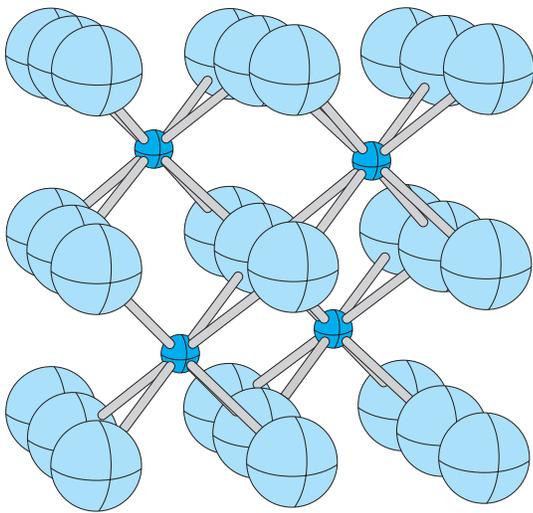


Fig. 4. Fluorite structure adopted by uraninite ( $\text{UO}_2$ ) and thorianite ( $\text{ThO}_2$ ).

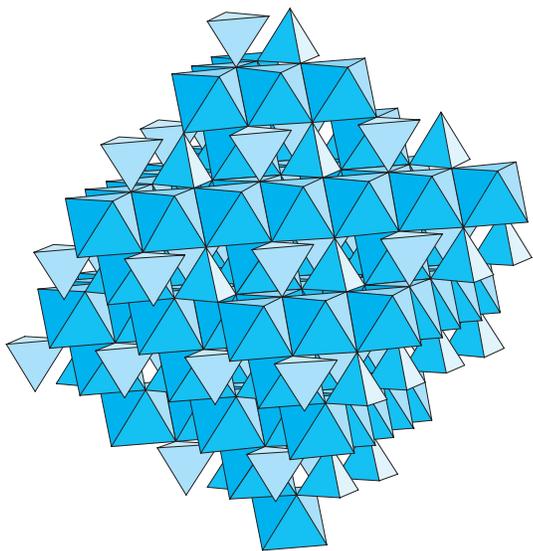


Fig. 5. Spinel structure adopted by  $\text{XY}_2\text{O}_4$  oxides. In the normal spinel structure (for example,  $\text{MgAl}_2\text{O}_4$ ), tetrahedral sites are occupied by divalent cations such as  $\text{Mg}^{2+}$  and octahedral sites are occupied by trivalent cations such as  $\text{Al}^{3+}$ . The inverse-spinel structure (for example,  $\text{Fe}_3\text{O}_4$ ) has trivalent cations in the octahedral sites and a mixture of divalent and trivalent cations in the octahedral sites.

cations, are in the octahedral sites. Electron hopping between  $\text{Fe}^{2+}$  and  $\text{Fe}^{3+}$  cations in the octahedral sites gives magnetite a high electrical conductivity. The most important geophysical property of magnetite is its ferrimagnetism, with a Néel temperature, the temperature at which an antiferromagnetic material becomes paramagnetic, of  $525^\circ\text{C}$  ( $980^\circ\text{F}$ ). As an igneous rock cools, the magnetic moments of individual magnetite domains align with the Earth's magnetic field. This preserves a record of the orientation of the rock relative to the Earth's magnetic field at the time of crystallization. These paleomagnetic signatures in rocks were used to confirm the hypothesis of sea-floor spreading and continental drift. Magnetite often contains significant amounts of other cations such as  $\text{Cr}^{3+}$  and  $\text{Ti}^{4+}$ . A complete solid solution be-

TABLE 2. Manganese(III,IV) oxide minerals

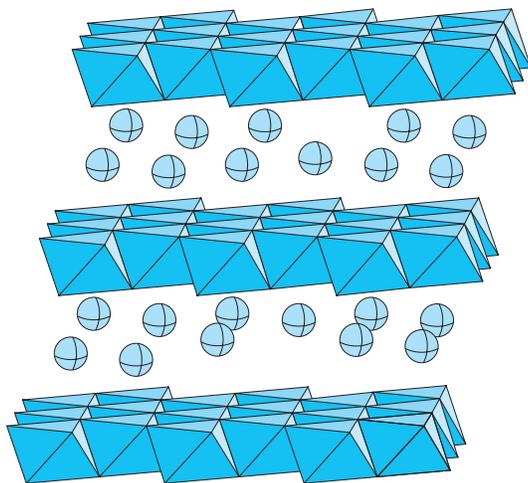
Mineral	Formula
<b>Chain Structures</b>	
Pyrolusite	$\text{MnO}_2$
Ramsdellite	$\text{MnO}_2$
<b>Tunnel Structures</b>	
Hollandite	$\text{BaMn}_8\text{O}_{16}$
Romanchite	$(\text{Ba}, \text{H}_2\text{O})_2\text{Mn}_5\text{O}_{10}$
Cryptomelane	$\text{K}_2\text{Mn}_8\text{O}_{16}$
Coronadite	$\text{PbMn}_8\text{O}_{16}$
Todorokite	$(\text{Na}, \text{Ca})_{0.5}(\text{Mn}, \text{Mg})_6\text{O}_{12} \cdot n\text{H}_2\text{O}$
<b>Layer Structures</b>	
Birnessite/Vernadite	$\text{Na}_{0.6}\text{Mn}_2\text{O}_4 \cdot 1.5\text{H}_2\text{O}$
Chalcophanite	$\text{ZnMn}_3\text{O}_7 \cdot 3\text{H}_2\text{O}$
Lithiophorite	$(\text{Li}, \text{Al})(\text{Mn}^{4+}, \text{Mn}^{3+})\text{O}_2(\text{OH})_2$

tween  $\text{Fe}_3\text{O}_4$  and  $\text{Fe}_2\text{TiO}_4$  (ulvospinel) is stable above  $600^\circ\text{C}$  ( $1100^\circ\text{F}$ ). See ANTIFERROMAGNETISM; CHROMIUM; FERRIMAGNETISM; GEOMAGNETISM; MAGNETIC SUSCEPTIBILITY; PALEOMAGNETISM; TITANIUM.

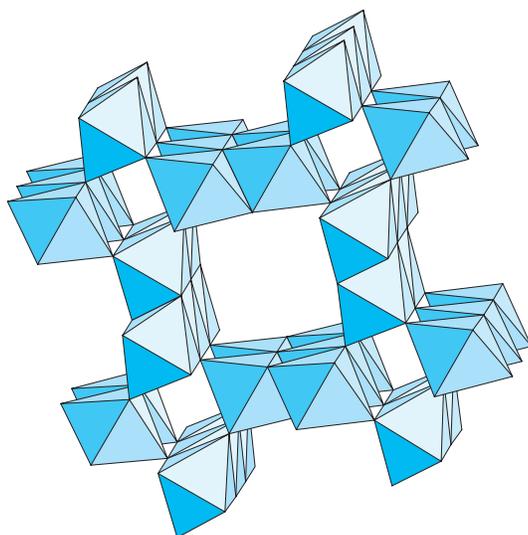
The structure of hausmannite ( $\text{Mn}_3\text{O}_4$ ), is a distortion of the spinel structure because of the Jahn-Teller effect for octahedrally coordinated  $\text{Mn}^{3+}$ . Hausmannite is found in high-temperature hydrothermal veins and in metamorphosed sedimentary manganese deposits but is not very common.

The spinel oxide, chromite ( $\text{FeCr}_2\text{O}_4$ ), is the dominant ore mineral of Cr. Chromite occurs in ultramafic rocks and in the serpentinites that are derived from them; significant ore deposits are found in Iran and Zimbabwe. Because of the high octahedral site preference energy of  $\text{Cr}^{3+}$ , chromite has a normal spinel structure. See COORDINATION CHEMISTRY; SERPENTINITE.

*Manganese(III, IV) oxides and oxide hydroxides.* Manganese hydroxides and oxides containing  $\text{Mn}^{4+}$  and  $\text{Mn}^{3+}$  form a variety of structures based on chains, tunnels, and sheets of  $\text{MnO}_6$  polyhedra (Table 2). The variations in the Mn oxidation state give variations in the charge of the  $\text{MnO}_6$  sheets and tunnel/chain frameworks. The layer and framework charges are compensated by the incorporation of cations (such as  $\text{K}^+$ ,  $\text{Ba}^{2+}$ , and  $\text{Pb}^{2+}$ ) in the interlayer and tunnel sites. Perhaps the most important example is birnessite (Fig. 6) which is a mixed-valence  $\text{Mn}^{4+}$ - $\text{Mn}^{3+}$  layer-structured oxide. This mineral, and the related phase vernadite (" $\delta$ - $\text{MnO}_2$ "; probably an incoherently stratified birnessite), are major phases in marine ferromanganese nodules and crusts which form on the sea floor. At least two structural modifications are present for birnessite due to the presence of cation vacancies in the sheets; ordering of the vacancies can lower the symmetry of the sheets from hexagonal to triclinic. In the interlayer, cations, such as Li, Al, and  $\text{Zn}^{2+}$ , will adsorb above the vacancy sites to give structures such as lithiophorite  $(\text{Li}, \text{Al})(\text{Mn}^{4+}, \text{Mn}^{3+})\text{O}_2(\text{OH})_2$  and chalcophanite  $(\text{Li}, \text{Al})(\text{Mn}^{4+}, \text{Mn}^{3+})\text{O}_2(\text{OH})_2$ . The former occurs in manganese nodules formed in acid soils. Chalcophanite is much less common and forms in the oxidized zone of Zn-Mn ore deposits (such as the



**Fig. 6.** Structure of birnessite. Between the MnO<sub>2</sub> layers are large exchangeable hydrated cations such as K<sup>+</sup>, Ca<sup>2+</sup>, Na<sup>+</sup>.



**Fig. 7.** The 2 × 2 tunnel structure adopted by hollandite and related manganese(IV, III) oxides. The tunnels can accommodate cations such as Ba<sup>2+</sup>, K<sup>+</sup>, Na<sup>+</sup>, and Pb<sup>2+</sup>. The same structure, but with Ca<sup>2+</sup> anions in the tunnel sites, is adopted by akaganeite (β-FeOOH).

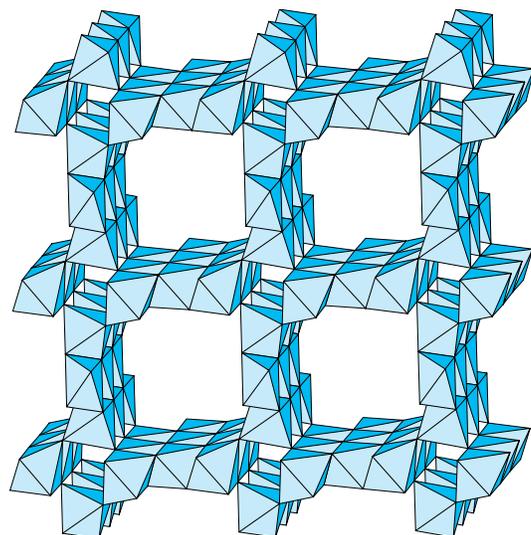
much-studied Franklin and Stirling Hill locality in New Jersey). See MANGANESE NODULES.

The simplest tunnel structure (Fig. 7) is based on double chains of MnO<sub>6</sub> polyhedra; this is adopted by hollandite and related phases with formula A<sub>0.2</sub>(Mn<sup>4+</sup>, Mn<sup>3+</sup>)<sub>8</sub>(O,OH)<sub>16</sub> (A = Ba, K, Pb, Na) [Table 2]. Todorokite, also found in marine manganese crusts and nodules, is a tunnel structure based on treble chains of MnO<sub>6</sub> polyhedra (Fig. 8). Incorporation of ions into manganese oxides must have important controls on the trace-element chemistry of the oceans.

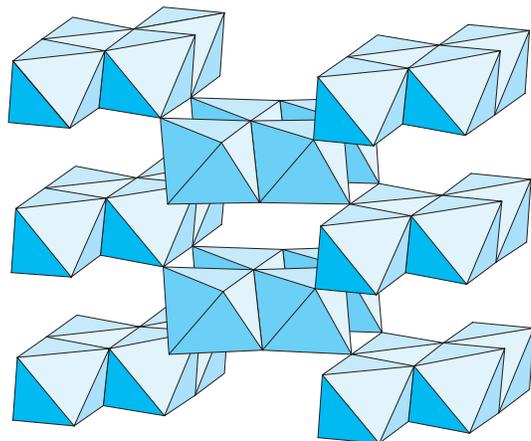
*Simple hydroxides.* The simplest hydroxide mineral is brucite [Mg(OH)<sub>2</sub>], the structure of which is based on Mg(OH)<sub>2</sub> layers that are held together by hydrogen bonds. Brucite forms during the alteration of magnesium silicates by hydrothermal fluids. It is not very common. The brucite structure also is adopted by hydroxides of other divalent cations such as Ca<sup>2+</sup>,

Fe<sup>2+</sup>, and Ni<sup>2+</sup>. The end member Fe(OH)<sub>2</sub> is easily partially oxidized, and the charge balance is maintained by the incorporation of various anions (such as CO<sub>3</sub><sup>2-</sup>, SO<sub>4</sub><sup>2-</sup>, and Cl<sup>-</sup>) between the (Fe<sup>2+</sup>, Fe<sup>3+</sup>) (OH)<sub>2</sub> layers to give “green rusts.” These minerals are probably important phases in subsonic and anoxic soils and sediments. The mineral gibbsite [Al(OH)<sub>3</sub>] has a structure based on layers of Al(OH)<sub>6</sub> octahedral with the layers held together by hydrogen bonds. See HYDROGEN BOND.

*MOOH oxide hydroxides and related minerals.* Trivalent cations, such as Fe<sup>3+</sup> and Al<sup>3+</sup>, form several oxide hydroxide structures. These minerals usually occur as clay-sized (<2 μm) particles in soils and sediments. Colloidal particles of oxide hydroxide minerals also are suspended in most natural waters. The surfaces of these minerals are quite reactive and strongly adsorb ions from aqueous solutions. In the environment, the aqueous concentrations of many trace micronutrients and toxic heavy metals are probably controlled by adsorption onto iron oxide hydroxide mineral



**Fig. 8.** Structure of todorokite. Within the 3 × 3 tunnels, exchangeable hydrated cations such as K<sup>+</sup>, Mg<sup>2+</sup>, and Ba<sup>2+</sup> are present.



**Fig. 9.** Structure of goethite (α-FeOOH) and diaspore (α-AlOOH). The MnO<sub>2</sub> polymorph ramsdellite (α-MnO<sub>2</sub>) also adopts this structure.

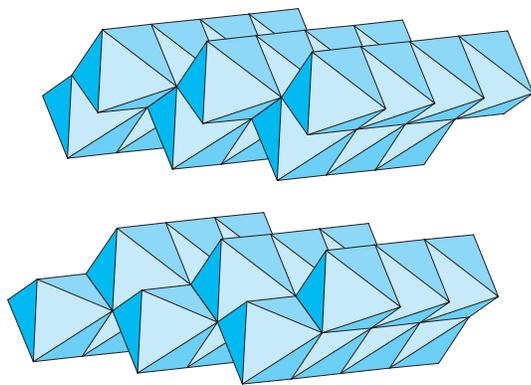


Fig. 10. Structure of lepidocrocite ( $\gamma$ -FeOOH) and boehmite ( $\gamma$ -AlOOH).

TABLE 3. MOOH and related minerals

Mineral	Formula
Goethite	$\alpha$ -FeOOH
Diaspore	$\alpha$ -AlOOH
Groutite	$\alpha$ -MnOOH
Lepidocrocite	$\gamma$ -FeOOH
Boehmite	$\gamma$ -AlOOH
Manganite	$\gamma$ -MnOOH
Akaganeite	" $\beta$ -FeOOH" = FeOOH · HCl
Schwertmannite	FeOOH · H <sub>2</sub> SO <sub>4</sub>
Ferrihydrite	FeOOH—Fe <sub>2</sub> O <sub>3</sub>

surfaces. The most common FeOOH phases are goethite ( $\alpha$ -FeOOH) and lepidocrocite ( $\gamma$ -FeOOH) [Figs. 9 and 10]. Goethite tends to form by hydrolysis of dissolved Fe<sup>3+</sup>, while lepidocrocite forms by oxidation of green rust. The aluminum analogues, diaspore ( $\alpha$ -AlOOH) and boehmite ( $\gamma$ -AlOOH), along with gibbsite (Al(OH)<sub>3</sub>) are the minerals that make up bauxite, the ore of Al formed by weathering of primary silicate minerals such as feldspar in tropical soils. There is only limited solid solution between the FeOOH and AlOOH (or MnOOH) phases. Several other FeOOH minerals are known, but they are less common. Akaganeite (" $\beta$ -FeOOH") forms by the hydrolysis of Fe<sup>3+</sup> in chloride-bearing solutions and is a minor phase in marine sediments. Its structure is similar to that of hollandite but the tunnels are occupied by Cl<sup>-</sup> anions. The mineral schwertmannite is believed to be similar to akaganeite, but with SO<sub>4</sub><sup>2-</sup> anions occupying the tunnels. This mineral forms in acid mine drainage, probably by bacterially mediated oxidation of dissolved Fe<sup>2+</sup>.

Perhaps the most ubiquitous FeOOH type mineral is ferrihydrite. This phase is poorly crystalline and forms by the rapid hydrolysis of dissolved Fe<sup>3+</sup>. This is facilitated by bacterial oxidation of dissolved Fe<sup>2+</sup> under less acidic conditions than those favoring schwertmannite. With time, ferrihydrite dissolves and recrystallizes to form the more stable phases goethite and hematite. Nevertheless, the extremely high, reactive surface area of ferrihydrite (up to 600 m<sup>2</sup>/gram) means that it can have a strong effect on aqueous geochemistry by sorbing dissolved ions.

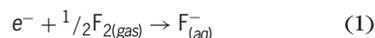
David M. Sherman

Bibliography. D. H. Lindsey (ed.), *Oxide Minerals*, Mineralogical Society of America, 1991; J. E. Post, Manganese oxide minerals: Crystal structures and economic and environmental significance, *Proc. Nat. Acad. Sci.*, 96:3447–3454, 1999; U. Schwertmann and R. M. Cornell, *Iron Oxides in the Laboratory: Preparation and Characterization*, 2d ed., Wiley-VCH, Weinheim, 2000.

## Oxidizing agent

A participant in a chemical reaction that absorbs electrons from another reactant. In the process a component atom of this substance undergoes a decrease in oxidation number. In this action as an oxidizing agent, the substance undergoes reduction.

A measure of the effectiveness of a reagent as an oxidizing agent is its reduction potential. This is, in electrochemical terms, the equivalent of the free-energy change for the reduction process. The element with the highest reduction potential (and, therefore, the strongest oxidizing agent) is fluorine, F<sub>2</sub>. The half-reaction in which fluorine absorbs an electron from another species in aqueous solution is shown by reaction (1). The reduction po-

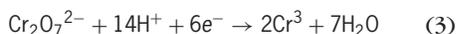


tential  $E$ , although strictly applicable only to thermodynamically reversible systems, is a very useful measure of the tendency of a reaction component to undergo oxidation or reduction. For the F<sub>2</sub>-F<sup>-</sup> pair, it is given at 25°C (77°F) by Eq. (2),

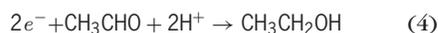
$$E = 2.87 - 0.1362 \ln \frac{a_{F^-}}{p_{F_2}^{1/2}} \text{ V} \quad (2)$$

where  $a_{F^-}$  is the activity of fluoride ion in the solution, and  $p$  is the pressure, in atmospheres of fluorine gas. Under standard conditions,  $a_{F^-} = 1$  and  $p_{F_2} = 1$ , the value for  $E$  is 2.87 V. This is  $E^\circ$ , the standard reduction potential at 25°C (77°F). The F<sub>2</sub>-F<sup>-</sup> pair constitute a reduction-oxidation couple.

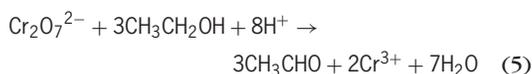
Thermodynamically, a substance is capable of oxidizing another if the corresponding couple has a lower reduction potential. When dichromate ion in acid solution acts as an oxidizing agent, it undergoes the reduction shown in reaction (3). The standard



reduction potential at 25°C (77°F) for this reaction is 1.33 V. As this reaction does not proceed reversibly, the standard reduction potential represents only a lower limit. For the reduction of acetaldehyde to ethanol, shown in reaction (4),  $E^\circ$  is 0.217 V. Dichro-

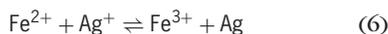


mate can, therefore, oxidize ethanol according to reaction (5). The corresponding electrochemical



potential is 1.11 V. The positive value for the resulting potential indicates a spontaneous process. Actually, in this example, the oxidation by dichromate will go beyond the aldehyde stage to produce the carboxylic acid. Dichromate ion in acid solution is a common oxidizing agent used in organic and analytical chemistry.

In oxidation-reduction (redox) systems composed of couples that are not widely separated, the actual reaction that occurs will depend upon the chemical activities in the system. For example, standard potentials for the  $\text{Fe}^{3+}-\text{Fe}^{2+}$  and  $\text{Ag}^+-\text{Ag}$  couples are 0.771 and 0.799 V, respectively. For the redox system shown in reaction (6), the resultant potential at 25°C (77°F) is given by Eq. (7). When the activity quotient



$$E = 0.028 - 0.1362 \ln \frac{a_{\text{Fe}^{3+}}}{a_{\text{Fe}^{2+}} a_{\text{Ag}^+}} \text{ V} \quad (7)$$

$Q = (a_{\text{Fe}^{3+}} + a_{\text{Fe}^{2+}} + a_{\text{Ag}^+}) < 1.23$ , chemical change will occur from left to right; with  $Q > 1.23$ , from right to left; and with  $Q = 1.23$ , the system is at equilibrium and no net chemical change will occur. The above considerations are based only on thermodynamics and allow no prediction concerning the actual rate of the oxidation-reduction process. The practical effectiveness of a given oxidizing (or reducing) agent will depend upon both the thermodynamics and the available kinetic pathway for the reaction process. See CHEMICAL THERMODYNAMICS.

Substances that are widely used as oxidizing agents in chemistry include ozone ( $\text{O}_3$ ), permanganate ion ( $\text{MnO}_4^-$ ), nitric acid ( $\text{HNO}_3$ ), as well as oxygen itself. Oxychlorine trifluoride,  $\text{OF}_2\text{Cl}$ , has been suggested and synthesized for possible use as an oxidizer in rocket engines. Organic chemists have empirically developed combinations of reagents to carry out specific oxidation steps in synthetic processes. Many of these utilize transition-metal oxides such as chromium trioxide,  $\text{CrO}_3$ , vanadium pentoxide,  $\text{V}_2\text{O}_5$ , ruthenium tetroxide,  $\text{RuO}_4$ , and osmium tetroxide,  $\text{OsO}_4$ . Oxidation by these species or by specific agents such as sodium perborate,  $\text{NaIO}_4$ , can often be restricted to a single molecular site.

The action of molecular oxygen as an oxidizing agent may be made more specific by photochemical excitation to an excited singlet electronic state. This can be accomplished through energy transfer from a dye, such as fluorescein, which absorbs light and transfers it to oxygen to form a relatively long-lived excited state. See PHOTOCHEMISTRY.

In an electrolytic cell, oxidation occurs at the anode, and the designation of a chemical oxidizing agent in the reaction is in general not possible. The electrode acts as an intermediary in the electron transfer. An important example of an anodic oxidation is the Kolbe reaction in which a carboxylic acid anion transfers an electron to the anode, forming a free radical [reaction (8)]. The radicals combine to form a hydrocarbon and  $\text{CO}_2$ , reaction (9).



See ELECTROLYSIS.

Enzymes in living organisms catalyze the exothermic oxidation of carbohydrate, fat, and protein material as a source of energy for life processes. Certain microorganisms are able to oxidize organic compounds at C-H bonds to produce alcohols. These have been suggested as of possible use in cleaning oil spills. See BIOLOGICAL OXIDATION; OXIDATION-REDUCTION.

F. J. Johnson

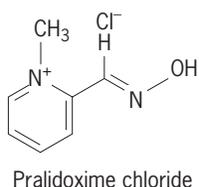
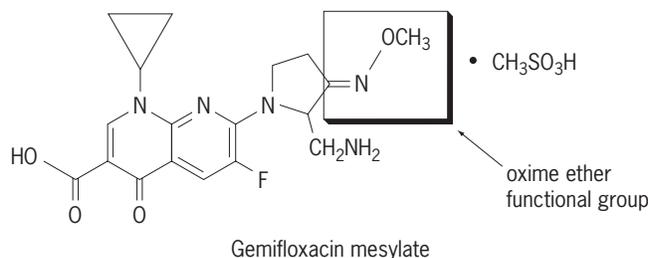
Bibliography. W. R. Adams, *Photosensitized Oxidations*, 1971; D. Benson, *Mechanisms of Oxidation by Metal Ions*, 1976; G. S. Fonken and R. A. Johnson, *Chemical Oxidations with Microorganisms*, 1972; T. E. King, H. S. Mason, and M. Morrison (eds.), *Oxidases and Related Redox Systems*, Advances in the Biosciences Series, vols. 33 and 34, 1982; M. Murari, *The Mechanism of Oxidation of Organic Compounds by Electronic Oxidants*, 1985.

## Oxime

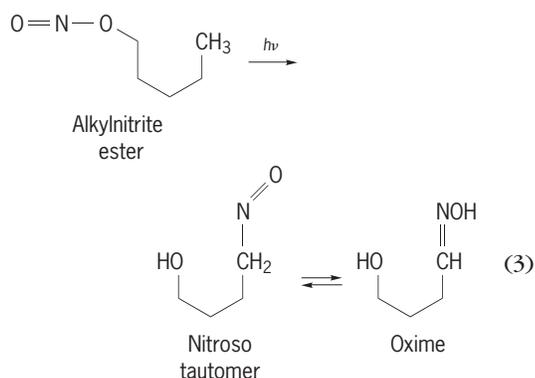
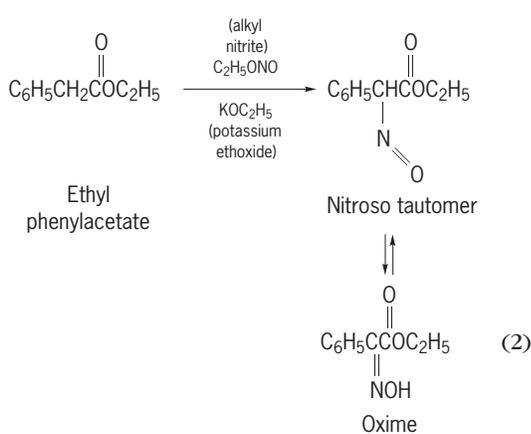
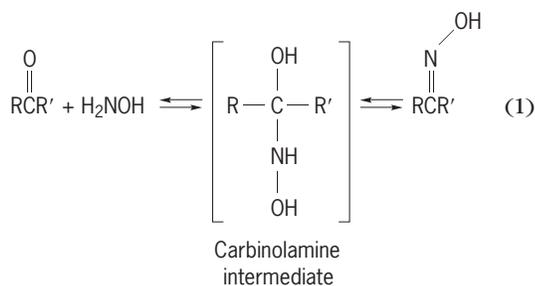
One of a group of chemical substances with the general formula  $\text{RR}'\text{C}=\text{N}-\text{OH}$ , where R and R' represent any carbon group or hydrogen. Oximes are derived from aldehydes (aldoximes,  $\text{RHC}=\text{NOH}$ ) and ketones (ketoximes,  $\text{RR}'\text{C}=\text{NOH}$ , where R and R' are not hydrogen). Oximes and oxime ethers ( $\text{RR}'\text{C}=\text{N}-\text{OR}''$ ) have important pharmaceutical and synthetic applications. The oxime (and oxime ether) functional group is incorporated into many organic medicinal agents, including some antibiotics, for example, gemifloxacin mesylate (see structure); and pralidoxime chloride (see structure) and obidoxime chloride are used in the treatment of poisoning by organophosphate insecticides malathion and diazinon. See ALDEHYDE; KETONE.

Hydroxylamine ( $\text{H}_2\text{NOH}$ ) reacts readily with aldehydes or ketones to give oximes. The rate of the reaction of hydroxylamine with acetone is greatest at pH 4.5. Oximes are formed by nucleophilic attack of hydroxylamine at the carbonyl carbon ( $\text{C}=\text{O}$ ) of an aldehyde or ketone to give an unstable carbinolamine intermediate [reaction (1)]. Since the breakdown of the carbinolamine intermediate to an oxime is acid-catalyzed, the rate of this step is enhanced at low pH. If the pH is too low, however, most of the hydroxylamine will be in the nonnucleophilic protonated form ( $\text{NH}_3\text{OH}^+$ ), and the rate of the first step will decrease. Thus, in oxime formation the pH has to be such that there is sufficient free hydroxylamine for the first step and enough acid so that dehydration of the carbinolamine is facile. See REACTIVE INTERMEDIATES.

Oximes can also be prepared by acid- or base-catalyzed oximation of active methylene compounds with alkyl nitrites [reaction (2)]. The nitroso compound that is first formed in this reaction rapidly rearranges to the more stable oxime tautomer. See TAUTOMERISM.

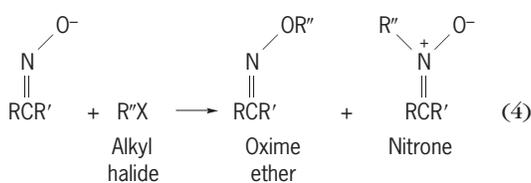


Photolysis of alkyl nitrite esters [reaction (3)] is useful in conformationally rigid systems, and leads to oxime formation at a  $\delta$ -carbon atom. Again, a nitroso tautomer rearranges to the oxime. This reaction



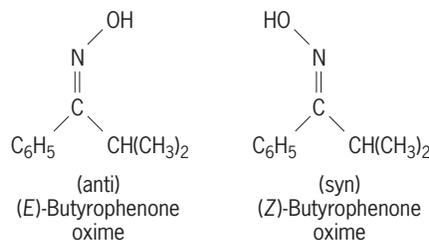
has been used to convert cortisone acetate into aldosterone acetate oxime. See NITRO AND NITROSO COMPOUNDS.

An oxime can act as both a weak acid and a weak base ( $\text{p}K_b \cong 12$  and  $\text{p}K_a \cong 11$ ). Oxime anions are capable of reacting at two different sites with alkyl halides ( $\text{R}'\text{X}$ ). Anions of this type are referred to as ambident (literally, two-sided tooth) nucleophiles. In the case of oxime anions, the reaction with an alkyl halide (an alkylation reaction) can give an oxime ether or a nitron [reaction (4)].



See ACID AND BASE.

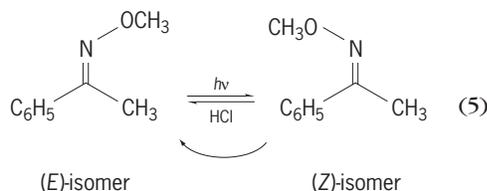
The *Z* and *E* isomers (formerly referred to as *syn* and *anti* isomers) of many oximes, such as butyrophenone oxime (see structure), are known. The *E*



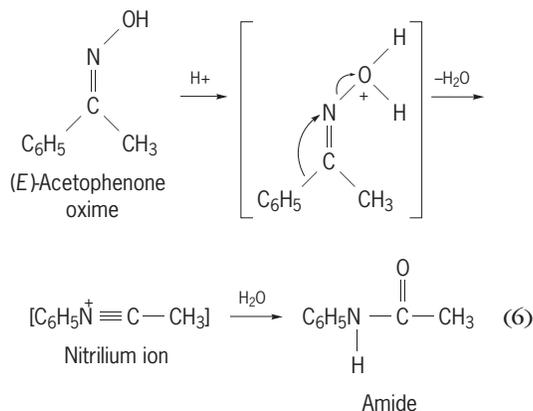
and *Z* isomers of oxime ethers ( $\text{RR}'\text{C}=\text{NOR}''$ ) are particularly resistant to thermal isomerization.

Many of the original (prior to 1921) assignments of configuration of the *Z* and *E* isomers were based on chemical reactions (the Beckman rearrangement) and were in error because of incorrect assumptions concerning the stereochemistry of these reactions. The assignment of configurations of oximes, when both isomers are available, can be made from  $^1\text{H}$  and  $^{13}\text{C}$  nuclear magnetic resonance spectra. See MOLECULAR ISOMERISM; STEREOCHEMISTRY.

The *O*-methyloxime prepared from reaction of acetophenone with hydroxylamine followed by alkylation with methyl iodide has the *E* configuration (OCH<sub>3</sub> and C<sub>6</sub>H<sub>5</sub> on opposite sides of the double bond) [reaction (5)]. Ultraviolet irradiation of benzene solution of the *E* isomer gives a mixture of the *E* and *Z* isomers from which the *Z* isomer is obtained by chromatography. Reversion to the more stable *E* isomer can be accomplished by acid-catalyzed isomerization (hydrogen chloride in dioxane) of the *Z* from [reaction (5)].



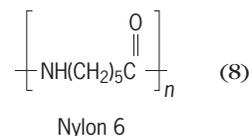
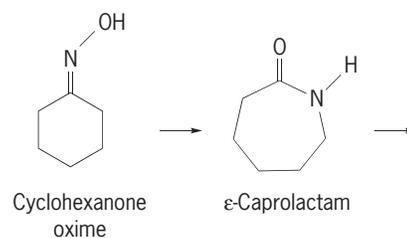
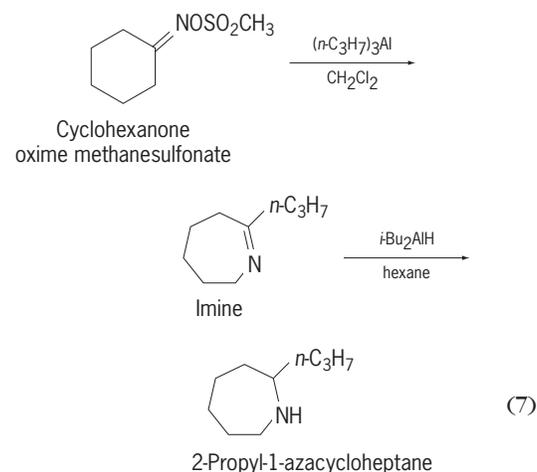
One of the best-known reactions of oximes is their rearrangement to amides. This reaction, the Beckmann rearrangement, can be carried out with a variety of reagents [such as phosphorus pentachloride (PCl<sub>5</sub>), concentrated sulfuric acid (H<sub>2</sub>SO<sub>4</sub>), and perchloric acid (HClO<sub>4</sub>)] that induce the rearrangement by converting the oxime hydroxyl group into a group of atoms that easily departs in a displacement reaction (a good leaving group) by either protonation or formation of a derivative. The Beckmann rearrangement has been shown to be a stereospecific process in which the group anti to the leaving group undergoes a 1, 2-shift to give a nitrilium ion. The nitrilium ion reacts with water to form the amide [reaction (6)].



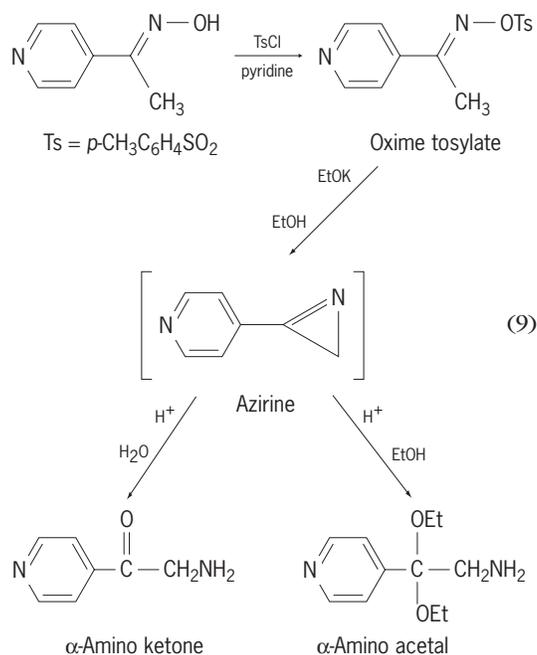
The Beckmann rearrangement can be carried out with a trialkylalane, which provides a Lewis acid for the rearrangement as well as a nucleophile to react with the intermediate nitrilium ion. For example, the tosylate of cyclohexanone oxime methanesulfonate reacts with tri-*n*-propylalane [(*n*-C<sub>3</sub>H<sub>7</sub>)<sub>3</sub>Al] to give an imine, which is reduced with diisobutylaluminum hydride to 2-propyl-1-azacycloheptane [reaction (7)].

The industrial synthesis of  $\epsilon$ -caprolactam is carried out by a Beckmann rearrangement on cyclohexanone oxime.  $\epsilon$ -Caprolactam is polymerized to the polyamide known as nylon 6, which is used in tire

cords [reaction (8)].

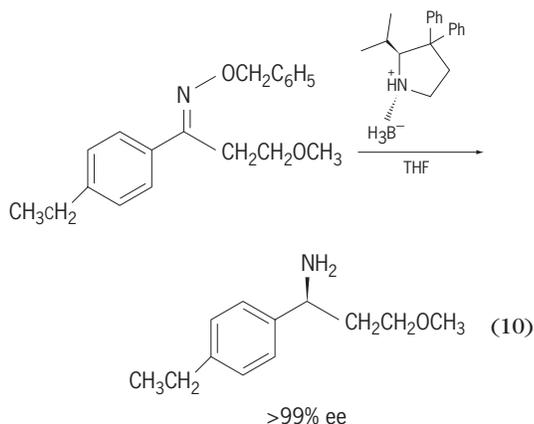


The conversion of oxime tosylates into  $\alpha$ -amino ketones is known as the Neber rearrangement [reaction (9)]. The intermediate azirine is usually not



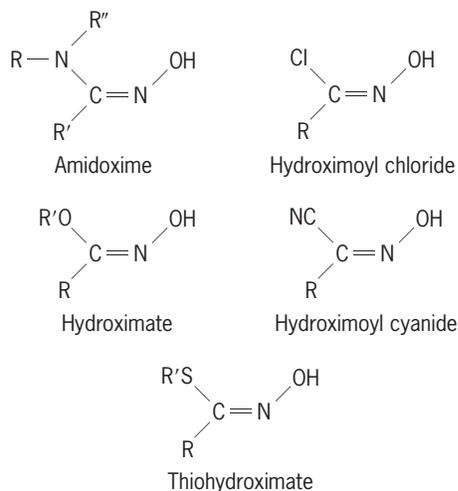
isolated and is reacted with aqueous acid to give the  $\alpha$ -amino ketone. Reaction of the azirine with acid in anhydrous ethanol gives an  $\alpha$ -amino acetal.

Aldoximes are dehydrated with acetic anhydride to nitriles ( $\text{RC}\equiv\text{N}$ ). Oxidation of oximes with trifluoroacetic acid at controlled pH gives nitro compounds ( $\text{RR}'\text{CHNO}_2$ ). Oximes can be reduced to amines with lithium tetrahydridoaluminate ( $\text{LiAlH}_4$ ) or by catalytic hydrogenation. Reductions of oxime ethers with a chiral oxazaborolidine give optically active amines with high enantioselectivity [reaction (10)]. Reduction of oximes with sodium cyanotrihy-



dridoborate ( $\text{NaBH}_3\text{CN}$ ) gives hydroxylamines. See HYDROGENATION.

There are many more complex functional groups that contain the oxime moiety, including:



See ORGANIC SYNTHESIS. James E. Johnson

Bibliography. B. R. Brown, *The Organic Chemistry of Aliphatic Nitrogen Compounds*, 1994; F. A. Carey and R. J. Sundberg, *Advanced Organic Chemistry, Part B: Reactions and Synthesis*, 4th ed., 2001; J. P. Freeman (ed.), *Organic Syntheses Collective Volume*, 8, pp. 568–572, 1993; J. P. Freeman (ed.), *Organic Syntheses Collective Volume* 7, pp. 149–152, 1990; R. E. Gawley, The Beckmann reaction: Rearrangements, elimination-additions, fragmentations, and rearrangement-cyclizations, *Org. React.*, 35:1-420, 1988; L. S. Hegedus (ed.), *Organic Syntheses*, 79:130–138, 2002; M. B. Smith and J. March,

*March's Advanced Organic Chemistry: Reactions, Mechanisms, and Structure*, 5th ed., 2001.

## Oximetry

Any of various methods for determining the oxygen saturation of blood. Oximeters are instruments that measure the saturation of blood in arteries by transilluminating intact tissue, thus eliminating the need to puncture vessels. Aside from the major advantage of being noninvasive and therefore free of risk, oximeters permit long-term monitoring in a variety of clinical settings, giving a continuous record of oxygen saturation.

**Blood color.** When dark venous blood passes through the pulmonary circulation or is exposed to air, it becomes bright red. In severe lung disease, the hemoglobin in the blood fails to become adequately oxygenated in its passage through the lungs and the mucous membranes lose their characteristic pink color. The mucous membranes then appear dusky or blue, a physical sign known as cyanosis. Oximeters determine oxygen saturation by making use of this color change; they compare photoelectrically the differences in light absorption between oxygenated and reduced hemoglobin. See HEMOGLOBIN.

**Light absorption.** The relationship between the concentration of a colored substance, length of the light path, and light absorption is expressed by the Lambert-Beer law; Eq. (1), where  $I$  is the amount

$$I = I_0 e^{-E'cd} \quad (1)$$

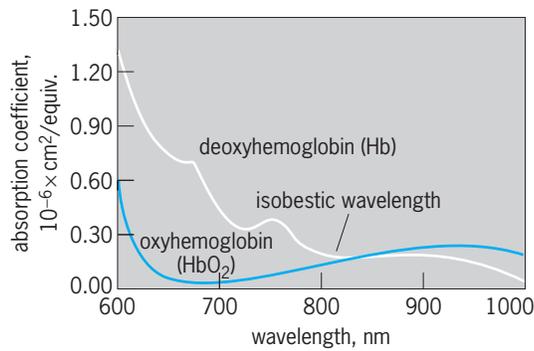
of transmitted light,  $I_0$  is the impinging light,  $c$  is concentration,  $d$  is the distance traveled by the light through the solution, and  $E'$  is a constant at a given wavelength of light for each substance, known as the extinction coefficient. The Lambert-Beer law may also be written in logarithmic form as Eq. (2) where

$$E = \frac{I}{cd} \log \frac{I_0}{I} \quad (2)$$

$\log(I_0/I)$  is referred to as optical density ( $D$ ). Thus, if incident light intensity and depth are held constant, the concentration of a substance in solution is a logarithmic function of the intensity of transmitted light.

As the wavelength of light increases from 300 to 1200 nanometers, the absorption of light by hemoglobin and oxyhemoglobin decreases, each following a slightly different pattern (**Fig. 1**). Between 600 and 800 nm, oxyhemoglobin absorbs less light than reduced hemoglobin, so that more red light is transmitted to produce the bright red color of oxygenated blood. In the infrared portion of the spectrum at 805 nm, there is an isobestic point, a wavelength at which reduced hemoglobin and oxygenated hemoglobin absorb identical amounts of light.

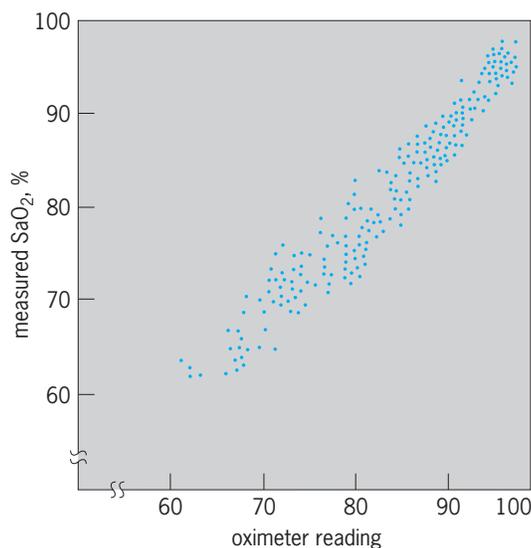
The Lambert-Beer relationship describes the absorption of monochromatic light passing at right angles through glass cuvettes containing homogeneous



**Fig. 1.** Absorption coefficients of oxyhemoglobin and deoxyhemoglobin in the red and near-infrared regions. (After S. Takatani, P. W. Cheung, and E. A. Ernst, *Ann. Biomed. Eng.*, 8:1–15, 1980)

solutions. While its validity has been demonstrated for hemoglobin solutions and hemolyzed blood in laboratory solutions, it may be inadequate to describe the more complex behavior of light in non-hemolyzed blood. Intact erythrocytes scatter, refract, and reflect incident light. When photometry is applied to hemoglobin in intact blood vessels, usually by transilluminating the helix or lobe of the ear, the situation is further complicated as the ear is an imperfect cuvette. Allowances must thus be made for absorption of light by tissue pigments such as melanin and bilirubin and for the variable depth of the cuvette (ear thickness) from subject to subject.

**Ear oximeters.** The original ear oximeters used a single wavelength of visible red light to monitor oxyhemoglobin levels. Later models used visible red and green light, allowing a distinction to be made between the effects of changes in hemoglobin concentration and saturation. The Millikan oximeter assessed the ear's blood thickness with an infrared cell,



**Fig. 2.** Comparisons of 223 samples of arterial blood  $O_2$  saturation ( $SaO_2$ ) using an ear oximeter. (After N. A. Saunders, A. C. P. Powles, and A. S. Rebeck, *Ear-oximetry: Accuracy and practicality in the assessment of arterial oxygenation*, *Amer. Rev. Resp. Dis.*, 113:745, 1976)

and arterial saturation with green light. The oximeter developed by Earl H. Wood developed the concept of making the ear “bloodless” by inflating a pressure capsule in the earpiece to create a “blank” cuvette, against which the perfused ear could be compared. In the 1970s, eight-wavelength oximeters were developed in which calibration was performed automatically against internal standards. These instruments were unaffected by variations in skin pigmentation and measured changes in arterial oxygen saturation almost immediately, continuously, and with considerable accuracy (Fig. 2). Unfortunately, the earpieces, as in earlier instruments, were cumbersome and uncomfortable to wear for prolonged periods of time.

Modern two-wavelength oximeters are known as pulse oximeters. These devices estimate oxygen saturation by the simple expedient of ignoring all non-pulsatile light absorbance. During the heart's systolic contractions, the inflow of arterial blood increases light absorbance across perfused tissues. Changes in incident light absorption from the nonflow baseline reflects absorbance by arterial blood. Incident light is emitted by two alternating light-emitting diodes—one in the visible red region (660 nm) and one in the infrared region (940 nm)—to determine the ratio of oxyhemoglobin to deoxyhemoglobin. Small light-emitting diodes are readily incorporated into a variety of probes; sites for transillumination include the earlobe, finger, toe, bridge of the nose and, in neonates, palms of the hand or foot. Pulse oximeters are regarded as accurate in the medically relevant range of 60–100% saturation.

**Applications.** Oximeters are used widely in modern hospital settings, where their ability to reflect oxygen saturation changes within 1–2 s can alert caregivers to the presence of life-threatening episodes of hypoxemia (inadequate oxygenation). In addition to monitoring the critically ill, oximeters are used routinely in the operating room to safeguard individuals having routine anesthetics. The growing awareness of sleep apnea and other disorders of breathing during sleep has led to widespread use of oximeters in sleep laboratories, where they are invaluable in the diagnostic process. The widespread use of accurate and continuous oxygen-monitoring devices has shown intermittent episodes of hypoxemia to occur much more frequently than was previously suspected. Guidelines for the prescribing of oxygen and other elements of care are being researched and revised in light of this wealth of new diagnostic information.

Kenneth R. Chapman; A. S. Rebeck  
**Bibliography.** K. R. Chapman, A. D'Urzo, and A. S. Rebeck, The accuracy and response characteristics of a simplified ear oximeter, *Chest*, 83(6):860–864, 1983; K. R. Chapman et al., Range of accuracy of two wavelength oximetry, *Chest*, 89:540–542, 1986; M. H. Kryger, T. Roth, and W. C. Dement (eds.), *Principles and Practice of Sleep Medicine*, 2d ed., 1994; J. W. Severinghaus and J. F. Kelleher, Recent developments in pulse oximetry, *Anesthesiology*, 76:1018–1038, 1992.

## Oxygen

A gaseous chemical element, O, atomic number 8, and atomic weight 15.9994. Oxygen is of great interest because it is the essential element both in the respiration process in most living cells and in combustion processes. It is the most abundant element in the Earth's crust. About one-fifth (by volume) of the air is oxygen. See PERIODIC TABLE.

1																	2
H																	He
3	4											5	6	7	8	9	10
Li	Be											B	C	N	O	F	Ne
11	12											13	14	15	16	17	18
Na	Mg											Al	Si	P	S	Cl	Ar
19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36
K	Ca	Sc	Ti	V	Cr	Mn	Fe	Co	Ni	Cu	Zn	Ga	Ge	As	Se	Br	Kr
37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54
Rb	Sr	Y	Zr	Nb	Mo	Tc	Ru	Rh	Pd	Ag	Cd	In	Sn	Sb	Te	I	Xe
55	56	71	72	73	74	75	76	77	78	79	80	81	82	83	84	85	86
Cs	Ba	Lu	Hf	Ta	W	Re	Os	Ir	Pt	Au	Hg	Tl	Pb	Bi	Po	At	Rn
87	88	103	104	105	106	107	108	109	110	111	112	113					
Fr	Ra	Lr	Rf	Db	Sg	Bh	Hs	Mt	Ds	Rg							
lanthanide series		57	58	59	60	61	62	63	64	65	66	67	68	69	70		
		La	Ce	Pr	Nd	Pm	Sm	Eu	Gd	Tb	Dy	Ho	Er	Tm	Yb		
actinide series		89	90	91	92	93	94	95	96	97	98	99	100	101	102		
		Ac	Th	Pa	U	Np	Pu	Am	Cm	Bk	Cf	Es	Fm	Md	No		

Oxygen is separated from air by liquefaction and fractional distillation. The chief uses of oxygen in order of their importance are (1) smelting, refining, and fabrication of steel and other metals; (2) manufacture of chemical products by controlled oxidation; (3) rocket propulsion; (4) biological life support and medicine; and (5) mining, production, and fabrication of stone and glass products.

Uncombined gaseous oxygen usually exists in the form of diatomic molecules, O<sub>2</sub>, but oxygen also exists in a unique triatomic form, O<sub>3</sub>, called ozone. See OZONE.

Under ordinary conditions oxygen is a colorless, odorless, and tasteless gas. It condenses to a pale blue liquid, in contrast to nitrogen, which is colorless in the liquid state. Oxygen is one of a small group of slightly paramagnetic gases, and it is the most paramagnetic of the group. Liquid oxygen is also slightly paramagnetic. Some data on oxygen and some properties of its ordinary form, O<sub>2</sub>, are listed in the table. See PARAMAGNETISM.

Properties of oxygen	
Property	Value
Atomic number	8
Atomic weight	15.9994
Triple point (solid, liquid, and gas in equilibrium)	-218.80°C (-139.33°F)
Boiling point at 1 atm pressure	-182.97°C (-119.4°F)
Gas density at °C and 10 <sup>5</sup> Pa pressure, g/liter	1.4290
Liquid density at the boiling point, g/ml	1.142
Solubility in water at 20°C, oxygen (STP) per 1000 g water at 10 <sup>5</sup> Pa partial pressure of oxygen	30

Practically all chemical elements except the inert gases form compounds with oxygen. Most elements form oxides when heated in an atmosphere containing oxygen gas. Many elements form more than one oxide; for example, sulfur forms sulfur dioxide (SO<sub>2</sub>) and sulfur trioxide (SO<sub>3</sub>). Among the most abundant binary oxygen compounds are water, H<sub>2</sub>O, and silica, SiO<sub>2</sub>, the latter being the chief ingredient of sand. Among compounds containing more than two elements, the most abundant are the silicates, which constitute most of the rocks and soil. Other widely occurring compounds are calcium carbonate (limestone and marble), calcium sulfate (gypsum), aluminum oxide (bauxite), and the various oxides of iron which are mined as a source of iron. Several other metals are also mined in the form of their oxides. Hydrogen peroxide, H<sub>2</sub>O<sub>2</sub>, is an interesting compound used extensively for bleaching. See HYDROGEN PEROXIDE; OXIDATION-REDUCTION; OXIDE; PEROXIDE; WATER.

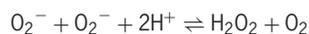
Arthur W. Francis, Sr.; Lawrence M. Sayre

Bibliography. C. E. Baukal (ed.), *Oxygen-Enhanced Combustion*, Springer, 1998; G. W. Everett and G. W. Everett, Jr., *Preparing and Studying Oxygen and Some of Its Compounds*, 1992; C. Foote et al. (eds.), *Active Oxygen in Chemistry*, Springer, 1996; D. MacKay et al., *Illustrated Handbook of Physical-Chemical Properties and Environmental Fate for Organic Chemicals: Oxygen, Nitrogen, and Sulfur Containing Compounds*, 1995; D. T. Sawyer, *Oxygen Chemistry*, 1991; J. S. Valentine et al. (eds.), *Active Oxygen in Biochemistry*, Springer, 1995.

## Oxygen toxicity

A toxic effect in a living organism caused by a species of oxygen. Oxygen has two aspects, one benign and the other malignant. Those organisms that avail themselves of the enormous metabolic advantages provided by dioxygen (O<sub>2</sub>) must defend themselves against its toxicity. The complete reduction of one molecule of O<sub>2</sub> to two of water (H<sub>2</sub>O) requires four electrons; therefore, intermediates must be encountered during the reduction of O<sub>2</sub> by the univalent pathway. The intermediates of O<sub>2</sub> reduction, in the order of their production are the superoxide radical (O<sub>2</sub><sup>-</sup>), hydrogen peroxide (H<sub>2</sub>O<sub>2</sub>), and the hydroxyl radical (HO·). See OXYGEN; SUPEROXIDE CHEMISTRY.

The intermediates of oxygen reduction, rather than O<sub>2</sub> itself, are probably the primary cause of oxygen toxicity. It follows that defensive measures must deal with these intermediates. The superoxide radical is eliminated by enzymes that catalyze the reaction below. These enzymes, known as superoxide



dismutases, have been isolated from a wide variety of living things; they have been found to contain iron, manganese, or both copper and zinc at their active sites. Elimination of superoxide dismutases from

oxygen-tolerant bacteria renders them intolerant of  $O_2$ , whereas exposure to elevated levels of  $O_2$  elicits an adaptive increase in the biosynthesis of these enzymes. Both of these responses suggest that  $O_2^-$  is a major cause of the toxicity of  $O_2$ .

Hydrogen peroxide ( $H_2O_2$ ) must also be eliminated, and this is achieved by two enzymatic mechanisms. The first of these is the dismutation of  $H_2O_2$  into water and oxygen, a process catalyzed by catalases. The second is the reduction of  $H_2O_2$  into two molecules of water at the expense of a variety of reductants, a process catalyzed by peroxidases. Plant peroxidases are hemoenzymes, that is, enzymes that contain heme as a prosthetic group; they can use ascorbate, phenols, or amines as the reductants of  $H_2O_2$ . A selenium-containing peroxidase that uses the tripeptide glutathione as the reductant of  $H_2O_2$  can be found in animal cells. See ENZYME.

The multiplicity of superoxide dismutases, catalases, and peroxidases, and the great catalytic efficiency of these enzymes, provides a formidable defense against  $O_2^-$  and  $H_2O_2$ . If these first two intermediates of  $O_2$  reduction are eliminated, the third ( $HO\cdot$ ) will not be produced. No defense is perfect, however, and some  $HO\cdot$  is produced; therefore its deleterious effects must be minimized. This is achieved to a large extent by antioxidants, which prevent free-radical chain reactions from propagating. For example, the human organism depends upon  $\alpha$ -tocopherol (vitamin E) to prevent such chain reactions within the hydrophobic core of membranes and depends upon ascorbate and glutathione to serve the same end in the aqueous compartments of cells. See ASCORBIC ACID; ANTIOXIDANT; CHAIN REACTION (CHEMISTRY); FREE RADICAL; PEPTIDE.

Some damage due to oxygen-derived free radicals is sustained on a continuing basis, in spite of the existing multilayered defenses, and must be repaired. Thus, there are enzymes that recognize and hydrolyze oxidatively damaged proteins, and there are other enzymes that repair oxidative damage to deoxyribonucleic acid (DNA). Indeed, analysis of human urine for oxidized pyrimidines, which are removed from DNA during such repair, indicates the occurrence of at least 1000 such events per cell per day.

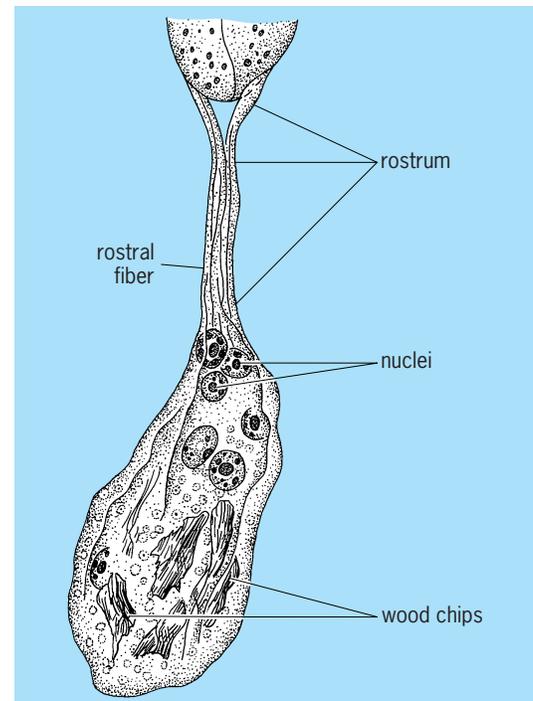
Thus, the apparent comfort in which aerobic organisms live in the presence of an atmosphere that is 20%  $O_2$  is due to a complex and effective system of defenses against this peculiar gas. Indeed, these defenses are easily overwhelmed, and overt symptoms of oxygen toxicity become apparent when organisms are exposed to 100%  $O_2$ . For example, a rat maintained in 100%  $O_2$  will die in 2 to 3 days.

Irwin Fridovich

Bibliography. I. B. Afanas'ev, *Superoxide Ion: Chemistry and Biological Implications*, 1989; I. Fridovich, Superoxide dismutases: An adaptation to a paramagnetic gas, *J. Biol. Chem.*, 264:7761-7764, 1989; H. Jonje, Genetic toxicology of oxygen, *Mutation Res.*, 219:193-208, 1989; A. Touati, Molecular genetics of superoxide dismutases, *Free Radical Biol. Med.*, 5:393-402, 1988.

## Oxymonadida

An order of class Zoomastigophorea in the phylum Protozoa. These are colorless flagellate symbionts in the digestive tract of the roach *Cryptocercus* and of certain termites. They are xylophagous; that is, they ingest wood particles taken in by the host. Seven or more genera of medium or large size have been identified, the organisms varying from pyriform to ovoid in shape. At the anterior end a pliable neck-like rostrum attaches the organism to the host intestinal wall, but they are sometimes free (see **illus.**). They can be either uni- or multinucleate. These organisms are termed karyomastigonts and each gives rise to two pairs of flagella in the unattached cells, two flagella to each blepharoplast. In the rostrum there is an axostylar apparatus, fibrils which pass to, and emerge at the posterior part of, the body. The nuclei contain long threadlike persistent chromosomes which appear to pass at random onto an intranuclear spindle. Isogamous (union of similar gametes) sexual processes have been described for three genera. In some cases, at least, these parallel the molting of the host.



An oxymonad, *Microrhopalodina inflata*.

A peculiar feature is the investment of the cell cuticle with bacteria and spirochetes, often resulting in a thick coating. The genera described have a global distribution. See CILIA AND FLAGELLA; PROTOZOA; SARCOMASTIGOPHORA; ZOOMASTIGOPHOREA.

James B. Lackey

Bibliography. L. R. Cleveland et al., The wood feeding roach *Cryptocercus*: Its protozoa and the symbiosis between protozoa and roach, *Mem. Amer. Acad. Arts Sci.*, 17:185-342, 1934; S. P. Parker (ed.), *Synopsis and Classification of Living Organisms*, 2 vols.,

1982; J. A. Pechenik, *Biology of the invertebrates*, 5th ed., McGraw-Hill Higher Education: Dubuque, 2005.

## Ozone

A powerfully oxidizing allotropic form of the element oxygen. The ozone molecule contains three atoms ( $O_3$ ), while the more common oxygen molecule has two atoms ( $O_2$ ).

Ordinary oxygen is a colorless gas and condenses to a very pale blue liquid, whereas ozone gas is decidedly blue, and both liquid and solid ozone are an opaque blue-black color, similar to that of ink. Even at concentrations as low as 4%, the blue color of ozone gas mixed with air or other colorless gas in a tube 1 in. (2.5 cm) or more in diameter and 4 ft (1.2 m) or more long can be seen by looking lengthwise through the tube.

**Properties and uses.** Some properties of ozone are given in the table. Ozone has a characteristic, pungent odor familiar to most persons because ozone is formed when an electrical apparatus produces sparks in air. Ozone is irritating to mucous membranes and toxic to human beings and lower animals. U.S. Occupational Safety and Health Administration standards for industrial workers exposed to ozone on a daily basis limit ozone concentration to 0.1 part per million on the average, with a maximum of 0.3 ppm for short exposures.

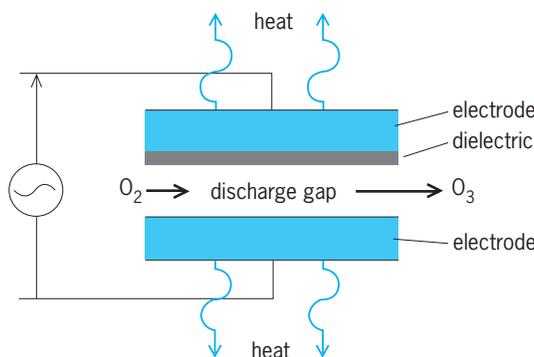
High ozone concentrations in liquid- and gas-phase mixtures can decompose explosively when initiated by an electric spark or other high-level energy source. Controlled decomposition to reduce ozone to desirable low concentrations can be accomplished catalytically.

Ozone is a more powerful oxidizing agent than oxygen, and oxidation with ozone takes place with evolution of more heat and usually starts at a lower temperature than when oxygen is used. In the presence of water, ozone is a powerful bleaching agent, acting more rapidly than hydrogen peroxide, chlorine, or sulfur dioxide. *See* OXIDIZING AGENT.

Ozone is utilized in the treatment of drinking-water supplies. Odor- and taste-producing hydrocarbons are effectively eliminated by ozone oxidation.

**Some properties of ozone**

Property	Value
Density of the gas at 0°C (32°F), 1 atm ( $10^5$ Pa) pressure	2.154 g/liter
Density of the liquid	
–111.9°C	1.354 g/ml
–183°C	1.573 g/ml
Boiling point at 1 atm ( $10^5$ Pa) pressure	–111.9°C (–169.4°F)
Melting point of the solid	–192.5°C (–314.5°F)
Wavelength range of maximum absorption in visible spectrum	560–620 nm
Wavelength range of maximum absorption in the ultraviolet spectrum	240–280 nm



**Fig. 1. Diagram of a generic corona cell.**

Iron and manganese compounds which discolor water are diminished by ozone treatment. Compared to chlorine, bacterial and viral disinfection with ozone is up to 5000 times more rapid. After treatment, the residual chlorine content leaves a characteristic undesirable taste and odor. In addition, chlorine may yield chloroform and other trihalomethane (THM) compounds which are potentially carcinogenic. *See* WATER TREATMENT.

Plants that use oxygen in aerobic digestion of sewage can add ozone treatment at reduced cost. Ozone can be produced more economically from pure oxygen. By proper integration of the facilities, oxygen not transformed into ozone in its generator passes through the ozonization tank into the aerobic digester with very high efficiency. *See* SEWAGE TREATMENT.

Ozone undergoes a characteristic reaction with unsaturated organic compounds in which the double or triple bond is attacked, even at temperatures as low as  $-150^{\circ}\text{F}$  ( $-100^{\circ}\text{C}$ ), with the formation of ozonides; these ozonides can be hydrolyzed, oxidized, reduced, or thermally decomposed to a variety of compounds, chiefly aldehydes, ketones, or carboxylic acids. Double ( $\text{C}=\text{C}$ ) bonds are almost always ruptured in this reaction. Commercially ozonolysis (ozonation followed by decomposition of the ozonide) is employed in the production of azelaic acid and certain chemical intermediates used in the drug industry. *See* OZONOLYSIS.

**Natural occurrence.** Ozone occurs to a variable extent in the Earth's atmosphere. Near the Earth's surface the concentration is usually 0.02–0.03 ppm in country air, and less in cities except when there is smog; under smog conditions in Los Angeles ozone is thought to be formed by the action of sunlight on oxygen of the air in the presence of impurities, and on bad days the ozone concentration may reach 0.5 ppm or more for short periods of time.

At vertical elevations above 12 mi (20 km), ozone is formed by photochemical action on atmospheric oxygen. Maximum concentration of  $5 \times 10^{12}$  molecules/ $\text{cm}^3$  (more than 1000 times the normal peak concentration at Earth's surface) occurs at an elevation of 19 mi (30 km).

Intercontinental air transports cruise at altitudes of 7.5 to 11 mi (12 to 17 km). On flights through northern latitudes, significant concentrations (up to

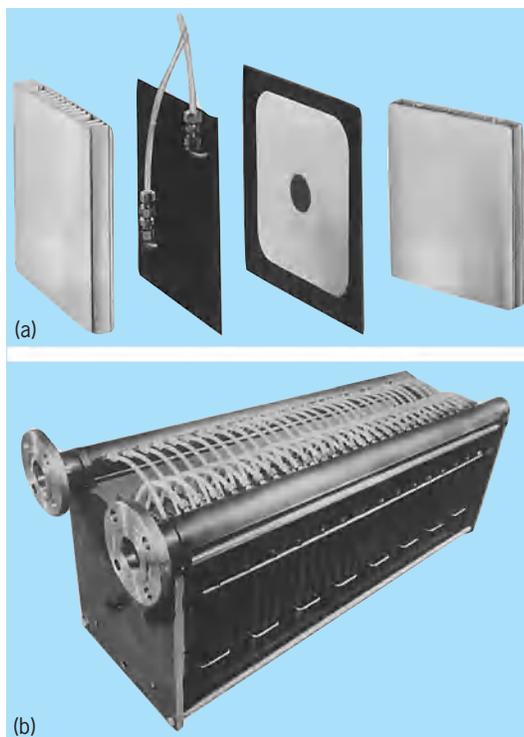


Fig. 2. Lowther cell for ozone generation: (a) expanded view of a single cell; (b) 30-cell module.

1.2 ppm) of ozone have been encountered. At these levels ozone can cause coughing and chest pains, especially for cabin attendants who are actively working. Carbon filters and catalytic ozone-decomposing equipment have been installed to eliminate the problem.

Absorption of solar ultraviolet radiation by ozone provides enough energy to raise the temperature of the stratosphere (6–30 mi or 10–50 km) significantly above that of the upper troposphere. This increase

of temperature with increasing height forms a stable layer resistant to vertical mixing. Gases injected into the stratosphere above 12 mi (20 km) may remain 2 years or longer.

By absorbing most of the short-wavelength light, the ozone layer protects human and other life forms. The layer is thinnest at the Equator, where it permits more ultraviolet radiation to reach ground levels in the torrid zone. This is believed to account for the high incidence of skin cancer in equatorial areas.

The dissociation of ozone to oxygen is catalyzed by several chemicals, especially nitrogen oxides and chlorine. Cosmic rays form nitric oxide in the stratosphere. As solar activity causes Earth's magnetic field to increase, cosmic rays are deflected away from Earth. Consequently, there is less nitric acid and more ozone immediately following the maximum phase of the solar activity cycle.

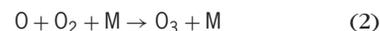
Volcanic eruptions and cosmic rays result in increased levels of chemicals which dissociate ozone. Above-normal levels of these natural events in previous geologic ages are believed to have reduced the ozone layer to 10% of normal. The resulting increase in ultraviolet radiation reaching the Earth's surface may account for the sudden extinction of some species.

Human activities also influence the ozone layer. Nuclear explosions in the atmosphere, and supersonic aircraft cruising at altitudes around 12 mi (20 km), inject nitric oxide into the stratosphere. A still larger effect may be developing from the release of certain relatively stable fluorocarbons, especially  $\text{CFCl}_3$  and  $\text{CF}_2\text{Cl}_2$ . This type of compound is believed to remain intact for many years in the atmosphere, permitting the gradual vertical transport from the surface into the stratosphere. Intense photochemical activity decomposes the fluorocarbon molecule, releasing chlorine atoms, each of which may destroy many ozone molecules. See FLUOROCARBON; STRATOSPHERIC OZONE.

**Preparation.** The only method used to make ozone commercially is to pass gaseous oxygen or air through a high-voltage, alternating-current electric discharge called a silent electric discharge. First, oxygen atoms are formed as in reaction (1). Some of



these oxygen atoms then attach themselves to oxygen molecules as in reaction (2). The excess energy



in the newly formed ozone is carried off by any available molecule (M) of gas, thus stabilizing the ozone molecule.

The corona discharge principle employed in all types of commercial ozone generators involves applying a high-voltage alternating current between two electrodes which are separated by a layer of dielectric material and a narrow gap through which the oxygen-bearing gas is passed (Fig. 1). The dielectric is necessary to stabilize the discharge over the entire electrode area, so that it does not localize as an intense arc.

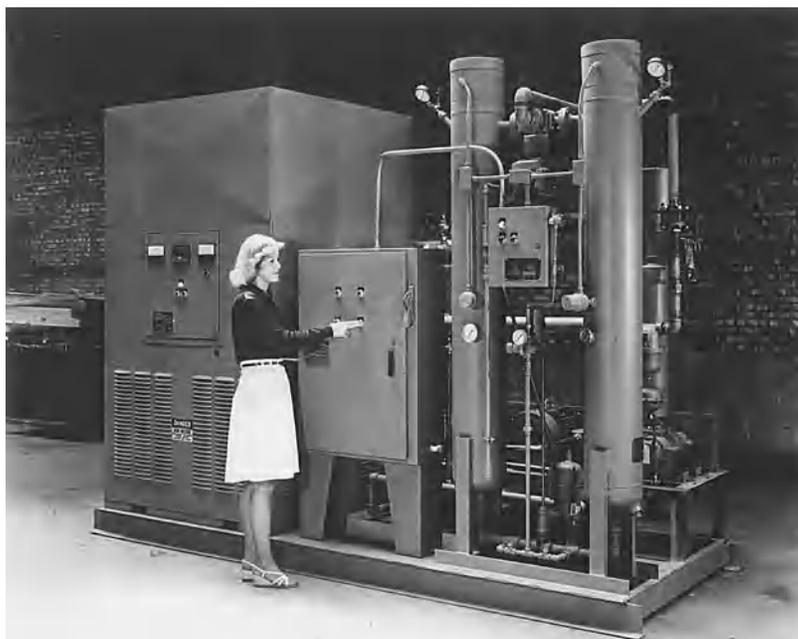


Fig. 3. Packaged ozone generator.

A substantial fraction of the electrical energy is converted to heat. The low volume of gas flowing between the electrodes does not have sufficient capacity to remove this heat. Some external heat sink is necessary, since the decomposition of ozone is accelerated by increasing temperature.

The Lowther cell (Fig. 2a) is an example of a modern, plate-type, air-cooled ozone generator. An individual cell is a gastight sandwich consisting of an aluminum heat dissipator, a steel electrode coated with a highly stable ceramic dielectric, a spacer to set the width of the discharge gap, a second ceramic-coated steel electrode with an oxygen inlet, and an ozone outlet passing through a second aluminum heat dissipator. Individual cells are stacked into 30-cell modules (Fig. 2b), which are grouped with power supplies and controls into packaged ozonators (Fig. 3). In the concentric-tube type the oxygen or air to be ozonized passes through the annular space (about 2–3 mm across) between two tubes, one of which must be made of a dielectric material, usually glass, and the other may be either glass or a metal which does not catalyze ozone decomposition, such as aluminum or stainless steel. The internal surface of the inner tube and the external surface of the outer tube, when made of glass, are in contact with an electrical conductor such as metal foil, an electrically conducting paint, or electrically conducting water; these conductors act as electrodes. Between 5000 and 50,000 V at a frequency between 50 and 10,000 Hz is then applied across the electrodes. In some commercial ozone generators the inner and outer tubes are both water-cooled. The latter represents, a simpler type of construction, but does not permit as high an input of electrical power as when both tubes are cooled.

The concentration of ozone in the gas stream leaving commercial ozone generators is usually 1–10% by weight. The yield of ozone is better when oxygen is used instead of air. Other factors which increase the yield of ozone in the silent electric discharge are thorough drying of the oxygen or air before it enters the ozonizer, refrigeration, increasing the pressure to a little above atmospheric, and increasing the frequency of the discharge.

A practical method has been developed for distribution of ozone in small quantities convenient for laboratory use. The ozone is dissolved in a liquefied fluorocarbon. The mixture is maintained at low temperature by a jacket filled with dry ice. Under these conditions, ozone decomposition proceeds very slowly, allowing sufficient time for transport to the user and a modest storage time. Rather high concentrations of ozone may be introduced safely into the laboratory in this manner.

**Analytical methods.** The analytical determination of ozone is usually carried out in the laboratory by bubbling the gas through a neutral solution of potassium iodide, acidifying the solution, and titrating the iodine thus liberated with standard sodium thiosulfate solution. Ozone in a gas stream may be determined automatically and continuously by passing the gas through a cell with transparent windows and measuring the absorption of either visible light

or of ultraviolet radiation beamed through the cell. See OXIDATION-REDUCTION; OXYGEN.

Arthur W. Francis

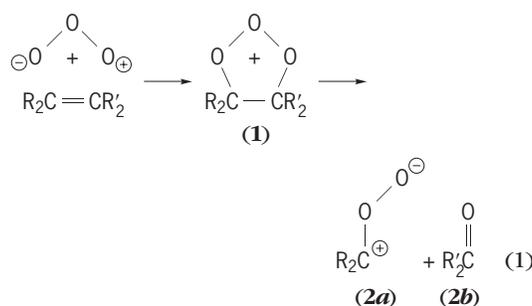
**Bibliography.** S. B. Majumdar and O. J. Sproul, Technical and economic aspects of water and wastewater ozonization, *Water Res.*, 8:253–260, May 1974; J. B. Murphy and J. R. Orr (eds.), *Ozone Chemical Technology*, 1975; National Academy of Science, *Protection Against Depletion of Stratospheric Ozone by Chlorofluorocarbons*, 1979; National Academy of Science, *Stratospheric Ozone Depletion by Halocarbons*, 1979.

## Ozonolysis

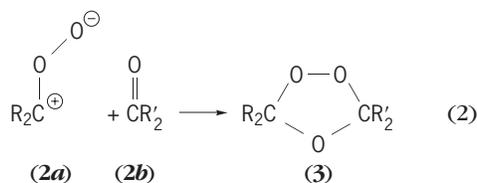
A process which uses ozone to cleave unsaturated organic bonds. General olefin cleavage was first extensively studied by C. Harries, beginning in 1903, as a technique to determine the structure of unsaturated compounds by identification of the cleavage products.

Generally, ozonolysis is conducted by bubbling ozone-rich oxygen or air into a solution of the reactant. The reaction is fast at moderate temperatures. Intermediates are usually not isolated but are subjected to further oxidizing conditions to produce acids or to reducing conditions to form alcohols or aldehydes. An unsymmetrical olefin is capable of yielding two different products whose structures are related to the groups substituted on the olefin and the position of the double bond.

The presently accepted mechanism of ozonolysis involves the initial formation of an unstable 1,2,3-trioxacyclopentane "primary ozonide" (1) by a 1,3-dipolar cycloaddition of ozone, as shown in one of its resonance structures, with an olefin, in reaction (1). Intermediate (1) readily decomposes to a zwitterion

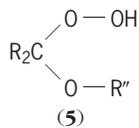
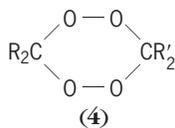


terion key intermediate, carbonyl oxide (2a), and a carbonyl (2b). An important reaction of intermediate (2a) is with ketones, for example (2b), to form a 1,2,4-trioxacyclopentane (3), called ozonide, as in reaction (2). The intermediate (2a) can also dimer-



ize to the diperoxide (4) or polymerize to polymeric ozonides and peroxides. The zwitterion produces

oxyperoxides of the general structure (5), where the



reaction media contains water ( $\text{R}'' = \text{H}$ ), ethanol ( $\text{R}'' = \text{OCH}_2\text{CH}_3$ ), or acetic acid ( $\text{R}'' = \text{OOCCH}_3$ ).

Before World War I, ozonolysis was applied commercially to the preparation of vanillin from isoeugenol. Today the only major application of the technique in the United States is in the manufacture of azelaic and pelargonic acids from oleic acid. See ALKENE; OZONE.

Robert K. Barnes

Bibliography. P. S. Bailey, *Ozonation in Organic Chemistry*, vol. 1, 1978; M. Horvath, L. Bilitzk, and J. Huttner, *Ozone*, 1985.



## Pacific islands — Paprika

### Pacific islands

A geographic designation that includes thousands of mainly small coral and volcanic islands scattered across the Pacific Ocean from Palau in the west to Easter Island in the east (see **illus.**). Island archipelagos off the coast of the Asian mainland, such as Japan, Philippines, and Indonesia, are not included even though they are located within the Pacific Basin. The large island constituting the mainland of Papua New Guinea and Irian Jaya is also excluded, along with the continent of Australia and the islands that make up Aotearoa or New Zealand. The latter, together with the Asian Pacific archipelagos, contain much larger landmasses, with a greater diversity of resources and ecosystems, than the oceanic islands, commonly labelled Melanesia, Micronesia, and Polynesia (see **illus.**). See AUSTRALIA; NEW ZEALAND; OCEANIC ISLANDS.

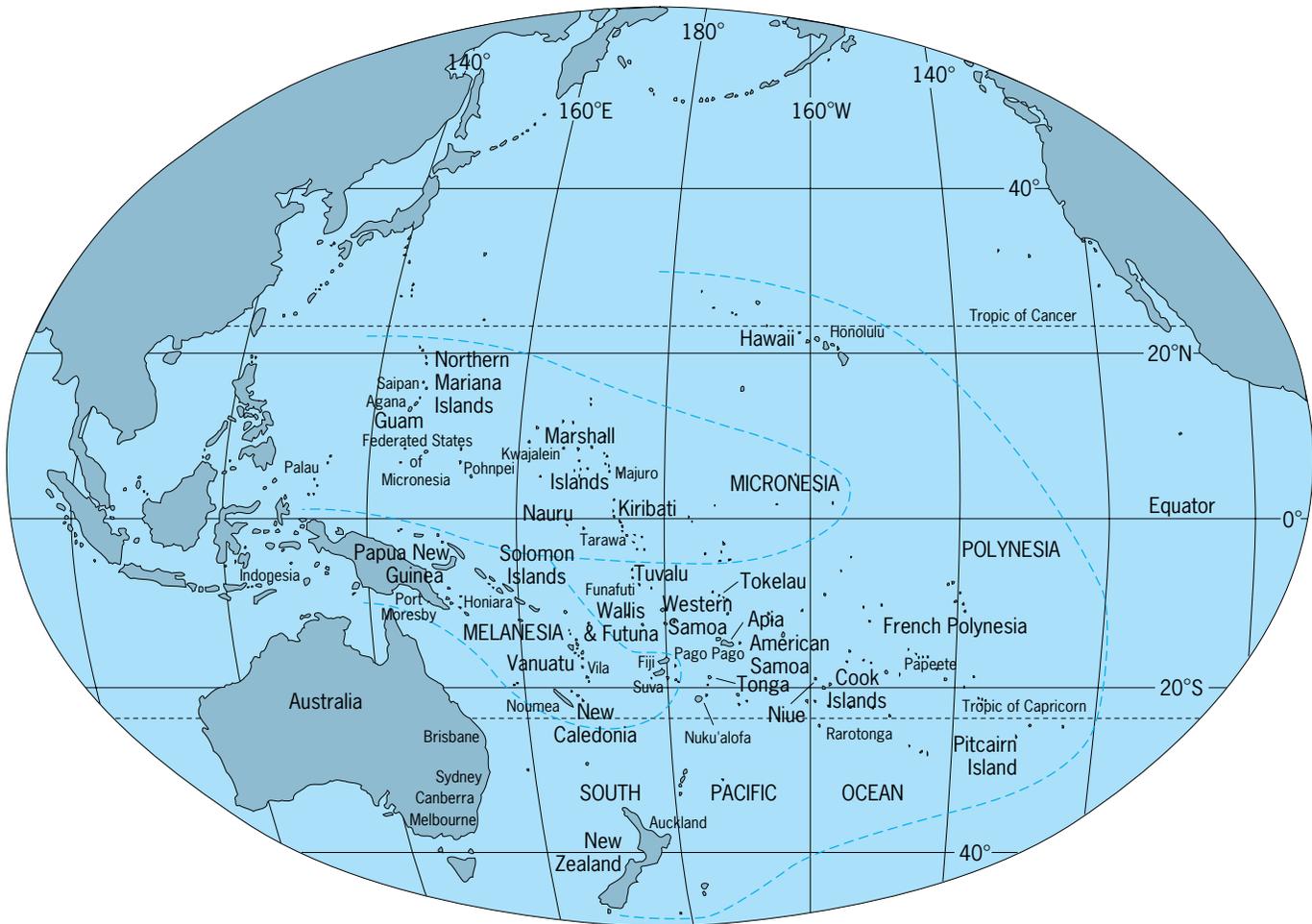
The great majority of these islands are between 4 and 4000 mi<sup>2</sup> (10 and 10,000 km<sup>2</sup>) in land surface area. The three largest islands include the main island of New Caledonia (6220 mi<sup>2</sup> or 16,100 km<sup>2</sup>), Viti Levu (4053 mi<sup>2</sup> or 10,497 km<sup>2</sup>) in Fiji, and Hawaii (4031 mi<sup>2</sup> or 10,440 km<sup>2</sup>) the big island in the Hawaiian chain. When the 80-mi (200-km) Exclusive Economic Zones are included in the calculation of surface area, some Pacific island states have very large territories (see **table**). These land and sea domains, far more than the small, fragmented land areas per se, capture the essence of the island world that has meaning for Pacific peoples. In this world of islands, small surface land areas are just one component of a large sea replete with opportunities for exploration. The inhabitants have produced generations of seafarers. See EAST INDIES.

**Island types.** Oceanic islands are often classified on the basis of the nature of their surface lithologies. A distinction is commonly made between the larger continental islands of the western Pacific, the volcanic basalt island chains and clusters of the east-

ern Pacific, and the scattered coral limestone atolls and reef islands of the central and northern Pacific. One problem with this classification is that islands in the Pacific cannot be grouped neatly by type on the basis of location. There are coral and volcanic islands on the continental shelf in the western Pacific, and

Countries and territories	Area	
	Sea Exclusive Economic Zone, 10 km <sup>3</sup>	Land, km <sup>2</sup>
American Samoa	2,390	197
Cook Islands	1,830	240
Micronesia <sup>a</sup>	3,050	710
Fiji	1,290	18,272
French Polynesia	5,030	3,265
Guam	220	541
Kiribati <sup>b</sup>	3,550	690
Marshall Islands <sup>c</sup>	206	180
Nauru	320	21
New Caledonia	1,740	19,103
Niue	390	259
Northern Marianas	1,870	471
Palau	616	460
Papua New Guinea	3,120	462,840
Pitcairn	800	5
Solomon Islands	1,340	28,369
Tokelau	290	10
Tonga <sup>d</sup>	680	747
Tuvalu	900	26
Vanuatu	680	11,880
Wallis and Futuna	300	255
Western Samoa	120	2,935
<b>Total</b>	<b>30,372</b>	<b>551,476</b>

<sup>a</sup>The Micronesia National Report describes its Exclusive Economic Zone as over 1 million square kilometers.  
<sup>b</sup>The Kiribati National Report reports its land area as 822.8 km<sup>2</sup>.  
<sup>c</sup>The Marshall Islands National Report describes its Exclusive Economic Zone as over 750,000 km<sup>2</sup>.  
<sup>d</sup>The Tonga National Report describes its total area as 747 km<sup>2</sup> with an inhabited area of 699 km<sup>2</sup>; the latter figure has been used in population density calculations. The EEZ is given as 677,021 km<sup>2</sup>.  
 SOURCE: After R. Thustlethwaite and V. Gregory, *Environment and Development: a Pacific Island Perspective*, Asian Development Bank (Manila), 1992.



Pacific islands. (After R. V. Cole and G. S. Dorrance, *Pacific Economic Bulletin*, vol. 8, no. 2, December 1993)

these two types of islands are found in both the eastern and the northern Pacific.

It has been suggested that a more useful distinction can be drawn between plate boundary islands and intraplate islands. The former are associated with movements along the boundaries of the great tectonic plates that make up the Earth's surface. Islands of the plate boundary type form along the convergent, divergent, or transverse plate boundaries, and they characterize most of the larger island groups in the western Pacific. These islands are often volcanically and tectonically active and form part of the Pacific so-called ring of fire, which extends from Antarctica in a sweeping arc through New Zealand, Vanuatu, Bougainville, and the Philippines to Japan.

The intraplate islands comprise the linear groups and clusters of islands that are thought to be associated with volcanism, either at a fixed point or along a linear fissure. Volcanic island chains such as the Hawaii, Marquesas, and Tuamotu groups are classic examples. Others, which have their volcanic origins covered by great thickness of coral, include the atoll territories of Kiribati, Tuvalu, and the Marshall Islands. Another type of intraplate island is isolated Easter Island, possibly a detached piece of a

mid-ocean ridge. The various types of small islands in the Pacific are all linked geologically to much larger structures that lie below the surface of the sea. These structures contain the answers to some puzzles about island origins and locations, especially when considered in terms of the plate tectonic theory of crustal evolution. See MARINE GEOLOGY; MID-OCEANIC RIDGE; PLATE TECTONICS; SEAMOUNT AND GUYOT; VOLCANO.

**Island landscapes.** Compared with continents, the geology of most oceanic islands is quite simple. They do not display the same degree of variation in rock types as do continental areas that are comparable in size, mainly because most oceanic islands originate as ocean-floor volcanoes. Their growth from the seabed is relatively unimpeded by structures formed in other ways. Coral limestone formations develop on and around these volcanic ridges and cones to form the fringing reefs and atoll lagoons and islets that are characteristic of Pacific islands.

The landscapes of oceanic islands can be grouped into four broad groups: continental, volcanic, high limestone, and low coral reef. The continental or plate boundary islands tend to have steep rugged interiors, often deeply incised by fast-flowing rivers

that have formed coastal plains subject to flooding in the wet season. The more complex geology of these islands has produced numerous mineral deposits, especially nickel, gold, copper, manganese, and bauxite. One of the world's largest nickel deposits forms the backbone of the economy of New Caledonia. Until 1989, the massive Bougainville copper mine at Panguna provided around 40% of Papua New Guinea's export revenues. Fiji's gold mine at Vatukoula has provided export revenues and employment since the 1930s. Mining is often difficult, given the rugged topography and heavy rainfall characteristic of islands in the western Pacific. *See* ORE AND MINERAL DEPOSITS.

The volcanic islands typically have sharp ridges and amphitheater-headed valleys. These valleys are characteristic of islands in the Hawaiian group, in the Society Islands (for example, Tahiti and Moorea), in the Cook Islands (Rarotonga), in Fiji (Viti Levua and Vanua Levu), and in American Samoa (Pago Pago harbor is a drowned amphitheater-headed valley). The principal variables involved in the formation of these spectacular landscapes are river downcutting, chemical weathering, and mass wasting, especially soil avalanching. *See* AVALANCHE; MASS WASTING; WEATHERING PROCESSES.

A common feature of many volcanic islands is a fringing coral reef. This type of reef forms the classical Pacific land-sea interface and is a major source of protein for island residents. A diverse range of fish and crustaceans are obtained from barrier reefs. The fringing coral can also comprise a raised platform (makatea) around the perimeter of volcanic cores, evidence either of sea-level change or tectonic uplift. *See* REEF.

There are also high limestone islands that have no volcanic core above the surface of the sea. Examples include Nayau and Tuvuca in eastern Fiji (Lau), Tongatapu (Tonga), and the isolated islands of Niue, Nauru, and Makatea (French Polynesia). Such islands commonly have a central plateau that dips slightly inland and a series of terraces cut into the plateau flanks. Some of these islands (such as Niue and Makatea) are raised atolls; others are emerged volcanoes draped with limestone. High limestone islands have also been the location of extensive organic phosphate deposits. The isolated oceanic islands of Nauru, Banaba (Ocean Island), and Makatea have been the resting places of migratory birds for thousands of years. Extraction of the resultant bird droppings has been an important mining activity on limestone islands. *See* PHOSPHATE MINERALS.

The low coral islands and reefs constitute a distinctive type of island environment and landscape. The land surfaces of these islands are literally at or near sea level; rarely do atolls exceed 10 ft (3 m) in elevation at their highest points. These islands owe their origin to coral growth, and the main control over coral growth is ocean-water temperature. Coral atolls are confined to a broad band between 20° north and south of the Equator. Atoll landscapes are distinctive because of the low topography, the absence of surface running water, and the amazing diversity in

island forms and lagoon shapes. Fresh water comes from the subsurface Ghyben-Herzberg lens, comprising fresh water that floats in hydrostatic balance on the salt water in the substrata of a coral atoll. This lens is the source of most fresh water on coral atolls, where there are no streams and very few lakes. This source of fresh water is becoming very contaminated on coral islands with dense human populations. *See* ATOLL.

**Climate.** The climate of most islands in the Pacific is dominated by two main forces: ocean circulation and atmospheric circulation. Oceanic island climates are fundamentally distinct from those of continents and islands close to continents, because of the small size of the island relative to the vastness of the ocean surrounding it. Because of oceanic influences, the climates of most small, tropical Pacific islands are characterized by little variation through the year compared with climates in continental areas.

A principal element in climate of Pacific islands near the Equator is the trade winds, which blow steadily for around 9 months of the year—from the northeast in the Northern Hemisphere and from the southeast south of the Equator. These winds moderate temperatures in the subequatorial zone. During the wet season (November to March in the Southern Hemisphere, May to September in the Northern Hemisphere) there is an increased frequency of tropical cyclones in the western Pacific Ocean, especially to the north of Australia and around Hong Kong and the Philippines. Tropical cyclones cause considerable damage to crops and built environments in the western Pacific during most years. They are much less frequent in the central and eastern Pacific. *See* CYCLONE; STORM; TROPICAL METEOROLOGY.

The pattern of tropical cyclones (and the incidence of drought) in the Pacific is affected considerably by variations in air pressure between the Indonesian and South Pacific regions. These variations are termed the Southern Oscillation. Unusually high air pressure in one region is accompanied by unusually low pressure in the other. A measure of the pressure difference between the two is termed the Southern Oscillation Index. When this index is negative, both atmospheric and ocean temperatures in the central and eastern Pacific are higher than normal and an El Niño condition prevails. The warmer sea surface favors the generation of tropical cyclones much farther east than normal, and weather in the eastern Pacific is much wetter and more variable. El Niño conditions also favor drought in the western Pacific. *See* AIR PRESSURE; DROUGHT; EL NIÑO.

When the Southern Oscillation Index is positive, temperatures in the eastern Pacific are cooler, easterly winds prevail, and there is little likelihood of tropical cyclones. This condition is known locally as La Niña. Climate variability under La Niña conditions is much less than that which prevails under El Niño conditions. A major concern for the future is the prospect of pressure difference between the western and eastern Pacific increasing as a result of global warming. Increased variability of climate, associated with this warming, will have profound

impacts on island ecosystems. See CLIMATE MODIFICATION.

On the larger high islands, especially those in the western Pacific, local topography can have an important impact on rainfall. Orographic effects are substantial in Fiji, New Caledonia, and Vanuatu and on the larger Solomon Islands. There are distinctive dry and wet zones, and these are reflected in densely forested landscapes on the windward sides and dry savanna landscapes on the leeward sides of islands, with rugged interiors. See HYDROMETEOROLOGY; SAVANNA.

**Natural hazards.** The major natural hazards in the Pacific are associated either with seasonal climatic variability (especially cyclones and droughts) or with volcanic and tectonic activity. The destruction of Rabaul on the island of New Britain in Papua New Guinea in September 1994 was a vivid reminder of the potency of volcanic activity in the western Pacific.

Global climate change has particular significance for islands in the Pacific. One thing that seems clear is the increasing variability in the distribution and incidence of extreme climatic events, especially cyclones and droughts. These, coupled with the degradation of natural ecosystems accompanying development and a scenario of longer-term sea-level rise, make island environments more vulnerable to natural hazards. See CLIMATE HISTORY.

**Biodiversity.** Oceanic islands have long been recognized as having unusual endemic or native biota, largely because of their remoteness from continental landmasses. The small size of many islands has resulted in less species diversity than is common on larger landmasses.

Notwithstanding the smaller range of species and a high level of speciation on oceanic islands in the Pacific, there is considerable interisland variation in biodiversity. The larger islands of western Melanesia show a greater ecosystem diversity than the smaller low-lying atolls of the eastern Pacific, which have perhaps the lowest ecosystem diversity in the world. There also seems to be an attenuation in marine species from west to east, although everywhere marine invertebrates and fish abound on or near reefs and in associated lagoons. Sea life is the chief source of protein for many island peoples. See ECOSYSTEM; MARINE FISHERIES; SPECIATION.

A prolific and highly varied bird life is another characteristic of many Pacific islands, especially those used as resting and nesting places for migratory species at different times of the year. Birds and fish feature significantly in island folklore and legend, much more than terrestrial fauna. There are few reptiles, although sea snakes are not uncommon, especially in the atoll territories.

**Human impact.** There is a complex history of human occupation of the Pacific islands, and settlement of parts of the western Pacific has been dated back to at least 30,000 years ago. Human impact on the island ecosystems has been extensive, although some scientists believe that changes by humans associated with the fashioning of landscape are superfi-

cial when compared to the changes in the landscape that occurred earlier in the Quaternary (prehuman). See QUATERNARY.

Humans have introduced a number of animals and plants to the islands. The pig and the rat have had a profound impact on indigenous flora and fauna. Plants such as sweet potato and tapioca (cassava or manioc) became staples in the subsistence diet. Some indigenous plants, especially the coconut palm, became the foundation of the monetary economy. On many of the smaller islands, a coconut overlay effectively destroyed quite diverse ecosystems. Cash crops such as coffee, cocoa, and livestock (especially cattle) contributed to significant changes in both island biota and the physical landscape. See COCONUT; STARCH; POTATO, SWEET. Richard D. Bedford

**Bibliography.** H. C. Brookfield, *Global change and the Pacific: Problems for the coming half-century*, *Contemp. Pacific*, 1(1):1-18, 1989; H. C. Brookfield and D. Hart, *Melanesia: A Geographical Interpretation of an Island World*, 1971; P. D. Nunn, *Oceanic Islands*, 1994; E. Waddell and P. D. Nunn (eds.), *The Margin Fades: Geographical Itineraries in a World of Islands*, Institute of Pacific Studies, University of the South Pacific, 1993; E. Waddell, V. Naidu, and E. Hau'ofa (eds.), *A New Oceania: Rediscovering Our Sea of Islands*, School of Social and Economic Development, University of the South Pacific, 1993; H. J. Wiens, *Atoll Environment and Ecology*, 1962.

## Pacific Ocean

The Pacific Ocean has an area of  $6.3 \times 10^7$  mi<sup>2</sup> ( $1.65 \times 10^8$  km<sup>2</sup>) and a mean depth of 14,049 ft (4282 m). It covers 32% of the Earth's surface and 46% of the surface of all oceans and seas, and its area is greater than that of all land areas combined. Its mean depth is the greatest of the three oceans, and its volume is 53% of the total of all oceans. Its greatest depths in the Marianas and Japan trenches are the world's deepest, more than 6 mi (10 km; **Fig. 1**).

**Surface currents.** The two major wind systems driving the waters of the ocean are the westerlies which lie about 40–50° latitude in both hemispheres (the “roaring forties”) and the trade winds from the east which dominate in the region between 20°N and 20°S. These give momentum directly to the west wind drift (flow to the east) in high latitudes and to the equatorial currents which flow to the west. At the continents, water flows from one system to the other, and huge circulatory systems result (**Fig. 2**).

The swiftest flow (greater than 2 knots or 1 m/s) is found in the Kuroshio Current near Japan. It forms the northwestern part of a huge clockwise gyre whose north edge lies in the west wind drift centered at about 40°N, whose eastern part is the south-flowing California Current, and whose southern part is the North Equatorial Current.

Part of the west wind drift turns north into the Gulf of Alaska, thence west again and into the Bering Sea, returning southward off Kamchatka and northern Japan, where it is called the Oyashio Current.



Fig. 1. Principal relief features of Pacific Ocean. 1 fathom = 1.8 m. (After F. P. Shepard, *The Earth Beneath the Sea*, Johns Hopkins, 1959)

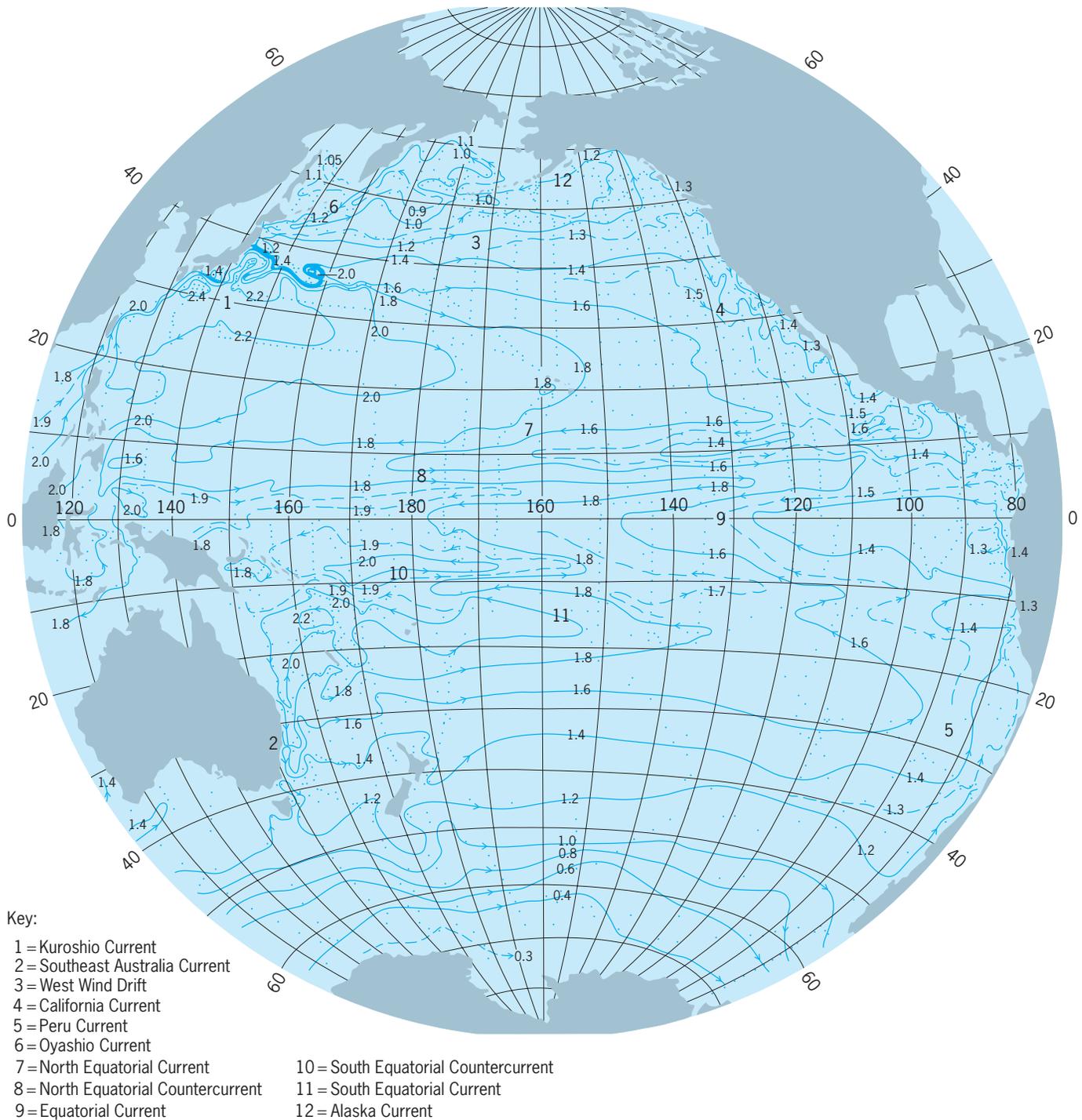
H. U. Sverdrup (1942) reported estimates of the flow in the North Pacific in the upper 4900 ft (1500 m). The Kuroshio carries about  $2.3 \times 10^9$  ft<sup>3</sup>/s ( $6.5 \times 10^7$  m<sup>3</sup>/s) at its greatest strength off Japan. The west wind drift in mid-ocean carries about  $1.2 \times 10^9$  ft<sup>3</sup>/s ( $3.5 \times 10^7$  m<sup>3</sup>/s), the California Current east of 135°W carries about  $5.3 \times 10^8$  ft<sup>3</sup>/s ( $1.5 \times 10^7$  m<sup>3</sup>/s), and the North Equatorial Countercurrent about  $1.6 \times 10^9$  ft<sup>3</sup>/s ( $4.5 \times 10^7$  m<sup>3</sup>/s).

A gyre corresponding to the Kuroshio-California-North Equatorial current gyre is found in the Southern Hemisphere. Its rotation is counterclockwise, with the highest speeds (about 2 knots) in the Southeast Australia Current at about 30°S. The current turns eastward and flows around New Zealand to South America, where it turns northward. Along this coast it is called the Humboldt or the Peru Current. It turns westward at the Equator and is known as the South Equatorial Current in its westward flow. It is to be remarked that the northwestern edge of this gyre is severely confused by the chain of islands extending southeastward from New Guinea to New Zealand. This island chain partly isolates the Coral and Tasman seas from the rest of the South

Pacific, so that the western equatorial edge of the gyre is not so regular in shape or so clearly defined as its northern counterpart. See SOUTHEAST ASIAN WATERS.

In the region of the west wind drift in the South Pacific the ocean is open to both the Indian and the Atlantic, although the eastward passage to the Atlantic through Drake Passage is narrower and shallower than the region between Australia and Antarctica. But part of the water flows around Antarctica with the wind behind it, and it receives more momentum than its northern counterpart. Total transport is several times greater.

G. E. R. Deacon (1937) estimated the transport to the east in the South Pacific part of the west wind drift as more than  $3.5 \times 10^9$  ft<sup>3</sup>/s ( $1.0 \times 10^8$  m<sup>3</sup>/s) in the upper 10,000 ft (3000 m). Sverdrup estimates only about  $1.0 \times 10^9$  ft<sup>3</sup>/s ( $3.5 \times 10^7$  m<sup>3</sup>/s) for the upper 4900 ft (1500 m) in the North Pacific west wind drift. The gyre which corresponds to the Oyashio-Gulf of Alaska gyre of the North Pacific is thus vaster in transport and area, since much of it passes around Antarctica. See ANTARCTIC OCEAN; INDIAN OCEAN.



**Fig. 2.** Principal currents of the Pacific Ocean. Contours give geopotential anomaly in meters at the sea surface with respect to the 1000-decibar (10-MPa) surface, from whose gradient the relative geostrophic flow can be calculated. The large-size numbers refer to the currents which are listed in the key. 1 m = 3.3 ft. (After J. L. Reid, Jr., *On the circulation, phosphate-phosphorus content and zooplankton volumes in the upper part of the Pacific Ocean*, *Limnol. Oceanogr.*, 7:287-306, 1962)

Between the two subtropical anticyclones (Kuroshio-California-North Equatorial Current and the Southeast Australia-Peru-South Equatorial Current) lies an east-flowing current between about 5 and 10°N, called the Equatorial Countercurrent. Sverdrup estimates flow of  $8.9 \times 10^8$  ft<sup>3</sup>/s ( $2.5 \times 10^7$  m<sup>3</sup>/s) for this current. There is also an eastward flow along 10°S from 155°E to at least 140°W.

Within the upper 3300 ft (1000 m) the flow of water is generally parallel to the surface flow, but slower. Certain important exceptions occur. Beneath the surface waters of the California and Peru currents, at depths of 660 ft (200 m) and below, countercurrents are found in which some part of the tropical waters is carried poleward. In the California Current this flow reaches to the surface in December,

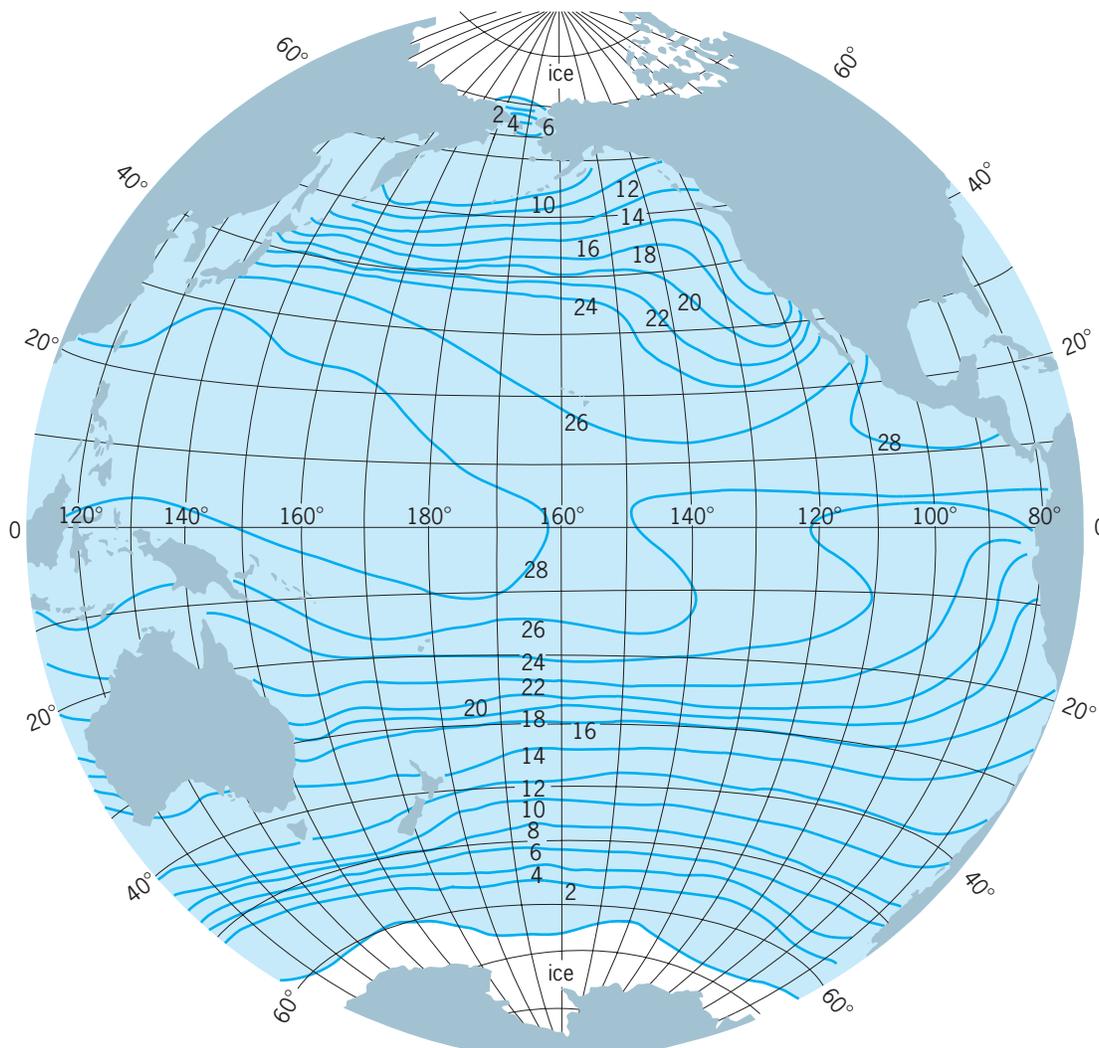


Fig. 3. Temperature at sea surface in August in °C. °F = (°C × 1.8) + 32. (After U.S. Navy Hydrographic Office, H.O. Publ. 225, 1948)

January, and February, and it is known as the Davidson Current north of 35°N. It is not known whether a surface poleward flow occurs in southern winter in the Peru Current.

Direct measurements at the Equator and 140°W have revealed that a subsurface flow to the east is found at least from 140°W to 90°W, with highest velocities at depths of about 3300 ft (100 m). Speeds as high as 2–3.5 knots (1–1.8 m/s) to the east were found at this level, while the upper waters (South Equatorial Current) were flowing west at 0.5–1.5 knots (0.25–0.75 m/s). This current has been called the Equatorial Undercurrent, or Cromwell Current after its discoverer. See OCEAN CIRCULATION.

**Temperature at the sea surface.** Equatorward of 30° latitude, heat received from the Sun exceeds that lost by reflection and back radiation, and surface waters flowing into these latitudes from higher latitudes (California and Peru currents) increase in temperature as they flow equatorward and turn west with the Equatorial Current System. They carry heat poleward in the Kuroshio and Southeast Australia currents and transfer part of it to the high-latitude cy-

clones (Oyashio–Gulf of Alaska Gyral and Antarctic Circumpolar Current) along the west wind drift. The temperature of the equatorward currents along the eastern boundaries of the subtropical anticyclones is thus much lower than that of the currents of their western boundaries at the same latitudes. Heat is accumulated, and the highest temperatures (more than 28°C or 82°F) are found at the western end of the equatorial region (Fig. 3). Along the Equator itself somewhat lower temperatures are found. The cold Peru Current contributes to its eastern end, and there is apparent upwelling of deeper, colder water at the Equator, especially at its eastern end [with temperatures running as low as 19°C (66°F) in February at 90°W] as a consequence of the divergence in the wind field.

Upwelling also occurs at the edge of the eastern boundary currents of the subtropical anticyclones. When the winds blow strongly equatorward (in summer) the surface waters are driven offshore, and the deeper colder waters rise to the surface and further reduce the low temperatures of these equatorward-flowing currents. The effect of these seasonal

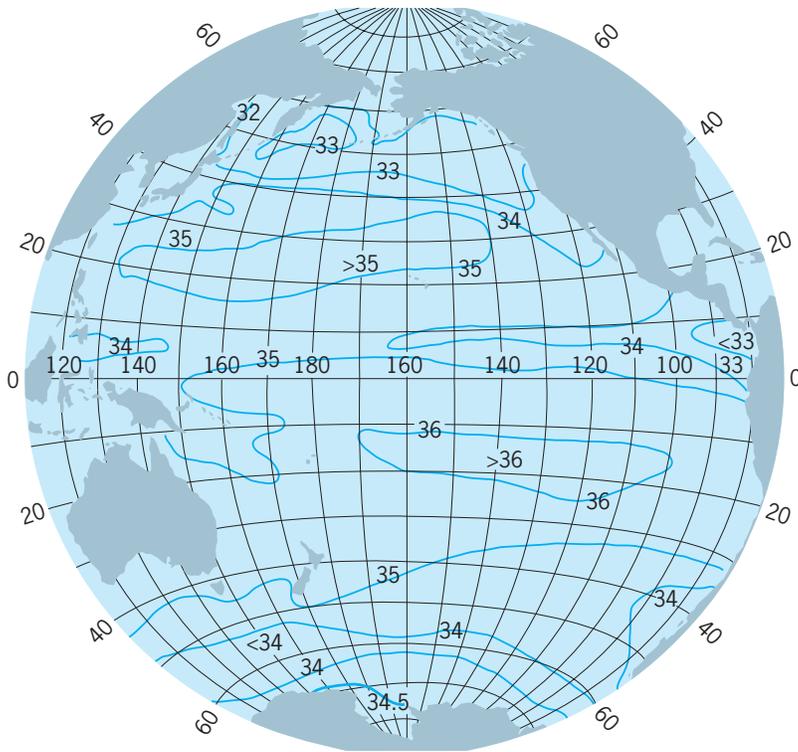


Fig. 4. Salinity at sea surface in northern summer, in parts per thousand by weight.

variations in the winds is thus to reduce the seasonal range of temperature, since the upwelling occurs in spring and summer. The seasonal range of nearshore surface temperature of the California Current at 35°N is less than 4°C or about 7°F (9–13°C or 48–55°F), though the latitudinal mean is about 10°C (50°F). The equatorward winds off Japan occur in winter and the poleward in summer, so that seasonal range is increased at that latitude to more than 16°C or about 29°F (10–26°C or 50–79°F).

The temperatures at the surface of the South Pacific Ocean have not been nearly so well documented, since most of the information comes from measurements made by merchant vessels, and the commercial shipping lines cover but a small part of its great extent. It may be reasoned that most of the temperature characteristics of the North Pacific will have analogies in the Southern Hemisphere. The damped seasonal variation of the California Current seems to occur in the Peru Current as well. But the temperatures of the Southeast Australia Current do not seem to vary so widely through the year as those of the Kuroshio, probably because of the restrictions upon the flow which are imposed by the islands.

The limiting temperature in high latitudes is that of freezing. Ice is formed at the surface at temperatures slightly less than –1°C (30°F) depending upon the salinity; further loss of heat is retarded by its insulating effect. The ice field covers the northern and eastern parts of the Bering Sea in winter, and most of the Sea in Okhotsk, including that part adjacent to Hokkaido (the northern island of Japan). Summer temperatures, however, reach as high as 6°C (43°F)

in the northern Bering Sea and as high as 10°C (50°F) in the northern part of the Sea of Okhotsk.

Pack ice reaches to about 62°S from Antarctica in October and to about 70°S in March, with icebergs reaching as far as 50°S. See BERING SEA; ICEBERG; SEA ICE.

**Salinity at the sea surface.** The highest values of salinity observed in the Pacific Ocean are slightly greater than 35.5 and 36.5 parts per thousand (‰), found respectively in the surface water of the centers of the north and south subtropical anticyclones (Fig. 4). These anticyclones cover the latitudes in which evaporation exceeds precipitation, and the overlying anticyclonic winds oppose outward flow at the sea surface. Three regions where precipitation most strongly exceeds evaporation are found poleward of 40° latitude and in the eastern tropical Pacific. The result is that the high-latitude cyclones are regions of low salinity (as low as 32.5‰ in the north, 33.8 in the south) which, through mixing with the anticyclones in the region of the west wind drift, contribute water of low salinity to the eastern boundary currents (off California and South America). The greater part of the effect of the eastern tropical precipitation is found at the surface of the North Equatorial Current and Countercurrent. Near Central America, values are less than 33‰ at 10°N, but they rise nearly to 34.5 near the Philippine Islands.

**Dissolved oxygen at the sea surface.** Above the thermocline the water is in continual overturn and is thus in constant contact with the atmosphere. Oxygen from the atmosphere dissolves in the water until equilibrium is established, and over most of the Pacific the upper layer is very close to saturation in oxygen content with typical values from about 98 to 103% of the saturation value.

The saturated value of dissolved oxygen rises as both the temperature and salinity fall, but the range of surface temperature in the ocean accounts for a wider variation in saturated value than does that of surface salinity, and it is principally variation in space and time of surface temperature which accounts for the oxygen values at the surface. Values greater than 7 ml/liter are found in cold waters of high latitudes and less than 5 ml/liter in warm regions near the Equator.

**Nutrients at the sea surface.** Nutrients such as inorganic phosphate-phosphorus, silicate-silicon, and nitrate generally increase from the sea surface downward, since photosynthesis and growth in the upper mixed layer tend to use such quantities as are there, and diffusion upward from the higher concentrations is limited by the great stability usually found immediately below the surface layer. At the surface phosphate-phosphorus varies from less than 0.25 microgram-atoms/liter in the centers of the anticyclones to more than 1.5 in the high-latitude cyclones (Fig. 5). High values are also found in the California and Peru currents. In a manner similar to the low temperatures found there, their high concentrations are partly the result of transport of mixed water from the cyclones and partly the result of upwelling at the coasts under equatorward winds.

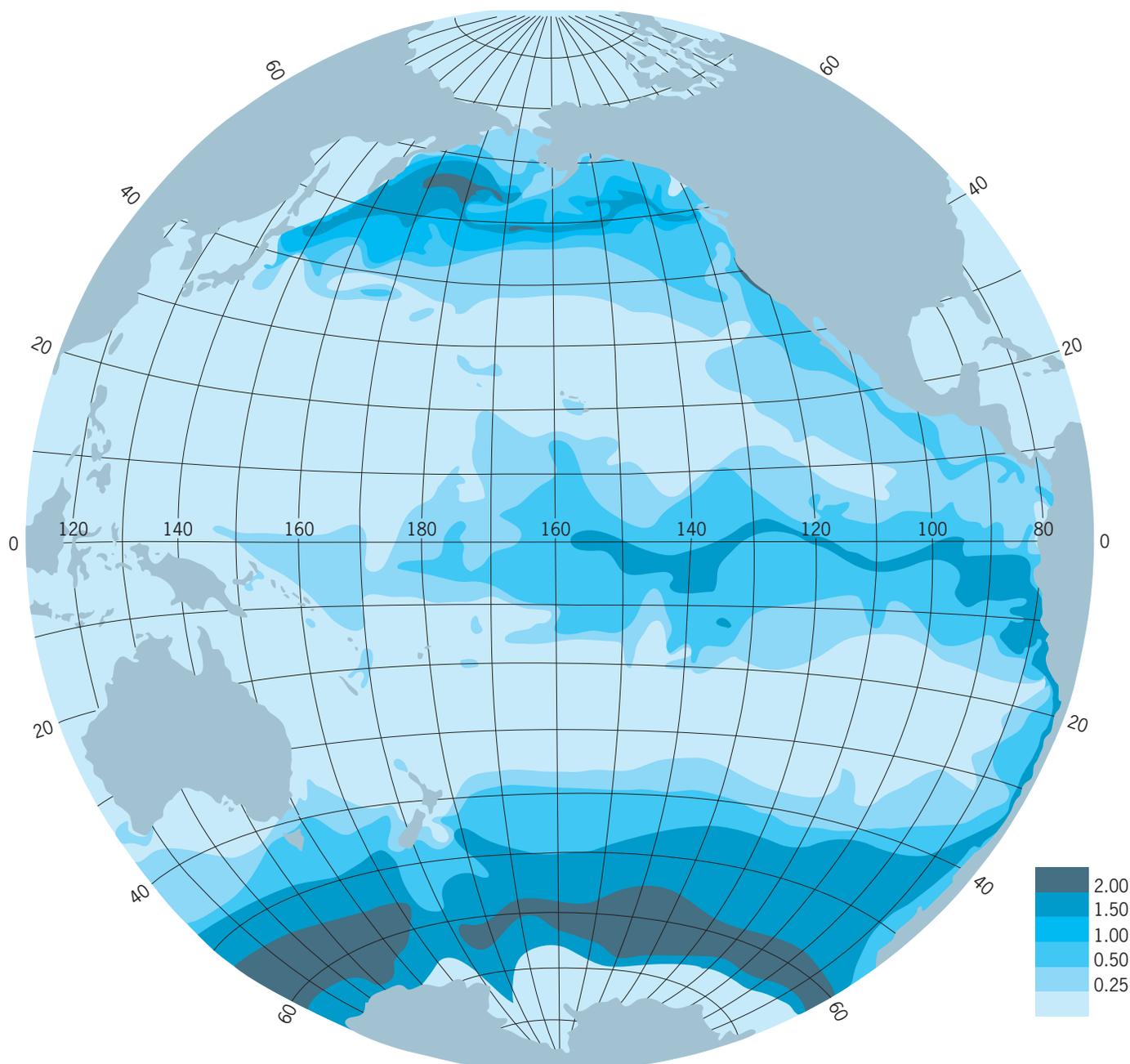


Fig. 5. Inorganic phosphate at sea surface, in microgram-atoms per liter.

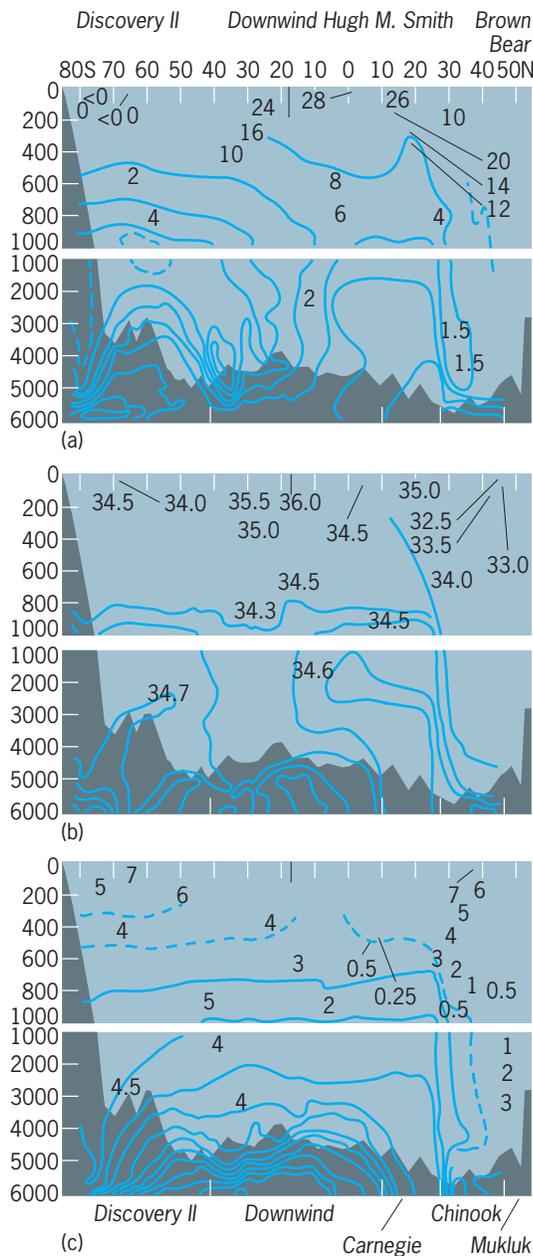
Values greater than  $1 \mu\text{g-atom/liter}$  are found in both areas during the summer period of upwelling. At the Equator in the eastern Pacific, upwelling raises values to more than 1 throughout the year.

Silicate-silicon has not been so extensively measured. It also is low in value at the surface and increases with depth. Surface values range from as high as  $40 \mu\text{g-atoms/liter}$  in the high-latitude cyclones and 12 in the upwelling regions of the California Current, to 4 or less in the center of the anticyclones and to values too small to be detected near the Equator in the eastern region. *See UPWELLING.*

**Distribution of properties with depth.** Surface waters in high latitudes are colder and heavier than those in

low latitudes. As a result, some of the high-latitude waters sink below the surface and spread equatorward, mixing mostly with water of their own density as they move, and eventually become the dominant water type in terms of salinity and temperature of that density over vast regions (Fig. 6).

The deep and bottom water of all oceans is believed to be formed in the Atlantic high latitudes off Greenland and in the Weddell Sea and arrives in the South Pacific with temperature less than  $2^\circ\text{C}$  ( $36^\circ\text{F}$ ), salinity about  $34.65\text{--}34.7\text{‰}$ , and oxygen about  $4.5 \text{ ml/liter}$ . The waters filling the Pacific below  $10,000 \text{ ft}$  ( $3000 \text{ m}$ ) retain these values of temperature and salinity everywhere, but in the northern part



**Fig. 6.** Vertical sections, central Pacific Ocean, approximately along meridian  $160^{\circ}\text{W}$ , from Alaska (right) to Antarctica (left). (a) Temperature in  $^{\circ}\text{C}$ .  $^{\circ}\text{F} = (^{\circ}\text{C} \times 1.8) + 32$ . (b) Salinity in parts per thousand. (c) Dissolved oxygen in milliliters per liter. Depths are in meters. Depth scale is expanded in upper 1000 m. 1 m = 3.3 ft. (After Carnegie expedition; NORPAC and EQUAPAC expeditions; Discovery expeditions; and Chinook, Mukluk, and Downwind expeditions)

the oxygen is reduced to values between 2.5 and 3.5. This would be consistent with depletion by decay and respiration during a slow movement north.

Over most of the North Pacific the temperature does not decrease all the way to the bottom, but beneath a minimum value at about 12,500–13,100 ft (3800–4000 m) it rises again. The increase in temperature is probably in close balance with that in pressure, since no evidence has been found of instability. Two possible explanations for the temperature maximum at the bottom are the flow of heat upward

from the ocean floor and an adiabatic rise in temperature as the water flows downward into the deeper basins.

The most conspicuous water masses formed in the Pacific are the Intermediate Waters of the North and of the South Pacific, which on the vertical sections include the two huge tongues of low salinity extending equatorward beneath the surface from about  $55^{\circ}\text{S}$  and from about  $45^{\circ}\text{N}$ . The southern tongue is higher in salinity and density and lies at a greater depth, since the surface waters of the high-latitude cyclone are more saline in the south than in the north.

It has been observed that the higher salinities are found at the surface in the anticyclones. The highest values are in the equatorward halves of the anticyclones. High values penetrate the base of the mixed layer, and tongues of high salinity extend in the thermocline toward the Equator from north and south.

Beneath the mixed layer the waters are not in contact with the atmosphere, and the oxygen that is consumed cannot be replaced directly from the atmosphere. The oxygen quickly falls below the saturated value. Even where high values of oxygen accompany the sinking of water masses, as in the South Pacific Intermediate Water, oxygen is not in saturation below 1000 ft (300 m).

Between saturated waters of the surface and cold bottom waters entering from the south lies a minimum value of oxygen. Beneath the tongue of high oxygen associated with the South Pacific Intermediate Water the minimum is only a little less than 4 ml/liter. Beneath the North Pacific Intermediate Water, however, no water as high as 3.5 is found, and in the minimum itself values less than 0.5 occur over large areas. Values of oxygen beneath the surface are the result of consumption by organisms and replenishment by mixing and renewal of the water. Since at any position the values are nearly constant in time, consumption must everywhere equal replenishment by both flow and diffusion.

Phosphate-phosphorus increases rapidly beneath the sea surface to a maximum which usually lies beneath the oxygen minimum. In regions of upwelling and divergence, higher values of phosphate and other nutrients from depth may from time to time be brought to the surface and thus made available to plants. Such regions (California and Peru currents, South Equatorial Current) are highly productive. The values gradually diminish beneath the maximum to about  $2.5 \mu\text{g-atom/liter}$  at the bottom in the north and slightly less than 2 in the south. The maximum value is greater than 3.5 in the north and between 2 and 2.5 in the south.

The Pacific Ocean is higher in phosphate concentration than the Atlantic or Indian oceans, exceeding the Atlantic by about  $1 \mu\text{g-atom/liter}$  on the average, though the values at the surface do not differ so much. This excess, like the lower value of oxygen, is undoubtedly related to deep circulation.

Nitrate-nitrogen is present in the ratio of approximately 8:1 by weight of phosphate-phosphorus in those areas where it has been measured, but the details of its distribution are not so well known.

Silicate-silicon has also not been adequately sampled. Its vertical distribution parallels that of phosphate-phosphorus except that beneath the level of the phosphate maximum it remains nearly constant, at about 170  $\mu\text{g-atoms/liter}$  in the center of the northern anticyclone, as high as 220 in the Bering Sea, about 130 in the center of the southern anticyclone. Silicate-silicon, like phosphate-phosphorus, is more highly concentrated in the Pacific than in the Indian and Atlantic oceans. See MARINE GEOLOGY; MARINE SEDIMENTS; OCEANOGRAPHY; SEAWATER; SEAWATER FERTILITY.

Joseph L. Reid

**Bibliography.** R. A. Barkley, *Oceanographic Atlas of the Pacific Ocean*, 1969; G. E. R. Deacon, The hydrology of the southern ocean, *Discovery Reports*, vol. 15, pp. 1–124, 1937; J. A. Knauss, Measurements of the Cromwell Current, *Deep-Sea Res.*, 6:265–286, 1960; H. W. Menard, *Marine Geology of the Pacific*, 1964; R. B. Montgomery and E. D. Stroup, *Equatorial Waters and Currents at 150°W in July–August 1952*, Hopkins Oceanographic Studies, vol. 1, 1962; R. A. Muller and T. M. Oberlander, *Physical Geography Today*, 3d ed., 1984; NORPAC Committee, *Oceanic Observations of the Pacific: 1955*, NORPAC Atlas, 1960; G. L. Pickard and W. J. Emery, *Descriptive Physical Oceanography: An Introduction*, 5th ed., 1990; J. L. Reid, Jr., *Intermediate Waters of the Pacific Ocean*, Johns Hopkins Oceanographic Studies, vol. 2, 1965; G. H. Sutton et al. (eds.), *The Geophysics of the Pacific Ocean Basin and Its Margin*, 1976; H. U. Sverdrup, M. W. Johnson, and R. H. Fleming, *The Oceans*, 1942; T. Teramoto (ed.), *Deep Ocean Circulation: Physical and Chemical Aspects*, 1993; W. S. Wooster and G. H. Volkmann, Indications of deep Pacific circulation from the distribution of properties at five kilometers, *J. Geophys. Res.*, 65:1239–1249, 1960.

## Packet switching

A software-controlled means of directing digitally encoded information in a communication network from a source to a destination, in which information messages may be divided into smaller entities called packets. Switching and transmission are the two basic functions that effect communication on demand from one point to another in a communication network, an interconnection of nodes by transmission facilities (see **illustration**). Each node functions as a switch in addition to having potentially other nodal functions such as storage or processing.

**Switching techniques.** Switched (or demand) communication can be classified under two main categories: circuit-switched communication and store-and-forward communication. Store-and-forward communication, in turn, has two principal categories: message-switched communication (message switching) and packet-switched communication (packet switching).

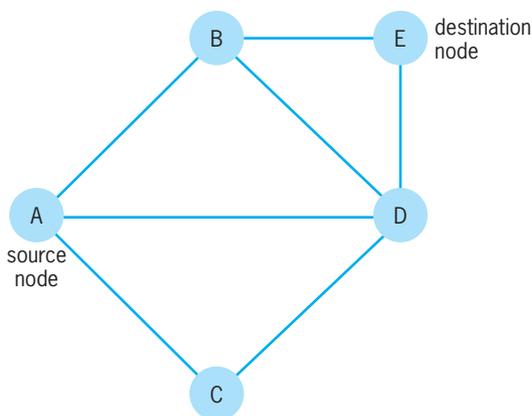
In circuit switching, an end-to-end path of a fixed bandwidth (or speed) is set up for the entire dura-

tion of a communication or call. The bandwidth in circuit switching may remain unused if no information is being transmitted during a call. In store-and-forward switching, the message, either as a whole or in parts, transits through the nodes of the network one node at a time. The entire message, or a part of it, is stored at each node and then forwarded to the next.

In message switching, the switched message retains its integrity as a whole message at each node during its passage through the network. For very long messages, this requires large buffers (or storage capacity) at each node. Also, the constraint of receiving the very last bit of the entire message before forwarding its first bit to the next node may result in unacceptable delays. Packet switching breaks a large message into fixed-size, small packets and then switches these packets through the network as if they were individual messages. This approach reduces the need for large nodal buffers and “pipelines” the resources of the network so that a number of nodes can be active at the same time in switching a long message, reducing significantly the transit delay. One important characteristic of packet switching is that network resources are consumed only when data are actually sent.

As a class of store-and-forward switching, packet switching can readily accommodate source and destination end points and transmission links operating at different speeds within the limits imposed by the network. The underlying link-layer protocol (discussed below) can also incorporate error detection and correction capability based on error detection by cyclic redundancy coding (CRC) and correction by retransmission of the packet in error. See INFORMATION THEORY.

**Packet format.** The International Organization for Standardization (ISO) has formulated a conceptual model, known as the open system interconnection (OSI) model, for exchange of information between dissimilar entities. The link-layer and the



Communication network with five switching nodes interconnected by transmission facilities. Information arriving at the originating or source node can be switched to the destination node through a variety of possible end-to-end paths, for example, ABE, ADE, ABDE, ACDE, ACDBE.

network-layer protocols represent the second and the third layers in this hierarchy. The link-layer protocol is responsible for ensuring the integrity of data across a single link, while the network-layer protocol is responsible for delivering statistically multiplexed data transparently across a network. During their passage through the network, the user data are contained within a packet which is itself contained within (link-layer) frames transmitted from one node to the next.

**Packet networks.** All public packet networks require that terminals and computers connecting to the network use a standard access protocol. Interconnection of one public packet network to others is carried out by using another standardized protocol.

*Services.* Packet networks can provide three services: switched virtual circuits, known as virtual calls; permanent virtual circuits; and datagrams. A virtual circuit is a bidirectional, transparent, flow-controlled path between a pair of logical or physical ports. (All bit patterns are transferred across a transparent path or connection without any mutilation.) A switched virtual circuit is a temporary association, while a permanent virtual circuit is a permanent association between two communicating end points of a packet network. Virtual circuits, switched or permanent, are also referred to as providing connection-oriented services. A datagram is a self-contained user data unit containing sufficient information to be routed to the destination without the need for a call to be established. Datagram-based services are referred to as connectionless services.

*Satellite and radio networks.* Packet-switched networks using satellite or terrestrial radio as the transmission medium are known as packet satellite and packet radio networks, respectively. Such networks are especially suited for covering large areas for mobile stations, or for applications that benefit from the availability of information at several locations simultaneously. The satellite acts as a repeater, amplifying the level of the signal it receives and broadcasting everything back to all the ground stations, including the one that sent the signal. If two or more stations try to transmit to the satellite at the same time, their signals collide, producing useless information. In such a situation, the transmitting stations can easily realize the occurrence of a collision by comparing the signal they received with the one they transmitted, and then can retransmit their information. The other ground stations also detect that the information they received was bad by relying on the check-sum bits that help to detect errors in the packet.

Packet radio networks have lower propagation delays and lower coverage range than satellite-based packet networks. They provide solutions that best fit situations with one or more of the following characteristics: Stations or end points are located in areas with poor telecommunications connectivity among themselves, such as in underdeveloped areas of the world; mobility of the stations or end points is important, such as involving vehicles or a moving data collection point; and stations have relatively infrequent need to communicate.

**Asynchronous transfer mode (ATM).** This is a type of packet switching that uses short, fixed-size packets (called cells) to transfer information. The term is in contrast to synchronous transfer mode, which is also known as time-division multiplexing, a familiar technique used in digital circuit switching and digital transmission. *See* MULTIPLEXING AND MULTIPLE ACCESS.

The ATM cell is 53 bytes long, containing a 5-byte header for the address of the destination, followed by a fixed 48-byte information field. The rather short packet size of ATM, compared to conventional packet switching, represents a compromise between the needs of data communication and those of voice and video communication, where small delays and low jitter are critical for most applications. Additionally, the ATM protocol takes advantage of high speed and cleaner transmission facilities, such as those provided by optical fibers, by minimizing protocol processing at the intermediate network nodes and relegating functions such as error correction to processing on an end-to-end basis rather than on a node-to-node basis.

**Applications.** Data communication (or computer communication) has been the primary application for packet networks. Computer communication traffic characteristics are fundamentally different from those of voice traffic. Data communication at speeds from several hundred bits to megabits per second is quite common. Data traffic is usually bursty, lasting from several milliseconds to several minutes or hours. The holding time for data traffic is also widely different from one application to another. These characteristics of data communication make packet switching an ideal choice for most applications. *See* DATA COMMUNICATIONS.

The principal motivation for ATM is to devise a unified transport mechanism for voice, still image, video, and data communication. Multimedia communication services are expected to benefit most from the ATM technology. *See* SWITCHING-SYSTEMS (COMMUNICATIONS).

**Internet.** The Internet is a global network, the largest packet-switched network in the world. The specific packet-switching protocol used by the Internet is TCP/IP (Transmission Control Protocol/Internet Protocol). Other packet-switching protocols are X.25 (standardized by the International Telecommunications Union) and Frame Relay, a link-layer protocol, commonly used to provide connectivity between enterprise networks (private networks that are designed to serve the communication needs of a particular corporation).

The architecture of the Internet Protocol (IP) is such that it separates applications from the underlying transmission medium. This makes the Internet Protocol suitable for carrying any digital content and, indeed, the Internet is increasingly used to carry any information in digital format. Voice over IP (VoIP) and the transport of visual information are examples of the Internet's growing usage. The precursor to the Internet is the ARPANET (Advanced Research Project Agency Network), funded by the U.S.

Department of Defense, which provided the initial test bed for packet-switching technology in 1969. In addition to the core technology of packet switching, other technological developments which led to the creation of the World Wide Web (WWW) are universal resource identifiers (URI's), the Hypertext Transfer Protocol (HTTP), and Hypertext Markup Language (HTML). See INTERNET; VOICE OVER IP; WORLD WIDE WEB.

Pramode K. Verma

**Bibliography.** T. Berners-Lee, *Weaving the Web*, HarperCollins, 1999; Special issue on evolution of Internet technologies, *Proc. IEEE*, vol. 92, no. 9, September 2004; Special issue on packet communication networks, *Proc. IEEE*, vol. 66, no. 11, November 1978; W. Stallings, *Data and Computer Communications*, 7th ed., Prentice Hall, 2003.

## Packing

A seal usually used for high pressure as in steam and hydraulic applications. The motion between parts may be infrequent as in valve stems, or continual as in pump or engine piston rods. There is no sharp dividing line between seals and packing; both are dynamic pressure resistors under motion.

Such diverse materials are used for packing as impregnated fiber, rubber, cork, or asbestos compounds. The form of the packing may be square, in ring or spiral form, trapezoidal, or V, U, or O ring in section. In packings, it is necessary that the surface finish of the contacting metal part be smooth for long life of the material. See PRESSURE SEAL. Paul H. Black

## Pain

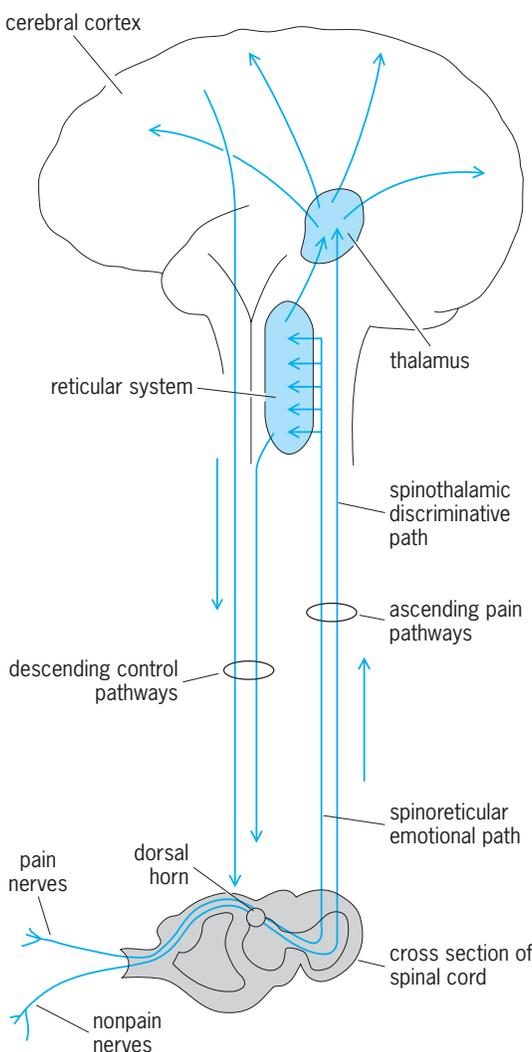
An experience of discomfort, primarily associated with tissue damage. Pain is more complex than other sensory systems such as vision or hearing because it not only involves the transfer of sensory information to the nervous system but produces suffering, which then leads to aversive corrective behavior. Unfortunately, sensations other than tissue damage or threatened tissue damage can lead to suffering and yet produce the language and behavioral responses as though tissue damage has occurred.

Pain, especially in its acute form, is usually a reflection of a tissue-damaging or potentially tissue-damaging stimulus. There is a transmission system that conveys this information to the central nervous system. This phenomenon is called nociception. In certain disease states, defects in this transmission system can of themselves generate false information to the nervous system, as though tissue damage were occurring in the periphery. An example of this is phantom limb pain, in which the individual often has a crushing type of pain in a foot that has been amputated. Both of these phenomena lead to a complaint of pain whereby the individual uses the language ("I hurt") and the behavior (rest, inactivity, taking medications, and so on) of pain. Pain causes suffering, but other aversive environmental problems can

also cause suffering such as interpersonal relationship problems, loss in its many forms, and such psychological problems as depression, anxiety, or guilt. Many patients with chronic pain complaints use the language and behavior of pain to express this suffering caused by non-tissue-damage problems. Learning is a very potent force, and if this pain behavior is reinforced by environmental contingencies, then the pain complaints can persist purely as a behavioral phenomenon, even though the tissue-damaging or other generators of suffering have ceased.

Acute pain such as occurs with broken bones and other significant injuries is almost inevitably accounted for by the phenomenon of nociception and is probably a purely neurophysiological event. However, the more pain becomes a chronic phenomenon, the more such influences as psychological factors and behavior become part of the expression of pain.

**Neurophysiology.** Acute pain is a useful warning system. There are specific nerve paths for conducting this sensation (see *illus.*). Pain receptors in the



**Neurophysiology of incoming pain.** Sensation from peripheral receptors travels along specific pain nerves, and is modulated throughout the spinal cord and brain.

skin and other tissues are nerve terminals which lack any special characteristics, and they are probably triggered by a chemical stimulus when potential tissue damage occurs. There appear to be two types of terminals: one responds to many types of painful stimuli, whereas the other specifically responds to either mechanical or thermal energy. When the terminals are stimulated, the pain (that is, nociception message) is carried along specific small sensory fibers called A-delta and C fibers. The A-delta fibers are larger and transmit the "first pain" or "fast pain." The smaller C fibers transmit a secondary dull continuous pain. These nerve fibers were traditionally believed to enter the spinal cord through the dorsal root, but it now seems that some sensory fibers may enter via the ventral route into the spinal cord.

Having entered the spinal cord, these fibers relay in the dorsal horn of the spinal gray matter, an area of considerable regulation and modulation of the incoming pain stimulus which is influenced by other incoming sensory stimuli; that is, touch or pressure sensations can suppress the transmission of signals in the small pain fibers. This helps to explain why when a person is hurting, the pain can be reduced by rubbing the affected part, and this phenomenon forms the basis of some of the treatment strategies of stimulation-produced analgesia. In addition, the incoming pain signal in the spinal cord is also modulated by descending signals from the brain. At times of anxiety, these pain signals may be augmented; however, if the individual is distracted by some interesting task, for example, playing competitive sports, then the pain signal may not be felt at all. From these relay stations in the dorsal horn, the pain signal is carried by two nerve paths up to the brain. The classical pathway is the spinothalamic tract, on the side of the spinal cord opposite to the incoming stimulus, and this leads to the posterior part of the thalamus in the brainstem, and from there nerve paths radiate the pain sensation to many parts of the cerebral cortex, where the pain is appreciated. In addition to this direct path, there is a diffuse ascending path known as the spinoreticular tract which relays to many of the basal ganglia in the brain, and from there to areas of the brain connected with motivational and affective behavior such as the hippocampus and the cingulate gyrus. It is possible that narcotic analgesics exert some of their action on this ascending spinoreticular tract because these drugs tend to reduce the suffering aspects of pain but still preserve many of the discriminative qualities so that individuals can still feel the pain, though it does not bother them so much.

Certain parts of the brainstem around the central canal appear to exert a strong inhibitory effect on incoming pain signals. Stimulation of these areas probably releases endorphins, which are morphinelike substances produced by the body and liberated at various sites on the incoming pain path to suppress these signals. *See* ENDORPHINS.

**Treatment.** Acute pain often has simple and effective treatment regimens. Chronic pain, however, rarely has simple solutions.

With acute pain, the main management approach is to treat the injured part by combating the infection, immobilizing injured tissues, draining abscesses, splinting broken bones, and so forth, until the natural healing processes resolve the problem. Pain-relieving medications such as anesthetic agents, narcotics, and aspirin can give satisfactory pain relief. *See* ASPIRIN; NARCOTIC.

There have been major improvements in treating acute pain, especially in the postoperative arena. It is now recognized that the patient is probably the best judge of how much analgesic medication is needed postoperatively, and as a result there has been an explosion of interest in and provision of patient-controlled analgesia. The patient is hooked up to an intravenous reservoir of what is usually a powerful narcotic and then pushes a button to deliver a small bolus of pain-relieving drug as needed. These machines, which usually have lock-out safety switches to prevent overdose, have proved most effective, and their use has become widespread.

The novel use of delivering a narcotic such as morphine directly to the spinal cord area has been successful in relieving pain associated with cancer and is becoming common in the treatment of postoperative and posttraumatic acute pain. A small tube catheter placed near the spinal cord can be used to provide analgesia both during surgery and for several days after in a closely monitored postoperative ward. Such systems have provided satisfactory pain relief for as long as 2 or 3 years. However, in most terminal cancer situations the duration of required analgesia is usually much shorter.

Pain that persists for months or even years is much more difficult to control, because such chronic pain is rarely fully explained by the tissue damage model and the neurophysiology of pain explained above. But it is inevitably accompanied by many psychological factors such as depression, anxiety, and personal and other environmental interactions coupled with a learning phenomenon.

Medications, immobilization, and rest, which are very effective for the treatment of acute pain, are usually ineffective and often counterproductive when used for chronic pain.

The potent narcotic analgesics all display the phenomenon of tolerance, and the longer they are used, the less pain-relieving effect they have. The side effects of these drugs, such as constipation and depression, produce secondary, often serious problems, which make them unsuitable for long-term use. The sedative hypnotic drugs such as barbiturates and many tranquilizers have similar disadvantages. If a patient has a fairly finite life span in, say, terminal cancer, then it is often justified to use these potent drugs long-term. But for noncancer chronic pain problems, they are indicated in only a small percentage (5%) of individuals. Aspirin-type drugs do not display tolerance and are useful for chronic noncancer pain problems. *See* ANALGESIC; BARBITURATES; TRANQUILIZER.

For those chronic pains due to a defect in the transmission system whereby nerves have been damaged through either trauma or infection, the antiepileptic

type of drugs are often helpful. They act by reducing the abnormal activity in the defective cells. Surgical section of nerve paths has a useful but limited application and usually in those patients with a finite life span (terminal illness). Sometimes nerves can be blocked by injecting alcohol or phenol into them, especially in those patients debilitated by terminal cancer who are not suitable for surgical operations. These radical forms of pain treatment are rarely (if ever) indicated in noncancer chronic pain because they tend to produce defects in the transmission system which can in themselves produce pain (like phantom pains) at a later stage.

As mentioned above, pain can be relieved by putting in an alternative stimulus such as touch or pressure, and this form of treatment has been used since antiquity with compresses, acupuncture, poultices, and such. There are now small electrical stimulating devices available which can effectively control pain in a certain percentage of patients. It is also feasible to implant electrically stimulating electrodes into the spine or deep areas of the brain such as the thalamus or the central gray matter. However, this approach is still at an experimental stage, and although inserting such systems into spinal areas is becoming increasingly common, the long-term effects are unknown at this time.

For those persons whose pain complaint is an expression of some underlying psychological problem, specific treatments often include the prescribing of antidepressants and anti-anxiety types of medications rather than directing therapeutic effort at the peripheral pain site. For those whose pain complaint is under control of behavioral environmental factors, it is often necessary to manipulate the environment; most of the pain clinics that are opening up around the world utilize behavioral modification techniques whereby the environmental sequelae of the pain complaints are changed, and individuals can "unlearn" their pain. This usually involves prolonged treatment. The cooperation of the immediate family is vital, and includes major efforts in physical and occupational therapy to improve activity coupled with very strict medication control; as such, these programs are quite effective at restoring these individuals to function and removing drug dependencies and reducing health care utilization. See NERVOUS SYSTEM (VERTEBRATE); PSYCHOSOMATIC DISORDERS; SOMESTHESIS. Terence M. Murphy

Bibliography. M. Brody and T. M. Murphy, Pain control and safety monitoring, *Semin. Anest.*, 7:301-306, 1988; R. D. Miller (ed.), *Anesthesia*, 5th ed., 2000; P. D. Wall and R. Melzack, *Textbook of Pain*, 4th ed., 1999; S. W. Weisel, *Pain*, 5 vols., 1989.

## Paint and coatings

Substances applied to other materials to change the surface properties, such as color, wear, and chemical or scratch resistance, without changing the bulk properties. The terms paint and coatings often are used interchangeably. However, it has become a com-

mon practice to use coatings as the broader term and to restrict paints to the familiar architectural and household coatings and sometimes to maintenance coatings for doors and windows, bridges, and tanks. Another common term for paint and coatings is finish.

Coatings are used for (1) decoration, (2) protection, or (3) some functional purpose. The low-gloss paint used on the ceiling of a room fulfills not only a decorative need but also functions to reflect and diffuse light for even illumination. The exterior coating on an automobile adds beauty, while protecting it from rusting. The public commonly thinks of house paint when talking about coatings; however, all kinds of coatings are important, as they make essential contributions to most high-tech fields.

Coatings may be described by their appearance (for example, clear, pigmented, metallic, or glossy) and their function (such as decorative, corrosion protection, abrasion protection, skid resistance). Coatings may be distinguished as organic or inorganic, although there is overlap. For example, many coatings consist of inorganic pigments dispersed in an organic matrix (binder).

The binder in an inorganic coating is a metal salt. Inorganic coatings also may contain organic additives such as lubricants. Electroplated copper, nickel, and zinc coatings will not be discussed, however; the discussion will be limited to coatings with organic binders that are applied purposefully to a substrate, with organic coatings restricted to materials that can be historically traced back to paints.

**Composition.** Organic coatings are complex mixtures of chemical substances that can be grouped into four broad categories: (1) binders, (2) volatile components, (3) pigments, and (4) additives. Most coatings contain several substances from each of the four categories, and each substance is usually a chemical mixture. The number of possible combinations is limitless, as are the possible applications.

**Binders.** These materials form the continuous film that adheres to the substrate, binds together the other substances in the coating, and provides an adequately hard outer surface. The binder governs, largely, the properties of the dried or cured film. The term vehicle usually means the combination of the binder and the volatile components of a coating. Today, most coatings, including waterborne coatings, contain at least some volatile organic solvents. Exceptions are powder coatings and radiation-cured coatings.

**Volatile components.** Solvents, water, or both are included in a majority of coatings. They play a major role in the process of applying coatings; that is, they make the coating fluid enough for application and they evaporate during and after application.

**Pigments.** Pigments are finely divided insoluble solids that are dispersed in the vehicle and remain suspended in the binder after film formation. Generally, the purpose of pigments is to provide color and opacity. However, they also have substantial effects on application characteristics and on film properties. While most coatings contain pigments, there are

important types of coatings that contain little or no pigment, commonly called clear coats or clears. Transparent varnishes and clear coats for automobiles are examples. *See* PIGMENT (MATERIAL).

**Additives.** These are all the materials that are included in small quantities to modify some property of the coating. Almost all coating formulations contain some type of additive. Some additives help to stabilize the coating before use, while others broaden its application properties or improve its durability. Examples are catalysts or initiators for polymerization reactions, stabilizers, and flow modifiers.

**Architectural coatings.** In general, these coatings are air-drying paints applied by brush, spray, or roller to architectural and structural surfaces for decorative and protective purposes. Materials are classified by formulation type as solvent-based and water-based paints.

**Trade sales paints.** These coatings are sold over-the-counter and usually applied in the field. Coatings intended for application to industrial structures are not included in this category.

**Industrial finishes.** Industrial finishes are either those coatings that are applied in the factory or those used for industrial maintenance.

**Enamel.** Enamel is a type of paint distinguished by its gloss. Enamels differ from flat paints by having a higher percentage of liquid binder, which usually makes them harder, smoother, less porous, more durable, and better able to withstand scrubbing. Commonly, they are either gloss or semigloss. The gloss enamels have a higher percentage of binder and a lower percentage of pigments. Flat enamels contain a flattening agent, a material used to eliminate gloss. Most enamels are solvent-based, but in recent years field-applied latex (water-based) semigloss enamels have been gaining in use.

**Lacquer.** Lacquer is a clear or colored solution coating that dries by evaporation alone. Drying may take place either at ambient temperature or at elevated temperature by applying heat. *See* LACQUER.

**Solvent-based coatings.** These form dry films through several mechanisms. Solvent-based coatings may be part of the lacquer or backed enamel families. An air-dry coating is any coating that dries or is chemically cured at ambient temperatures, either in the field or in a controlled factory environment. Shellac is a natural product that is usually dissolved in alcohol and is commonly used as varnish. Coatings based on nitrocellulose, other cellulose derivatives, and acrylic resins are usually called lacquers. *See* POLYACRYLATE RESIN; SHELLAC; SOLVENT; VARNISH.

Solvent-based paints, which dry essentially by evaporation, rely on a fairly hard resin as the vehicle. While the dried coating produces a tack-free and hard film, the coating is sensitive to solvents and may be dissolved or removed by heat or solvent. Such a coating is said to be thermoplastic. Coatings that require higher temperature to accelerate the solvent evaporation also are in this class. Resins for these coatings include shellac, cellulose derivatives, polyurethanes, acrylic, vinyl, and polyester resins. *See* CELLULOSE;

POLYESTER RESINS; POLYURETHANE RESINS; POLYVINYL RESINS.

Air-drying finishes once were the conventional factory-applied finishes. These lacquers and varnishes are still used in the finishing of furniture. Until the adoption of baked acrylic and urea finishes, automobile manufacturers used lacquers and air-drying enamels.

Coatings that dry by a combination of air oxidation and solvent evaporation usually are based on unsaturated oils, which are primarily plant-based. An oleoresinous varnish made from a drying oil and phenolic or modified phenolic resin once was commonly used where a hard finish was required for trim or interior surfaces. Such formulations gave way to new resins which cure in shorter times and with improved properties. Modifications of these oils with other chemicals, such as phthalic anhydride and glycerin, produce alkyd resins which can be cured oxidatively at room temperature or at higher temperatures. The oils, such as linseed oil, tung oil, and their alkyd derivatives, usually contain driers to accelerate drying and film formation. The consumption of alkyd- and drying oil-based coatings in United States is declining due to their replacement by environmentally friendly, waterborne coatings. *See* DRIER (PAINT); DRYING OIL; FAT AND OIL; GLYCEROL; OXIDATION-REDUCTION; PHENOLIC RESIN.

**Baking finishes.** A baking coating is a coating composition, either clear or pigmented, that requires elevated temperature for chemical reactions to occur; that is, curing (crosslinking) the film to a hard, insoluble form. Coatings that require crosslinking are called thermosetting coatings. Solvent-based coatings can also be formulated to form a crosslinked film at room temperature.

Urea and melamine resins polymerize when heated and are used in baking finishes where extreme hardness, chemical resistance, and color retention are required, such as on kitchen and laundry appliances. Acrylic resins crosslinked with melamine-formaldehyde resins and polyisocyanates now dominate the automobile finish market. The reaction of epoxy resins with carboxylic acids and anhydrides requires a minimum temperature of 285°F (140°C) to produce a durable coating. Certain phenolic resins are crosslinked by heat to produce finishes with excellent water and chemical resistance. *See* ACID ANHYDRIDE; FORMALDEHYDE; POLYETHER RESINS; UREA-FORMALDEHYDE RESINS.

**Plural component coatings.** These coatings begin forming a film by chemical reaction after mixing two or more components. The reactions can proceed at either room or elevated temperature. In theory, neither component reacts by itself; that is, all the components are required to yield a dry film. These coatings generally possess superior durability, adhesion, and corrosion protection properties.

**Waterborne coatings.** Waterborne coatings are similar to conventional coatings, except that the resin (binder) has been suspended in water through the use of surfactants and the majority of the solvents have been replaced by water. The term waterborne

coating applies to latex, water-reducible (water-thinned), and emulsion paints. Materials formed by emulsion polymerization are described as latex, and the formulated product is called latex paint. The most common latexes used in interior house paint are copolymers of vinyl acetate, butadiene, and acrylic monomers. Exterior house paint usually contains acrylic and styrene polymers. *See* EMULSION; EMULSION POLYMERIZATION; POLYMER; SURFACTANT.

Waterborne coatings can either be air-dried or chemically crosslinked. While latex is said to air-dry, the mechanism of drying is more complicated than evaporation. Ordinary latex coatings used as architectural paints dry by coalescence, forming impermeable films that are no longer soluble in water. Chemical crosslinking, water-reducible, and latex coatings are used as basecoats (color coat) for some automobile finishes. Because they contain lower amounts of volatile organic compounds (VOCs), waterborne coatings are environmentally friendlier than solvent-based paints. Other polymers, such as epoxy, polyurethane, alkyds, and polyesters, can be made into waterborne coatings.

**Powder coatings.** These coatings are solids. The fine particles of a one-part coating are applied to metals, wood, and some plastics by an electrostatic spray. *See* ELECTROSTATICS.

There are thermoplastic and thermosetting powder coatings. Thermoplastic powder coatings are based on nylon, ethylene, and vinyl polymers. Thermoplastic powders are sprayed onto preheated parts (or dipped into clouds of powder in a fluidized bed) followed by further baking in an oven at 350–450°F (176–232°C). Upon cooling, a tough film much like a conventional coating is obtained. The average film thickness is about 200 micrometers. Typical applications of these powder coatings include office equipment, shopping carts, and refrigerator shelving.

Thermosetting powder coatings include epoxy, polyurethane, polyester resins, and any combination thereof. The fine powder is usually sprayed electrostatically and baked at 275–450°F (135–232°C). The average film thickness is about 75  $\mu\text{m}$ . Since powder coatings are virtually VOC-free, they are more environmentally friendly than solvent-based and waterborne coatings. Unlike liquid coatings, the over-spray is usually reclaimed and reused. Thermosetting powder coatings are applied to a variety of metal, plastic, and wood substrates. Powder coating is used in automotive industry by certain manufacturers as a primer and clear coat.

Thermosetting powder coatings can also be cured by a combination of infrared and ultraviolet radiation. Some heat-sensitive substrates, such as wood and plastics, are ideal for this type of powder coating.

**Radiation cured coatings.** Coatings for which crosslinking is initiated by radiation, instead of heat, have the potential advantage of being indefinitely stable when stored in the dark. After application, crosslinking occurs rapidly at ambient temperature on exposure to radiation. Most radiation-cured coat-

ings use either acrylic or aliphatic epoxy materials. Generally, the formulations contain no solvent, and emissions are negligible. Rapid curing at ambient temperature is particularly useful for heat-sensitive substrates, such as paper, some plastics, and wood. The two classes of radiation-cured coatings are (1) ultraviolet (UV) where the initial step is the absorption of UV-visible electromagnetic radiation by the photoinitiator to generate a reactive species (ion or radical) which initiates a chemical reaction, and (2) electron beam (EB) in which the initial step is ionization of the coating resin by high-energy electrons. *See* CHARGED PARTICLE BEAMS; FREE RADICAL; PHOTOCHEMISTRY; ULTRAVIOLET RADIATION.

In the case of clear acrylic UV-cured coatings, the curing time required is a fraction of a second. The capital cost for UV curing is low and the energy requirement is minimal. There are limitations and disadvantages. Radiation curing is most applicable to flat sheets or webs for which the distance to the UV source or the window of an EB unit is approximately constant. For UV curing, pigments can limit the thickness of films that can be cured. Radiation cured coatings are essential to the production of computer chips, optical fibers, credit cards, and printed circuit boards, and are used in a variety of other economically important applications.

**Electrodeposition.** Electrodeposition is a method in which metal objects are virtually plated with resins and color by suspending them in a water bath containing resins, electrolytic stabilizers, and pigments. The objects in the bath become either the anode or cathode, depending on the chemistry of coating, while the anode (cathode) is either the tank or some suitable object suspended in the tank. Resins and pigments for this process were formerly limited to water-soluble epoxies and alkyd-modified phenolic resins and iron oxides, but the selection and color range of the resins have been extended. *See* ELECTROCHEMISTRY.

**Corrosion-resistant coatings.** In general, these are tough, highly crosslinked coatings that are applied in multiple layers and contain various corrosion-inhibiting pigments and additives. Three basic coating techniques for controlling corrosion are barrier protection (paint), inhibitive primer (coating) protection, and cathodic protection (sacrificial zinc-galvanized steel). *See* CORROSION; INHIBITOR (CHEMISTRY).

When paint is used as a barrier, the substrate is isolated [for example, from salt (conductive electrolyte)] by the physical, insulating barrier of the crosslinked film and by the use of flat platy pigments, such as micaeous iron oxide, which provide another barrier. In this case, the adhesion of coating to the substrate is of paramount importance. Coating compositions containing epoxy and phenolic resins generally provide the best corrosion resistance.

Another approach is to use primers that contain special inhibitive pigments that dissolve in water. Unlike the ions of corrosive salt, these materials react with the substrate to form a protective film that further protects steel from corrosion. *See* STEEL.

A somewhat different technique is to alter the surface of the substrate. One way to prevent or reduce corrosion is to use a material that is a sacrificial coating; that is, it dissolves instead of the iron or steel. Galvanized steel has a zinc coating that plays this part. The zinc becomes the anode (a larger one) and slowly dissolves, but the steel (the cathode) does not. A variation of this method exists in which powdered zinc is incorporated into a coating. The sacrificial material used in organic coatings is zinc dust (90–95% of the total weight of the dry coating). Metal contact between the steel and the zinc particles is essential in the early stages of the exposure; but due to the corrosion products, good protection is provided even after the contact has been lost. *See* GALVANIZING; ZINC.

By varying degrees, all organic coatings are permeable to water, salt, and oxygen. To control corrosion, several layers of coatings are used, each having a different function. A base or prime coat is applied to a surface that has been prepared to a specified degree of cleanliness and roughness by abrasive blasting or another method. The prime coat provides adhesion to the substrate for the entire coating system by mechanical anchorage, atomic attractive forces, chemical reaction, or a combination of these. Prime-coat materials once were predominantly a red pigment dispersed in linseed oil and cured by oxidation. Today, the prime coat material commonly specified for application to steel is a dispersion of zinc in a suitable inorganic or organic vehicle. In some circumstances, such as protection of aluminum for aircraft or for substrates such as steel, a one- or two-part epoxy with corrosion-inhibiting pigments is used.

An intermediate coat in a corrosion-resistant system is not always required. When used, intermediate coatings usually are higher solids versions of the same material specified for the topcoat, and are applied only to increase the dry film thickness to increase the barrier properties.

In a corrosion-resistant system, the topcoat may be selected from a variety of formulations, such as a two-part acrylic polyurethane cured with polyisocyanates, chlorinated rubber, epoxy/phenolics, and in some cases epoxy resins cured with polyfunctional amines and polyamides. Topcoat materials are required to have good sealing properties, as well as high resistance to corrosion, erosion, and ultraviolet degradation. *See* METAL COATINGS.

**High-temperature coatings.** These coatings withstand services temperatures higher than 750°F (400°C). Applications include aircraft engine components, mufflers, heat exchangers, missile technology, chemical reaction vessels, aerospace vehicles, turbines, furnaces, and barbecue grills.

Typical coating formulations, such as alkyds, polyurethanes, and epoxies, have limited service temperatures, well below 500°F (260°C). Compared to other organic polymers, silicone resins have higher thermal stability and greater resistance to UV radiation and oxidation. Coatings based on methyl silicone resins and pigmented with high levels (up to

3 lb/gal or 360g/L) of aluminum are capable of withstanding temperatures greater than 1200°F (650°C). Coatings using some ceramic frits or ferrites give service up to 1500°F (815°C). White and colored silicone-based coatings are suitable for services up to 550–600°F (290–315°C). These decorative coatings are used for incinerators, space heaters, reflectors, clothes driers, and stoves, where requirements for color and gloss retention are combined with the need for high-temperature resistance. *See* ALUMINUM; CERAMICS; CERMET; FERRITE; SILICONE RESINS.

Other high-temperature coatings include zinc dispersed in a suitable vehicle, serviceable to 750°F (400°C); a phosphate bonding system in which ceramic fillers are mixed with an aqueous solution of monoaluminum phosphate, serviceable to 2800°F (1538°C); and a ceramic gold coating used on jet engine shrouds, serviceable to up to 1000°F (538°C). For outer-space coating applications, research continues for higher temperatures and radiation effects.

**Smart coatings.** A smart coating senses its environment and makes an appropriate response, and is designed so that function and response can be switched on or off. Applications include hygienic coatings and coatings to detect and destroy chemical and biological warfare agents, as well as healable, antifouling, radar-transparent, and conductive coatings. The sensing agent in the coating may be an additive, pigment, or polymer itself. Smart coatings respond to various types of stimuli, such as heat, pressure, pH, impact, vibration, light, viruses, electromagnetic fields, and chemicals.

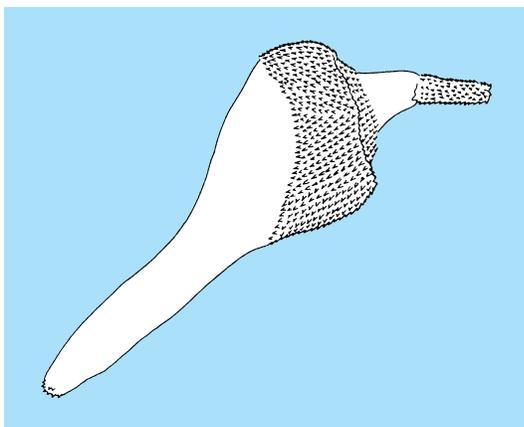
Jamil Baghdachi

**Bibliography.** C. H. Hare, *Protective Coatings: Fundamentals of Chemistry and Composition*, Technology Publishing Company, Pittsburgh, 1994; G. P. Turner, *Introduction to Paint Chemistry and Principles of Paint Technology*, 4th ed., 1997; Z. W. Wicks, Jr., F. N. Jones, and P. S. Pappas, *Organic Coatings: Science and Technology*, Wiley-Interscience, New York, 1999.

## Palaeacanthocephala

An order of the Acanthocephala, the adults of which are parasitic worms found in fishes, aquatic birds, and mammals. They have the following characteristics. The nuclei of the hypodermis are fragmented and the chief lacunar vessels are lateral. The males have usually two to seven cement glands. The ligament sac in the female breaks down so that the eggs develop in the body cavity. Proboscis hooks occur in long rows and spines are present on the body of some species. Species which commonly occur in vertebrates are *Leptorhynchoides thecatus* and *Corynosoma*.

**Leptorhynchoides thecatus.** This is one of the most common species of acanthocephalan in North American fresh-water fish. The body is a creamy color, long, slender, and devoid of spines. Both ends of



*Corynosoma reductum*. (After H. J. Van Cleave, *Acanthocephala of North American Mammals*, University of Illinois Press, 1953)

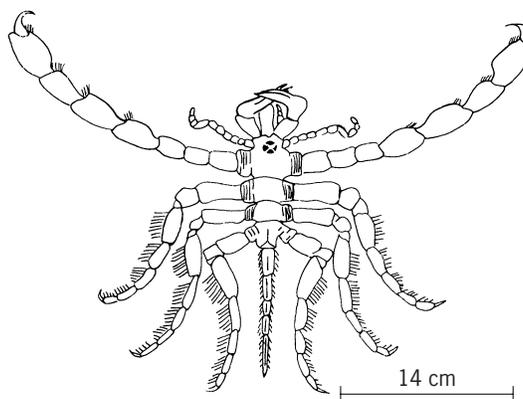
the body are curved ventrally. The females are 0.2–1.0 in. (6–26 mm) and males 0.1–0.5 in. (3–12 mm) in length. Females usually are thicker than the males. The proboscis is long, slightly enlarged in the middle, and bent ventrally at an angle to the body. Proboscis hooks are quincuncial in arrangement with 12 longitudinal rows of 12–16 hooks each. The base of the hook is enveloped by a prominent cuticular sheath. Genital organs of the male occur in the posterior half of the body. Eight cement glands usually are arranged in three levels. The eggs are spindle-shaped with the middle membrane or shell thicker because of polar enlargements. The outer membrane is thick and refractive. The intermediate host is an amphipod, *Hyaella azteca*. Adults have been recorded from the ceca and small intestine of 79 species of fish.

**Corynosoma spp.** Individuals of the genus *Corynosoma* reach sexual maturity in the small intestine of birds and mammals with aquatic habitats. The body is club-shaped and 0.1–0.4 in. (2–10 mm) long; the anterior end is thickened as an inflated bulb, whereas the hind trunk is narrow and cylindrical (see **illus.**). The body is provided with spines on the anterior extremity of the trunk, which extend farther along the ventral surface than the dorsal. In some species the trunk spines extend the full length of the trunk on the ventral surface. The body spines are of various forms, often sigmoidal, and the tip of each is commonly invested by a cuticular fold. The proboscis is directed ventrally. Proboscis hooks occur in longitudinal rows, and the individual hooks increase in thickness from anterior to posterior. The male organs are restricted to the posterior half of the body with the testes rounded or slightly elongate. Six cement glands, pyriform to clavate in shape, are present. Eggs are spindle-shaped with a short axial prolongation of the middle membrane. The species of this genus are distributed through all continents. Aquatic mammals (seals) and water birds (ducks) are the normal definitive hosts, although fishes serve as the second intermediate or transport host and various crustaceans as the first intermediate host. See ACANTHOCEPHALA.

Donald V. Moore

## Palaeoisopus

A peculiar arthropod, evidently related to the Pycnogonida and represented by a number of well-preserved fossils from the Devonian Hunsrück shales. It was formerly considered by a number of paleontologists to have an anterior jointed proboscis and a bulbous terminal abdomen, but studies of material under ultraviolet light have demonstrated that the bulbous abdomen is in fact a pair of robust, well-developed chelae (see **illus.**). Furthermore,



*Palaeoisopus problematicus*.

additional appendages (palps and ovigers) not apparent in previously studied material have been discerned, bringing its complement of anterior appendages into agreement with that of the Pycnogonida. As reoriented, *Palaeoisopus* has anterior flattened appendages that separate it from extant families of Pycnogonida. See PYCNOGONIDA.

Joel W. Hedgpeth

Bibliography. M. W. Lehmann, Neue Entdeckungen an *Palaeoisopus*, *Paleontol. Z.*, 33:96–103, 1959; R. C. Moore (ed.), *Treatise on Invertebrate Paleontology*, pt. P, 1955.

## Palaeonemertini

A rarer order of the class Anopla in the phylum Rhynchocoela, characterized by an unarmed proboscis, a thin gelatinous dermis, and either a two-layered (outer circular and inner longitudinal strata) or three-layered (outer circular, median longitudinal, and inner circular strata) body musculature. Many members (such as *Tubulanus* = *Carinella*) show primitive features, such as a peripherally located nervous system and the absence of ocelli, ciliated grooves, and intestinal diverticula. Cerebral organs, if present, are generally simple. See ANOPLA; ENOPLA; HETERONEMERTINI; RHYNCHOCOELA. J. B. Jennings

## Palaeonisciformes

A large, extinct assemblage of primitive ray-finned (actinopterygian) fishes known from the Silurian to Cretaceous, which were most numerous in

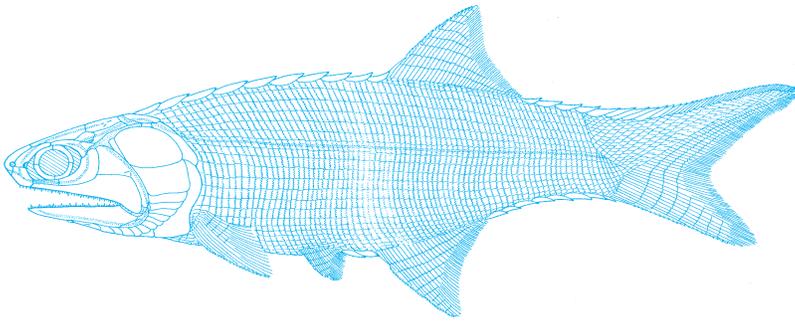


Fig. 1. *Mimia toombsi* (Gardiner & Bartram, 1977), Late Devonian, Gogo, Australia. One of the earliest palaeonisciforms, reaching 15 cm (6 in.) long, and known by complete fossils. (Reprinted with permission from B. G. Gardiner, *The relationships of the palaeoniscid fishes, a review based on new specimens of Mimia and Moythomasia from the Upper Devonian of Western Australia*. *Bull. Brit. Mus. (Nat. His.), Geology*, 37(4):173–428, 1984)

Carboniferous and Permian times. They are known as fossils from all continents except Antarctica. Approximately 40 families are recognized but collectively they are not a natural group; some are genealogically more closely related to derived ray-finned fishes (holosteans and teleosts) than to other palaeonisciforms. See ACTINOPTERYGII; HOLOSTEI; TELEOSTEI.

**Morphology.** Primitive palaeonisciforms are mostly small [5–30 cm (2–12 in.) long], with a streamlined body covered with thick, shiny scales and head bones. The scales are rhomboid and covered with a substance called ganoine (in older literature these and a few other fishes are called ganoids). The ganoine consists of a superficial enamellike layer overlying dentine. In turn, the ganoine overlies a thin bone layer—the latter is all that remains in the scales of the vast majority of modern bony fishes. Over most of the body, each scale abuts with its neighbor above and below by means of a peg-and-socket joint. This type of scale, found as isolated remains, is the first

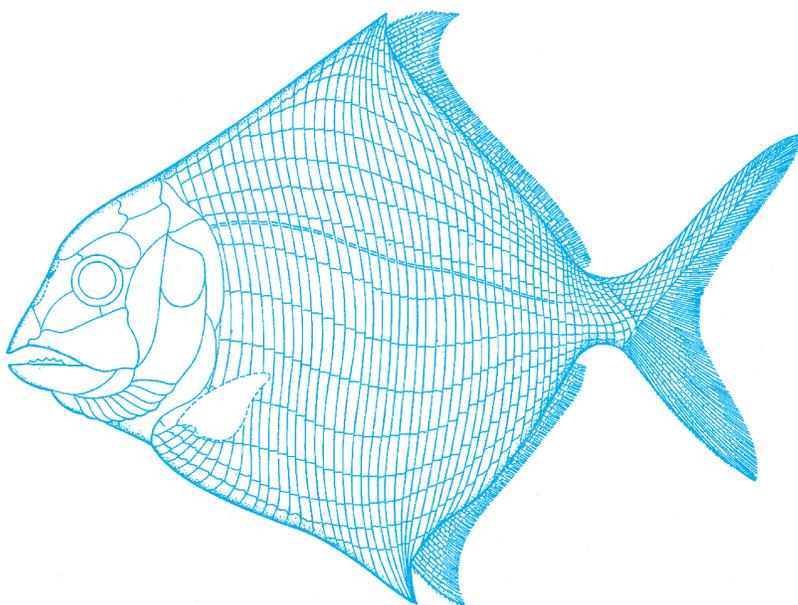


Fig. 2. *Chirodus granulosis* (Young, 1866) was a deep-bodied palaeonisciform that grew to 10 cm (4 in.) and lived in Carboniferous coal swamps of Britain. The teeth formed cutting blades that may have been used to crop vegetation. This reconstruction is taken from Traquair (*Transactions of the Royal Society of Edinburgh*, Vol. 29, Plate 5, Fig. 1, 1880)

evidence of ray-finned fishes in late Silurian rocks, and it gives its name to the group (palaeonisciform literally means “ancient scale”).

The body tapers to a gently upturned, long, asymmetrical tail, in which all the fin rays are inserted along the ventral edge; there is a single dorsal fin placed midway along the back. The paired fins (pectoral and pelvic) are located close to the ventral midline, and in more primitive palaeonisciforms, the pelvic fin has a broad base of insertion into the body. All fins are supported by finely segmented and branched fin rays, and the leading edges are covered with specialized pointed scales (fulcral scales) that form a cutwater. There are also enlarged scales along the back.

The head tapers to a blunt snout, and the large eyes (placed well forward) suggest that palaeonisciforms relied on sight for capturing prey, such as small crustaceans and other fishes, as well as to avoid predators. The nostrils are tiny and located right at the tip of the snout. The jaws are long and equipped with many tiny pointed teeth that were used to grab and impale rather than cut or grind prey. The tip of each tooth is formed by a glassy cap made of a special type of hard tissue called acrodin; this is characteristic of the teeth of all ray-finned fishes. Thick, heavily ornamented, ganoine-covered bones cover the head and cheek, and most are tightly sutured together to form an inflexible armor. The upper jaw bone (maxilla) is rigidly attached to the other cheek bones so that it is incapable of independent movement. Therefore, the only way to open the mouth is to raise the head at the same time as dropping the lower jaw. The articulation of the lower jaw with the palate is located well behind the level of the eye, such that the jaw suspension from the braincase is markedly oblique. This implies some restriction on the ability of the muscles to close the jaw.

**Evolution.** Evolutionary trends within palaeonisciforms include thinning of the scales by a reduction of the ganoine layer; development of a more upright jaw suspension and shortening of the jaws that allowed a more efficient jaw closure; a shortening of the base of insertion of the pelvic fin, allowing more flexibility; shortening of the axial lobe of the tail which was then able to generate a more horizontal thrust; and a reduction in the ratio of the dorsal, anal, and caudal fin rays to their respective internal supports, which led to more subtle control over fin movements.

**Devonian.** Most Devonian palaeonisciforms are small, slender fishes, such as *Mimia*, and lived in shallow seas (Fig. 1). *Cheirolepis*, a freshwater representative, is exceptional in that it grew to 70 cm (28 in.) and had tiny studlike scales, very similar to those of acanthodians, but the pattern of skull bones is typically palaeonisciform. See ACANTHODII.

**Carboniferous.** During the Carboniferous, many separate lineages evolved to occupy both marine and freshwater habitats. They were particularly common in deltas and rivers associated with coal forest swamps, and fishes such as *Rbadinichthys* must have swam in large schools. Body forms became more

diverse, and there were several that foreshadowed the range of body forms of modern ray-fins. Some, such as *Chiroodus* (Fig. 2) and *Platysomus*, are deep-bodied and laterally compressed. In these Palaeonisciforms, the dorsal and anal fins are very elongated and the scales on the flank are greatly enlarged, while the teeth developed as cutting and grinding plates. It is likely that these palaeonisciforms were poor at swimming in open water but highly adept at maneuvering among water vegetation and Palaeozoic coral reefs. The body form, represented by *Tarrasius*, is characterized by an elongated eel-shaped body surrounded by a continuous median fin; the scales are reduced and pelvic fins are absent, as in modern eels.

*Late Carboniferous to Late Permian.* Typical-shaped palaeonisciforms survived until the late Palaeozoic, during which *Palaeoniscum*—the fish after which the group is named—was common in Late Permian seas. Other palaeonisciforms showed several deviations in body form. In haplolepidids, which were common in Late Carboniferous freshwaters, the snout is reduced in deference to the huge eyes. On the body, the scales are thin and the fin rays are unbranched—a very unusual feature in fishes. Most significantly, the jaw suspension is upright and the cheek bones are reduced.

*Triassic.* Redfieldiids were common inhabitants of Triassic freshwaters. In these fishes, the lobe supporting the tail is shortened, and the dorsal and anal fins are large and placed well back on the body. The snout is prominent and ornamented with tubercles and may have supported a soft lip.

Deep-bodied palaeonisciforms are represented in the Triassic by the bobasatranids. These fishes had small jaws equipped with rounded, crushing teeth. Unlike earlier Carboniferous deep-bodied palaeonisciforms, the pectoral fin of *Bobasatrania* is located on the flank and was probably used more for maneuverability than for generating lift. This type of body form is taken to an extreme in the Late Permian *Dorypterus*, in which the pelvic fin is located well forward and the dorsal fin is produced as a long filamentous leading edge, analogous to that seen in the modern-day John Dory (*Zeus faber*).

*Late Triassic to Cretaceous.* Most palaeonisciforms had died out by the end of the Triassic; however, one small group, represented by *Coccolepis*, survived into the Cretaceous. In *Coccolepis*, most of the body scales had been reduced to very thin plates covered with thornlike denticles, while the scales on the axis of the tail were elongated. The snout protrudes so that the jaw is decidedly underslung. It is possible that *Coccolepis* is closely related to modern-day acipenseriforms.

**Classification.** Classification of palaeonisciforms is complicated because most are known only from flattened and fragmentary fossils. Although many separate groups, often called families, are recognized, the relationships of these families to each other and to other fishes are not well understood. Three modern fish groups must have originated from palaeonisciform fishes: the Polypteriformes, or bichirs; the

Acipenseriformes, or sturgeons and paddlefishes; as well as fishes classified as holosteans and teleosteans. See ACIPENSERIFORMES; HOLOSTEI; OSTEICHTHYES; POLYPTERIFORMES; TELEOSTEI. Peter L. Forey

**Bibliography.** R. L. Carroll, *Vertebrate Paleontology and Evolution*, Freeman, New York, 1988; P. Janvier, *Early Vertebrates*, Clarendon Press, Oxford, 1996; J. A. Moy-Thomas and R. S. Miles, *Palaeozoic Fishes*, 2d ed., Chapman & Hall, London, 1971.

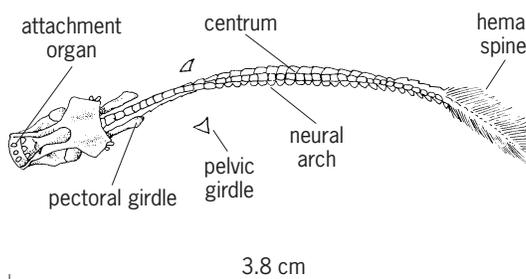
## Palaeospondylus

A tiny fossil fish found only in rocks of Middle Devonian age within a small geographic region of Caithness, Scotland, and principally at one site, Achanaras Quarry, where it is surprisingly abundant. One of the most enigmatic of all fossil fishes, *Palaeospondylus gunni* was discovered in 1890 by Marcus and John Gunn and described by Ramsay Traquair. Its original environment was a large, deep, freshwater lake system with a broad diversity of fishes. Unusually for a tiny fish, *Palaeospondylus* is typically found in sediments associated with the deepest part of the lake.

**Morphology.** Specimens of *Palaeospondylus* never exceed 6 cm (2.36 in.) in length. Usually, the head, vertebral column, and a dorsal tail fin are preserved, along with occasional traces of limb girdles (see **illustration**). (*Palaeospondylus* means “ancient vertebrae”.) The front of the head includes a strange structure identified as the nasal apparatus or jaws. A characteristic feature of the back of the head is a pair of elongated, backwardly projecting dorsal ribs.

**Classification.** *Palaeospondylus* has often been taken to be an adult because the head and backbone appear to be well ossified. However, the individual elements have been much transformed in the process of fossilization and often appear fused together. Past attempts to identify *Palaeospondylus* with any known fishes, living or fossil, have always failed, although it has variously been identified as a relative of the lampreys, a placoderm, a lungfish, a shark, a larval amphibian, and even a herring. Debates over its identity became particularly intense in the 1920s and 1930s, as different authors used it to promote rival theories of early vertebrate evolution.

Digital reconstructions of a new series of microscopic thin sections show that *Palaeospondylus* is a larval lungfish, probably the immature stage of



**Skeleton of Middle Devonian *Palaeospondylus*, shown dorsally.**

the common Caithness lungfish, *Dipterus valenciennesi*. This conclusion is supported by the presence of dorsal ribs, which is a morphological feature unique to living and fossil lungfishes. The structure on the front of the head is a sucker or attachment organ (see illustration) similar to that seen in another type of primitive (but not directly related) fish, the living garfish *Lepisosteus*. The geological environment suggests that the larvae lived in marginal lake environments, where they attached onto rocks or plants and fed on detritus before undergoing metamorphosis into the adult stage. Periodically, swarms of larvae were swept into the middle of the lake where they perished. *Palaeospondylus* is currently the oldest known vertebrate larval form. Keith Thomson

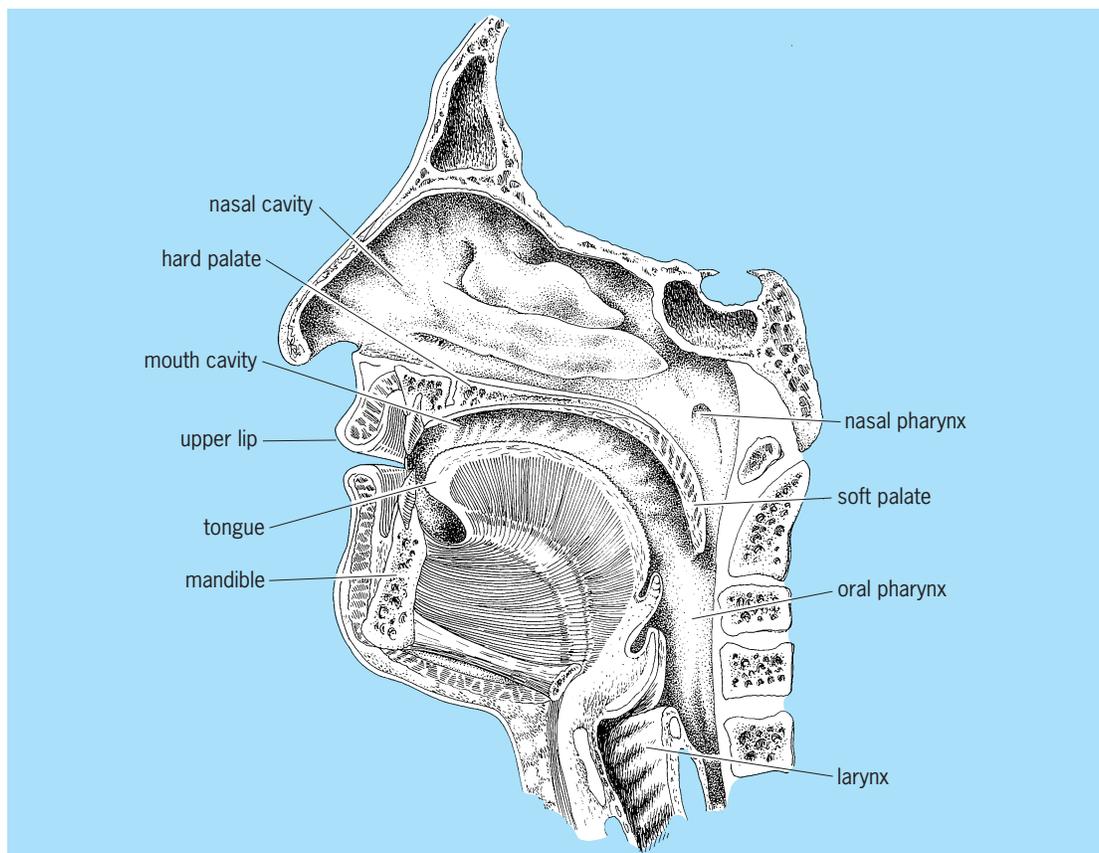
Bibliography. J. A. Moy-Thomas and R. S. Miles, *Palaeozoic Fishes*, Saunders, 1971; K. S. Thomson, A palaeontological puzzle solved? *Ameri. Sci.*, 92: 209-211, 2004; K. S. Thomson, The puzzle of *Palaeospondylus*, *Ameri. Sci.*, 80:216-219, 1992.

## Palate

The roof of the mouth in those vertebrates whose mouth cavity and nasal passages are wholly or partially separate. A communication between each nasal sac and the mouth cavity first appears in the specialized lungfishes. The two passages open into the front part of the mouth by apertures called the internal

nares, or primitive choanae, and the nose then becomes respiratory as well as olfactory in function. A similar adaptation to air breathing when the mouth is closed occurs in amphibians where the premaxillary, median region of the upper jaw serves to separate the nasal cavities from the mouth in front; elsewhere the mouth is an unpartitioned, common chamber. A distinct advance toward a complete palate distinguishes reptiles and birds, with both groups possessing a pair of bony flanges separated by a median cleft so that the incomplete palate provides something of a conduit for the free passage of air. This inverted trough extends from the primitive choanae back to the pharynx. Crocodiles and mammals create wholly separate channels by interposing a complete palate between the air passages and the mouth cavity. The definitive communication of the air channels with the nasal pharynx is known as the secondary choanae. The premaxillary palate of amphibians and the paired unfused palatal shelves of reptiles and birds correspond to progressive stages occurring during the development of the definitive palate in all mammalian embryos.

Teeth usually occur on the roof of the mouth of bony fishes, amphibians, and reptiles. Many mammals, and especially hoofed and carnivorous forms, have the palate set with transverse ridges of cornified epithelium which aid in the manipulation of food. The toothless whales elaborate these structures into sheets of fringed "whalebone" that serve to strain out minute organisms to be eaten. See NOSE; TOOTH.



Median section of the human head showing the palate (composed of hard and soft parts) and its relations with nearby structures. (After M. W. Woerdeman, *Atlas of Human Anatomy*, vol. 2, McGraw-Hill, 1950)

**Hard palate.** The palate of mammals consists of two portions (see *illus.*). The hard palate, more anterior in position, underlies the nasal cavity, whereas the soft palate hangs like a curtain between the mouth and nasal pharynx. The hard palate has an intermediate layer of bone, supplied anteriorly by paired palatine processes of the maxillary bones, and posteriorly by the horizontal part of each palate bone. The oral surface of the hard palate is a mucous membrane, covered with a stratified squamous epithelium. Anteriorly in humans there are four to six transverse palatine ridges; these diminish in prominence between fetal life and old age. A submucosal layer bears pure mucous glands and binds the membrane firmly to the periosteum of the bony component. Above the bone is the mucous membrane that constitutes the floor of the nasal cavity. There is a falsely stratified, ciliated epithelium underlaid by mixed seromucous glands. Nearest to the periosteum of the bone is a layer of elastic fibers.

**Soft palate.** The soft palate is a backward continuation from the hard palate. Its free margin connects on each side with two folds of mucous membrane, the palatine arches, enclosing a palatine tonsil. In the midline the margin extends into a fingerlike projection named the uvula. Both the hard and soft palate bear a seam, or raphe, along the midline. The oral side of the soft palate continues as the covering of the hard palate, and the submucosa contains pure mucous glands. The intermediate layer is a sheet of voluntary muscle, to which several palatal muscles contribute. The nasal side continues the structures described for the hard palate, but posteriorly, near the free margin, the epithelium becomes stratified because it makes contact at times with the nasopharynx.

Besides separating the nasal passages from the mouth, the hard palate is a firm plate, against which the tongue crushes and manipulates food. The soft palate, at rest, is pendant. In sucking, swallowing, or vomiting, it is raised to separate the oral from the nasal portion of the pharynx. This closure prevents food from passing upward into the nasopharynx and nose. The closing action also occurs in speech, except for certain consonants requiring nasal resonance. The soft palate can also be lowered into contact with the root of the tongue. The palate develops from lateral folds of the primitive upper jaw that meet the fuse in the midline. Bone differentiates in the front half and muscle in the remainder. When the process of fusion fails to any degree along its course, there results a malformation known as cleft palate. See CLEFT LIP AND CLEFT PALATE; SPEECH.

Leslie B. Arey

## Paleobiochemistry

The study of chemical processes used by organisms that lived in the geological past. Most information on the nature of life in the geological past comes from the study of fossils; a record of biochemical processes that occurred can be found in the organic molecules

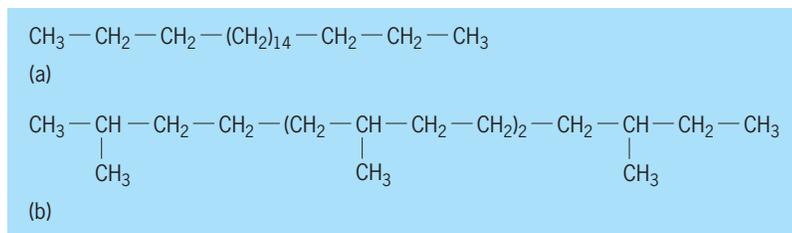


Fig. 1. Alkane isomers in sedimentary rock. (a) Normal alkane (*n*-eicosane). (b) Isoprenoid alkane (phytane).

of sedimentary rocks and fossils. The organic matter in fossil fuel deposits (coal, petroleum, and oil shale) and finely dispersed in shales and limestones represents the debris of cells which have been chemically altered to a more stable form. The relatively reactive organic constituents of cells are subject to bacterial and chemical transformation on the death of the organism. Most organic matter of living organisms is consumed in the metabolism of other organisms and returned to the seas and to the atmosphere. A small amount is deposited in the sediments, where it is preserved or converted to more stable entities. A comparison of the molecular structure of these preserved organic compounds with that of components of living cells enables the researcher to identify similarities and dissimilarities between past and present biochemistry. See PETROLEUM.

**Alkanes.** Saturated hydrocarbons (alkanes) are among the most stable organic compounds and presumably represent transformation products of the lipid fraction of cells. Two major classes of alkanes in sedimentary rocks are the normal hydrocarbons with linear arrays of carbon-to-carbon bonds, and the isoprenoid hydrocarbons with combinations of five-carbon branched-chain units (Fig. 1). See ALKANE.

Normal and isoprenoid carbon chains are used preferentially in present-day lipids. There are many thousands of possible spatial arrangements (isomers) of alkanes ( $\text{C}_{20}\text{H}_{42}$ ); yet those two structural types are the predominant ones both in living cells and in sedimentary rocks. Thus the evidence indicates that this mode of biochemical synthesis has been in use for over  $1.5 \times 10^9$  years. See LIPID METABOLISM.

**Tetrapyrrole compounds.** Molecules with tetrapyrrole ring structures, such as chlorophyll (Fig. 2) and hemin, are used by all living organisms as either photosynthetic or respiratory pigments. In a sedimentary rock, such compounds are transformed into more stable metalloporphyrins and have been found in rocks as old as  $1.1 \times 10^9$  years. The use of tetrapyrrole-containing molecules in energy-transformation processes obviously evolved very early in the history of terrestrial life. See CHLOROPHYLL; HEMOGLOBIN.

**Asymmetrical molecules.** The ability to synthesize compounds with molecular asymmetry is a characteristic property of living organisms and is due to the use of enzymes as catalysts in biochemical processes. Such asymmetrical compounds display optical activity; that is, their solutions can rotate the plane of polarization of light. Optically active hydrocarbons

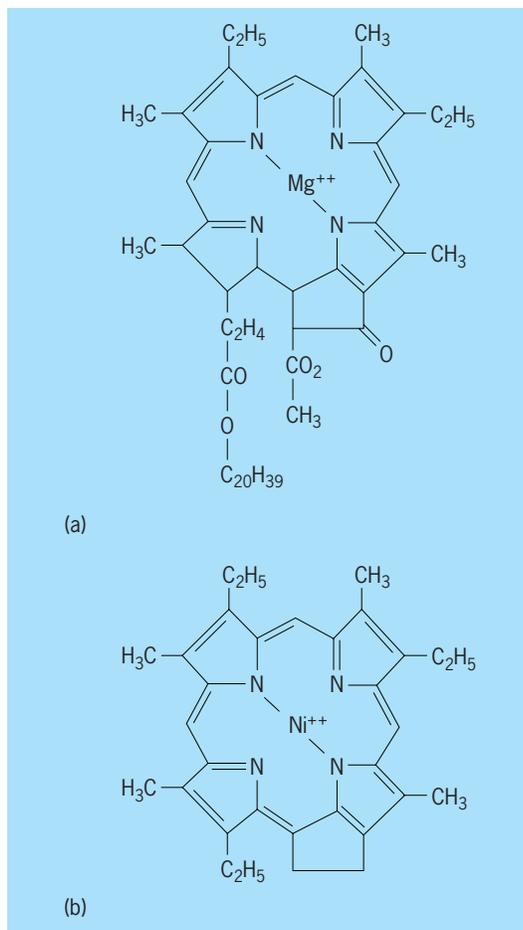


Fig. 2. Structural formulas of two pigment molecules. (a) Chlorophyll a. (b) Metalloporphyrin.

have been detected in petroleum as old as  $4 \times 10^8$  years. It is believed that the optically active preserved hydrocarbons are the steranes, which have been derived from steroids of cells (Fig. 3). If correct, this indicates that the use of enzymes and steroids is an ancient feature of terrestrial life. See ENZYME; OPTICAL ACTIVITY; STEROID.

**Amino acids.** The structural units of proteins are amino acids, of which there are about 25. While the

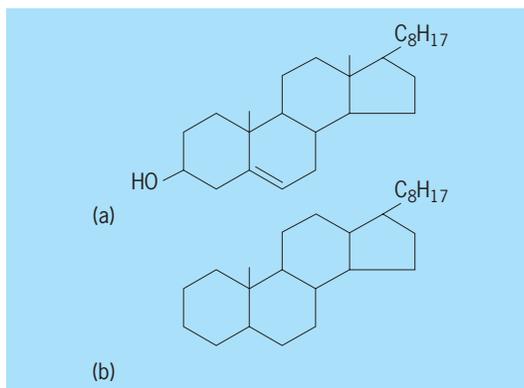


Fig. 3. Structural formulas for a steroid compound and for its possible sterane derivative. (a) Cholesterol (steroid). (b) Cholestane (sterane).

proteins and most of the amino acids are not very stable in the geological environment, seven of the amino acids are stable, can persist for great lengths of time, and have been found in fossils several hundred million years old. These amino acids are alanine, glycine, glutamic acid, leucine, isoleucine, proline, and valine. If any of the other common amino acids, such as serine, threonine, and arginine, which are not very stable, were to be detected in a very ancient fossil, contamination by recent organic matter would be suspected. Such contamination is a serious problem in paleobiochemical studies. See AMINO ACIDS; PROTEIN.

Paleobiochemical studies have shown that a number of the common chemical processes used by living organisms today have been in use for a very great length of time. Paleobiochemical techniques will be used on materials returned from extraterrestrial sources to determine whether life exists outside of Earth, and if so, to see how it may differ from life on Earth. See FOSSIL; PALEOECOLOGY; PALEONTOLOGY.

Thomas C. Hoering

Bibliography. P. H. Abelson (ed.), *Researches in Geochemistry*, vol. 2, 1967; K. A. Kvenvolden (ed.), *Geochemistry of Organic Molecules*, 1980.

## Paleobotany

The study of fossil plants of the geologic past. A paleobotanist is a plant historian who carefully pieces together the geologic history of the plant kingdom. Other organisms, including fungi and various types of microscopic plankton, are also studied by paleobotanists. Paleobotany is a branch of paleontology that requires a knowledge of both plant biology (botany) and the geological sciences. See BOTANY; FOSSIL; PLANT KINGDOM.

Materials used by paleobotanists to reconstruct plant life through geologic time include fossilized remains preserved in the rock layers of the Earth. Materials such as fossil leaves, seeds, fragments of wood, fruits, and flowers are used to interpret the biology and evolution of ancient plants. In addition, the products and distribution of plants are recorded in the rock record in the form of coal, resins, various chemicals, and other substances produced by plants. Some plant fossils, such as pollen grains and spores, are studied by palynologists; palynology has been especially important in mineral and petroleum exploration and in correlating rock layers that are widely separated geographically. From all of these materials, paleobotanists attempt to reconstruct the habit, structure, and biology of plants that grew on the Earth millions of years ago. See PALYNOLOGY.

Paleobotanists are also concerned with the myriad interactions that existed in ancient communities and ecosystems. An examination of fossil plants throughout geologic time points to a succession of organisms that inhabited the Earth, from relatively simple forms that are preserved in rocks about 3.6 billion years old to forms that are progressively more complex and suited to different types of habitats. Paleobotany not

only involves the collection, description, reconstruction, and naming of fossil plants but is also concerned with the evolution of major groups, relationships that exist between fossil and living forms, how ancient plants functioned and reproduced, what type of environment they lived in, how they were fossilized, and many other biological and geological topics.

**Plant fossilization.** In modern ecosystems, plant litter generally decomposes as a result of fungal and microbial degradation and mechanical disintegration. However, a very small percentage of all plant materials that inhabit the Earth at any one time become buried in accumulating sediments, and are thus preserved as fossils. Plants are preserved in a variety of ways, and various combinations of physical and chemical processes are involved at the time of preservation. For a plant part to become fossilized, several unique circumstances must occur. First, the item must be close to a site where sediments are accumulating; second, it must be rapidly buried. For example, a leaf that falls from a tree has little fossilization potential if it is transported in a stream a long distance, since along the way it will become torn, abraded, or otherwise destroyed. Rapid burial is also necessary to ensure that the biological activities of various microbes do not destroy the tissues of the leaf. *See* BIODEGRADATION.

Plant parts are best preserved in very fine grained silts and shales, which are the lithified muds of ancient deposits. Such sediments generally yield excellent fossils because the small grain size preserves minute details of the leaf; coarser-grained sediments such as sands generally do not reproduce delicate features. Plant parts composed of thick-walled cells or those having fibers within their tissues have a better chance of being preserved (fossilization potential) than more delicate tissues composed of thin-walled cells. The aboveground parts of most plants are covered by a layer of cutin or wax that is generally resistant to decay and aids in the preservation of some tissues.

In addition to the many variables that contribute to the processes of fossilization, the site of burial is critical as to whether fossil remains will ever be discovered. For example, sediments deposited in certain types of lakes, rivers, and swamps may be easily lost as the depositional system changes over time, or as the sediments are eroded away as younger river systems cut into older fossil-bearing rocks. As a result of all of these biological and physical constraints, relatively few organisms become fossils, and an even smaller number are ever found by paleobotanists. In addition, the number of fossils observed is reduced due to the fact that most of the sediment making up the crust of the Earth has never been examined for evidence of ancient plants. It is from this incomplete fossil record that the paleobotanist attempts to reconstruct the flora at a particular point in time and space.

Plants can be fossilized in a variety of ways, each yielding different types of information about the ancient flora. Thus, the paleobotanist must employ different techniques to extract the maximum amount

of information from each fossil. There may be more techniques used to study fossil plants than there are ways in which the plants are preserved. There is no single "best" technique. The questions that are asked and the type of preservation will dictate how to examine a particular fossil plant. *See* FOSSIL SEEDS AND FRUITS.

*Impressions.* Impressions probably represent the most common type of plant fossil. They occur when a plant part is covered by sediment and the water is squeezed from the cells and tissues. Cells that make up the plant are eventually degraded, perhaps as a result of microbial activities, so that only a shallow negative, or imprint, of the plant organ (for example, a fern leaf) remains. The paleobotanist determines the shape, venation, and other structural features of the leaf and notes any characteristics that make it unusual. A leaf forms a fossil impression much like a modern leaf on the surface of wet concrete. The leaf is partially impressed into the surface and results in a negative outline of its shape. In a short time, the leaf is gone, degraded by microorganisms, but the impression remains until it is weathered or abraded away.

Although impressions lack organic material, sometimes when the rock is composed of exceedingly fine grains it is possible to see the outlines of leaf cells and surface features such as hairs. This can be accomplished by using a light microscope or by preparing a latex replica of the surface of the fossil; this replica can be examined in a scanning electron microscope, where greater magnification and resolution are available.

*Compressions.* The same processes that form impressions are responsible for the formation of compressions. However, compressions result in the surface of the fossil being covered by a thin film of carbonaceous material (**Fig. 1**). This film of carbon represents the original plant, in which the cells have been highly compressed. Most of the best-preserved compressions are found in clay and shale, although some have been discovered in volcanic ash deposits. The most attractive compressions are often preserved in a light-colored matrix that makes the dark carbonaceous film easy to study and photograph. In a few compressions the remains of the plant are exceptionally well preserved and may contain various cell organelles. In some plants the preservation is so good that details can be examined with a transmission electron microscope. In other compressions all that remains is the cuticle that covered the surface of the leaf. Specimens with cuticle can be prepared to show the distribution and types of cells and stomata present on the surface of the leaf. The ultrastructure and chemistry of the cuticle have been observed in some fossil plants. With some compressions the entire fossil can be immersed in acid to dissolve away the rock matrix, leaving the cuticle, which is resistant to the acid, available for detailed examination. Because of the fragile nature of cuticle this technique can be used only with relatively small plants.

Coal is a type of compression fossil. In coals that have not been subjected to extensive heat and



Fig. 1. Compression specimen of the Triassic leaf type *Dicroidium*.

pressure, it is possible to recognize plant materials. Pollen grains (Fig. 2), spores, and fragments of plant cuticle can be macerated (disintegrated by cutting into very small fragments) from coals and provide information about the kinds of plants that lived when the coals were being formed. They can also be examined with the aid of certain types of electron microscopes. There are a few examples where the entire coal seam is made up of the cuticle envelopes of plants. After washing and staining, these “paper coals” can be mounted directly on microscope slides for examination. See COAL; COAL PALEOBOTANY.

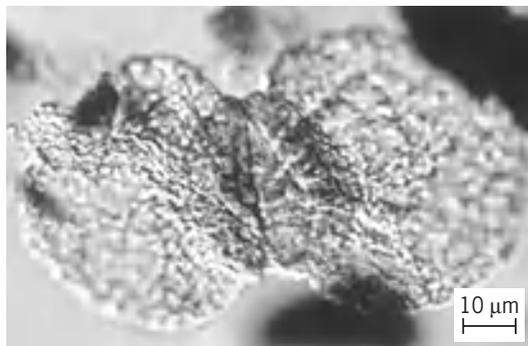


Fig. 2. Permian pollen grain macerated from Antarctic coal.

*Molds and casts.* Often, three-dimensional plant parts such as stems and seeds become buried in sediment but are not flattened in the process (as in impressions). After a period of time, the plant tissues disintegrate, leaving a negative or mold of the plant organ. Occasionally, the mold becomes filled with new sediment and forms a three-dimensional replica of the original plant part, termed a cast. Molds and casts do not contain any original plant material; however, they show the external form and the original shape of the fossil plant. This type of fossilization process is especially helpful in reconstructing large plants, for example in determining the morphology of underground organs, and understanding the position of large branches.

*Permineralizations.* Although permineralization is relatively rare, this type of fossilization process provides perhaps the greatest amount of information about the cellular detail and tissue systems in some fossil plants. Permineralizations form when a plant part becomes immersed in solutions containing concentrations of dissolved minerals that permeate the tissues, filling the cell lumens and intercellular spaces. As the minerals crystallize, the plant part becomes entombed by the rock, and thus is not distorted from its original shape as in the case of compression fossils.

The most common entombing minerals in permineralizations are carbonates and silicates, but oxides may also be present. In this fossilization process the cell walls remain organic; however, there apparently is some type of chemical change to the cellulose and lignin. Cell organelles such as nuclei, various membrane systems, and even chromosomes have been described from some permineralizations. Some of the best-known permineralizations are Carboniferous coal balls, and silicified peat blocks collected from Antarctica. To study permineralizations the surface of the rock can be etched with acid so that the plant cell walls are slightly elevated from the surface of the rock. The surface then can be covered with acetone and a thin sheet of cellulose acetate rolled on to the surface. The sticky acetone/acetate mixture flows in and around the elevated cell walls and dries. This thin slice of plant material now embedded in the acetate can be “peeled” from the surface, pulling with it the cells of the permineralized plants. After slightly grinding the surface of the rock again with an abrasive (silicon carbide) powder, a new peel can be prepared. Thus, permineralized plant parts can be sectioned much like that done with extant plants. Peels then can be mounted on microscope slides and photographed (Fig. 3). See COAL BALLS.

Petrifactions is another type of fossilization which is similar to permineralization. In petrifactions, however, the original cell walls are completely replaced by other minerals. Woods from the Petrified Forest in Arizona, the Cerro Cuadrado Petrified Forest in Patagonia, and Rhynie chert in Scotland are examples of this type of fossilization process. To study petrifactions a thin slice of the petrified plant is cemented to a glass slide and ground with abrasive powder until it transmits light (Fig. 4). This thin section is the only

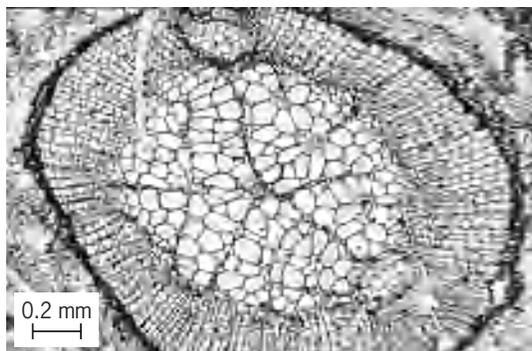


Fig. 3. Cross section of a permineralized stem from a Carboniferous coal ball.

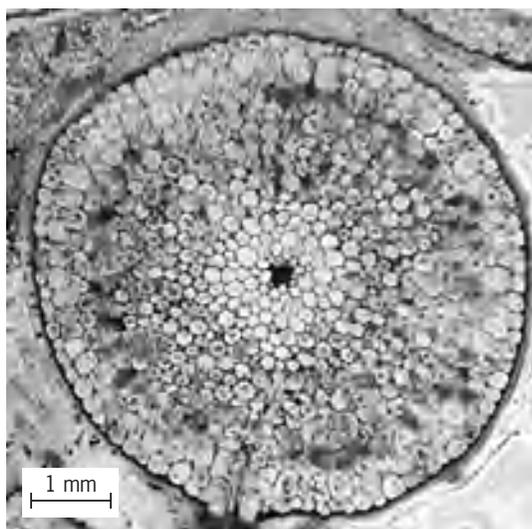


Fig. 4. Thin section of a petrified Rhynie chert stem.

means available to examine the cells that are petrified. See PETRIFICATION; PETRIFIED FORESTS.

*Unaltered plant material.* In special situations, plants may be preserved in an almost unaltered form. Examples are pollen grains and spores, various forms of plant exudate (such as amber), the silica frustules of diatoms, and even certain types of algae that cause calcium carbonate to be formed.

Various chemical signatures that remain in the rock long after the organism is gone represent still another form of unaltered plant material. Chemical compounds such as lignin, sterols, and carboxylic acids are examples of chemical fingerprints that can be detected in rock layers millions of years old. In some instances, it is possible to relate the chemical constituents of the fossils to those found in closely related living plants, and thus to suggest the relationship between living and fossil organisms more clearly. There are several reports of chemical traces in Precambrian rocks, some as old as 3.8 billion years, but the nature of these rocks is being debated. Such "fossils" require critical analysis since contamination from recent organisms is a major problem. A case in point is the deoxyribonucleic acid (DNA) reported from an exceptionally well-preserved fossil leaf of Miocene age. Some believe that this DNA was a con-

taminant from younger or modern organisms, perhaps even from recent bacteria rather than that of the fossil leaf.

Amber comprises fossilized plant resins of a variety of different types and can be found in the fossil record as early as the Carboniferous. It is an example of unaltered plant material. Phytochemical and x-ray diffraction techniques have been useful in determining the kinds of plants responsible for producing the different forms of amber. Because of its sticky nature, amber also serves as a fossilization matrix for other organisms, including insects. Small flowers, fungi, algae, and various wind-borne plant parts, including pollen grains, spores, and seeds, have also been discovered from samples of amber throughout the geologic column. See AMBER.

Other types of preservation are less frequently encountered in the rock record. It is rare that a single fossil is preserved in only one way. Impressions sometimes contain a bit of carbonaceous film, making part of the fossil a compression; and on the outside of permineralized coal balls is often found coal, a type of compression.

Regardless of how plants are fossilized, the most important factor is determining what techniques should be used by the paleobotanist to extract the maximum amount of information. For many fossils (both plant and animals) a particular specimen is the only available example of that organism. Thus, a critical step in paleobotany is first evaluating how the organism is preserved, and then determining how to examine it based on the types of questions to be asked.

Still other paleobotanists use biomechanical principles and morphometric approaches to study fossil plants.

**Use of fossil plants.** The geological record of past floras can be traced to the present. Fossil plants have provided an enormous body of knowledge about the evolution of plants from the earliest Precambrian unicellular forms to the complex multicellular flowering plants used for food and shelter today. Tracing plants through geologic time has provided the basis for examining how and when the first plants became adapted to a terrestrial environment, when

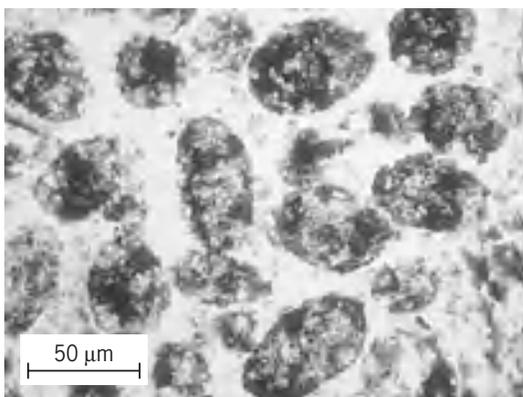


Fig. 5. Several coprolites found within the tissues of a Carboniferous plant.

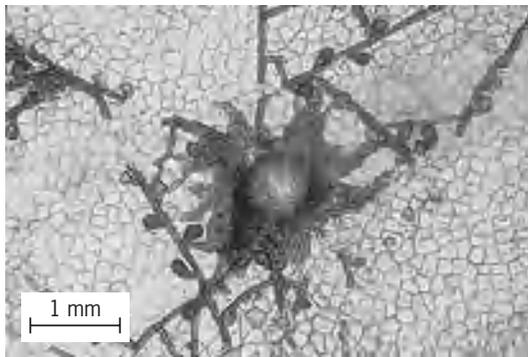


Fig. 6. Cuticle of a fossil (Eocene) leaf showing the outline of the epidermal cells and fungus with specialized hyphae.

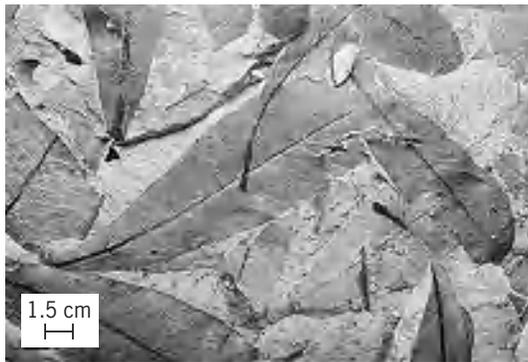


Fig. 7. Several *Glossopteris* leaves from the Permian of Antarctica. Leaves of this type were some of the first evidence that at one time the southern continents were all physically linked into a major continental plate called Gondwana.

the angiosperms first appeared, why seeds evolved, and why some plants acquired the ability to produce wood. Once plants are identified, the composition of the community and ecosystem can be examined, including the types of interactions that existed with other organisms; for example, certain types of herbivores can be identified in permineralized tissues based on coprolites (Fig. 5) or wound responses. Fungi (Fig. 6) and various types of galls can be identified on fossil leaves that in turn can provide information about the ecosystem. Fossil plants also demonstrate that certain groups that once flourished (for example, the cycads) are now geographically restricted and probably heading, like other plants in the geologic past, toward extinction. See PLANT GEOGRAPHY.

The physical environment where the plants lived is an important component of paleobotany, as is the geologic setting. How were the plants preserved, and does the assemblage of fossils that the paleobotanist has collected represent an accurate depiction of the total community? How was preservation affected by seasonality and the types of plants? What types of plants lived together, and how were they distributed in time and space? Another set of questions that can be answered by the paleobotanist relates to biostratigraphy. For example, plant fossils have been used to correlate rock layers that are widely separated geographically. Thus, where the geologic time range of

a plant or assemblage of plants has been determined in rocks whose relative age is known, the presence of the same plants in other rocks will indicate an equivalent age (Fig. 7). See STRATIGRAPHY.

Fossil plants can also be valuable in reconstructing climates of the past. For example, modern woody trees in temperate climates produce different sizes and numbers of cells throughout their growing season (Fig. 8). This periodicity results in the production of annual rings in the wood that can be related to climate signals such as moisture, light, temperature, and even the presence of fire. Permineralized fossil wood sometimes shows similar ring features that can be used to interpret the climatic conditions under which the plant was growing many millions of years ago. For example, trees more than 30 m (100 ft) tall grew at approximately 85° South during the Permian and Triassic in what is now Antarctica and produced wide growth rings (Fig. 9). These rings are quite different from those produced in temperate climates today. While modern tree rings form in response to changes in temperature and moisture, those in the Antarctic woods formed in response to changing light levels. The overall morphology and various structural features on certain fossil leaves also can be used to interpret climates of the past. For

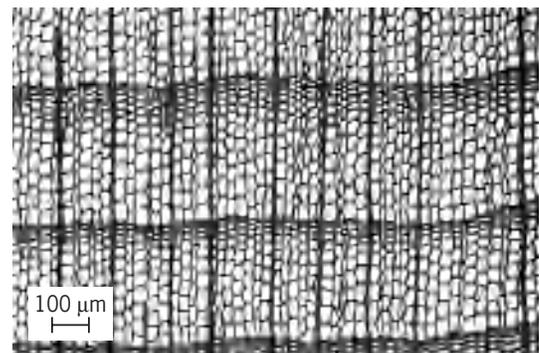


Fig. 8. Section showing tree rings from a temperate plant, showing change in cell size and wall thickness.

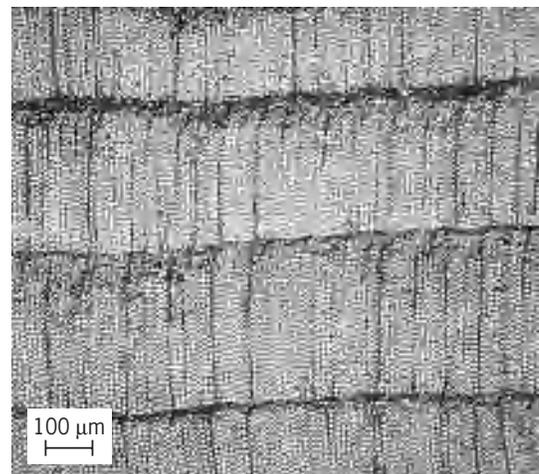


Fig. 9. Section showing rings from a tree with uniform cell size and wall thickness, but growing at approximately 82° South during the Triassic.

example, the number and distribution of stomata on living angiosperm leaves can be measured against known carbon dioxide concentrations in the atmosphere today. This stomatal density can then be calculated for closely related fossil plants. The differences produce a proxy record that is used to infer changes in carbon dioxide concentration through geologic time. See PALEOCLIMATOLOGY.

In modern floras, there are certain species of plants that are restricted to specific climatic regimes. Analyzing the margin of the leaves has proven to be a useful method of relating modern and fossil plants, and hypothesizing the climatic conditions under which the plants once grew. For example, the discovery of palms, laurels, magnolias, and cycads in the early Tertiary of southeastern Alaska provides strong evidence of subtropical to warm-temperate, lowland conditions which existed at that time. See PALEONTOLOGY; POSTGLACIAL VEGETATION AND CLIMATE.

Thomas N. Taylor

Bibliography. T. P. Jones and N. P. Rowe, *Fossil plants and Spores: Modern Techniques*, Geological Society, 1999; K. J. Niklas, *The Evolutionary Biology of Plants*, University of Chicago Press, 1997; W. N. Stewart and G. W. Rothwell, *Paleobotany and the Evolution of Plants*, 2d Ed., Cambridge, 1993; T. N. Taylor and E. L. Taylor, *The Biology and Evolution of Fossil Plants*, Prentice Hall, 1993; B. A. Thomas, *The Evolution of Plants and Flowers*, Eurobook Ltd., 1981; B. A. Thomas and R. A. Spicer, *The Evolution and Palaeobiology of Land Plants*, Croom Helm, 1987; K. J. Willis and J. C. McElwain, *The Evolution of Plants*, Oxford, 2002.

## Paleoceanography

The study of the history of the ocean with regard to circulation, chemistry, biology, and patterns of sedimentation. The source of information is largely the biogenous deep-ocean sediments, so the field may be considered a branch of sedimentology or paleontology. However, there are also strong links to geophysics, marine geochemistry, and mathematical modeling. Geophysical sciences are called upon for reconstruction of geography (position of continents, horizontal motion of the ocean floor), topography (changing depth patterns in the ocean, general subsidence of any given piece of sea floor), and dating (radioisotopes, magnetic patterns on the sea floor and in sediments). Geochemical analyses deliver information on sediment composition (stable and unstable isotopes; major and minor components such as carbonate, opal, and trace elements). Such information is useful in correlating sedimentary sequences and, when combined with geochemical arguments, yields insights about the dynamics of carbon and nutrient cycles. Mathematical modeling introduces a strong quantitative element. It draws upon the knowledge reservoir of modern oceanography and climatology. See CLIMATOLOGY; GEOCHEMISTRY; GEOPHYSICS; PALEONTOLOGY; SEDIMENTOLOGY.

**Origins and development.** The study of marine sediments on land is as old as geology itself. Modern paleoceanography is set apart by the study of sediments recovered from the ocean, especially the deep ocean, and by the use of concepts developed by oceanographers (controls on ocean currents and upwelling, vertical stratification, heat budget, nutrient and carbon cycles, pelagic biogeography and water masses). Important early studies were on cores raised by the German *Meteor* Expedition (1925–1927) to the central and southern Atlantic, by the Südpolar Expedition (*Gauss*, 1901–1903) in the Antarctic, and by the United States cable ship *Lord Kelvin* (1936) in the North Atlantic. The glacial debris zones noted in the *Kelvin* cores are now commonly referred to as Heinrich layers; they are witness to sporadic input of iceberg armadas during the last glacial epoch. See GLACIAL EPOCH.

Comparisons of climate-related changes between major ocean basins became possible through the systematic recovery of long cores by the circumglobal Swedish Deep Sea Expedition (1947–1949) on the research vessel *Albatross*. Using new technology developed by Börje Kullenberg in Gothenburg, the expedition retrieved cores up to 15 m (45 ft) long, with records reaching back 500,000–1,000,000 years. Many fundamental paleoceanographic concepts were established by the geologists who analyzed these cores, including the role of trade winds in promoting glacial-age upwelling, the changing supply of North Atlantic Deep Water, the cyclicity of

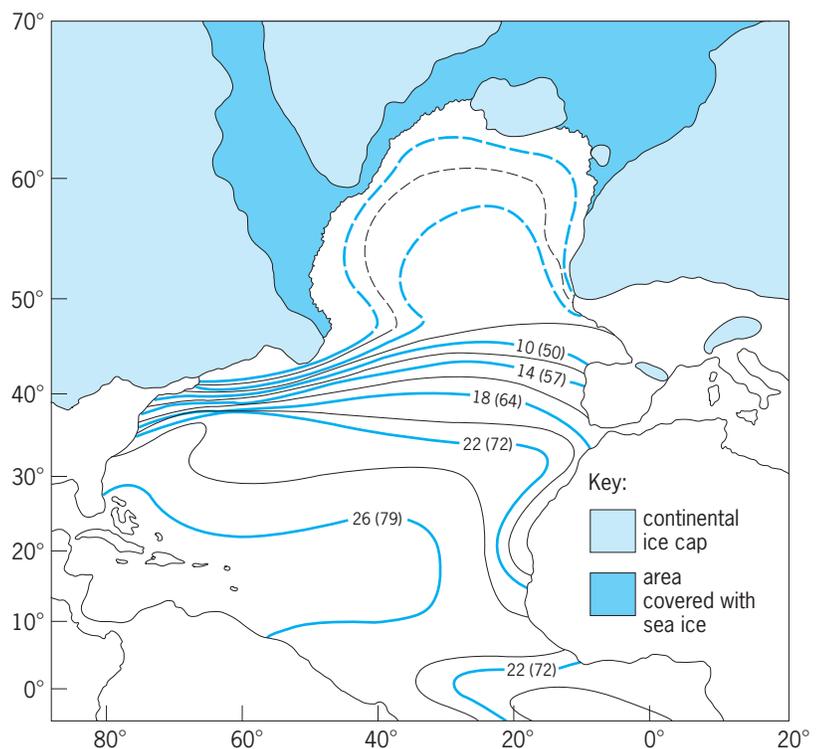


Fig. 1. August temperature of sea surface waters [isotherms in °C (°F) in the North Atlantic during the height of the last glacial 18,000 years before present] broken lines represent extrapolated isotherms (after A. McIntyre and N. G. Kipp, *Geol. Soc. Amer. Mem.*, 145:61, 1976). Recent research suggests that the coastal waters off Norway were open during summer.

climatic change in the late Quaternary, and large-scale shifts in biogeographic boundaries.

A quantum jump in paleoceanographic research resulted from the initiation of deep-sea drilling using the research vessel *Glomar Challenger* (1968). Enormous blank regions on the world's map suddenly became accessible for detailed exploration far back into geologic time, that is, into the Early Cretaceous. Highlights of the first decade of drilling results include the documentation of cooling steps as the planet moved into the present ice age; the reconstruction of long-term fluctuations in the carbonate compensation depth; the documentation of large-scale salt deposition in an isolated Mediterranean basin; and the discovery of temporary anoxic conditions in the Cretaceous deep sea.

Other important developments, in the decades following the *Albatross* expedition, were associated with the collection of large numbers of piston cores from many parts of the ocean, the improvement of analytical instrumentation, and the introduction of mathematical statistics to the interpretation of fossil assemblages. In the 1970s, a different emphasis on climatic change arose, stemming from concerns about the impact of human activities on climate, that is, the expectation that the increase in carbon dioxide in the atmosphere would produce global warming. A temperature map of the glacial ocean (the so-called 18,000-year map) resulted from the efforts of the CLIMAP group (Fig. 1), which provided an opportunity to test the application of climate models to conditions that are funda-

mentally different from those of today. See CLIMATE MODELING.

**Quaternary research.** A useful time scale for the ice-age cycles of the Quaternary first emerged from the combination of magnetic reversal stratigraphy and oxygen isotope measurements in the same piston cores. Based on these data, it was possible to show that changes in deep-ocean sedimentation (and hence climate) are influenced (or governed) by orbital forcing, that is, changes in seasonal insolation intensity in high latitudes. However, the physical mechanisms remain unknown by which Milankovitch forcing (changes of up to 8% in the amount of summer sunlight reaching the high latitudes of the Northern Hemisphere due to changes in the geometry of the Earth's orbit with time) is translated into changes of ice mass (and hence of sea level). In particular, it is not clear why there is a pronounced change in the response to forcing 900,000 years ago, when obliquity-dominated cycles gave way to much longer cycles soon displaying a period of about 100,000 years (100 kyr) [the Mid-Pleistocene climate shift; Fig. 2]. See CHEMOSTRATIGRAPHY; GEOLOGIC THERMOMETRY; INSOLATION; MAGNETIC REVERSALS; PALEOCLIMATOLOGY.

One likely origin for the 100-kyr cycles of the late Quaternary is a general increase of ice mass throughout the previous 2 million years, together with sinking and erosion of the ice-bearing continental crust. An increased amount of ice was thus based below sea level during glacial periods. Upon a minor rise of sea level, at times of exceptionally warm summers

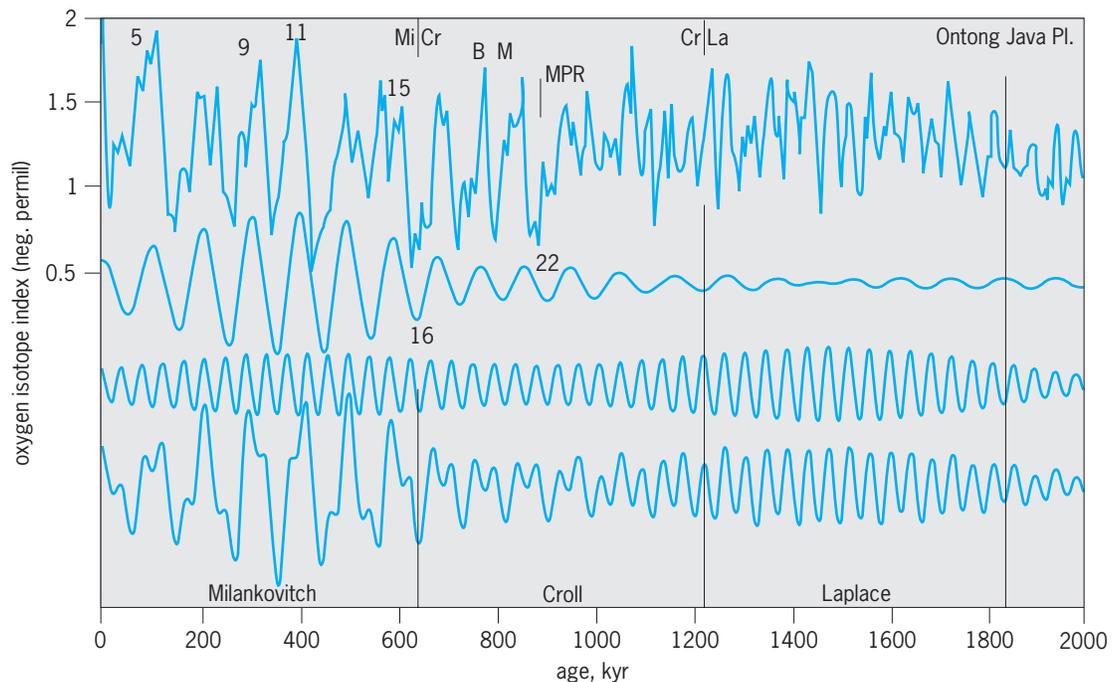
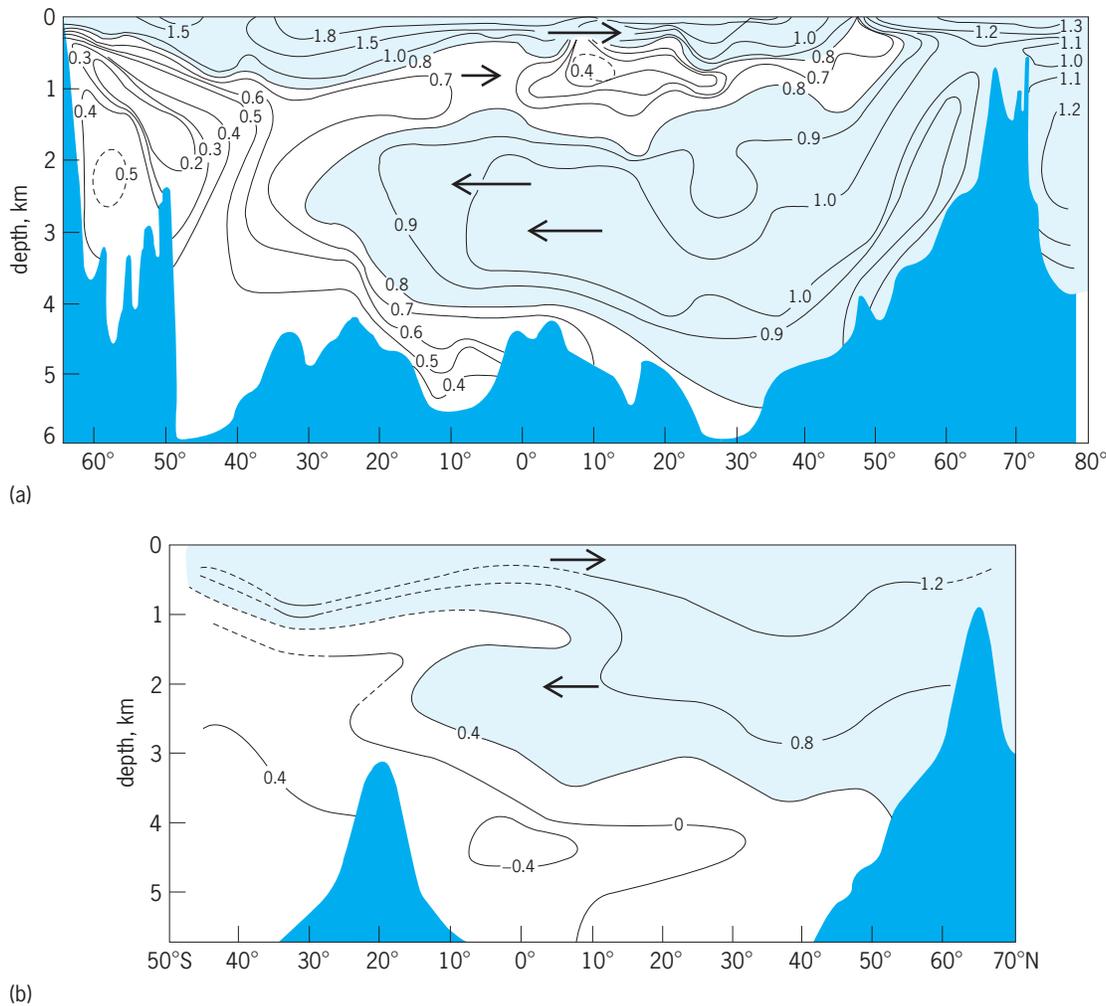


Fig. 2. Oxygen isotope values of the planktonic foraminifera *Globigerinoides sacculifer* from Ontong Java Plateau (Site 806, ODP Leg 130) are shown in the top curve. The two curves in the middle show 100-kyr (eccentricity) and 41-kyr (obliquity) cycles extracted from the isotope curve. The bottom curve is a combination of the two middle curves. The good agreement between top and bottom curves is an indication for orbital forcing of climate variability. The last 2000 kyr can be divided into three periods: Milankovitch period dominated by eccentricity, Laplace period showing obliquity-dominated variations and a transition period (Croll). MPR = mid-Pleistocene revolution. (Modified from W. H. Berger and G. Wefer, *Naturwissenschaften*, 79:541–550, 1992)



**Fig. 3.** Deep-water patterns and flow in the South Atlantic (a) present and (b) last glacial maximum at about 20,000 years ago. Stratification is exhibited by the pattern of  $\delta^{13}\text{C}$  values of dissolved inorganic carbon (‰ parts per thousand). During the last glacial period, the bottom-near layer of cold Antarctic Water tended to thicken, as the overlying North Atlantic Deep Water flow was reduced. (Modified from G. Wefer et al., *The South Atlantic: Present and Past Circulation*, Springer, 1996)

in northern latitudes, such marine-based ice can become unstable and move out to sea. As a result, support is removed from adjacent land-based ice, which then collapses in surges. Such ice collapse, in combination with changes in ocean circulation providing for changes in heat budget of high northern latitudes, could produce the terminations (that is, sudden ice wasting following maximum glaciations) that dominate the late Quaternary climate history. Thus, the 100-kyr cycles are best understood as an internal oscillation based on ice-shield instability, forced by the Milankovitch mechanism (summer insolation in high northern latitudes). Direct evidence for inherent instability comes from the quasi-cyclic collapse of modestly sized ice masses in Canada (Heinrich events). These iceberg floods occur roughly every 7000 years. See QUATERNARY.

**Heat conveyor and abrupt climatic change.** An important concept concerning the role of the ocean in abrupt climatic change, such as seen during deglaciation, is the Atlantic heat conveyor, which is associated with anti-estuarine circulation in the North Atlantic. Water from the South Atlantic moves north

across the Equator at shallow depths to replace North Atlantic Deep Water, which forms in the seas adjacent to Greenland due to cooling of salty surface waters, and moves south across the Equator at depth. At present the Atlantic heat conveyor is very effective in warming western Europe and the northern regions. During glacial times, North Atlantic Deep Water formation was greatly reduced (Fig. 3). A slowing down or shutting down of this conveyor (for example, from meltwater input) greatly changes the heat budget of the North Atlantic. Modeling suggests that sudden warming or a large input of fresh water in high latitudes could cause the conveyor to stop, with the oceanic system moving into another equilibrium. The speed of this transformation will depend on the feedback processes involved, which generally take place within a few centuries. According to the circulation models, however, it is possible that strong forcing could change the deep circulation completely or collapse the conveyor in less than 10 years. Once the conveyor was stopped, self-stabilization effects could hinder the resumption of the circulation, even if the induced forcing was no

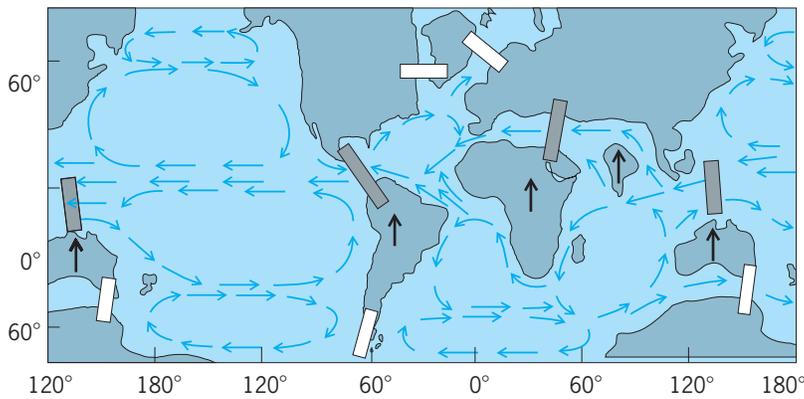


Fig. 4. Geography of the middle Eocene (45 Ma) and major critical valve points for ocean circulation. Tropical valves are closing (shaded rectangles) and high-latitude valves are opening (white rectangles) throughout the Cenozoic. Arrows indicate direction of circulation. (After B. U. Haq, in E. Seibold and W. H. Berger, *The Sea Floor*, Springer, 1993)

longer effective. These modeling results, combined with the paleoceanographic evidence, have important implications for future global warming. It is possible that the Atlantic deep circulation will decrease significantly in the twenty-first century within the framework of the increased greenhouse effect. Although this prognosis is subject to large uncertainties (mainly due to the low resolution of the ocean models), historical experience leaves no doubt that the scenario is a possible one. See GREENHOUSE EFFECT.

**Cenozoic research.** The Cenozoic Era (Age of Mammals) comprises the Paleogene Period (65–25 million years ago, or Ma) and the Neogene Period (25 Ma to present). During the Cenozoic, the Tethys seaway, which allowed free exchange between the major ocean basins in tropical regions, closed, and the main mixing region of the ocean's water masses was transferred to the circumpolar Southern Ocean, ensuring a general cooling trend in deep-water temperatures (Fig. 4). Marine life in the early Paleogene (Paleocene) was marked by recovery from the end-of-Cretaceous catastrophe. The deep ocean was as warm as ever, but cooled from about 50–40 Ma on. The end of the Paleocene witnessed a remarkable warming spike, apparently linked to a pulse of methane from the large-scale melting of methane clathrates. The event produced the widespread extinction of benthic foraminifers. In the Eocene, diversity of marine organisms reached a maximum. Around 40 Ma, there was a major change in the ocean. Zones of high productivity developed along the Equator and in coastal regions, extracting silicate and phosphate from the ocean at large, thus impoverishing the pelagic realm. Sea level dropped markedly, and much carbonate was transferred from shelves to the deep ocean. Deep-water sources moved from subtropical regions to higher latitudes. Diversity of both plankton and benthic organisms reached a low point within the Oligocene (the last period of the Paleogene, 36–25 Ma). In the middle to late Oligocene, the first large glaciers grew in Antarctica, as seen in ice-rafted debris around the continent. See EOCENE; HYDRATE; PALEOCENE.

In the Neogene, the ocean acquired thoroughly modern characteristics, such as cold deep waters, a strong thermocline, vigorous surface currents, and increased upwelling along the Equator and in eastern boundary coastal oceans. Major geographic events included the opening of the Drake Passage, closing of the Indonesian Passage, isolation of the Mediterranean, and closing of the Panama gateway. Deep-water temperatures decreased further. This cooling occurred in steps, presumably as a result of critical changes in geography (mountain building), as well as climatic threshold events (snow and ice buildup first on and around Antarctica and later in northern latitudes, with attendant positive feedback on cooling from albedo increase). The possibility of positive feedback on cooling from the carbon cycle is of interest, involving increased removal of carbon dioxide from the atmosphere from weathering of fresh rocks (mountain buildup), from increased organic carbon deposition in upwelling areas (thermocline effect), from increased burial in terrigenous marine sediments, and from increased uptake by the ocean (alkalinity increase from transfer of carbonate deposition from shelves to deep ocean, general cooling of the ocean). See CENOZOIC.

**Cretaceous research.** The ocean of the Cretaceous Period (140–65 Ma) was fundamentally different from the present one, because of geography and a generally warmer climate. Large areas of the continents were flooded during much of the period, so marine sediments (especially carbonate) tended to accumulate in shallow seas and the deep ocean was starved. Deep waters were relatively warm (perhaps comparable to today's Mediterranean). The oxygen supply was modest, because of the warm temperature and a more sluggish circulation. Productivity (it may be assumed) was greatly reduced because of a shortage of nitrate (which is used as a source of oxygen by certain bacteria, when oxygen is in short supply) and perhaps a shortage of trace metals (very little windblown material entered the ocean).

Of special interest are organic-rich deposits that occur in thin layers within sequences of clay or chalk, quite commonly in cyclic fashion suggesting orbital forcing. The most widespread black shale deposits of this kind are found during the Aptian and early Albian (110–100 Ma) in the South Atlantic. At the same time, this basin was a long narrow seaway, providing a cul-de-sac opening to the south with opportunities for pulsed estuarine circulation pattern. Black shales also occur elsewhere, however, so some deposits imply a global factor rather than a regional one. A global signal, apparently, is contained in the record of the differences in carbon-13 ratios ( $\delta^{13}\text{C}$ ) of pelagic marine limestones, with major excursions in the Aptian and at the Cenomanian-Turonian boundary. It has been suggested that major volcanic activity (related to mantle plumes reaching the Earth's surface and producing flood basalts such as Ontong-Java Plateau) provided for conditions favorable for the accumulation of organic-rich sediments in the deep ocean. If true, volcanism might be ultimately responsible for a large portion of the marine



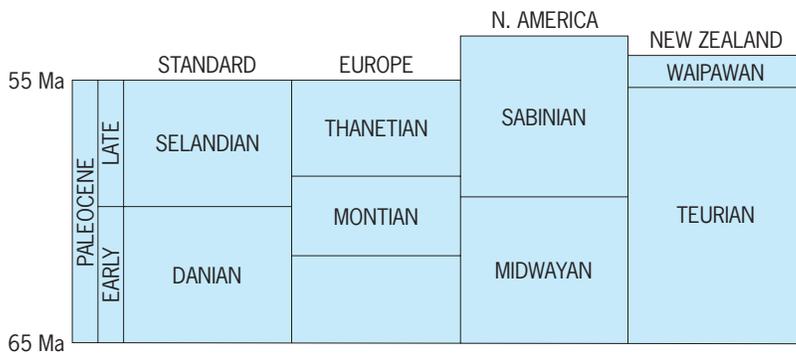


Fig. 1. Standard Paleocene stages and their temporal equivalents in Europe, North America, and New Zealand.

for the stage. Later workers have amended the concept of standard Selandian to include all of the temporal equivalent interval of the Late Paleocene series, although the debate about the lithostratigraphic continuity within these sections continues. This stage is, nevertheless, preferable to the previously used upper Paleocene standard stage of Thanetian, which spans only the very upper part of Late Paleocene.

Regional subdivisions of the Paleocene include the Montian Stage, proposed by G. Dewalque in 1868, with a type locality near Mons, Belgium. Like the Danian, the Montian Stage has also been a subject of much discussion. It is now considered to be temporally equivalent to the youngest part of Danian and older part of Selandian (Fig. 1). The Thanetian Stage in Britain was based on the marine Thanet Sands of the Isle of Thanet in Kent. It was first raised to a stage level by E. Renevier in 1873 and is time-equivalent to the upper part of Selandian. The Landenian Stage was proposed by A. Dumont in 1839. It also has its type locality in Belgium and is often used in northwestern Europe as a stage younger than the Montian. Detailed studies have shown the Belgian Landenian also is equivalent to the British Thanetian. In the United States the Gulf coast Paleocene subdivisions include the Midwayan and the older part of the Sabinian stages. The base of the Midwayan shows a major faunal and lithological hiatus. In California, local stratigraphers have often used the Ynezian Stage to represent an informal equivalent of the upper Paleocene. In the former Soviet Union, the Kachinian Stage has been shown to encompass much of the Paleocene, extending slightly into the earliest Eocene. In New Zealand, the Teurian Stage spans most of the Paleocene, with the overlying Waipawan Stage ranging into the youngest Eocene. In Australia, the Late Paleocene strata are sometimes ascribed to the regional Wangerripian Stage. Paleontologists working with vertebrate fossils have often used their own informal age classifications for assemblage associations. In North American, they subdivide the Paleocene into the Puercan, Torrejonian, and Tiffanian ages, whereas in Europe the terms Dano-Montian and Cernyasian have often been used to span the Paleocene in vertebrate paleontological literature. See STRATIGRAPHY.

**Tectonics, oceans, and climate.** Several major tectonic events that began in the Mesozoic continued into the Paleocene. For example, the Laramide Orogeny that influenced deformation and uplift in the North American Rocky Mountains in the Mesozoic continued into the Paleocene as a series of diastrophic movements, which ended abruptly in the early Eocene. See OROGENY.

On the ocean floor the most notable tectonic events were the separation of the Seychelles from the rapidly northward-moving Indian plate in the early Paleocene, and the initiation of sea-floor spreading in the Norwegian-Greenland Sea, between North America and Greenland in the late Paleocene. The Indian plate had broken away from eastern Gondwanaland in the Late Cretaceous and started moving relatively rapidly northward some 80 m.y. ago. In the early Paleocene, the spreading ridge between Madagascar and India jumped northeastward toward India. This initiated the spreading between India and the Seychelles Platform and the formation of the Chagos-Laccadive transform ridge.

During the Paleocene Epoch, the Indian plate continued its rapid flight across the eastern Tethys Ocean, the ancestral seaway that occupied the position of the modern northern Indian Ocean, to eventually collide with the Asian mainland in the mid-Eocene around 50 m.y. ago. In the Paleocene, the eastern Tethys was also characterized by another very active transform, the Ninetyeast Ridge. This ridge was formed when the Indian plate moved over a fixed mantle plume hot spot in the Late Cretaceous. See PLATE TECTONICS.

The lower-latitude shallow-water seas, as exemplified by the Tethys seaway, received thick deposits of calcium carbonates during the Paleocene. However, the seaway became progressively narrower and shallower, and the nature of carbonate accumulation changed correspondingly.

The sedimentary record of the Paleocene in the northwestern Atlantic indicates that a relatively calm regime of predominantly calcareous sediments typical of the Late Cretaceous was largely replaced by facies that represents more vigorous bottom waters that packed great erosive power due to increased convective overturn and dynamic ocean bottom currents. Although the deep-water connection between the North and South Atlantic was already established by the Early Paleocene, the South Atlantic sedimentary record indicates that in this basin the extensive erosion on the sea floor did not begin until the Late Paleocene. Sedimentary hiatuses representing this dynamic change in bottom-water regime during various times in the Paleocene are also common in other oceans. Thus, the Paleocene deep ocean can be said to have been characterized by extensive erosion and redeposition of sediments on the deep sea floor, reflecting expanded bottom-water activity, compared to the Cretaceous. See BASIN.

As a whole, the Cenozoic Era is characterized by a long-term withdrawal of the seas from coastal and inland oceans. In the latest Cretaceous, the global sea level had already begun to fall from the all-time

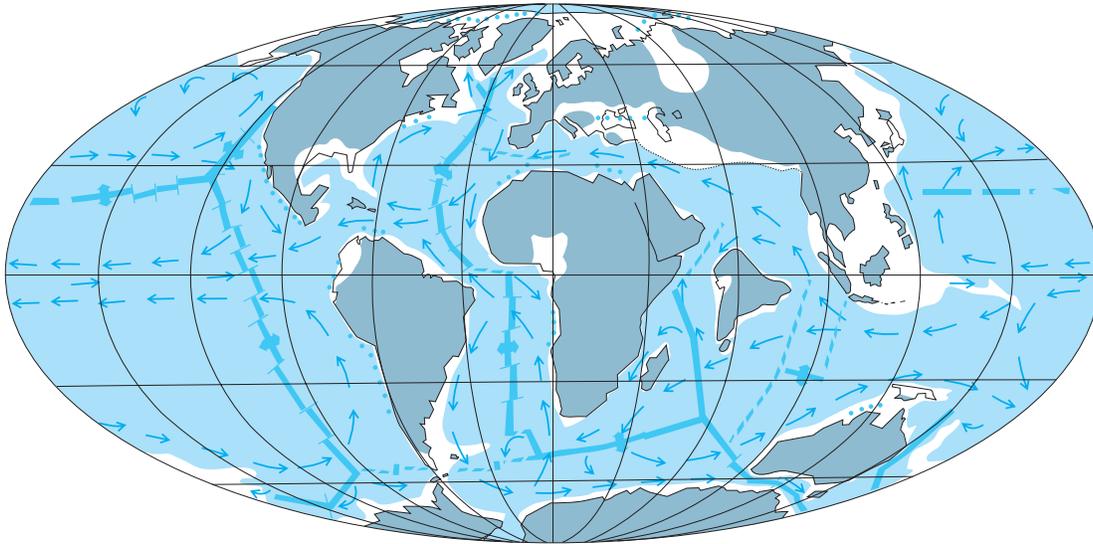


Fig. 2. Paleogeography, oceans, and circulation patterns of the Paleocene. (After B. U. Haq and F. W. B. van Eysinga, *Geological Time Table*, Elsevier, Amsterdam, 1998)

Mesozoic high of the mid-Cretaceous. The Cretaceous-Tertiary boundary event is marked by a general transgression in the Danian, following a sea-level fall in the latest Cretaceous. This trend toward overall regressing seas was further accentuated in the later part of the Paleocene. Whereas the Danian is typified by relatively high sea levels, it is followed by a major fall in the early part of the Selandian that has been recorded in many parts of the world. The restriction of the seas in the Selandian is reflected in the sediment facies of the Paleocene, which changed from more carbonate-dominated during the relatively higher sea levels of the Danian, to more terrigenous and siliciclastic in the Selandian. Latest Paleocene once again saw a sea-level rise that continued into the early Eocene. The Selandian sea-level fall, which is estimated at around 120 m, was large enough that it could have exposed extensive areas of the continental margins. Major coastline retreats are followed by stream incision of the shelf, and accordingly there is evidence that in the Selandian the sea-level fall is associated with major episodes of canyon formation, cut by rivers that migrated to the shelves during the eustatic drop. *See* CONTINENTAL MARGIN.

Paleogeographic-oceanographic considerations of the Paleocene record (Fig. 2) suggest that the western Tethyan seaway between Europe and Africa was open and a circumglobal Tethys current flowed through it, dominating the tropical oceanographic scene of the Paleocene. Toward the north a major epicontinental seaway, the Ural Sea, separated Asia from Europe through much of the Paleocene and Eocene. Epicontinental gulfs also existed on the Asian, African, and North and South American continents. These shallow interior seas were sites of extensive carbonate deposition, and also may have been prone to high production and preservation of organic matter that is important for hydrocarbon source-rock accumulation.

The establishment of deeper connections between the North and South Atlantic in the Paleocene facilitated enhanced deep-water flow from the northern to the southern basin. Similar to the Cretaceous, the source of deep water in the Paleocene was most likely in the warm low and middle latitudes, rather than the cooler higher latitudes as in the Neogene. Warm saline bottom water was characteristic of this epoch. In the south, the Drake Passage between South America and Antarctica was still closed, although Australia had already separated from Antarctica by Paleocene time. The lack of circum-Antarctic flow precluded the geographic isolation of Antarctica and the development of cold deep water from a southern source. *See* PALEOCEANOGRAPHY; PALEOGEOGRAPHY.

The isotopic and paleontological climatic proxy indicators all point to an overall rise in global temperatures in the Paleocene that led to a period of peak warming in the latest Paleocene and Early Eocene. A mean sea-surface temperature of around 10°C (50°F) in the higher latitudes is indicated by oxygen isotopic analysis of marine plankton. A prominent, relatively cooler interval that coincides with the major lowering of sea level has been recorded in the Late Paleocene. The marine microplankton (foraminifera and calcareous nannoplankton) exhibit latitudinal migratory patterns that are consistent with major climatic fluctuations indicated by the Paleocene isotopic record. The Selandian sea-level lowering and concomitant climatic cooling is accompanied by a migration of high-latitude microplankton assemblage toward low latitudes in the Atlantic Ocean, while the latest Paleocene is characterized by a poleward migration of warm, low-latitude assemblages. *See* GEOLOGIC THERMOMETRY; MARINE SEDIMENTS.

Terrestrial floras and faunas corroborate the peak warming in the latest Paleocene and Early Eocene and suggest that the warm tropical-temperate belt may have been twice its modern latitudinal extent.

The temperate floral and faunal elements extended to 60°N, which has been used as an argument to invoke a very low angle of inclination of the Earth's rotational axis in the Paleocene-Eocene. Alternatively, the mild, equable polar climates and well-adapted physiological responses of plants and animals of those times to local conditions may be enough to explain the presence of a rich vertebrate fauna on Ellesmere Island in arctic Canada. See CLIMATE HISTORY; PALEOBOTANY; PALEOCLIMATOLOGY.

At the close of the Paleocene Epoch, a prominent carbon-isotopic ( $\delta^{13}\text{C}$ ) shift occurred in the global carbonate reservoir that coincides with the peak warming at this time. Recent studies have ascribed this to the dissociation of sediments on continental margins that contained methane hydrates. When hydrates dissociate due to reduced hydrostatic pressure or increased temperature on the sea floor, large quantities of methane can be released into the water and atmosphere. In the latest Paleocene, bottom-water temperature increased rapidly with a coincident negative shift of  $\delta^{13}\text{C}$  in the global carbon reservoir by 2.5%. This isotopic change was accompanied by important biotic changes in the oceanic microfauna and was synchronous in the oceans and on land. The rapid (<10,000 years) and prominent  $^{12}\text{C}$  enrichment of the global carbon reservoir cannot be ascribed to increased volcanic emissions of carbon dioxide, changes in oceanic circulation, and/or terrestrial and marine productivity. The increased flux of methane from gas-hydrates into the ocean-atmosphere system and its subsequent oxidation to carbon dioxide is held responsible for this isotopic excursion in the inorganic carbon reservoir. High-resolution data support the gas-hydrate connection to the latest Paleocene is abrupt climate change. Evidence from two widely separated sites, from the low-latitude and southern high-latitude Atlantic Ocean, indicates multiple injections of methane with global consequences during the relatively short interval at the end of the Paleocene. See HYDRATE.

**Life.** The Paleocene Epoch began after a meteorite struck the Earth, causing massive extinctions at the end of Cretaceous and decimating a large percentage of the terrestrial and marine biota. On land the last of the dinosaurs are the most familiar casualty of this event. In the oceans, all ammonites, genuine belemnites, rudistids, most species of planktonic foraminifera and nannoplankton, and marine reptiles disappeared at the close of the Cretaceous Period. Even though some groups, such as squids, octopus, nautilus, and a few species of marine plankton, survived, the genetic pool was relatively small at the dawn of the Tertiary Period. The recovery of the marine biota was, however, fairly rapid after the mid-Paleocene due to overall transgressing seas and ameliorating climates. By the Late Paleocene, the biota was well on its way to explosive evolutionary proliferation and high diversification of the Eocene. In the Paleocene, endemism in marine and terrestrial biota increased. For example, the larger foraminifera, *Nummulites*, thrived in the shallow

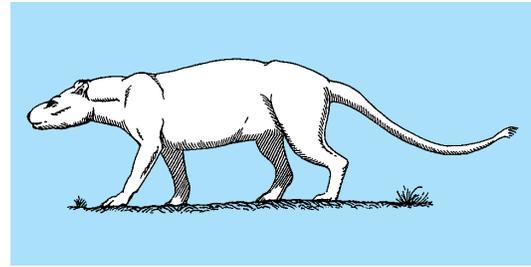


Fig. 3. Paleocene condylarth *Tetraclaenodon*. This five-toed ungulate was near the ancestry of the first horses. (After R. A. Stirton, *Time, Life and Man*, John Wiley, 1959)

seas of the Tethyan margins, but were excluded from the marginal seas of the New World. Planktonic microfauna also show increasing hemispheric and latitudinal provincialism. A new group of warm-water phytoplankton, the discoasters, made their first appearance in the Late Paleocene, and soon thereafter proliferated in the Eocene. The end of the Paleocene Epoch saw marked changes in deep-water circulation of the world ocean that resulted in a massive extinction of the benthic marine species. See EXTINCTION (BIOLOGY).

On land the large dinosaurs, which had been on the decline for over 20 m.y, died out at the close of the Cretaceous Period. However, smaller reptiles, including alligators and crocodiles, and some of the land flora escaped extinction and continued into the Paleocene. The Paleocene saw the first true radiation of mammals. The mammals of this epoch were characteristically primitive and small in size (50 cm or 20 in. or less). Their principal record is to be found in terrestrial deposits in Asia and North America. A typical example of Paleocene mammals is provided by the condylarths (odd-toed ungulates), which evolved from a Late Cretaceous mammalian lineage. These small animals (Fig. 3) were ancestral to many carnivores as well as the important group of perissodactyls, which include horses, rhinos, and tapirs. The appearance of the first true grasses at the close of the Paleocene may have helped the later radiation of these animals. Ancestral insectivores, rodents, and primates also had their beginning in the Paleocene. During the early part of this epoch, North and South America were connected temporarily, allowing free migration of animals, but were separated soon thereafter. This led to evolutionary isolation of South America and survival of an archaic fauna dominated by anteaters, armadillos, and opossums, until the Pliocene when the isolation finally ended following the connection of the two continents at the Panama isthmus. Similarly, as the continent of Australia became more isolated geographically, its mammalian fauna, such as the marsupials, became sequestered and more specialized. See DINOSAUR; MAMMALIA; PALEONTOLOGY. Bilal U. Haq

Bibliography. B. U. Haq and F. W. B. van Eysinga, *Geological Time Table*, Elsevier, Amsterdam, 1998; R. D. Norris and U. Roehl, Carbon cycling and chronology of climate warming during the Paleocene/Eocene transition, *Nature*, 401:775-778, 1999;

C. Pomeroy, *The Cenozoic Era: Tertiary and Quaternary*, 1982.

## Paleoclimatology

The study of ancient climates. Climate is the long-term expression of weather; in the modern world, climate is most noticeably expressed in vegetation and soil types and characteristics of the land surface. To study ancient climates, paleoclimatologists must be familiar with various disciplines of geology, such as sedimentology and paleontology, and with climate dynamics, which includes aspects of geography and atmospheric and oceanic physics. A compelling need exists for societies to be able to predict climate change. Understanding the history of the Earth's climate system greatly enhances the ability to predict how it might behave in the future.

Information about ancient climates comes principally from three sources: sedimentary deposits, including ancient soils; the past distribution of plants and animals; and the chemical composition of certain marine fossils. These are all known as proxy indicators of climate (as opposed to direct indicators, such as temperature, which cannot be measured in the past). In addition, paleoclimatologists use computer models of climate that have been modified for application to ancient conditions. *See* GEOLOGY; PALEONTOLOGY.

**Climate dynamics.** Like modern climatologists, paleoclimatologists are concerned with boundary conditions, forcing, and response. Boundary conditions are the limits within which the climate system operates. The boundary conditions considered by paleoclimatologists depend on the part of Earth history that is being studied. For the recent past, that is, the last few million years, boundary conditions that can change on short time scales are considered, for example, atmospheric chemistry. For the more distant past, paleoclimatologists must also consider boundary conditions that change on long time scales. Geographic features—that is, the positions of the continents, the location and orientation of major mountain ranges, the positions of shorelines, and the presence or absence of epicontinental seaways (large areas of the continents flooded during times of high sea level)—are important for understanding paleoclimatic patterns. In addition, the solar constant and the rotation rate of the Earth have changed through geologic time and also must be considered in studying paleoclimates, particularly before about  $4 \times 10^8$  years ago (shelly fossils first appeared in the record about  $6 \times 10^8$  years ago). Forcing is a change in boundary conditions, such as continental drift, and response is how forcing changes the climate system. Forcing and response are cause and effect in paleoclimatic change. Paleoclimatic change during the history of the Earth has taken place on time scales that are very short (thousands of years) and very long (hundreds of millions of years). *See* CONTINENTAL DRIFT; CONTINENTS, EVOLUTION OF; PALEOGEOGRAPHY; PLATE TECTONICS.

**Modeling ancient climates.** The record of climate in sediments and rocks is remarkable because it illustrates a multitude of abrupt and slow changes in climate throughout Earth history. The record is also incomplete. Many older rocks have been subject to erosion or metamorphism, and farther back in time the record becomes increasingly sparse. In addition, paleoclimatologists do not have true “paleo” barometers or thermometers. Rather, paleoclimate is studied indirectly through its imprint on sediments and organisms. Thus, ancient climates cannot be reconstructed on a global scale. However, the available data are useful if climate is modeled; the data can then be used to test aspects of the model predictions, and thereby the model itself.

Although many aspects of paleoclimatic change are actively studied, some topics are particularly important in that the studies are challenging traditionally accepted ideas. These topics include the role of mountain ranges in determining global climate patterns, the climate systems of supercontinents, warm polar climates, the role of ocean circulation in global climate, and the role of carbon dioxide in long-term climate change.

The study of paleoclimates has benefited greatly from the application of comprehensive computer models. The development of atmospheric and oceanic models based on fundamental physical laws of fluid motion, energy, and momentum was at the leading edge of research in the atmospheric and oceanic sciences during the 1980s. Very rapid development in predictive model capability has occurred because of increased supercomputer power, advances in programming, and a growing inventory of observations about the modern climate system. *See* CLIMATOLOGY; MODEL THEORY; SIMULATION.

**Paleoclimatic indicators.** Proxy indicators of paleoclimate are abundant in the geologic record. Many were first discovered during the nineteenth century and helped scientists and naturalists appreciate the dynamism of Earth history. Fossil palm leaves and crocodiles in the Canadian Arctic Islands and thick coal seams in Antarctica are among the many seemingly anomalous occurrences that indicate that climate has changed. Many of the changes are explained by continental drift, but other changes represent fundamental differences between modern climatic patterns and ancient ones.

**Sedimentary indicators.** Important sedimentary indicators forming on land are coal, eolian sandstone (ancient sand dunes), evaporites (salt), tillites (ancient glacial deposits), and various types of paleosols (ancient soils), such as bauxite (aluminum ore) and laterite (some iron ores). Coals may form where conditions are favorable for growth of plants and accumulation and preservation of peat, conditions that are partly controlled by climate, especially seasonality of rainfall. Eolian sandstones and evaporites are indicative of arid climates. Paleosols record water-table levels and intensity of weathering, both of which are dependent on temperature or rainfall. Certain marine sediments also provide important information about ancient atmospheric and oceanic

circulation patterns. These are bedded chert, which is formed from the siliceous shells of microscopic marine plants and animals, and phosphate, both of which form in upwelling zones, that is, zones of high biologic productivity in the oceans. At present, these zones owe their distribution to global wind patterns, and thus geologic evidence of upwelling helps in the interpretation of ancient atmospheric circulation. Extensive deposits of sediments rich in organic substances provide information about stagnation, which may be related to temperature, and productivity in the oceans. In addition to specific types of sedimentary rock, patterns of sedimentation also provide information about paleoclimates. Layering in lake deposits, for example, may be controlled by seasonal changes that are, in turn, controlled by climate. Much of the understanding of more recent climatic change has come from such deposits. See BAUXITE; CHERT; COAL; DEPOSITIONAL SYSTEMS AND ENVIRONMENTS; LATERITE; MARINE GEOLOGY; MARINE SEDIMENTS; PALEOSOL; SALINE EVAPORITES; UPWELLING.

**Fossil indicators.** Fossils provide information about climate mostly by their distribution (paleobiogeography), although a few specific types of fossils may be indicative of certain climatic conditions. The latter are usually fossils from the younger part of the geologic record and are closely related to modern species that have narrow environmental tolerances. Paleobiogeographic patterns, on the other hand, commonly follow climatic gradients such as temperature and rainfall; they are useful qualitatively, and sometimes quantitatively, for delineating those gradients. Paleobiogeographic patterns are especially useful in the earlier part of Earth history, because the evidence required for other methods usually was not preserved or is difficult to interpret.

In addition to indicator species and paleobiogeographic patterns, fossils, particularly plants, commonly show morphological features that are related to climate. The morphology of leaves of woody angiosperms (flowering plants) is controlled by mean annual temperature, mean annual range of temperature, and rainfall (Fig. 1). In addition, the thickness and structure of the cuticle (waxy outer layer) of angiosperm and other types of leaves are partly de-

pendent on water supply. Leaf-margin analysis of angiosperms has provided a quantitative tool for the estimation of mean annual temperature; the proportion of smooth-margined versus toothed-margined leaves is highly correlated with this climatic parameter. Finally, growth rings in fossil wood are sensitive to frosts, rainfall, and temperature during the growing season. These methods are useful only for the latter part of Earth history, after trees and flowering plants evolved. See BIOGEOGRAPHY; DENDROCHRONOLOGY; FOSSIL; PALEOBOTANY.

**Geochemical indicators.** The third type of information available for documenting paleoclimatic patterns and change is stable isotope geochemistry of fossils and certain types of sedimentary rock. Many elements that are used by organisms to make shells, teeth, and stems occur naturally in several different forms, known as isotopes. The most climatically useful isotopes are those of oxygen (O). Most oxygen occurs as  $^{16}\text{O}$ , an atom with eight protons and eight neutrons in the nucleus; however, some oxygen occurs as  $^{18}\text{O}$ , which has two extra neutrons. The rate at which the two isotopes are cycled through the oceans and atmosphere and taken up by organisms differs in response to a number of factors, including temperature; the biology of the organisms also has a strong effect, so that only a few types of organisms can be used for oxygen isotopic studies. In general, organisms living in tropical waters have a higher proportion of  $^{16}\text{O}$  in their shells compared with those living in polar waters. In addition, the proportions of the two isotopes available in the global system depends on the amount of ice on the Earth's surface. Although the effects of temperature change and ice volume change can be difficult to distinguish, the analysis of oxygen isotopes has provided a powerful quantitative tool for the study of both long-term temperature change and the history of the polar ice caps. Isotopic studies are less reliable for the earlier part of Earth history, because shell material from older rocks tends to be poorly preserved and because the biology of very ancient organisms is unknown. See GEOCHEMICAL PROSPECTING; ISOTOPE.

**Causes of paleoclimatic change.** The rarest of paleoclimatic changes is that which can be attributed confidently to a single cause. A great deal of research in paleoclimatology has been devoted to understanding the causes of climatic change, and the overriding conclusion is that any given shift in the paleoclimatic history of the Earth was brought about by multiple factors operating in concert. Nevertheless, some forcing factors have been identified as being particularly important. The most important forcing factors for paleoclimatic variation are changes in paleogeography and atmospheric chemistry and variations in the Earth's orbital parameters.

Paleogeographic forcing operates on relatively long time scales, millions to tens of millions of years, and occurs in a variety of modes. Changes in continental positions may close off or open up new routes for ocean currents, which can change the distribution of temperature over the Earth's surface. For example, the opening of the Drake

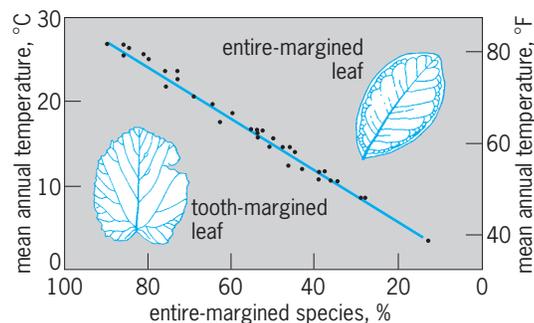


Fig. 1. Correlation of leaf-margin shape with mean annual temperature. (After J. A. Wolfe, *Temperature parameters of humid to mesic forests of eastern Asia and relation to forests of other regions of the Northern Hemisphere and Australasia*, U.S. Geol. Surv. Prof. Pap. 1106, 1979)

Passage between South America and Antarctica is regarded as a contributing factor to the formation of the Antarctic ice cap. While land connection between the two continents existed, the circumpolar current was deflected equatorward along the western side of the coast while relatively warm water was deflected poleward along the eastern side. With the opening of the Drake Passage, however, the circumpolar current became continuous, isolating Antarctica from the warm, low-latitude waters. Although the paleogeographic change that brought about this isolation—the drifting of Antarctica and South America away from each other—occurred slowly, the climatic change that resulted was relatively rapid. This is an example of a paleoclimatic threshold; the climate system seems to resist change until finally the forcing becomes too great, and then the adjustment takes place quickly.

Sea-level changes also may have brought about paleoclimatic change. Sea level is partly controlled by changes in ice volume at the poles, and thus may sometimes be an effect, rather than a cause, of climatic change. However, sea level also changes in response to mountain-building activity and continental drift, and these changes may be forcing factors for paleoclimatic change. In general, times of high sea level were times of warmth. However, this pattern is not necessarily true in detail, indicating that other factors also must play a role. Among these factors are the location, size, and orientation of mountains and epicontinental seaways. In addition, geochemical models suggest that times of high sea level were also times of high carbon dioxide content in the atmosphere, so the direct paleoclimatic forcing might have been carbon dioxide, not sea level.

Most paleoclimatologic research on atmospheric chemistry has been directed toward modeling the variations in carbon dioxide content of the atmosphere. Carbon dioxide is well known as a so-called greenhouse gas that traps heat at the Earth's surface. Carbon dioxide content is thought to have varied in the past, and strong evidence for such variations during the recent past (the last few thousand years) has been gathered by using detailed records of atmospheric bubbles in ice cores and stable isotopes (Fig. 2). In addition, the carbon cycle is relatively well known, and changes in carbon dioxide content in the atmosphere must have taken place, because large fluctuations are observed in the fossil record for other reservoirs for carbon. Finally, carbon dioxide changes appear to be the only mechanism that can account for some global temperature changes. See ATMOSPHERIC CHEMISTRY; BIOGEOCHEMISTRY; CARBON DIOXIDE; GREENHOUSE EFFECT.

Characteristics of the Earth's orbit around the Sun vary regularly and lead to cyclical variations in the amount and distribution of solar radiation reaching the Earth. The three characteristics most important to paleoclimatic studies are precession, obliquity, and eccentricity. Precession of the equinoxes determines the time of year when the Earth is closest to the Sun; in the present geological era, this occurs in January. Precession varies with a period of

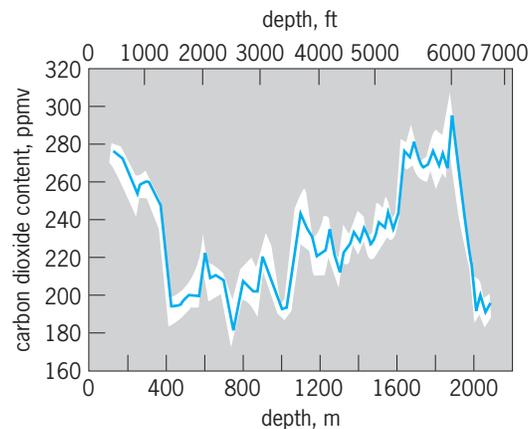


Fig. 2. Carbon dioxide (CO<sub>2</sub>) from bubbles in the Vostok ice core from Antarctica. The core spans about 160,000 years. White area represents the data envelope. (After J. M. Barnola et al., *Vostok ice core: A 160,000-year record of atmospheric CO<sub>2</sub>*, *Nature*, 329:404–414, 1987)

about 19,000–23,000 years. Obliquity is the tilt of the Earth's axis relative to the plane of the Earth's orbit around the Sun; at present, obliquity is about 23.5°. Obliquity varies with a period of about 41,000 years. Eccentricity is the shape of the Earth's orbit around the Sun, which varies from circular, allowing solar input to be constant throughout the year, to oval, which creates strong annual variability. Eccentricity varies with a period of about 100,000 years. These are known as Milankovitch cycles, after the Yugoslavian astronomer who developed mathematical models of the cycles. Because the different cycles possess different periods, the paleoclimatic response to Milankovitch cyclicity has been complicated. Nevertheless, Milankovitch cyclicity has been described from sedimentary rocks of many parts of Earth history. See EARTH ROTATION AND ORBITAL MOTION; PRECESSION OF EQUINOXES.

**Paleoclimatic history.** The paleoclimatic history of the Earth can be divided into several stages, based on events and on the increasing resolution in the geologic record toward the present. See GEOLOGIC TIME SCALE.

*Pre-Phanerozoic climate.* Paleoclimatic history before the appearance of the first shelly fossils is poorly known because many of the strongest types of evidence for paleoclimatic studies do not exist in pre-Phanerozoic rocks. It is clear, however, that the interval was punctuated by extensive glaciations, as indicated by the abundance of tillites. See CONGLOMERATE.

*Pre-Pangaeian Paleozoic.* During the early part of the Phanerozoic, land consisted of Gondwana, which would later break up into the modern continents of the Southern Hemisphere and pieces of southern Asia, and numerous small continents, which were to become the modern Northern Hemisphere continents. Sea level was high and temperatures generally warm through much of the interval, although sea level and temperatures were low at the very beginning of the interval and large ice sheets formed near the South Pole at the end of the Ordovician Period and again during the early Carboniferous. Reefs

were widespread during the Ordovician through Devonian, and peat swamps were widespread during the Carboniferous. See BOG; CARBONIFEROUS; DEVONIAN; ORDOVICIAN; PEAT; REEF.

**Pangaean climate.** By the end of the Paleozoic, nearly all the continents had come together to form the supercontinent, Pangea. At the same time, global climate changed dramatically. Abundant evidence exists for aridity or strongly seasonal rainfall. These changes were brought about by the influence that the large landmass had on atmospheric circulation, creating an intensely seasonal climate. Sea level was very low at this time, increasing the size of the exposed continent and augmenting its climatic influence. Temperature is perhaps most difficult to determine for this time because biogeographic patterns, so useful in the previous interval, are extremely weak, and oxygen isotopic determinations of paleotemperatures are still unreliable. See PALEOZOIC.

**Late Mesozoic to early Tertiary climate.** With the breakup of Pangea, sea level rose and climate again changed, becoming wetter in many parts of the globe and probably cooler. Leaf-margin analysis is possible from this time forward, because flowering plants had evolved and dispersed across the globe; and relatively reliable and abundant oxygen-isotope analyses also begin to be possible for about the same time. The middle Cretaceous through early Tertiary was warm relative to later paleoclimates, but it was punctuated by cooling events. In addition, coal, phosphate, and marine rocks rich in organic substances were abundant at various times during this interval. Particularly impressive was a short-lived but extremely intense episode of deposition of organic-rich rock worldwide. See CRETACEOUS; TERTIARY.

**Late Tertiary.** Climate cooled substantially in several steps through the latter part of the Tertiary Period (Fig. 3). Ice caps at both poles were formed, first in the Southern Hemisphere. Rapid and significant changes in oceanic biogeographic patterns took place, and the foundations of the modern vegetation patterns were established.

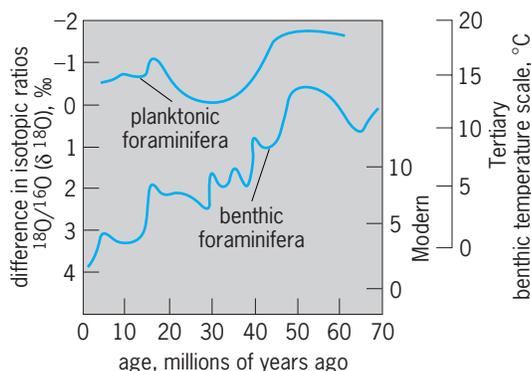


Fig. 3. Temperature curve for the Tertiary Period, based on oxygen isotopes in foraminifera. The modern scale takes into account the effect of ice caps on isotope ratios.  $^{\circ}\text{F} = (^{\circ}\text{C} \times 1.8) + 32$ . (After S. M. Savin, *The history of the Earth's temperature during the past 100 million years*, *Annu. Rev. Earth Planet. Sci.*, 5:319–355, 1977)

**Quaternary.** The Quaternary Period includes the major Northern Hemisphere glaciations of the last  $2 \times 10^6$  years. Although there is no reason to regard these as different in scale from earlier glacial epochs in Earth history, the proximity of these events to the present day has permitted paleoclimatologists to study them in far greater detail than is possible for the earlier events. Much of the understanding of the dynamics of continental glaciations and paleoclimate has come from studying the global patterns of the Quaternary. See CLIMATOLOGY; GLACIAL EPOCH; GLACIOLOGY; QUATERNARY.

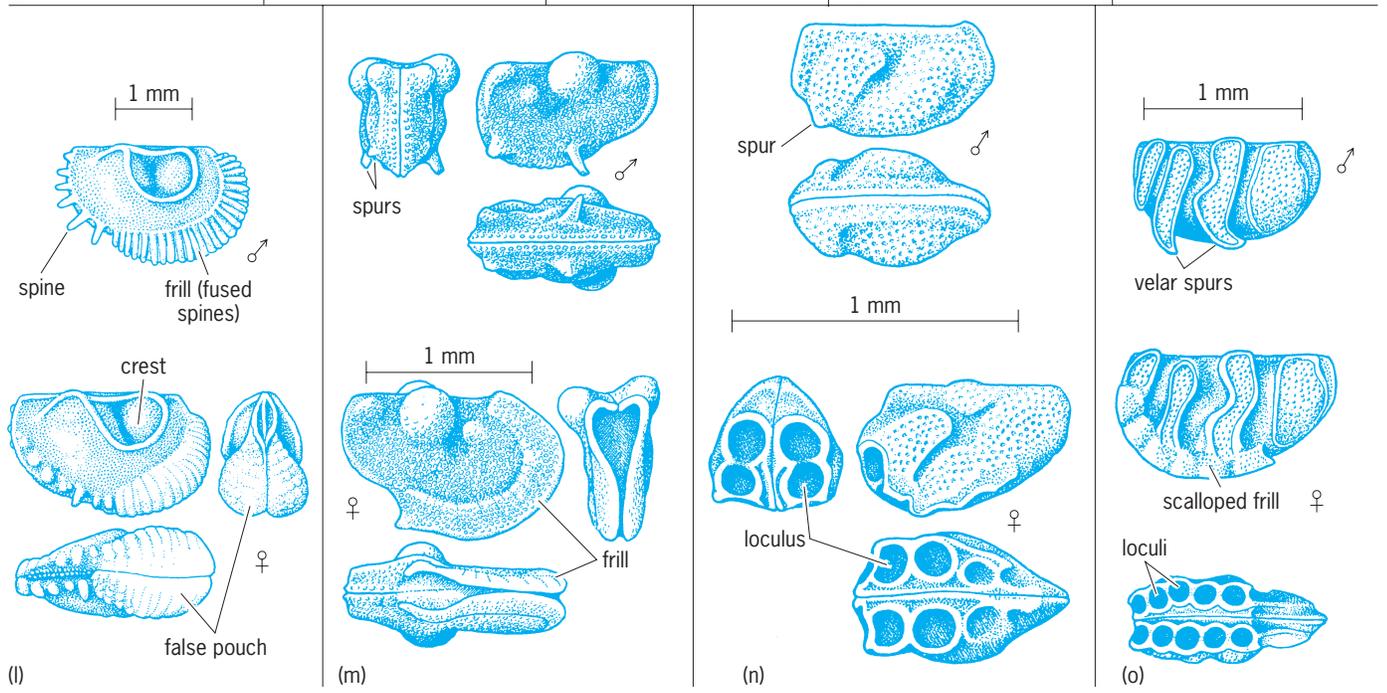
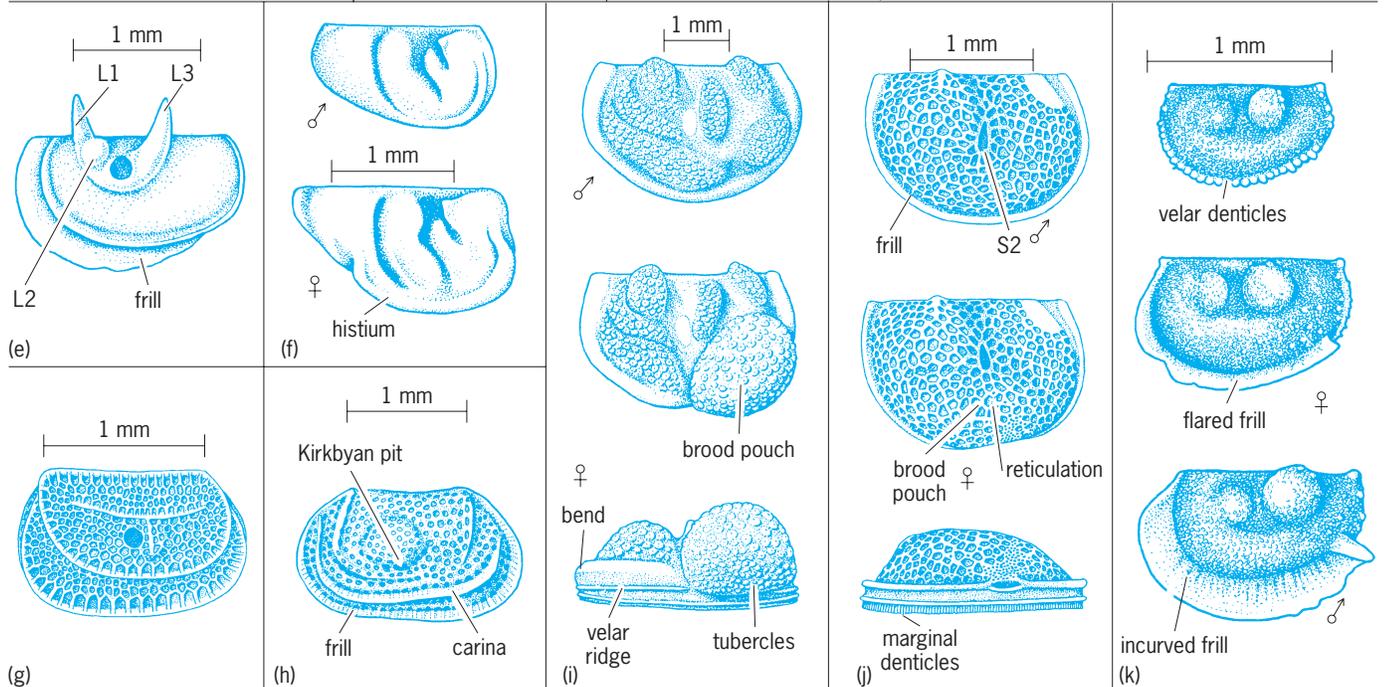
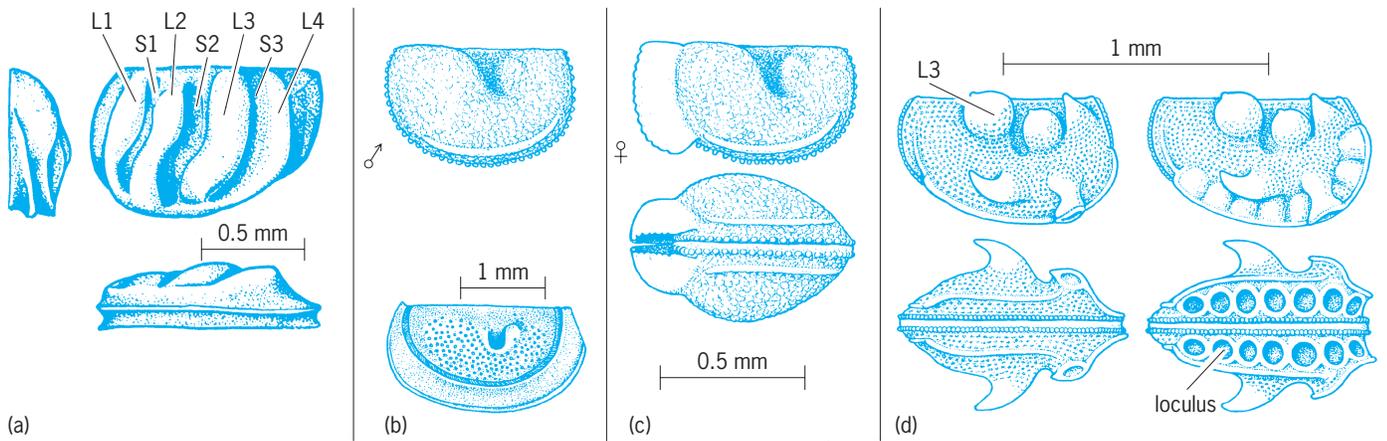
Judith Totman Parrish; Eric J. Barron

**Bibliography.** A. Berger et al. (eds.), *Milankovitch and Climate*, NATO AST Ser., vol. 126, 1984; T. J. Crowley, *The geologic record of climatic change*, *Rev. Geophys. Space Phys.*, vol. 21, 1983; L. A. Frakes, *Climates Throughout Geologic Time*, 1979; J. E. Kutzbach, *Modeling of paleoclimates*, *Adv. Geophys.*, vol. 28A, 1985; J. T. Parrish, *Climate of the supercontinent Pangea*, *J. Geol.*, vol. 101, 1993; P. V. Rich et al., *Evidence for low temperatures and biologic diversity in Cretaceous high latitudes of Australia*, *Science*, vol. 242, 1988; J. A. Wolfe and G. R. Upchurch, Jr., *North American nonmarine climates and vegetation during the Late Cretaceous*, *Palaeogeog. Palaeoclim. Palaeoecol.*, vol. 61, 1987.

## Paleocopa

Paleocopa, also called Palaeocopida, is an extinct order of the crustacean class Ostracoda. The order is divided into nine superfamilies, one of which, Barychilinaea, is only tentatively assigned to the order. A principal feature of the species in the order is their long, straight hinge that extends along the dorsal margin of the carapace and joins the two valves together (see *illus.*). As is true of most benthic ostracodes, the palaeocopes lack a frontal opening through which to extend their walking legs. Unlike modern ostracodes, however, they have no calcified inner lamella, and their muscle-scar patterns, which are quite useful in the taxonomy of

**Paleocopa.** (a) *Ogmoopsis nodulifera*, left valve; from the Ordovician of Sweden. (b) *Eurychilina subradiata*, right valve showing wide complete frill; Ordovician of Minnesota. (c) *Sulcicuneus porrectinatum*, male and female carapaces; Devonian of Michigan. (d) *Abditoloculina pulchra*, male and female carapaces; Devonian of New York; the bulbous L3 is typical of the family Hollinidae. (e) *Dicranella bicornis*, left valve; Ordovician of Minnesota. (f) *Sigmoopsis platyceras*, right valves, with histial dimorphism; Ordovician of Estonia. (g) *Amphizona asceta*, right valve, a species of the Arcyzonidae, the Devonian forerunners of the Amphissitidae. (h) *Amphissites marginiferus*, left valve, Carboniferous of North America. (i) *Beyrichia tuberculata*, right valves; Silurian of Europe. (j) *Hibbardia lacrimosa*, right valves; Devonian of New York. (k) *Hollinella dentata*, juvenile, left valves; Carboniferous of North America. (l) *Piretella acmaea*; the carapace of the dimorphic false pouch is formed by incurved, fused, hollow spines; Ordovician of Estonia. (m) *Falsipollex laxivelatus*, carapaces; Devonian of Michigan. (n) *Tetrascalculus mirabilis*, carapaces; Carboniferous of Illinois. (o) *Ctenoloculina cicatricosa*, carapaces; Devonian of Michigan.



many other groups of ostracodes, are very poorly known. See OSTRACODA.

The carapaces of many palaeocopes are marked by ornamentation in the form of lobes and sulci, which are designated L1, S1, L2, S2, . . . , to L4. Sulcus S2 is the most prominent and is present in all lobate forms. S2 is the external manifestation of the adductor muscle scar, which is located internally in the same spot. Appendages of the palaeocopes are largely unknown except for a few rather poorly preserved specimens.

**Sexual dimorphism.** In general, Ostracoda are characterized by pronounced sexual dimorphism. The males tend to have more elongated carapaces to accommodate their comparatively gigantic copulatory apparatuses. The carapaces of the females are likely to have a ventral swelling that provides space for carrying the eggs and brooding the young. Not all ostracodes show these differences between the sexes, but most do.

One of the most remarkable morphological features of many species of palaeocope ostracodes is that their sexual dimorphism is quite pronounced and is carried far beyond the rather simple sexual dimorphism that is described above. In typical instances, the males resemble the instars (immature forms), and the females have developed strongly modified morphology that is associated with reproduction—especially the development of pouches for carrying eggs and brooding the young.

The sexual dimorphism of palaeocopes can be of three general kinds: external, internal and distinct, or internal and indistinct. External brood pouches do not open into the interior of the carapace where the animal lives, an area referred to as the domicilium (illus. *b-d, k-o*). These external brood pouches are enigmatic because they are positioned on the anteroventral portion of the carapace. In this position it seems that the walking legs would be likely to disrupt the instars in the brood chamber or that the eggs or young might fall out when the valves are opened. Nevertheless, this plan seems to have worked well, for ostracodes with this kind of dimorphism were present for much of the late Paleozoic Era, an interval of several tens of millions of years.

In some instances the brood pouches open directly into the domicilium and are not separated from the domicilium by an internal partition. This indistinct internal dimorphism is shown in illus. *a* and *f*.

By far the most fascinating brood pouches, however, are those that are internal and distinct; that is, they open into the domicilium and are walled off from the domicilium except for a comparatively small opening (illus. *i* and *j*). For a long time the function of the brood pouches was unknown, and which dimorph was the male and which the female was much debated. Careful study, however, has revealed the presence of instars within the brood pouches, and since the time of that discovery the morphs with the pouches have been regarded as females and the others as males.

Given the prominence of sexual dimorphism among some of the palaeocops, it is surprising that

some of the superfamilies are not dimorphic at all, apparently not even to the extent described for the ostracodes in general (illus. *g* and *h*). Much more research on the palaeocopes is needed to establish conclusively this surprising evolutionary development and what may have brought it about.

**Ontogeny.** Arthropods, including the ostracodes, grow by molting, a process that involves shedding the external skeleton, growing quickly, and then secreting a new skeleton. Growth occurs only during molting. Ostracodes are nearly unique among the marine arthropods in having determinate growth. The trilobites and such other crustaceans as the crabs, lobsters, and shrimps continue to grow throughout their lives. The ostracodes, however, grow through only a set, genetically determined number of instars. When an ostracode reaches the adult stage and sexual maturity, it stops growing. The number of growth stages is constant at rather high taxonomic levels—usually at the superfamily level—and varies from seven to nine, or perhaps a few more.

This mode of growth provides the ostracodes and other arthropods with some distinct advantages and disadvantages. During molting, when a palaeocope passes from one instar to another, it is possible for it to undergo a pronounced change in its morphology. Most palaeocopes do not, in fact, alter their morphology radically during ontogeny except for the development of brood pouches by the females, but the potential is there. A disadvantage of growth by molting is that the ostracode is quite vulnerable to predation after it has shed one skeleton and before it has secreted another.

The study of the ontogeny of the palaeocopes presents paleontologists with a unique opportunity to learn about the pathways and mechanisms of evolution. The palaeocopes lend themselves to this sort of study because they molt, have a predetermined number of growth stages, and have pronounced sexual dimorphism that allows one to determine the sex of individuals and the sex ratios of populations. Thus, research on the palaeocopes can reveal the pathways their evolution has followed in a way that is possible for few other kinds of fossil organisms.

When it molts, a palaeocope ostracode roughly doubles its volume before it secretes a new carapace. The cube root of 2 (from doubling) is 1.26. This means that if an ostracode doubles in volume, it should increase in length, height, and breadth by 1.26. This relationship has helped paleontologists determine to which instar an ostracode carapace belongs, although the rule about doubling is only approximate and the exact size can be greatly affected by the environmental conditions.

**Paleoecology and biostratigraphy.** As is true of other benthic ostracodes, the paleocopes do not have a planktonic larval stage. As a result, they are quite limited in their means of dispersal, and few species are biogeographically widespread. Indeed, most forms are limited primarily to a single basin of deposition. Thus, they are not useful for long-distance, intercontinental correlation, but they are well suited for the study of paleobiogeography. If the sedimentary rocks

in two areas have paleocope faunas that comprise the same species, it is fairly certain that the areas were close enough together, at the time of deposition of the sediments that formed the rocks, to allow the paleocopes to walk from one area to the other or to be transported by fishes or marine currents.

A great deal of work remains to be done to decipher details of the paleoecology of the paleocopes, but they are a promising group for study of organism-sediment interactions because of their small size and intimate association with the sediment substrate. *See* PALEOECOLOGY.

Roger L. Kaesler

Bibliography. R. H. Bate, E. Robinson, and L. M. Sheppard (eds.), *Fossil and Recent Ostracods*, Ellis Horwood Limited, Chichester, 1982; R. L. Kaesler, Superclass Crustacea, in R. S. Boardman, A. H. Cheetham, and A. J. Rowell (eds.), *Fossil Invertebrates*, pp. 241-258, Blackwell Science, Cambridge, MA, 1987; R. C. Moore (ed.), *Treatise on Invertebrate Paleontology*, Part Q: *Arthropoda 3, Crustacea, Ostracoda*, Geological Society of America and University of Kansas Press, 1961; R. C. Whatley and C. Maybury, *Ostracoda and Global Events*, Chapman & Hall, New York, 1990; R. C. Whatley, D. J. Siveter, and I. D. Boomer, Arthropoda (Crustacea: Ostracoda), in M. J. Benton (ed.), *The Fossil Record 2*, Chapman & Hall, New York, 1993.

## Paleoecology

Ecology of prehistoric times, extending from about 10,000 to about  $3.5 \times 10^9$  years ago. Although the principles of paleoecology are the same as those underlying modern ecology, the two fields actually differ greatly. Paleoecology is a historical science that must rely on empirical data from fossils and their enclosing sedimentary rocks to make inferences about past conditions. Experimental approaches and direct measurement of environmental parameters, which are critical components of modern ecology, are generally impossible in paleoecology. Furthermore, distortion and loss of information during fossilization means that fossil assemblages and distributions are rarely congruent with living communities. Hence, the resolution of ancient ecosystems must remain relatively imprecise. The lack of precision is compensated for by the fact that paleoecology deals with processes occurring over vast spans of time that are unavailable to modern ecology. Long-term changes in communities (replacement) may be discerned and related to patterns of environmental change. More significantly, overall patterns of ecological change in the global biosphere may be documented; evolutionary paleoecology focuses on recognition and interpretation of long-term ecological trends that have been critical in shaping evolution.

Among the goals of paleoecology are the reconstruction of ancient environments (primarily depositional environments), the inference of modes of life for ancient organisms from fossils, the recognition of recurring groupings of ancient organisms that de-

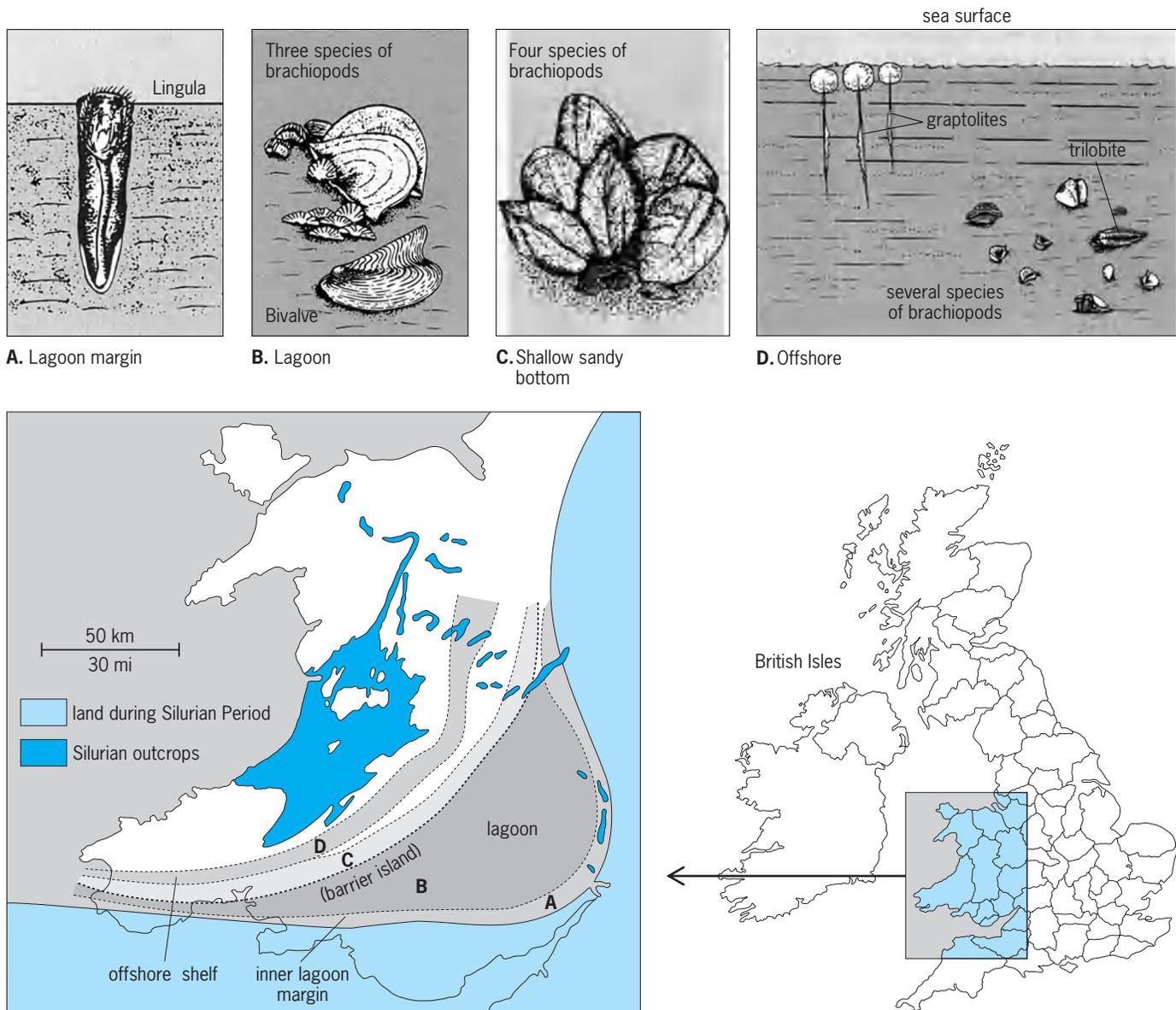
fine relicts of communities (paleocommunities), the reconstruction of the interactions of organisms with their environments and with each other, and the documentation of large-scale and long-term patterns of stasis or change in ecosystems. *See* ECOSYSTEM.

### Paleoenvironmental Interpretations

To reconstruct ancient marine environments, many different parameters must be inferred, such as temperature, water salinity, oxygen levels, nutrient concentrations, and water movements and depth (Fig. 1). In this regard, paleoecology interfaces directly with the fields of sedimentology and stratigraphy, including study of modern depositional environments. *See* DEPOSITIONAL SYSTEMS AND ENVIRONMENTS; STRATIGRAPHY.

**Taxonomic uniformitarianism.** One of the most useful, but also potentially misused, aspects of paleoecological application is known as taxonomic uniformitarianism. This concept relies on studies of modern organisms to determine limiting environmental factors, such as salinity tolerance, temperature preference, or depth ranges. Fossils of the same or closely related species are then inferred to have had similar environmental preferences, and their occurrence is judged to indicate that particular strata were deposited under a comparable range of environmental conditions. Such an approach is valid for very closely related organisms in relatively recent geologic time. Species and even genera may have relatively uniform environmental ranges through time, but the same cannot always be said of higher taxa such as families. At the level of order or class, only the broadest uniformitarian generalizations apply. For example, it is probably valid to consider fossil nautiloids or echinoderms as indicators of normal marine salinities, as all living representatives of these taxa have very limited abilities for osmotic regulation and therefore are restricted to near-normal salinity. Similarly, the restriction of photosynthetic organisms (such as algae) to the euphotic zone may be useful in determining relative depth. However, the precision and reliability of taxonomic uniformitarianism breaks down in increasingly ancient samples.

**Morphologic features.** Certain features of the morphology of fossils may be useful in making environmental inferences without reliance on evolutionary relationships. For example, the presence of entire margins and drip tips on leaves of plants is indirectly related to humid, warm climates, and so the proportions of leaf floras with entire margins and drip tips have been used as an index of paleolatitudinal zonation. Growth forms of colonial organisms relate to environmental factors such as turbulence and sedimentation rates. Flexible or articulated skeletons or flat encrusting form in colonial marine animals are associated with highly turbulent shallow-water environments where streamlining becomes important. Delicately branched, inflexible colonies typify quiet areas and, commonly, areas of high turbidity where a branching skeleton may shed sediment more readily than a flat or globose form. Such ecologically related morphology may transcend taxonomic boundaries.



**Fig. 1.** Zonation of Silurian (425 million year old) marine communities in the Welsh Basin, United Kingdom. Marine animal communities were typically arranged in belts parallel to shoreline and related to water depth. A classic study using fossil communities in basin analysis (Ziegler, 1965) mapped out the distribution of different communities of marine fossils, primarily brachiopods, to show the contouring of belts of ancient environments from the shoreline in the southeast to deep basinal environments. (Reprinted with permission from S. Stanley, *System History*, 2d ed., Freeman Co., 2005)

**Skeletal mineralogy and geochemistry.** The microstructure and geochemistry of organism skeletons may provide clues about ancient environments. For example, the presence of growth banding in skeletons provides evidence for seasonal variability in climates. The skeletons of fossil organisms, if they are well preserved, also encode valuable environmental information in the form of trace elements and isotopic signatures. For example, the calcium carbonate skeletons of marine invertebrates incorporate trace elements whose proportion is related both to physiology and environmental factors such as temperature and salinity. The isotopic composition of oxygen or carbon within carbonate skeletons is a function of isotopic composition of the seawater in which the skeleton was secreted as well as of water temper-

ature. If temperature can be determined independently, the ratio of  $^{18}\text{O}$  to  $^{16}\text{O}$  (often expressed as a deviation from a standard and referred to as  $\delta^{18}\text{O}$ ) can be used to determine whether a shell was secreted in water of normal (35%) or abnormal salinity. Conversely, if a given shell can be assumed, on independent evidence, to have come from a normal marine environment and is unaltered, then  $\delta^{18}\text{O}$  may be used to determine paleotemperature. In general, carbonate secreted at lower temperatures is preferentially enriched with respect to  $^{18}\text{O}$ , and so,  $\delta^{18}\text{O}$  is useful for temperature determination. See GEOLOGIC THERMOMETRY.

**Comparative taphonomy.** Taphonomy, which deals with processes and patterns of fossil preservation, has a critical dual role with respect to paleocology.

On the one hand, preservational processes impose distinct biases on the fossil record that must be considered carefully in any attempt at paleoecological reconstruction. On the other hand, the bodies and skeletons of dead organisms constitute biologically standardized sedimentary particles whose orientations, sorting, and general preservational condition bear the imprint of environmental processes active in the depositional environment. Comparative taphonomy uses the differential preservation of fossils as a source of paleoenvironmental information. The degree of preservation of fossils reflects biostratigraphic processes, such as current-wave transport, decay, disarticulation, fragmentation and corrosion of skeletons, and fossil diagenetic factors acting after final entombment of the remains in sediment.

Evidence of mode of death of organisms may also provide critical details. For example, layers of beautifully preserved fish or reptile carcasses signify mass mortalities that involved changes in the water column itself. But such mass mortalities can be recorded only if they were also timed with burial events.

Soft tissues can be preserved only by exceptionally rapid burial in anoxic sediments followed by very early coating or impregnation by minerals. Such deposits not only yield important data on the paleobiology of organisms but also provide detailed insights into depositional environments.

Usually, however, only skeletal remains are preserved. Skeletons composed of bivalved shells (for example, brachiopods and pelecypods) or, particularly, of multiple articulated elements (for example, echinoderms, arthropods, and vertebrates) are sensitive indicators of episodic burial rates. Experimental studies have demonstrated the rapidity of disarticulation under normal marine conditions; most starfish, for example, disintegrate into ossicles in a few days. Hence, intact preservation of these organisms signals episodic burial events. *See* FOSSIL.

Individual skeletons, or parts of skeletons, may become physically fragmented, chipped, or abraded. Such evidence reflects the general degree of turbulence of a particular depositional environment. Similarly, the degree to which skeletal remains are size- or shape-sorted may signify the extent of current and wave processing. Skeletal destruction by bioerosion, physical abrasion, or chemical solution is generally a good indicator of residence time of skeletons on the sea floor prior to burial. The orientation of fossils may yield specific clues as to the extent and types of environmental energy. Pavements of convex-upward valves typically are associated with persistent current reworking, whereas abundant concave-upward shells may signify an episode of stirring of the shells from the sea bottom and resettlement during storms. Furthermore, alignment of elongated shells may provide data on the orientation of unidirectional currents or the propagation direction of oscillatory waves. Vertically embedded specimens of ammonoids are typical of water areas with depths less than 30 ft (10 m). Finally, the early diagenetic features of fossils reflected in solution, compaction, and mineralization may yield information about sed-

iments and bottom water geochemistry, water pH and oxygen content.

Various aspects of biostratigraphic and diagenetic fossil preservation can be combined to form predictive models of taphonomic facies or taphofacies. Certain suites of quantifiable preservational conditions, for example, characterize particular environments, and so their recognition by paleoecologists may help to "fingerprint" those environments. *See* SEDIMENTOLOGY; TAPHONOMY; TRACE FOSSILS.

### Paleoautecology

Paleoautecology, the interpretation of modes of life (broadly, niches) of ancient organisms, involves a multidisciplinary approach. Although ancient modes of life cannot be determined completely, paleoecologists can often assign fossils to generalized guilds in terms of types of feeding, substrate preference, and degree of activity.

A thorough understanding of the biology of closest modern analogs is particularly important in any attempt to reconstruct paleoautecology. If the species or a closely related species is extant, then its mode of life, general physiology, and even behavior can be inferred with some confidence through the use of taxonomic uniformitarianism, provided that the biology of living relatives is well understood. "Living fossils," or relict extant taxa, such as *Nautilus*, sclerosponges, horseshoe crabs, and modern stalked crinoids provide valuable clues for interpreting the paleobiology of extinct organisms. *See* LIVING FOSSILS.

**Functional morphology.** For extinct organisms that have no adequate modern analogs, alternative approaches, particularly functional morphology, provide some hints as to life modes. Comparative morphology seeks analogies between the anatomical features of fossil skeletons and those in living forms for which the function can be determined. In some cases, structures in unrelated organisms have evolved convergently, and their function may be interpreted by analogy. When no biological analog exists, a physical or mechanical model, or paradigm, may provide clues to interpreting structures in extinct organisms.

An experimental approach to functional morphology may also provide useful insights. Models of ammonite shells, for example, have been tested in flumes in which artificial currents are produced to determine frictional drag effects of shell shape and sculpture. Certain shell shapes were found to be more hydrodynamically streamlined and probably correspond to more rapidly swimming modes of life. Testing the resistance of different brachiopod shell architectures to crushing, as by predators, has shown that certain features of shell architecture, such as ribbing and deflections of shell margins, can increase shell rigidity.

**Fossil data.** Certain natural experiments also shed light on the paleobiology of extinct organisms. The fact that oysters encrusted the shells of living ammonites has enabled paleontologists to calculate the buoyancy compensation capabilities of those

ammonoids. Encrustation and boring of cephalopod shells by bryozoans and barnacles that grew preferentially aligned toward currents has demonstrated a predominance of forward swimming motion in these extinct hosts.

Remnants of soft parts, muscle scars, gut contents, and associated trace fossils all provide information useful in the reconstruction of ancient ecological niches. Rare occurrences of rapidly buried fossils in unusual positions can be interpreted as original life positions. Unusual associations with substrates or other organisms also provide insights. Finally, the consistent association of poorly understood fossil species with other fossils whose modes of life are well known or with sediments that indicate particular environments may help to establish the habits and environmental ranges of extinct forms.

**Population studies.** Certain properties of species, such as mortality patterns, birth rates, and numbers of individuals per age class, can be studied only at the level of populations. Despite the difficulties of studying fossil populations, it is still possible to make some inferences about population parameters. For example, the distribution of individuals of a particular species into different age or size classes may yield some indirect data on the age-frequency distribution of a population that can be used to construct crude mortality curves showing age-at-death relationships. Some species may display a high juvenile mortality, a feature typically associated with stressed environments and rather opportunistic species; others, in stable environments, may display delayed mortality.

Of particular ecological importance is the population strategy of a given species of organisms. Two end-member conditions have been recognized: opportunistic species, sometimes termed r-selected forms, and equilibrium, or k-selected, species. Opportunistic organisms are typically rather generalized in habit and habitat preferences, are commonly stress-adapted, and display exceedingly high rates of reproductive maturation and fecundity. Extremely opportunistic, or "weedy," species of this sort are recognizable in the fossil record by their widespread distribution and occasional presence in extremely dense, monospecific populations on single beds of rock that are otherwise barren of fossils. Equilibrium species, on the other hand, tend to occur in moderate or small numbers in a narrow range of environments commonly associated with diverse assemblages of other species, such as in reef environments. The distinction between equilibrium and opportunistic mode of life may have important implications for understanding the distribution and evolutionary patterns of fossil taxa as well as for interpreting the stability of particular ancient environments. See POPULATION ECOLOGY.

#### **Paleosynecology**

The study of interrelationships within organism communities that coexist in time and space is known as synecology. At the most basic level of synecology are the interacting pairs of organisms that coexist in a particular environment. Paleosynecology also in-

volves study of ancient community structure and dynamics.

**Organism interactions.** Interactions range from tolerance to symbiosis, which involves highly dependent and coevolved species pairs. Although interactions are very difficult to determine with fossils, there are cases where strong clues are observed. In some instances, organism interaction may be very indirect. For example, the accumulation of shells on a sea floor may lead to colonization by other organisms that require hard substrates to encrust onto or bore into, a process referred to as positive taphonomic feedback. Conversely, armoring of muddy sea bottoms with shell debris will inhibit burrowing organisms (negative taphonomic feedback).

Shallow-burrowing nuculid bivalves convert muddy sea bottoms into a water-rich pelleted floc. The high turbidity and instability of this fluid substrate may inhibit the settlement of many epifaunal suspension-feeding organisms. Such negative feedback is referred to as trophic group amensalism.

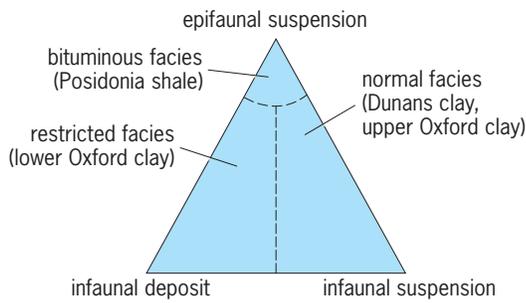
Mutualism, involving symbiotic algae called zooxanthellae, is inferred for many fossil reef-dwelling organisms, including various corals, sponges, and even some bivalves, based on taxonomic uniformitarianism as well as morphology and evidence for prolific skeletal growth. Such mutualism is difficult to substantiate for extinct groups, although distinctive patterns of carbon isotopes within skeletal carbonates may prove a useful fingerprint of secretion aided by zooxanthellae.

Many marine organisms use the skeletons of other living organisms as substrate or to obtain an elevated feeding position without having any effect on the hosts. Evidence of this type of commensal interaction is abundant in the fossil record.

Parasitic interactions are very difficult to observe in fossils because they normally involved the soft tissues of the host. However, rare evidence for paleopathology (fossil diseases) can be documented from Paleozoic times onward in certain organisms such as echinoderms or vertebrates that have an internal skeleton, or endoskeleton. For example, malformations in fossil crinoids may record parasitism.

Fossil evidence for competition is best seen in cases of spatial competition. For example, certain types of bryozoans appear to overgrow other species preferentially. Many aspects of fossil distribution have been attributed to the effects of competition or the evolutionary response for reducing competition by niche partitioning. Examples include the subdivision of many marine communities into distinct feeding groups based on vertical height (tiering) above and below the sediment-water interface. Some researchers have claimed that competition is a primary motor of evolutionary change, often alluding to Darwin's analogy of the wedge, in which more and more species are packed into a particular ecospace by increasingly finely divided specialized niches.

Predation or carnivory is probably one of the most significant ecological interactions in any environment. Direct predator-prey links are difficult to

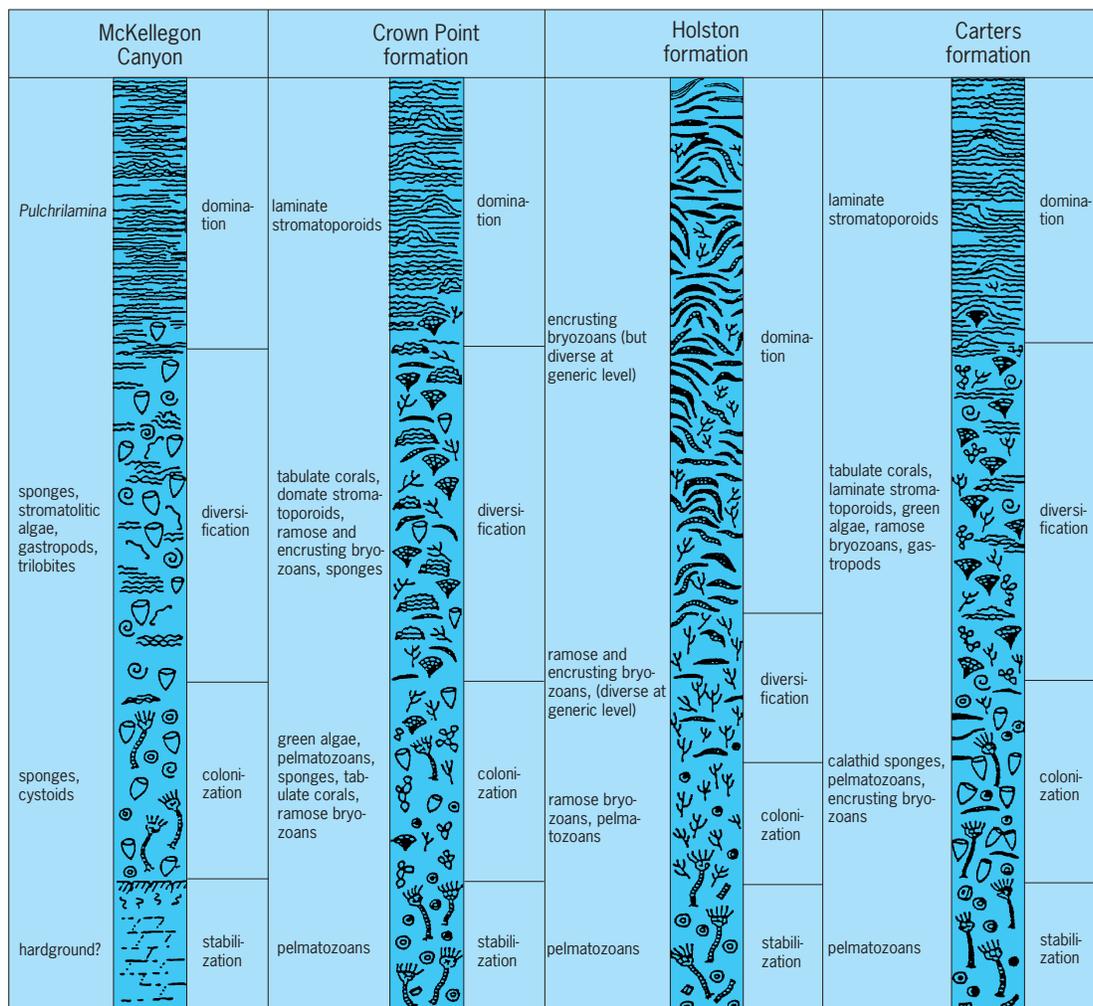


**Fig. 2.** Classification of shales in terms of percentages of three benthic habitat groups: epifaunal suspension-feeding bivalves (those living at the sediment-water interface and filtering seawater), infaunal (living within and feeding on the sediment), and suspension-feeding groups. (After K. A. Morris, *Comparison of major sequences of organic-rich mud deposition in the British Jurassic*, *J. Geol. Soc. London*, 137:157-170, 1980)

establish in the fossil record, but there are several lines of evidence that may be used. Bite marks of particular types, such as tooth marks of mosasaurs on ammonoids, provide one line of evidence, so do boreholes of predatory snails on particular prey species and remnants of prey shells preserved in

the stomach contents or coprolites (fossilized feces). In turn, numerous morphological trends may signify antipredatory adaptations. The fossil record of predation extends back to the Early Cambrian, as evidenced by bite marks in trilobites, which commonly show a preference for the right side of the prey. The frequency of healed and unhealed predatory fractures in some shells increases significantly in the Paleozoic in concert with the rise of fossil evidence for shell-crushing predators. The earliest shell-drilling snails appear to have been Ordovician in age, but the habit of drilling shells for predation probably evolved independently at least four times in different groups of gastropods.

**Paleocommunities.** The fossil record contains highly biased remnants of past communities or paleocommunities. Paleocommunities are generally recognized as recurring associations of fossil species. Multivariate statistical techniques such as cluster analysis and ordination analysis are commonly employed to aid in discerning the recurrent groupings of fossil species, or persistent gradients of species composition. Such analyses are based upon field studies in which data on the presence, absence, or



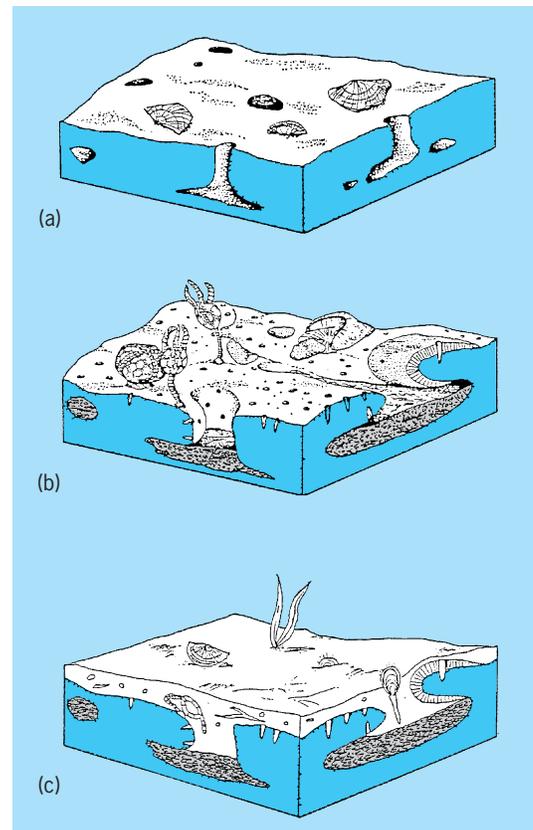
**Fig. 3.** Comparison of four developmental stages in four ancient reef masses. (After K. R. Walker and L. P. Alberstadt, *Ecological succession as an aspect of structure in fossil communities*, *Paleobiology*, 3:238-257, 1975)

relative abundance of fossil taxa have been recorded in a large number of samples, typically from many stratigraphic levels.

**Taphonomic biases.** Statistically defined groupings may or may not represent real ecological entities. For example, in most offshore marine environments, the transport of skeletons between environments is minimal. However, because of differential preservation, the proportions of organisms in the living assemblages (biocoenoses) are not always faithfully reproduced in the death assemblages of skeletal remains (taphocoenoses). Nearly all soft-bodied organisms are lacking in the death assemblage, and those with fragile skeletons tend to be underrepresented. Moreover, because of the accumulation of skeletons over extended periods of time, death assemblages commonly display mixtures of organisms that inhabited slightly differing environments at different times, a phenomenon referred to as time averaging. Fossil assemblages actually may be more diverse than living assemblages of skeletonized organisms at any one time. They record a very biased and averaged-out view of communities that existed over a long period of time.

**Relationship to environments.** In most studies of paleocommunities, recurrent groups can be related to environments, as inferred from independent evidence such as rock type, sedimentary structures, taphonomy, trace elements, and isotopic studies. Classic studies modeled paleocommunity distribution patterns on relative bathymetry or distance from shoreline, but many later studies emphasized the control of paleocommunity distribution by multiple factors. Depth-related factors such as turbulence, light penetration, and oxygen level are clearly important controls in many cases. However, sedimentation-related factors such as rates of deposition, turbidity, and substrate consistency may be equally important, giving rise to a much more complex array of paleocommunities (Fig. 1).

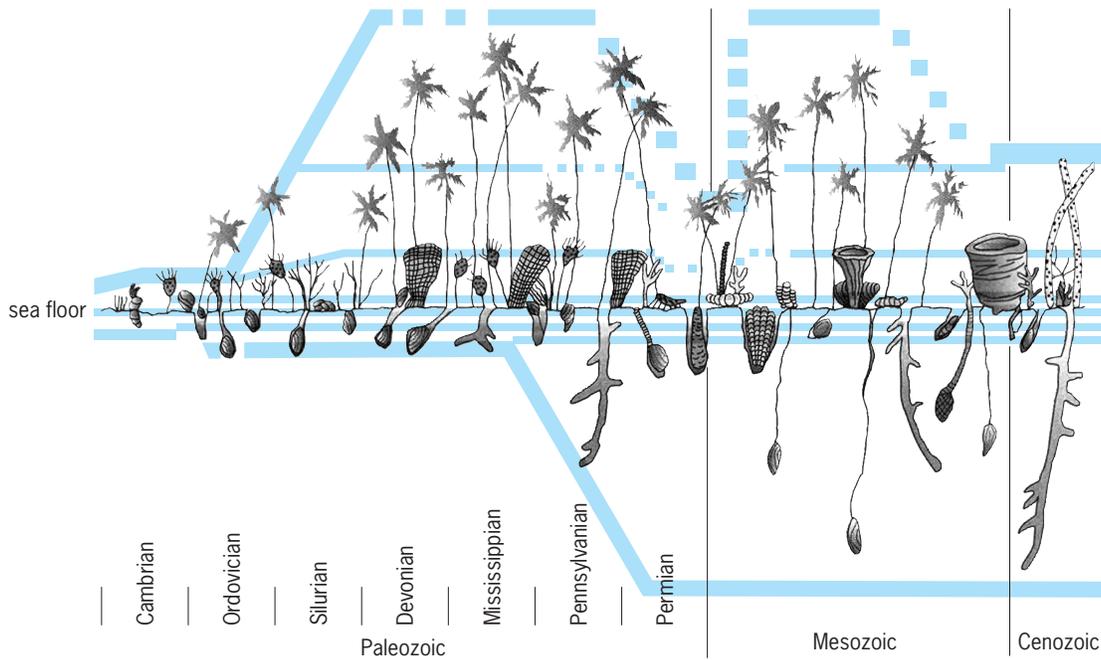
After recurring associations of fossils are recognized, they are generally analyzed in terms of organism interactions and trophic (feeding) relationships. Primary producers of ancient ecosystems, such as algae, are likely to be poorly preserved or absent from the fossil record, but some links in ancient food chains may be recognizable. One aspect of paleocommunity structure that is commonly analyzed is the proportion of different feeding (trophic) and life-habit guilds. Certain marine paleocommunities are dominated by skeletons of infaunal deposit feeders, others by epifaunal suspension feeders. Unfortunately, live-dead comparisons in modern communities suggest that the original proportions of various life habitat groups are not preserved in the fossil record. But the biased trophic proportions of the taphocoenoses may still relate in a meaningful way to the original environment. Consistent differences in the proportions of infaunal suspension-feeding, infaunal deposit-feeding, and epifaunal suspension-feeding bivalves have been detected in differing ancient oxygen-restricted facies (Fig. 2). See FOOD WEB.



**Fig. 4.** Formation and burial sequence of Bobcaygeon hardground. (a) Soft-bottom community of strophomenid brachiopods and infaunal burrowers inhabiting carbonate mud. (b) Hardground community consisting of boring and encrusting organisms. (c) Post-hardground community inhabiting muds that blanketed the hardground. (After C. E. Brett and W. D. Liddell, *Preservation and paleoecology of a Middle Ordovician hardground community*, *Paleobiology*, 4:329–348, 1978)

**Temporal changes.** Communities and paleocommunities are not static entities in time, but undergo important structural changes on at least three different time scales: succession, replacement, and evolution. Because it operates on a very short time scale, from decades to centuries, ecological succession can be resolved only in a few fossil samples (Fig. 3). Some instances of supposed ecological succession, such as encrusting communities upon shells, may in fact reflect taphonomic feedback. Allogenic succession represents changes in communities induced by physical environmental change. Good examples are seen in many hardgrounds, areas of early lithified sea floors (Fig. 4). See ECOLOGICAL SUCCESSION.

Longer-term changes in community composition, encompassing thousands of years, are not truly succession, but instead record allogenic effects such as sea level or climate variations. These changes are properly termed community replacement, and involve wholesale migration or restructuring of communities at particular locations due to changing environments. In many instances, particular fossil assemblages appear to track shifts in preferred environments and facies within sedimentary cycles. Habitat tracking may provide important clues to deciphering patterns of environmental fluctuations



**Fig. 5. Tiering or vertical depth stratification in marine communities through time. Note the rapid rise of the highest-tier (level) organisms (crinoids or sea lilies, bryozoans) from a few centimeters to over a meter above the sea floor during the Ordovician Period. Branching bryozoans and shorter-stemmed crinoids took advantage of an intermediate tier 10–20 cm above the sea bottom, while burrowing clams and worms dug down to a tier about the same distance into the sediment. In the middle Paleozoic, still much deeper burrowing forms evolved the ability to mine sediments down to nearly a meter. Note minor readjustment of the tiers associated with mass extinctions at the Paleozoic-Mesozoic and Mesozoic-Cenozoic era boundaries. (Modified from W. I. Ausich and D. J. Bottjer, *J. Geol. Educ.*, 1991)**

such as transgressive-regressive cycles. On a scale of millions of years, communities show evolutionary changes because their component species have evolved. See ECOLOGICAL COMMUNITIES.

**Larval ecology and evolution.** Ecological patterns such as larval type affect overall patterns in life history. Larval ecology of marine animals controls their geographic distribution. Species with long-lived larvae may be dispersed much more widely than forms with short-lived planktonic phases or direct development from eggs. In turn, geographic distribution, whether localized or cosmopolitan (global), undoubtedly plays an important role in their tendency toward speciation as well as extinction. Thus, it may be possible to develop models to better explain evolutionary patterns in different groups in relation to paleoecology.

### Evolutionary Paleoecology

Organisms evolve within the context of other organisms, not in a vacuum. There is substantial fossil evidence to indicate increasing complexity of organism interactions through time. This escalation in the intensity of predatory interactions, for example, may have important implications for evolutionary change. For example, trends of increased spinosity, greater shell thickness, increasingly restricted apertures, and other antipredation adaptations may reflect the intensification of predatory behavior by shell-boring and crushing predators.

Increased vertical stratification or tiering in marine-level bottom communities through time has

been recognized (**Fig. 5**). Cambrian communities possessed mainly low-lying suspension-feeding and scavenging organisms that lived mostly just above or below the sediment-water interface. By mid-Paleozoic time, crinoids extended up to heights of several feet or more off the sea floor, and various burrowers extended downward a couple of feet or more into the sediment. The Mesozoic rise of deep-burrowing clams and other infauna increased the infaunal tier to over 3 ft (1 m). The increased vertical structuring of these communities may represent a response to increasing crowding. By feeding at different levels in the water and substrate, organisms were able to further subdivide the resources of a given environment.

Marine animals form a hierarchy of ecological units through the Phanerozoic time interval. These range from blocks of relative stability at time scales of a few million years, to broader intervals of general stability of faunas, to three great evolutionary faunas.

First, at a scale of a few million years, groups of species may show considerable ecological stability punctuated by episodes of abrupt change. Brett and Baird (1995) introduced the concept of “coordinated stasis” to describe a pattern of approximately concurrent long-term stability and abrupt change in many taxa. During a large proportion of geologic time a majority of genera and, in some cases, species show little or no change in morphology. Moreover, general groups of communities or “biofacies” also may be similar throughout blocks of stability referred to as “ecological-evolutionary units and subunits.” These

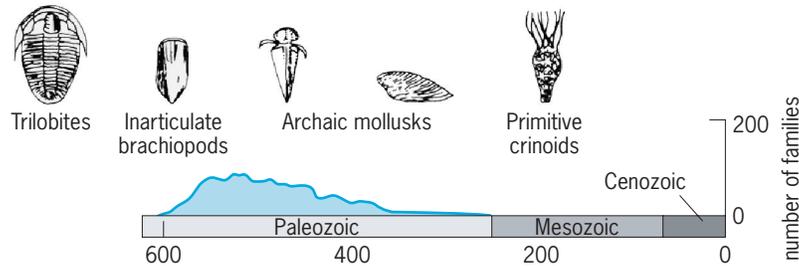
relatively stable intervals, spanning up to several million years, are punctuated by much shorter intervals, perhaps a few tens of thousands of years, of abrupt change across many biofacies, local extinction of many long-standing lineages, immigration and emigration from the local basin, and general faunal turnover. The original example of the Silurian–Devonian (380 to 440 million year old) fossil assem-

blages of eastern North America—in particular, the Middle Devonian Hamilton Group—features examples of assemblages, separated by up to 5 million years, with nearly identical composition and similar guild structure and even relative abundance. Consideration of a larger number of case studies ranging in age from Cambrian to modern suggests that this original example represents one end member in an array of conditions ranging from similar cases but some with somewhat more species level variability, to examples of nearly continual change in species composition, and biofacies ecological structure. This variability probably depends on local environmental variability. The observed pattern of similarities between samples from cases of coordinated stasis could imply a form of stable, lock-step tracking of certain well-organized “communities.” However, this pattern could also be the result of recurrence of a similar assemblage due to persistence of environmental gradients and because species do not drastically change their habitat preferences through time. The retention of habitat preferences by species is perhaps the most important aspect of ecological stasis. It would appear that under appropriate conditions species can simply track shifting preferred environments for millions of years rather than adapt to local change.

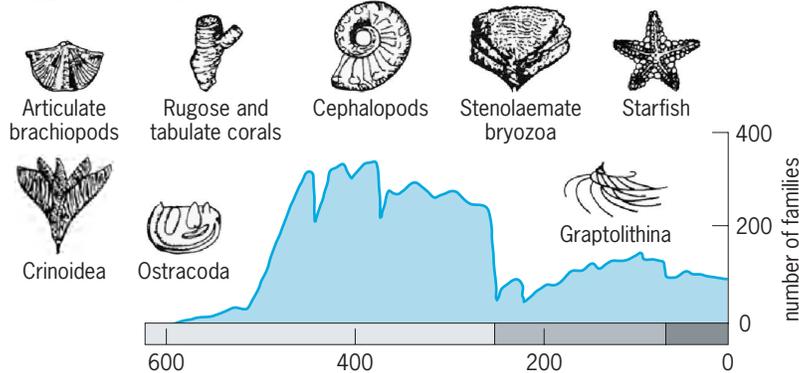
At the next larger level, marine communities appear to show strong similarities of family- and genus-level composition, as well as general ecological structure (guilds, trophic structure, diversity, and so on) for tens of millions of years. These blocks of relative stability, termed ecological-evolutionary units (EEUs) by Boucot (1990), were terminated by major extinctions. Raup and Sepkoski also recognized five major mass extinctions—the “Big Five” (Late Ordovician, Late Devonian, Permian-Triassic, Late Triassic, and Cretaceous-Tertiary) that stand out from background rates of extinction. These and lesser mass extinctions played critical roles in restructuring the ecology of the biosphere, including changes in guild structure and tiering patterns. Ecological-evolutionary units are bounded by major biotic turnover events, involving widespread extinctions including the “Big Five” mass extinctions. Again, the EEU concept implies that the ecological history of life was not one of continuous, gradual change. Rather, it was characterized by extended periods of near equilibrium that were interrupted by much shorter periods of crisis and major ecological restructuring.

The largest scale of faunal pattern consists of “evolutionary faunas.” By analyzing patterns of marine family and genus level diversity using a large database, Sepkoski (1981) recognized three such units through the past 540 million years of the Phanerozoic Eon, each characterized by a different pattern or trajectory of diversification (Fig. 6). The “Cambrian fauna”—typified by trilobites, lingulid brachiopods, and certain primitive groups of mollusks and echinoderms—appeared during the earliest Paleozoic, diversified in the Cambrian, and then began to decline as the second or “Paleozoic fauna” diversified. The latter was characterized by

### CAMBRIAN FAUNA



### PALEOZOIC FAUNA



### MODERN FAUNA

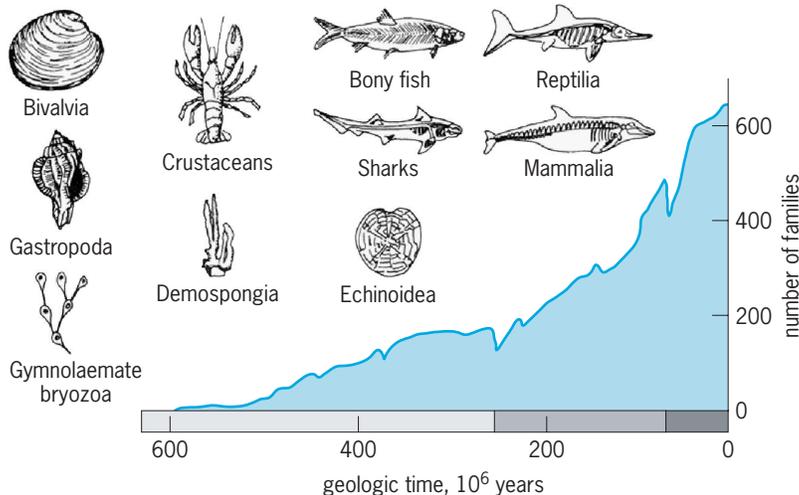


Fig. 6. The three great marine evolutionary faunas of the Phanerozoic. The Cambrian fauna, composed of trilobites, primitive groups of brachiopods, and mollusks arose early during that period to a diversity of about 50 families, then dwindled during the later Paleozoic as more archaic groups migrated offshore and were replaced by the “Paleozoic fauna,” typified by rugose and tabulate corals, brachiopods, bryozoans, crinoids, and graptolites. The latter diversified rapidly in the Ordovician Period to over 300 families and then fluctuated around this level until the end of the Paleozoic Era. The great Permian-Triassic extinction reduced the “Paleozoic fauna” and may have favored the rise of the “Modern fauna” during the Mesozoic and Cenozoic eras with diversities as high as 620 families, including especially mollusks, crustaceans, and both sharks and bony fishes. (Modified from J. J. Sepkoski, Jr., *Paleobiology*, 1981)

rugose and tabulate corals, articulate brachiopods, bryozoans, and crinoids which formed the major faunas of shallow seas from the Ordovician to the Permian Period and displayed a relatively stable “platform” of family diversity. Finally, the “Modern fauna,” characterized by mollusks and crustaceans, arose in nearshore environments during the early Paleozoic, but expanded greatly following the end Permian mass extinctions. Sepkoski and Sheehan (1983) recognized that evolutionary innovations tended to arise first in shallow, nearshore environments. Through time the newly arising groups typical of the “Paleozoic” and then the “Modern” faunas tended to spread offshore, while more archaic forms were displaced to deep ocean “refugia.” This is one of the most profound of all paleoecological patterns, and the explanation of this pattern remains imperfectly understood. It may imply that stressed nearshore settings favor evolution of new life strategies and/or that there has been a general intensification of energy utilization through time such that more archaic “Cambrian” or “Paleozoic” faunas were relatively “low energy” and had less competitive ability than physiologically more sophisticated, “high-energy” Modern faunas. See ECOLOGY; PALEOCLIMATOLOGY; PALEOGEOGRAPHY; PALEONTOLOGY.

Carlton E. Brett  
Bibliography. A. J. Boucot, *Principles of Benthic Marine Paleocology*, 1981; R. J. Dodd and R. J. Stanton, Jr., *Paleoecology, Concepts and Applications*, 2d ed., 1990; J. J. Sepkoski, Jr., A factor analytic description of the Phanerozoic marine fossil record, *Paleobiology*, 7:36–53, 1981; M. J. P. Tevesz and P. L. McCall (eds.), *Biotic Interactions in Recent and Fossil Benthic Communities*, 1983; G. J. Vermeij, *Evolution and Escalation: An Ecological History of Life*, 1987.

## Paleogeography

The geography of the ancient past. Paleogeographers study the changing positions of the continents and the ancient extent of land, mountains, and shallow-sea and deep-ocean basins. The Earth’s geography changes because its surface is in constant motion due to plate tectonics. The continents move at rates of 2–10 cm/yr (0.75–4 in./yr). Though this may seem slow, over millions of years continents can travel across the globe. As the continents move, new ocean basins form, mountains rise and erode, and sea level rises and falls. The best way to illustrate these changes is through a series of paleogeographic maps. See CONTINENTS, EVOLUTION OF; GEOGRAPHY; PLATE TECTONICS.

Paleogeographic maps are necessary in order to understand global climatic change, migration routes, oceanic circulation, mountain building, and the formation of many of the Earth’s natural resources, including oil and gas. The maps in this article show the ancient mountains (dark tint), land areas (medium tint), and shallow seas (light tint), as well as the present-day coastline for reference. Past plate bound-

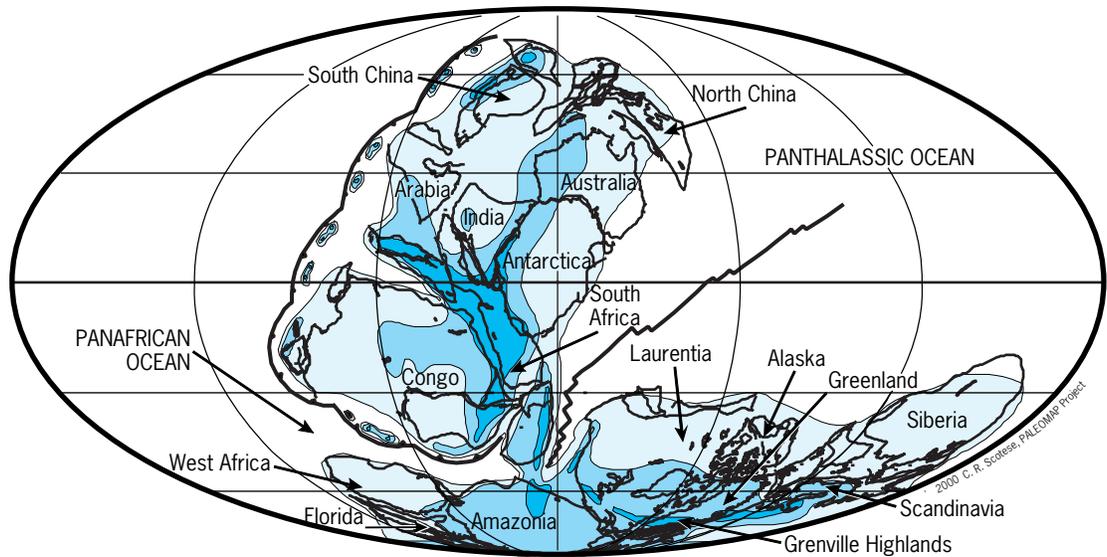
aries are also shown. The jagged lines in the center of ocean basins are mid-ocean ridges, the curving lines along the edges of continents are deep-sea trenches; the triangular teeth point in the direction of subduction. See BASIN; MID-OCEANIC RIDGE; SUBDUCTION ZONES.

**Late Precambrian, 650 Ma.** In the late Precambrian the continents were colliding to form supercontinents, and the Earth was locked in a major ice age (Fig. 1). About 1100 million years ago (Ma), the supercontinent of Rodinia was assembled. Though Rodinia’s exact size and configuration are not known, it appears that North America formed the core of this supercontinent. At that time, the east coast of North America was adjacent to western South America, and the west coast of North America lay next to Australia and Antarctica. Rodinia split into halves approximately 750 Ma, opening the Panthalassic Ocean. North America rotated southward toward the South Pole. The other half of Rodinia, composed primarily of Antarctica, Australia, India, Arabia, and China, rotated counterclockwise, northward across the North Pole. Between these two halves lay the Congo continent, made up of much of north-central Africa. The oceans between these three continents were completely subducted by the end of the Precambrian, and the three parts of Rodinia came together to form the supercontinent of Gondwana(land). This major continent-continent collision is known as the Pan-African orogeny. See OROGENY; PRECAMBRIAN; PROTEROZOIC; SUPERCONTINENT.

The climate was cold during the late Precambrian. Evidence of glaciation is found on nearly every continent. Why cold conditions were so widespread during the late Precambrian has long puzzled geologists. Some scientists believe the Earth was completely frozen like a snowball. However, as Fig. 1 shows, many continents lay close to the North and South poles at that time. See CLIMATE HISTORY; GLACIAL EPOCH; GLACIOLOGY; PALEOCLIMATOLOGY.

**Early and middle Paleozoic, 545–360 Ma.** The supercontinent that formed at the end of the Precambrian Era, approximately 600 Ma, broke apart at the beginning of the Paleozoic Era (Fig. 2). A new ocean, the Iapetus, widened between the ancient continents of Laurentia (North America), Baltica (northern Europe), and Siberia. Gondwana, which was considerably larger than any of the other continents, stretched from the Equator to the South Pole. See ORDOVICIAN; PALEOZOIC.

Approximately 400 Ma, the Iapetus Ocean closed, crashing Laurentia and Baltica together. This continental collision, preceded in many places by the collision of marginal island arcs in Maritima Canada and New England, resulted in the formation of the Caledonide Mountains in Scandinavia, northern Great Britain, and Greenland, and the Northern Appalachian Mountains along the eastern seaboard of North America. It is also likely that by middle Paleozoic times North China and South China had rifted away from the Indo-Australian margin of Gondwana, and were headed northward across the Paleo-Tethys Ocean (Fig. 3). See CONTINENTAL DRIFT.



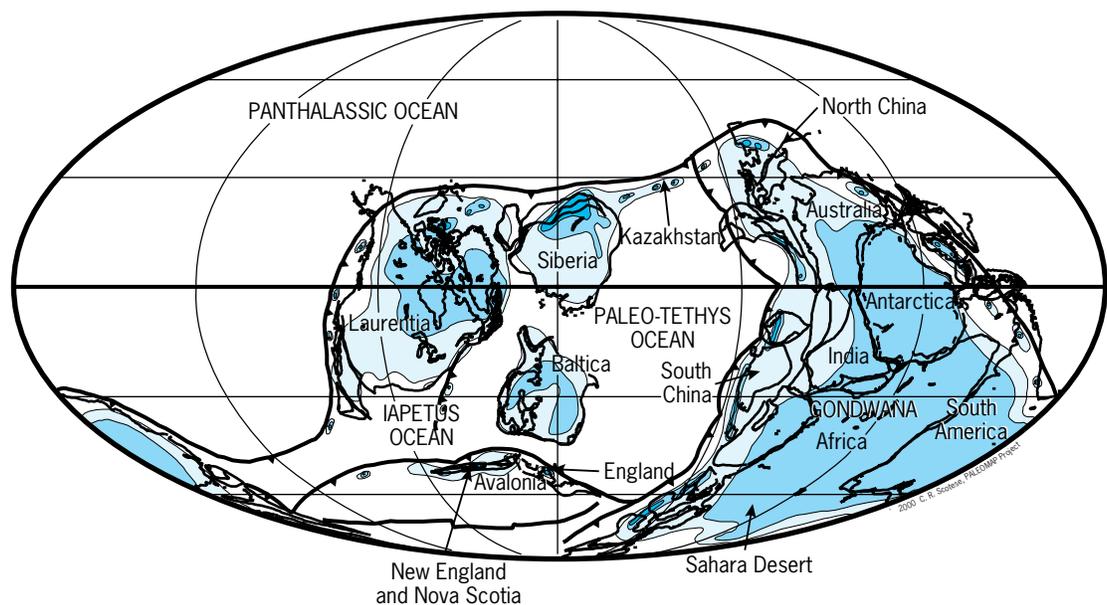
Key: ■ shallow sea ■ land ■ mountains, uplands

**Fig. 1. Vendian (or Ediacaran), 650 Ma.** (Paleogeographic maps by Christopher R. Scotese, PALEOMAP Project, University of Texas at Arlington)

**Late Paleozoic, 360–245 Ma.** At the Paleozoic Era, the continents collided to form the supercontinent of Pangea (Fig. 3). Centered on the Equator, Pangea stretched from the South Pole to the North Pole, and separated the Paleo-Tethys Ocean to the east from the Panthalassic Ocean to the west. Though the supercontinent that formed at the end of the Paleozoic Era is called Pangea (literally, “all land”), this supercontinent probably did not include all the landmasses that existed at that time. In the Eastern Hemisphere, on either side of the Paleo-Tethys Ocean, there were continents that were separated from Pangea. These continents were North and South China, and a long,

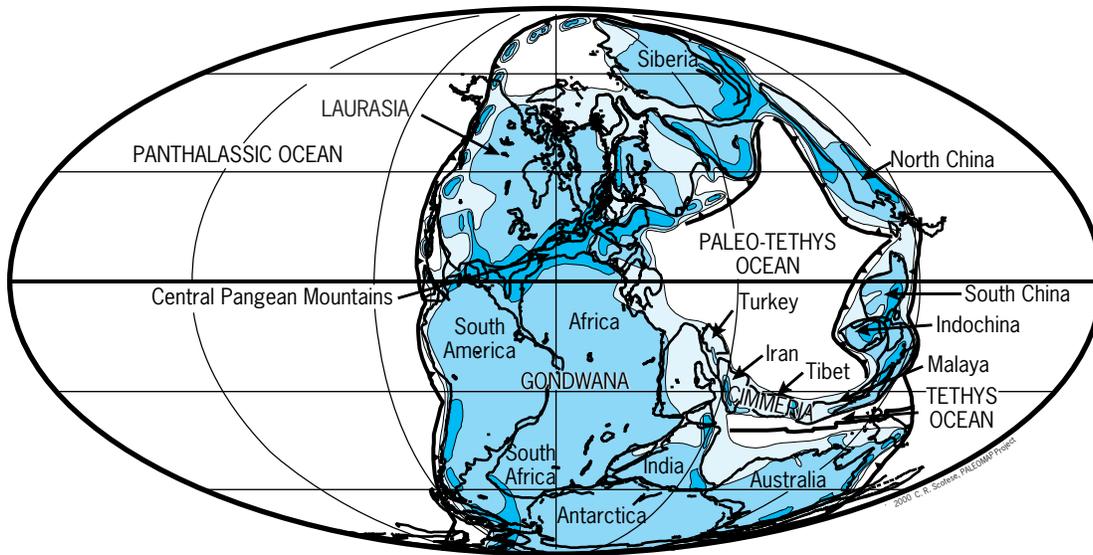
narrow continent known as Cimmeria. Cimmeria consisted of parts of Turkey, Iran, Afghanistan, Tibet, Indochina, and Malaya. It appears to have rifted away from the Indo-Australian margin of Gondwana during the late Carboniferous–early Permian. Together with the Chinese continents, Cimmeria moved northward toward Eurasia, ultimately colliding along the southern margin of Siberia during the Late Triassic. See PERMIAN.

**Early Mesozoic, 245–144 Ma.** The supercontinent of Pangea did not rift apart all at once, but in three main episodes. The first episode of rifting began in the Middle Jurassic, about 180 Ma when North America



Key: ■ shallow sea ■ land ■ mountains, uplands

**Fig. 2. Late Ordovician, 458 Ma.** (Paleogeographic maps by Christopher R. Scotese, PALEOMAP Project, University of Texas at Arlington)



Key: ■ shallow sea ■ land ■ mountains, uplands

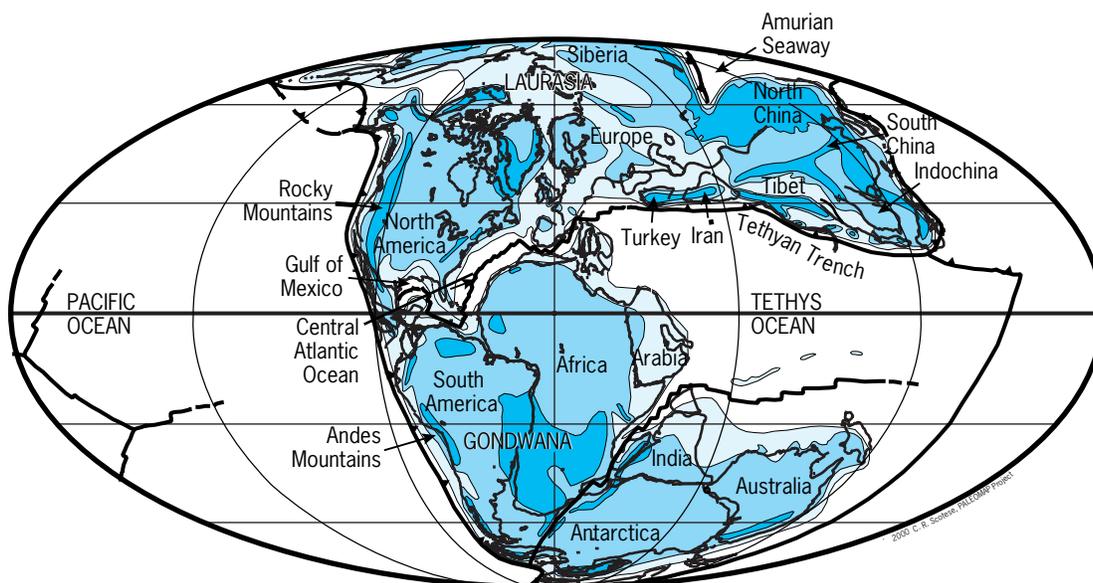
Fig. 3. Late Permian, 258 Ma. (Paleogeographic maps by Christopher R. Scotese, PALEOMAP Project, University of Texas at Arlington)

rifted away from northwest Africa, opening the Central Atlantic (Fig. 4). This movement also gave rise to the Gulf of Mexico. At the same time, on the other side of Africa, extensive volcanic eruptions along the adjacent margins of east Africa, Antarctica, and Madagascar heralded the formation of the Western Indian Ocean. See JURASSIC; VOLCANO; VOLCANOLOGY.

During the Mesozoic, North America and Eurasia were one landmass, called Laurasia. As the Central Atlantic Ocean opened, Laurasia rotated clockwise, sending North America northward and Eurasia southward. This clockwise, seesaw motion of Laurasia also led to the closure of the wide V-shaped ocean,

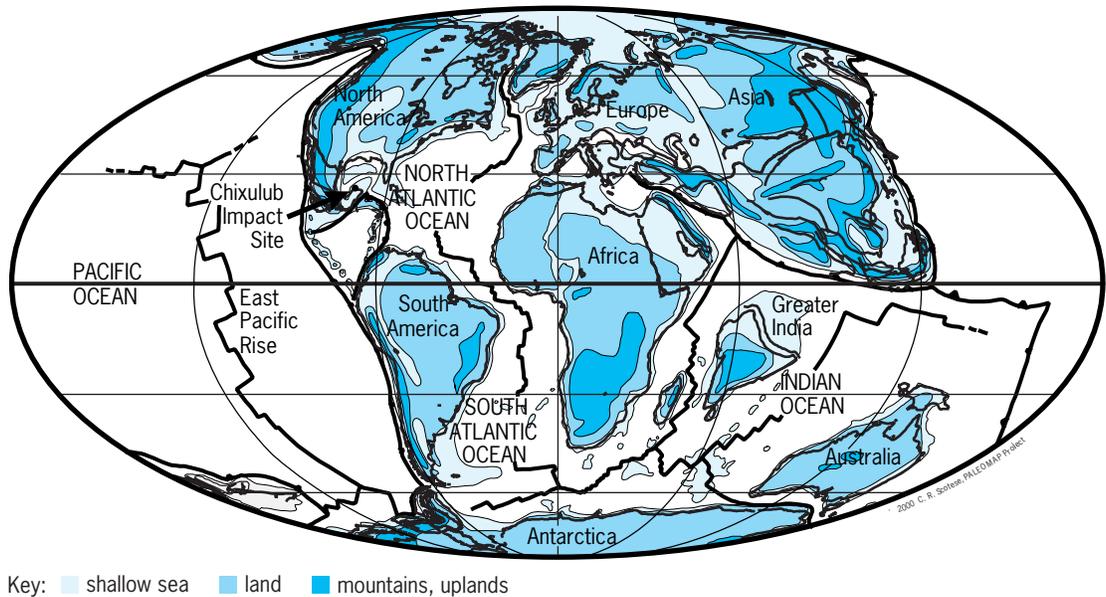
Tethys, that separated Laurasia from the fragmenting southern supercontinent, Gondwana. See MESOZOIC.

**Late Mesozoic, 144–66 Ma.** The second phase in the breakup of Pangea began in the Early Cretaceous, about 140 Ma. Gondwana continued to fragment as South America separated from Africa, opening the South Atlantic, and India together with Madagascar rifted away from Antarctica and the western margin of Australia, opening the Eastern Indian Ocean (Fig. 5). The South Atlantic did not open all at once, but progressively “unzipped” from south to north. The initiation of rifting between North America and



Key: ■ shallow sea ■ land ■ mountains, uplands

Fig. 4. Late Jurassic, 152 Ma. (Paleogeographic maps by Christopher R. Scotese, PALEOMAP Project, University of Texas at Arlington)



**Fig. 5.** Cretaceous-Tertiary (K/T) boundary, 66 Ma. (Paleogeographic maps by Christopher R. Scotese, PALEOMAP Project, University of Texas at Arlington)

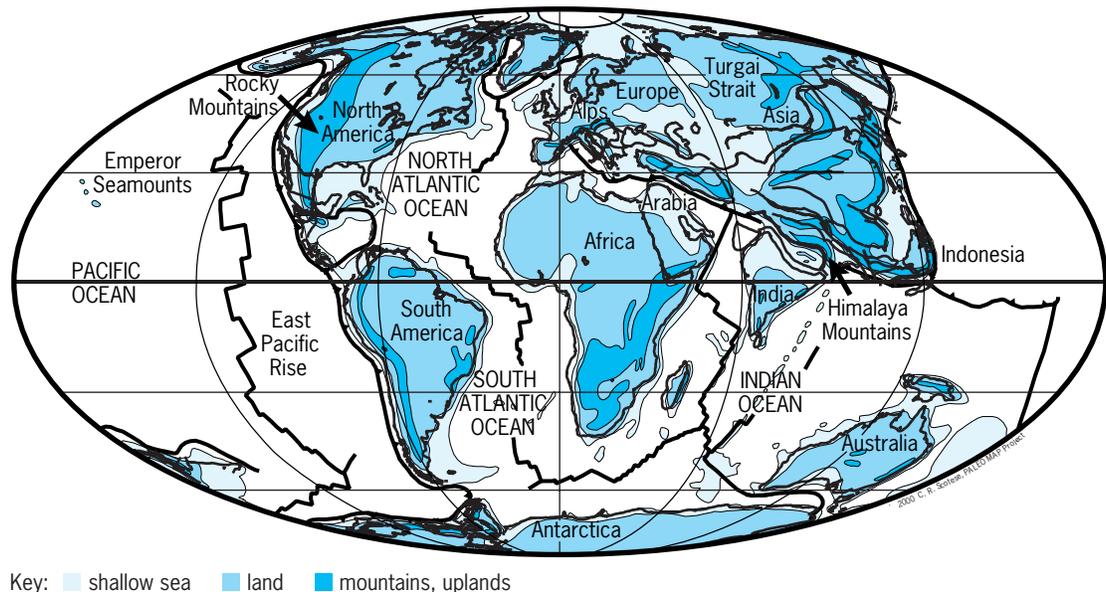
Europe, the counterclockwise rotation of Iberia from France, the separation of India from Madagascar, the uplift of the Rocky Mountains, and the arrival of exotic terranes (Wrangellia, Stikinia) along the western margin of North America also took place during the Cretaceous. *See* CRETACEOUS; TERTIARY.

During the late Mesozoic, shallow seaways covered the continents. Higher sea level was due in part to the creation of new rifts in the ocean basins that displaced water onto the continents.

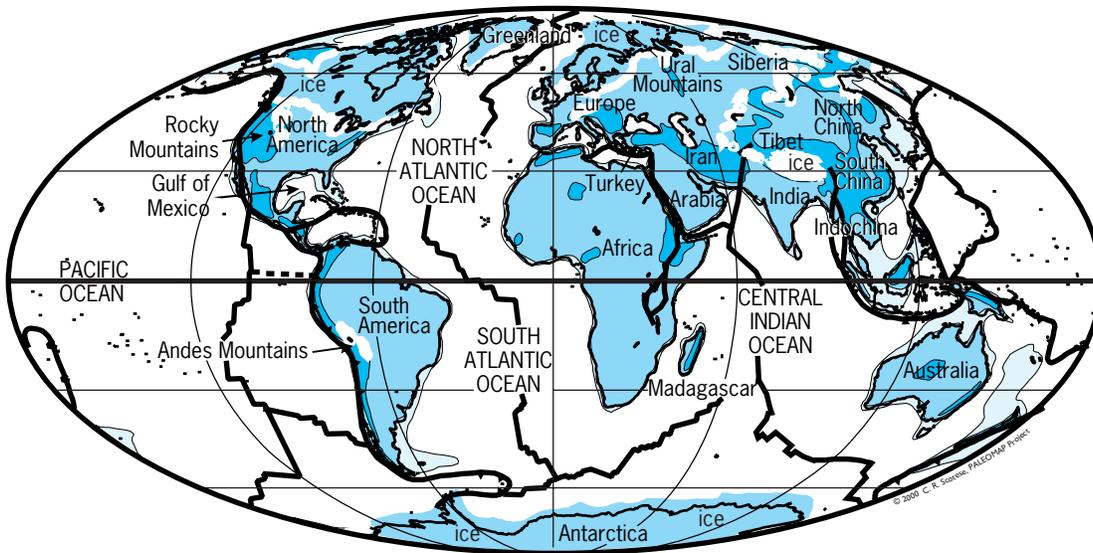
**Cenozoic Era, 66–0 Ma.** The third and final phase in the breakup of Pangea took place during the early Cenozoic. North America and Greenland split away from Europe, and Antarctica released Australia

(**Fig. 6**). Additional, important rifting events have taken place during the last 20 million years of the Cenozoic Era. They include the rifting of Arabia away from Africa, opening the Red Sea; the creation of the East African Rift System; the opening of the Sea of Japan; and the northward motion of California and northern Mexico away from North America. *See* CENOZOIC; EOCENE.

Though several new oceans have opened during the Cenozoic, the last 66 million years of Earth history are better characterized as a time of intense continental collision. The most significant of these collisions was that between India and Eurasia, which began about 50 Ma. During the Late Cretaceous,



**Fig. 6.** Eocene, 50 Ma. (Paleogeographic maps by Christopher R. Scotese, PALEOMAP Project, University of Texas at Arlington)



Key: ■ shallow sea ■ land ■ mountains, uplands

Fig. 7. Last glacial maximum, 18,000 years ago. (Paleogeographic maps by Christopher R. Scotese, PALEOMAP Project, University of Texas at Arlington)

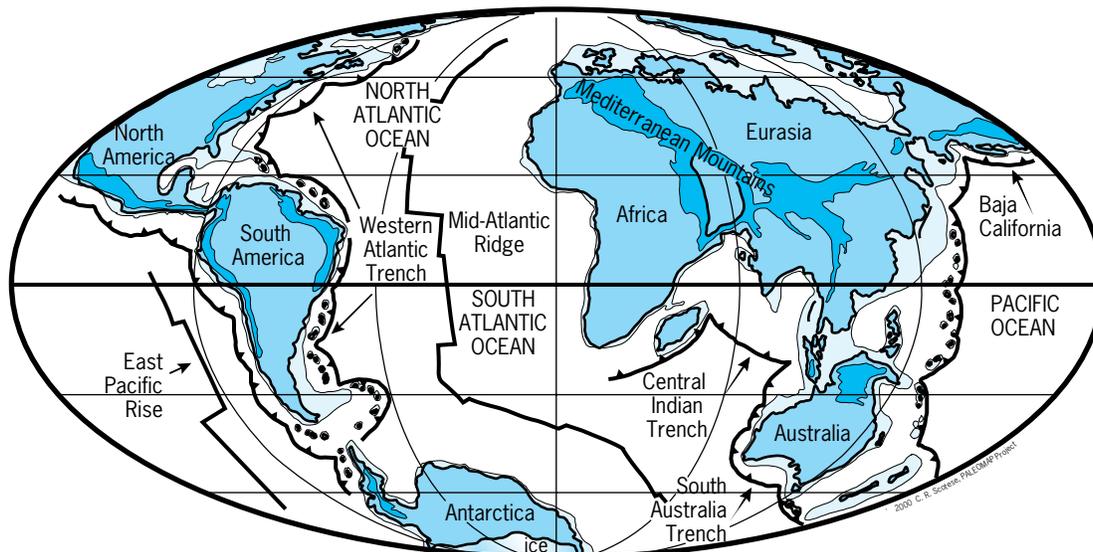
India approached Eurasia at rates of 15–20 cm/yr (6–8 in./yr)—a plate tectonic speed record. After colliding with marginal island arcs in the Late Cretaceous, the northern part of India, called Greater India, began to be subducted beneath Eurasia, raising the Tibetan Plateau.

The collision of India with Asia is just one of a series of continental collisions that has all but closed the great Tethys Ocean. From east to west these continent-continent collisions involved Spain with France, forming the Pyrenees mountains; Italy with France and Switzerland, forming the Alps; Greece and Turkey with the Balkan States, forming the Hellenide and Dinaride mountains; Arabia with Iran,

forming the Zagros mountains; India with Asia; and finally, Australia with Indonesia.

**Modern world.** About 18,000 years ago, all of Antarctica and much of North America, northern Europe, and the mountainous regions of the world were covered by glaciers and great sheets of ice (Fig. 7). These ice sheets melted approximately 10,000 years ago, giving rise to familiar geographic features such as Hudson’s Bay, the Great Lakes, the English Channel, and the fjords of Norway.

The continental rifting and collisions that began in the late Cenozoic are continuing today. Notable are the opening of the Red Sea and Gulf of Aden, the rifting of East Africa, the opening of the Gulf of



Key: ■ shallow sea ■ land ■ mountains, uplands

Fig. 8. In 50 million years hence. (Paleogeographic maps by Christopher R. Scotese, PALEOMAP Project, University of Texas at Arlington)

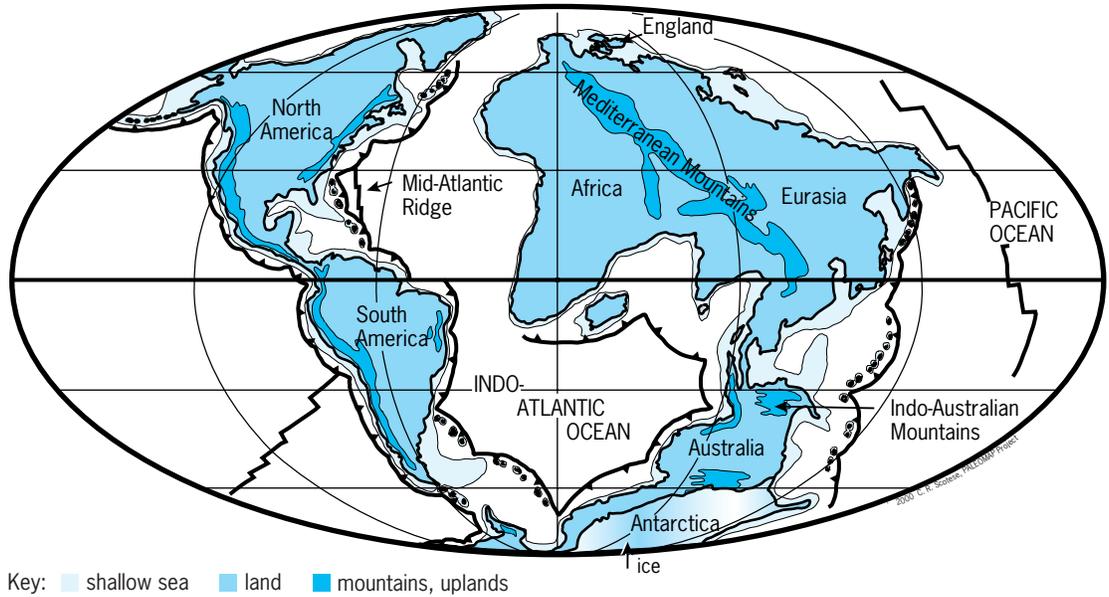


Fig. 9. In 150 million years hence. (Paleogeographic maps by Christopher R. Scotese, PALEOMAP Project, University of Texas at Arlington)

California and the northward translation of California west of the San Andreas Fault, and the incipient collision of Australia with Indonesia giving rise to the mountain ranges of New Guinea. See CENOZOIC.

**Future world.** Though there is no way of knowing the future geography of the Earth, it is possible to project current plate motions and make an educated guess. In the final three maps presented here, the Atlantic and Indian oceans continue to widen until new subduction zones recycle the ocean floor in these ocean basins and bring the continents back together in a new Pangean configuration some 250 million years hence.

The reconstruction of the world 50 million years

in the future looks slightly askew (Fig. 8). North America is rotated counterclockwise, while Eurasia is rotated clockwise bringing England closer to the North Pole and Siberia down to warm subtropical latitudes. Africa has collided with Europe and Arabia, closing the Mediterranean Sea and the Red Sea. Similarly, Australia has beached itself on the doorstep of Southeast Asia, and a new subduction zone encircles Australia and extends westward across the Central Indian Ocean. It is interesting that current plate trajectories suggest that the East African Rift will not grow into a wide ocean.

Though the Atlantic Ocean has widened, extensions of the Puerto Rican Trough and the Scotia Arc

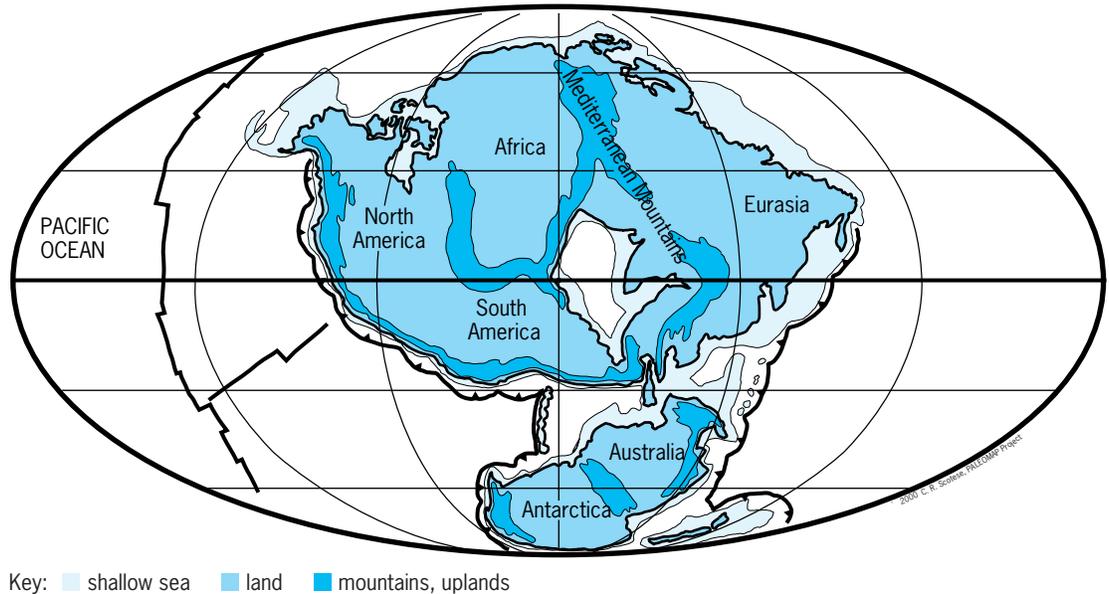


Fig. 10. In 250 million years hence. (Paleogeographic maps by Christopher R. Scotese, PALEOMAP Project, University of Texas at Arlington)

have started a new subduction zone along the eastern edge of the Americas. In time, this subduction zone will consume the North and South Atlantic oceans.

The map of the world 150 million years hence shows the contraction of the Atlantic and Indian oceans. Antarctica has collided with Australia, and the Mid-Atlantic Ridge has nearly been subducted beneath the Eastern American subduction zone (Fig. 9).

In the final reconstruction of future plate motions (Fig. 10), the Indo-Atlantic Ocean is completely closed and a new supercontinent, "Pangea Ultima," was formed. North America has collided back against Africa, but in a more southerly position that juxtaposes Miami with Cape Town. South America wraps around the southern tip of Africa, with Patagonia in contact with Indonesia enclosing a remnant of the Indian Ocean. Antarctica is once again at the South Pole, and the Pacific has grown wider, encircling half the Earth.

Christopher R. Scotese

**Bibliography.** I. W. D. Dalziel, Pacific margins of Laurentia and East Antarctica-Australia as a conjugate rift pair: Evidence and implications for the Eocambrian supercontinent, *Geology*, 19:598–601, 1991; W. S. McKerrow and C. R. Scotese, Palaeozoic Palaeogeography and Biogeography, *Geol. Soc. London Mem.*, no. 12, 1990; C. R. Scotese et al., Paleozoic base maps, *J. Geol.*, 87:217–277, 1979; C. R. Scotese and W. W. Sager (eds.), Mesozoic and Cenozoic Plate Reconstructions, *Tectonophysics*, vol. 155, no. 1–4, Elsevier, Amsterdam, 1989; A. G. Smith, D. G. Smith, and B. M. Funnell, *Atlas of Mesozoic and Cenozoic Coastlines*, Cambridge University Press, 1994; R. Van der Voo, *Paleomagnetism of the Atlantic, Tethys, and Iapetus Oceans*, Cambridge University Press, 1993.

## Paleoindian

The oldest archeological cultures of the New World, the ancestors of modern Native Americans, are termed Paleoindian. These colonizing populations of the Americas were *Homo sapiens sapiens* who arrived during the late Pleistocene (Ice Age) from Asia, though precisely when, and whether in a single or multiple pulses of migration, are not yet known.

**Arrival in America.** The presumed entryway was the Bering Straits, which emerged as dry land during glacial periods, when as much as 5% of the world's water was frozen in massive continental ice sheets (see **illustration**). As glaciers expanded, sea levels dropped 100–150 m below their present levels, exposing shallow parts of the continental shelf worldwide, including that beneath the Bering Sea. The Bering Land Bridge (or Beringia) existed most recently between 25,000 and 11,000 years before present (B.P.), during the last major episode of the Pleistocene. Even at earlier (and later) unglaciated times when Beringia was under water, crossings on foot may have been possible over winter pack ice.

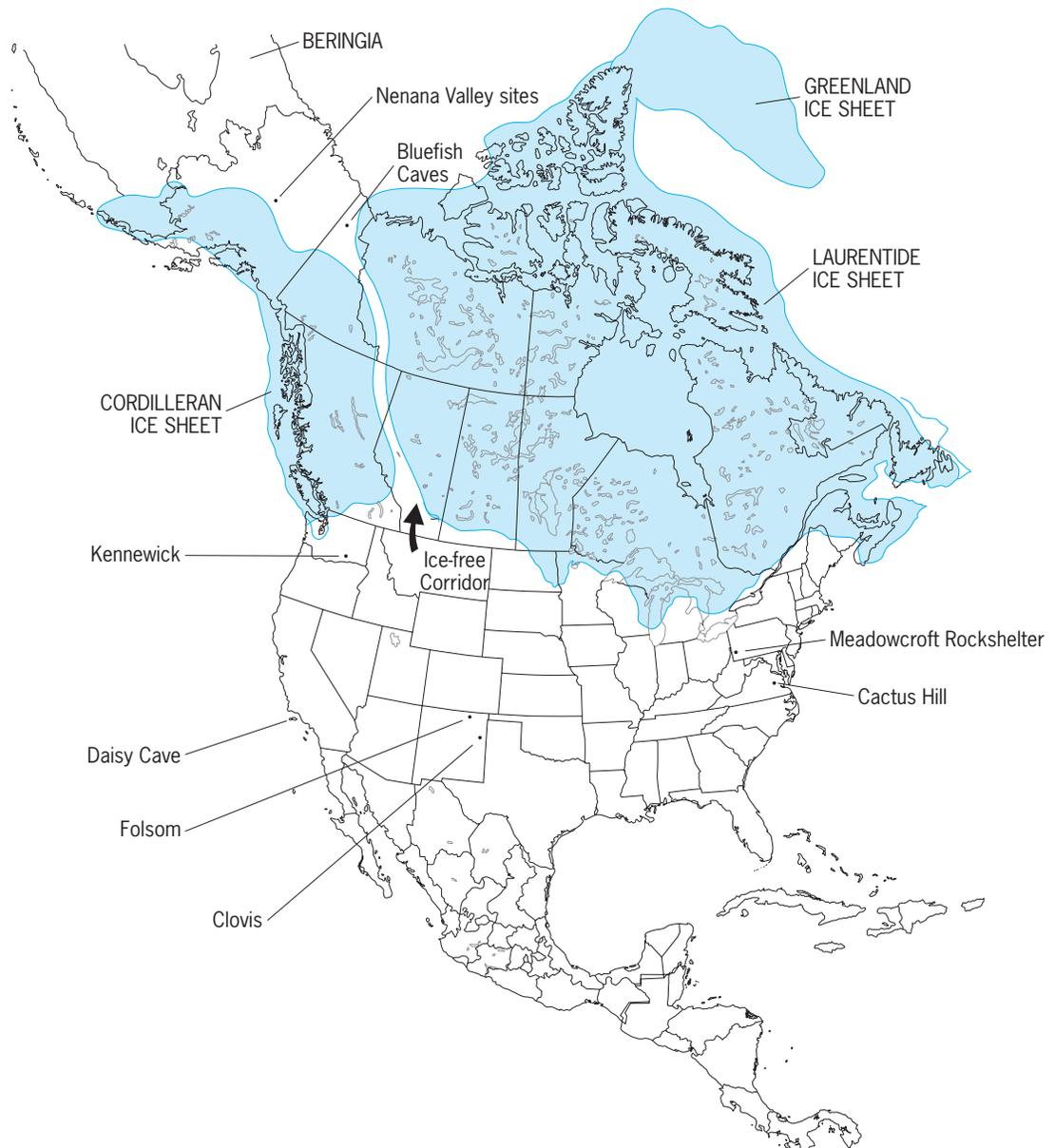
There were no significant geographic or ecological barriers to cross-Beringian traffic. The land bridge was about 1500 km (930 mi) wide, mostly flat, cold, relatively dry, and free of glaciers, except in the mountains on its edges. The absence of barriers explains why the late Pleistocene animal communities of Siberia and Alaska were essentially identical.

The conditions that linked Siberia to Alaska may have simultaneously hindered migration south from Alaska. Groups headed in that direction had two potential routes (broadly defined): down the Pacific coast, or via the continental interior along the eastern flank of the Rocky Mountains. During the late Pleistocene the vast North American glaciers, the Cordilleran and Laurentide, covered much of the Northwest Territories and Canada (at their maximum at 18,000 B.P., extending in western North America from 75° to 48°N), and at times effectively blocked both routes.

Current evidence indicates that a coastal route was largely unglaciated and traversable on foot before 23,000 and after 14/13,000 B.P. (the timing of deglaciation is complex and poorly known). In the millennia in between, 20–50-km-wide tongues of heavily crevassed glacial ice extended down to the coast, obstructing passage. Groups in watercraft may have skirted down the coast, but there is no evidence of such craft, or of a group with the adaptation necessary to survive in a maritime environment (the land being either under ice or resource-poor). By 14,000 B.P., the outer coast was becoming suitable for human colonization: a partly vegetated environment of open tundra with grasses, sedges, and dwarf willows began to develop, and became forested by 12,000 B.P.

For groups headed south via the continental interior, the northern approaches to the eastern slopes of the Rocky Mountains were blocked by the Laurentide glacier, which lapped against the eastern flanks of the Richardson and Mackenzie mountains as early as 30,000 B.P., and remained there until around 11,500 B.P. Groups could have skirted south of those ranges and emerged east of the Rocky Mountains via one of the southern passes (such as along the Liard River), but whether or when such routes would have been traversable is yet unknown. Once the Cordilleran and Laurentide glaciers retreated and opened an ice-free corridor along the Rocky Mountain front, it was 2000–3000 years before the deglaciated landscape was recolonized by plants and animals. Only then, after approximately 11,500 B.P., did the corridor provide sufficient food resources for migrating humans.

There is no evidence yet as to which routes might have been taken. The earliest archeological evidence of a human presence along the Pacific Northwest coast dates to about 9500 B.P., and in the interior corridor to 10,500 B.P.—both well past the presumed entry time. To complicate the picture, the Pleistocene coastline, and any early archeological sites that might be on it, is now mostly underwater following postglacial sea-level rise.



Location of some important North American Paleoindian sites during the late-Pleistocene ice age.

**Antiquity.** So far, the earliest archeologically confirmed dates put human groups in the Lena Basin and Lake Baikal region of northeast Asia at about 39,000 B.P., in subarctic Siberia by 25,000 B.P., but not in western Beringia (such as Kamchatka) until 14,000 B.P. Humans were in eastern Beringia (Alaska) soon after 12,000 B.P., and present south of the ice sheets in North America by at least 11,500 B.P.—the latter represented by the Clovis culture. Yet, the earliest accepted archeological evidence puts human groups in South America earlier still, by at least 12,500 B.P. at the site of Monte Verde, Chile.

There are no obvious historical or technological affinities between Clovis and the Monte Verde materials, suggesting that the two may represent populations with distinct archeological traditions and separate migratory pulses: a later one (Clovis) that came south through the ice-free corridor soon after

it became viable for travel, and an earlier population that perhaps moved along the Pacific coast and reached South America without, so far at least, any traces being found in North America. A hypothesis of multiple migrations, however, must remain tentative, given the small number of early South American sites known (Monte Verde represents one of the only sites of this age, so broader cultural patterns in artifacts or adaptations have yet to be recognized), the liabilities of comparing archeological assemblages so widely separated in time and space, and the evidence from some genetic studies which indicates all Native Americans are descended from a single group. Nonetheless, the possibility of multiple migrations is reasonable, given the absence of barriers to cross-Beringian traffic, and the long period in which potential source populations were present in northeast Asia. There were likely repeated movements from Asia to America and back.

**Progenitors.** None of the artifact complexes evident among late Pleistocene northeast Asian groups appears similar to those of New World Paleoindians. There are, however, archeological assemblages in Alaska (the Nenana Complex) that slightly predate Clovis and are argued to be similar in form and technology and thus historically linked. Yet, these assemblages lack the diagnostic hallmark of the Clovis technology—fluted projectile points. Fluted points (though not Clovis fluted points) are present (if rare) in Alaska, but none of these specimens has been dated earlier than Clovis, and many appear younger than those farther south.

The absence of an obvious Asian (or Alaskan) predecessor has led some to argue that Clovis is derived from Upper Paleolithic groups in western Europe, known as Solutrean, who supposedly crossed the North Atlantic to reach America. However, current archeological evidence, as well as evidence from genetics, human osteology and teeth, and linguistics, provides no support for such a link (or, for that matter, for a link to any other non-Asian source population). The morphological differences between a few ancient human skeletal remains found in the Americas (such as at Kennewick, WA) and those of modern Native Americans have been seen as supporting the hypothesis that the Americas may have been peopled from continents other than Asia. Yet, most likely all these individuals are descended from northeast Asians, the differences in their form over time the result of long periods of geographic isolation, mutation and genetic drift, and evolutionary change among descendant populations.

**North and South America.** Clovis and Clovis-like materials are concentrated in North America, and the northern reaches of Central America. Clovis is a widespread entity that first appears on the western Plains and southwest at 11,500 B.P. and in eastern North America at 10,600 B.P. That Clovis and related groups apparently expanded across the continent in what may have been less than 1000 years is all the more remarkable given that they spread at a time of geologically rapid environmental and climatic change marked by continental deglaciation, the extinction of nearly 36 genera of large mammals (megafauna), and the dissolution of long-standing biotic communities. Yet Clovis groups seemingly coped with such adaptive challenges with ease: Their stylistically distinctive projectile points and tool kits—often including bifacial knives, a variety of unifacial scrapers, occasional blades and flake tools, and (more rarely) bone and ivory implements—are surprisingly similar across the continent. These were highly mobile groups who relied on high-quality stone often obtained from geological sources hundreds of kilometers from the sites where the stone was used and discarded. Their rapid radiation, broadly similar tool kits, and long-distance movement bespeak a cultural “founders effect,” suggesting that their access to large areas of North America was largely unrestricted—these groups were moving across an essentially empty landscape.

Some argue the similarity in the Clovis tool kit and the rapidity of their movement bespeak a uniform adaptation—big-game hunting. Others take the argument a step further, suggesting that Clovis predation on mammoth and mastodon and other megafauna drove these animals to extinction in the late Pleistocene. Yet, the archeological record for big-game hunting is limited to a dozen sites on the Plains and southwest. In most other areas, there is no evidence that big game was exploited. Instead, Clovis subsistence, it appears, more often involved less risky and smaller prey—and presumably plants, though remains of such are rarely preserved in the archeological record of this period.

Although the timing varies by area, by 10,500 B.P. the Clovis tradition was replaced by regional Paleoindian variants, which generally (though not always) have reduced settlement mobility (relative to Clovis), and include new technologies, prey-specific strategies for hunting and processing (such as the intensive use of bison on the Plains), increasing use of local resources, and distinctive stylistic elements and functional artifact forms. This shift from broad and overall homogeneity in Clovis times to narrower and more regionally restricted complexes in later Paleoindian times likely reflects the setting in of colonizers to specific areas or habitats, and increasingly region-specific adaptations.

The South American Paleoindian record, by contrast, does not evince any artifact forms that dominate the archeological landscape as Clovis does. Instead, this period is marked by more diverse unifacial and bifacial stone tool technologies, often made of stone acquired locally (and not necessarily of superior quality), and includes forms such as bolas—modified spherical stones used in slings or as hand missiles. Projectile points tend to be less common in assemblages here than in North America, and show considerable stylistic variety. While Clovis points *per se* are absent from South America, some point forms are fluted, which is often cited as evidence of a historical link between the continents. However, there is growing debate about whether that similarity is merely a case of technological convergence.

South American Paleoindians utilized a wide range of animals, and early on even made occasionally heavy use of plants. This is especially evident at Monte Verde, which yielded (in 12,500 B.P. deposits) nearly 70 species of plants, most locally available but some acquired from the distant highlands and coast. Many had food, medicinal, or economic value. An early (pre-10,000 B.P.) use of plants across the continent was followed within just a few millennia by the emergence of domesticated plants (which occur much earlier here than in North America). Generalized foraging, with occasional big-game hunting, is not unexpected, given the considerable ecological variability and richness of South America. In part, this may reflect the fact that glaciation was not nearly so extensive on this continent—its effects were limited to high altitudes and high latitudes. While that prevented early colonization of those areas, in the remainder of the continent early

immigrants could move freely throughout environments that were fairly dynamic in some areas, relatively stable in others, and in places ecologically rich. There was also use of nonterrestrial environments: early use of maritime resources is evident at several sites on the Atlantic and (especially) the Pacific coasts. Such adaptations played a key role in establishing early sedentary lifestyles.

Once the founding population dispersed across South America (over an unknown length of time), subgroups became geographically isolated relatively quickly. From the earliest known site at 12,500 B.P. (Monte Verde) until the end of the Pleistocene (10,000 B.P.), there is a continuing diversification in tool forms and technology, evidently reflecting less mobility, increasing heterogeneity and regional mosaics in culture and adaptations, and less expansive social networks and territories.

All told, it is a different trajectory from the one that unfolded in North America—testimony that the earliest colonization of the two continents, though ultimately derived from the same northeast Asian source, may have taken place at different times under very different circumstances. See ARCHEOLOGY; PLEISTOCENE; PREHISTORIC TECHNOLOGY.

David J. Meltzer

Bibliography. J. M. Adovasio and J. Page, *The First Americans: In Pursuit of Archaeology's Greatest Mystery*, Random House, New York, 2002; R. Bonnichsen and K. Turnmire (eds.), *Clovis: Origins and Adaptations*, Center for the Study of the First Americans, Oregon State University, Corvallis, 1991; T. D. Dillehay, *The Settlement of the Americas: A New Prehistory*, Basic Books, New York, 2000; N. Jablonski (ed.), *The First Americans: The Pleistocene Colonization of the New World*, California Academy of Sciences and University of California Press, Berkeley, 2002; D. J. Meltzer, *Search for the First Americans*, Smithsonian Books, Washington, DC, 1993; F. H. West (ed.), *American Beginnings*, University of Chicago Press, 1996.

## Paleolithic

The prehistoric period when people made stone tools exclusively by chipping or flaking. John Lubbock proposed and defined the term Paleolithic, or Old Stone Age, in 1865, and also defined a subsequent stage, Neolithic or New Stone Age, during which some stone tools were formed by polishing or grinding. Later archeologists altered these definitions; to many today the Paleolithic is the period during which human beings lived entirely by hunting and gathering, while the Neolithic is the following interval during which plant and animal domestication was introduced. To other archeologists, the Paleolithic is simply a time interval, roughly equivalent to the Pleistocene Epoch, while the Neolithic comprises the early part of the succeeding Holocene (or Recent) Epoch. See GEOLOGIC TIME SCALE; NEOLITHIC.

The different definitions sometimes lead to contradictory results, as when hunter-gatherers making chipped stone tools are found to have survived into the Holocene (even to historic times) or when polished stone tools appear to have been made by people who did not know agriculture. To eliminate the contradictions or to refine the definitions, some specialists inject a transitional Mesolithic between the Paleolithic and Neolithic. It is impossible, however, to devise a rigorous, global definition of the Paleolithic or any other cultural stage, because artifact technology and economic practices have changed independently at different times in different parts of the world. Hence, Paleolithic will be used here informally to refer to the time interval between the earliest appearance of stone tools, more than 2.5 million years before present (m.y. B.P.) and the end of the last glacial period, 12,000–10,000 years B.P. (Fig. 1).

**Artifacts.** The oldest artifacts found so far come from sites in Ethiopia, Kenya, and Tanzania where they are dated to between 2.6 and 1.6 m.y. B.P. They comprise crude flakes and the modified pebbles and stone chunks from which the flakes were struck (Fig. 2). Collectively they are often assigned to the Oldowan Industrial Complex, named after Olduvai Gorge (Tanzania). Some researchers believe that the Omo industrial tradition, found in a few places in East Africa, preceded the Oldowan. These “tools” are little more than smashed nodules, requiring less technical skill to make than those of the Oldowan. Older artifacts such as the Omo type may be difficult to find even if they exist elsewhere, because their makers might not yet have developed the uniquely human habit of accumulating refuse at repeatedly occupied sites with good archeological visibility.

While it is not certain which type of australopithecine, the earliest of human ancestral forms, may have produced the earliest stone tools, the principal makers of Oldowan artifacts were probably members of the subsequent human species, *Homo habilis*. Damage marks on associated animal bones show that Oldowan tools were sometimes used for butchering, while wear traces observed on the tools themselves show that some were used to cut meat or reeds, and others functioned to scrape or saw wood. Any wooden artifacts that resulted perished long ago in the ground.

Approximately 1.6–1.5 m.y. B.P., at least some people in East Africa, most likely a *Homo erectus* type, began to manufacture the bifacially flaked tools known to archeologists as hand axes. The flaking of the first hand axes was crude, and they differ only subtly from earlier Oldowan flaked pebbles and hunks, which continued to be made. As time passed, however, their flaking tended to become more refined, and the thinness and bilateral symmetry of some later hand axes may reflect esthetic as well as functional considerations.

By 1 m.y. B.P. hand-ax makers had spread through most of Africa and the Near East, and by 900,000–600,000 years B.P. they had reached Europe. All African, Near Eastern, and European artifact assemblages with hand axes are conventionally placed in

the Acheulean Industrial Complex, after the site of St. Acheul, northern France, discovered in the 1850s (Fig. 3).

Acheulean artifacts were made by *Homo ergaster*/*H. erectus* before 500,000–400,000 years B.P., and by early *H. sapiens* after this time. The Acheulean Complex does not appear to have extended to the Far East, where *H. erectus* and perhaps early *H. sapiens* produced artifact assemblages broadly similar to Oldowan ones. It does not follow that the Far East lagged behind Africa and Europe technically, however, since hand ax-like bifacial artifacts have now been recorded at several east Asian early Paleolithic sites, and many important early Paleolithic assemblages in the west also lack hand axes. Early peoples of the Far East simply appear to have employed different means of solving their technological problems than the producers of Acheulean artifacts did. See FOSSIL HUMANS.

**Lower Paleolithic.** In sub-Saharan Africa the Oldowan and Acheulean complexes are frequently placed together in the Early Stone Age, while elsewhere Acheulean and contemporaneous assemblages are commonly referred to the Lower Paleolithic. The Early Stone Age/Lower Paleolithic apparently persisted until sometime between 200,000 and 130,000 years B.P., the exact time perhaps depending on the region.

**Middle Paleolithic.** In sub-Saharan Africa the Early Stone Age was succeeded by the Middle Stone Age, while in North Africa and Eurasia the Lower Paleolithic was followed by the Middle Paleolithic, also commonly known as the Mousterian, after the cave of Le Moustier, southwestern France, where rich Middle Paleolithic levels were first excavated in the 1860s.

Most Middle Stone Age/Middle Paleolithic assemblages lack hand axes; when hand axes are present they tend to be far smaller and more delicate than Acheulean ones. The principal Middle Stone Age/Middle Paleolithic tools are well-made stone flakes, often modified by edge flaking (“retouch”) into types called sidescrapers, knives, denticulates (serrate-edged pieces), and so forth (Fig. 4). The typology is based on form alone, and wear traces on some tools show that pieces assigned to a single type may have been used for different purposes. Wear traces also reveal that unmodified flakes were often used as well.

In Europe, Middle Paleolithic people were the Neandertals, *H. (sapiens) neanderthalensis*. Neandertals also lived in the Near East during the early Middle Paleolithic. Neandertal occupation of the region apparently overlapped with that of the some of the first anatomically modern people (*H. sapiens sapiens*) who disappear from the archaeological record for some time, but then return and supplant the Neandertals after ?70,000–60,000 years B.P. Outside of Europe and the Near East, bones representing Middle Paleolithic/Middle Stone Age people are rare, very fragmentary, or both. The few known African fossils suggest that earlier Middle Stone Age people belonged to an archaic variety of *H. sapiens*, while later

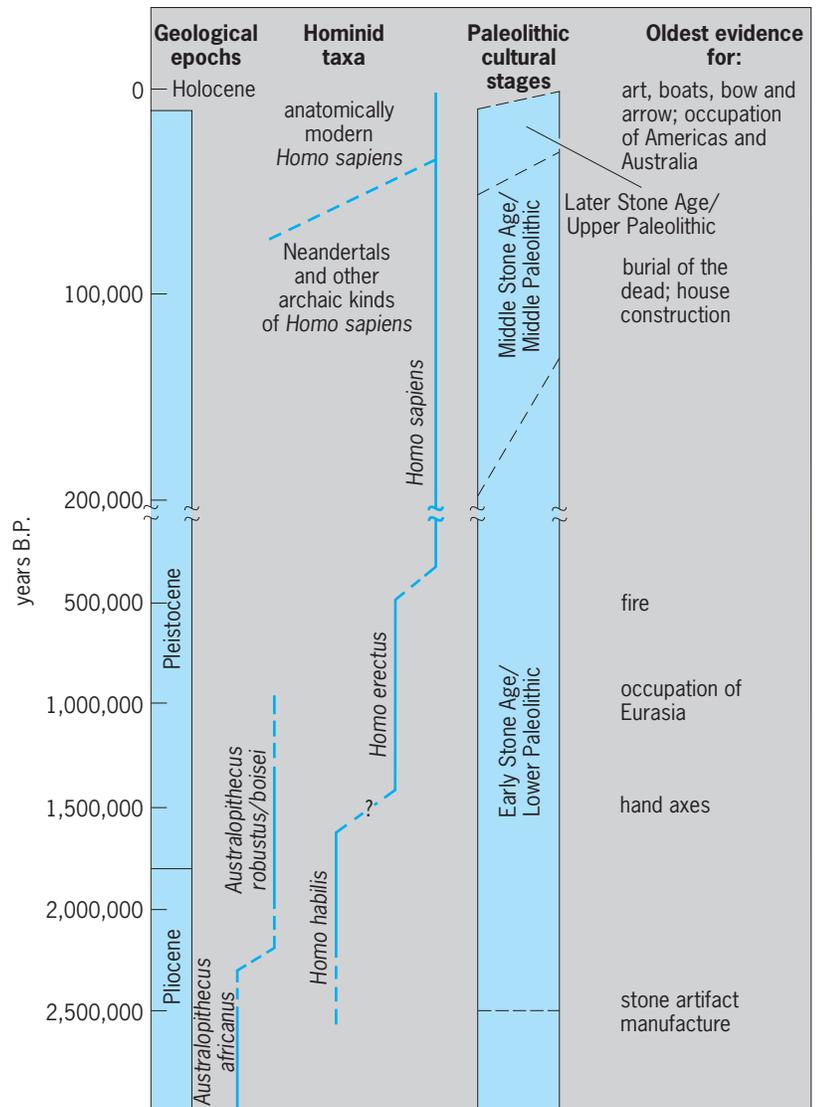


Fig. 1. Relationship between time, hominid taxa, and major cultural developments during the Paleolithic.

Middle Stone Age people (after ?70,000–60,000 years B.P.) may have been anatomically modern. See NEANDERTALS.

**Upper Paleolithic.** In Europe, the Near East, and North Africa, the Middle Paleolithic was followed by the Upper Paleolithic, perhaps 50,000 years B.P. in the Near East, adjacent North Africa, and eastern Europe, and beginning around 40,000 years B.P. in western Europe. The time difference within Europe

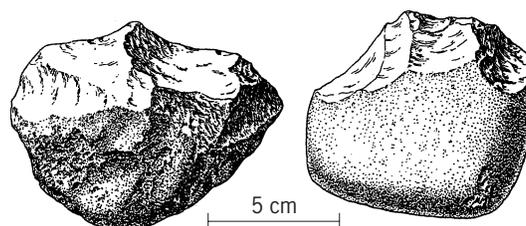


Fig. 2. Oldowan pebble tools from Bed I, Olduvai Gorge. (After M. D. Leakey, *Olduvai Gorge*, vol. 3, p. 27, Cambridge University Press, 1971)

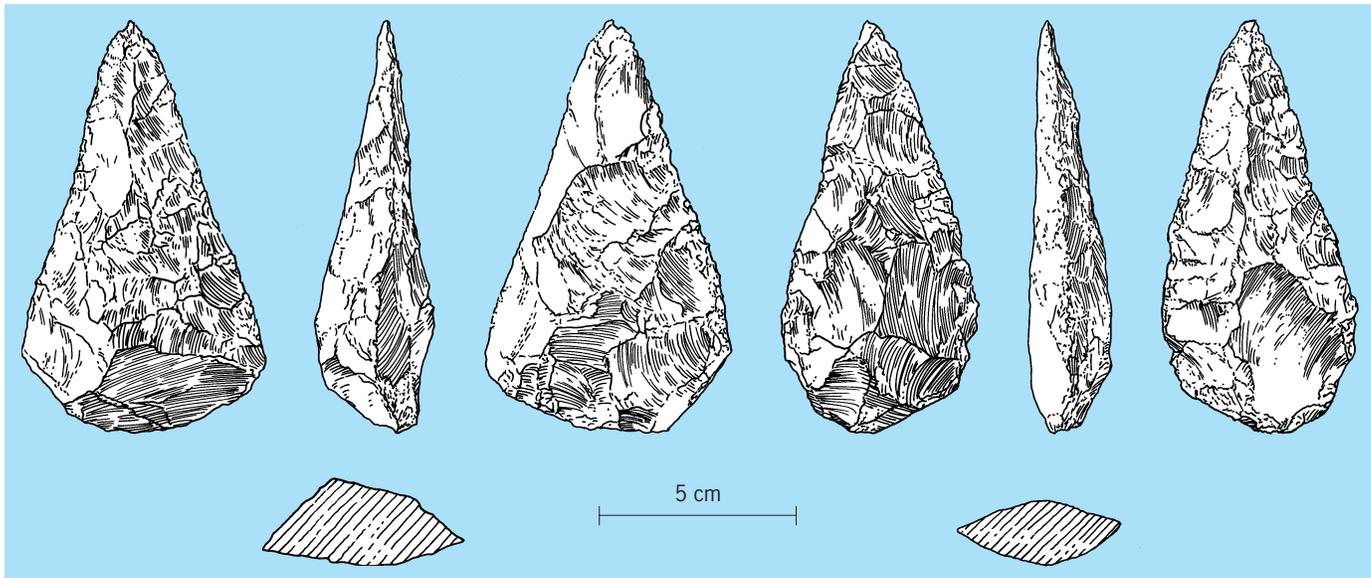


Fig. 3. Late Acheulean hand axes from England. (After D. A. Roe, *The Lower and Middle Palaeolithic Periods in Britain*, Routledge Kegan Paul, 1981)

may reflect the movement of technology, people, or both from east to west. In sub-Saharan Africa, the Middle Stone Age was replaced by the Later Stone Age between 50,000 and 35,000–30,000 years B.P., the precise time again perhaps depending on the region. With the exception of one individual from a site in France, all known Upper Paleolithic/Later Stone Age people were anatomically modern. However, human skeletal remains (archaic or modern) dating to the time of cultural and technological transition itself are extremely rare, and the noted exception is a Neandertal from a level that not all specialists would in fact assign to the Upper (versus the Middle) Paleolithic. The disagreement is part of an ongoing debate about whether the Upper Paleolithic (and modern people) were intrusive into western Europe or evolved there more or less independently.

The Upper Paleolithic and Later Stone Age are difficult to characterize artifactually, because wherever they are well known, they exhibit great variability in time and space. It is often said that Upper Paleolithic “cultures” shared a tendency to manufacture tools on flakes that are at least twice as long as wide (such pieces are called blades) and that the principal Upper Paleolithic tool types were end scrapers and burins (Fig. 5). While most Upper Paleolithic assemblages do exhibit these features, there are some that do not, and even among those that do, there is extraordinary variability in the actual frequency of long flakes (blades) and in the kinds of end scrapers and burins that were made. The amount of artifactual change through time and space during the Upper Paleolithic/Later Stone Age far exceeds that in earlier periods and suggests an ability to innovate that earlier people did not exhibit.

Among the most important Upper Paleolithic/Later Stone Age innovations were the spearthrower, the bow and arrow, and tailored clothing. These projectile technologies help to explain why Upper Paleolithic/Later Stone Age people were more effective

hunters than their predecessors. The invention of tailored clothing helps explain how Upper Paleolithic people managed to colonize northeastern Europe and northern Asia (Siberia), which were apparently uninhabited earlier. See PREHISTORIC TECHNOLOGY.

Upper Paleolithic/Later Stone Age people were also the first to manufacture standardized, formal tool types from bone, antler, and ivory. These include objects that archeologists call points, awls, and needles. In addition, they were the first people to produce what anthropologists all agree is art, including both painted and engraved wall art and numerous portable items—figurines, pendants, beads, and other objects (Fig. 6). These were often made of ivory, bone, and antler and their appearance may be connected to Upper Paleolithic/Later Stone Age realization that such organic material could be carved, ground, or otherwise shaped into a wide variety of forms.

Conventionally, the Upper Paleolithic is said to end with the end of the Last Ice Age, 12,000–10,000 years B.P., although Upper Paleolithic/Later Stone Age artifact types and economic practices continued for many millennia throughout much of Eurasia and sub-Saharan Africa.

**Economy.** Throughout the long Paleolithic time span, all human beings lived by hunting and gathering wild resources. Only from the very end of the Paleolithic, about 12,000–10,000 years B.P., is there evidence that some people domesticated animals, plants, or both. See DOMESTICATION.

Like historic hunter-gatherers, most Paleolithic people collected edible wild plants. In lower and middle latitudes, plant foods probably supplied the bulk of Paleolithic diets, but direct evidence is lacking, due to the perishability of plant tissues. In contrast, animal bones, often representing food debris, are preserved at many sites, and have led archeologists to focus on the meat component of the diet out of proportion to its probable importance.

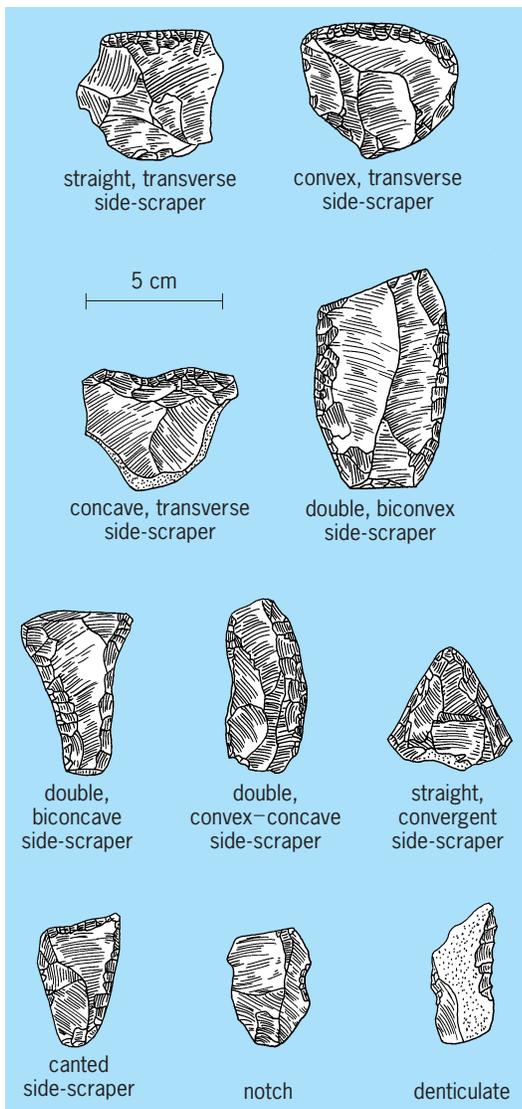


Fig. 4. Middle Paleolithic (Mousterian) flake tools from France. (After R. G. Klein, *Ice-Age Hunters of the Ukraine*, University of Chicago Press, 1973)

Reconstructing the diet of Early Stone Age/Lower Paleolithic people is complicated not only by preservation bias, but also by the geomorphic setting of nearly all the known sites. These are open-air (versus cave) occurrences located on the margins of ancient streams and lakes. At many sites, water action and perhaps kicking by drinking animals have displaced both artifacts and animal bones from their original positions. As a result, spatial associations that might have reflected ancient human butchering activity have been destroyed. There is the further problem that coprolites (fossil feces) and gnawed bones found at many sites show that carnivores were present. At no site can it be simply assumed that people killed all the animals represented or even that they scavenged the carcasses. The problem is to estimate the relative roles of people and other agents. For the moment, the best guess is that as time passed, meat became progressively more important in Lower Paleolithic diets and that human scavenging perhaps declined relative to hunting, but this cannot be demonstrated.

Demonstrating the importance of meat in Middle Stone Age/Middle Paleolithic and Later Stone Age/Upper Paleolithic diets is less problematic, because many of the relevant sites are caves in which people were the principal or sole bone accumulators. This is indicated by the abundance of artifacts, fossil hearths, and other traces of human activity (including cut animal bones) and by the rarity or absence of carnivore coprolites and gnawed bones. From such sites, it is possible to show that the people ate mainly medium-sized ungulates, including antelopes and zebra in Africa, and deer and wild horses in Eurasia.

It is also possible to show that Later Stone Age/Upper Paleolithic people probably ate more meat than Middle Stone Age/Middle Paleolithic people. Bone assemblages from southern African sites indicate that only Later Stone Age people regularly caught fish and flying birds and that they obtained truly dangerous prey such as wild pigs much more frequently than their predecessors did. It is probably not coincidental that only Later Stone Age sites have provided artifacts directly interpretable as fishing and fowling implements along with pieces that were almost certainly parts of arrows. Armed first with the spear-thrower, then later with the bow and arrow, Later Stone Age hunters could have attacked prey from a greater distance, increasing their chances of success and reducing their personal risk.

**Other aspects of culture.** The oldest reasonably secure evidence for human use of fire comes from the famous Peking man (*H. erectus*) site of Zhoukoudian in north China, tentatively dated to 500,000–240,000 years ago. More equivocal evidence for older or equally old controlled use of fire has been found in Kenya, South Africa, and Europe. Unequivocal fireplaces are found in many sites occupied by European Neandertals and their near-modern African

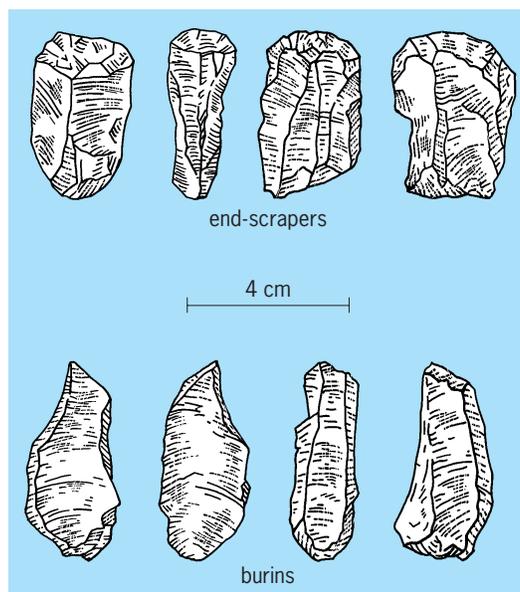


Fig. 5. Upper Paleolithic flake and blade artifacts from the Ukraine. (After R. G. Klein, *Ice-Age Hunters of the Ukraine*, University of Chicago Press, 1973)

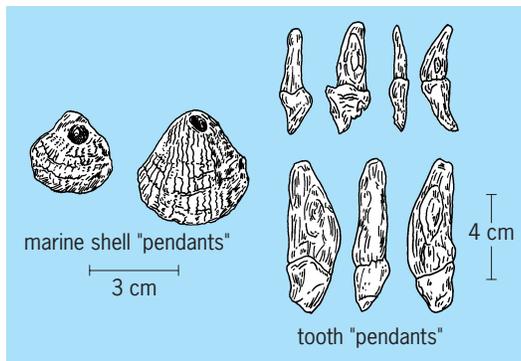


Fig. 6. Upper Paleolithic art objects from the Ukraine. (After R. G. Klein, *Ice-Age Hunters of the Ukraine*, University of Chicago Press, 1973)

contemporaries after 127,000 years B.P. The paucity of evidence for controlled use of fire during the Lower Paleolithic does not necessarily indicate a more limited ability to make fire. Rather, the Lower Paleolithic evidence is mostly from open-air sites near ancient water sources where charcoal fragments and concentrations that mark hearths are much less likely to survive than they are in cave sites, whence comes most of the Middle Paleolithic/Middle Stone Age evidence post 127,000 years ago.

Concentrations of rocks or other debris that may mark the positions of ancient structures such as wind breaks and flimsy tentlike shelters have been uncovered at several Lower Paleolithic sites in Africa and across Eurasia, but the oldest widely accepted "ruins" date from the Middle Paleolithic. However, it is not until 50,000–40,000 years ago that evidence for housing becomes compelling, with such features as fireplaces and substantial wall supports being commonplace.

Upper Paleolithic sites frequently contain spreads of large bones, artificially excavated depressions, or post holes marking the spot of ancient dwellings. Hearths, sometimes elaborate, are always present, and cache or storage pits for ivory, art objects, and other valuable items are common. At some sites, there are multiple ruins, which, if occupied simultaneously, would imply social groups probably far larger than any which existed earlier. Carefully constructed, well-heated dwellings were probably important in permitting Upper Paleolithic people to colonize northeastern Europe and Siberia for the first time.

Finally, the oldest undeniable evidence for burial of the dead comes from the Middle Paleolithic of Europe and the Near East. The deceased were Neandertals, and their skeletons sometimes exhibit pathologies or deformities that must have incapacitated the owners long before death. Perhaps these people could only have survived with care and aid from other members of their group.

Upper Paleolithic people also buried their dead, and the pathological or senile state of some skeletons again suggests a characteristically human care for sick or aged comrades. In addition, unlike Middle Paleolithic graves, Upper Paleolithic ones often contain special objects, such as hundreds of ivory beads,

carved pendants, or other body ornaments, perhaps the deceased's personal belongings or items to facilitate the transition to an afterlife. Upper Paleolithic graves, together with Upper Paleolithic art, provide the oldest available evidence for the intangible part of culture called ideology or religion.

**Population expansion.** Human population growth during the Paleolithic is reflected in the ever-wider area that was occupied. In addition, population probably tended to increase in those areas that had long been inhabited.

Problems in finding and dating Lower Paleolithic/Early Stone Age sites make it impossible to compare their number per unit time to the numbers of later sites, but there are places where the number of Middle Paleolithic sites can be compared to the number of Upper Paleolithic ones. In each instance, Upper Paleolithic sites are much more numerous per unit time, suggesting larger Upper Paleolithic populations. This probably reflects Upper Paleolithic advances in technology, especially related to hunting.

Upper Paleolithic expansion to Siberia was important not only in its own right, but also because it was essential for the colonization of the Americas, which almost certainly occurred across a dry land bridge linking Siberia and Alaska during the Upper Paleolithic. This bridge owed its existence to the drop in sea level caused by the growth of the great continental ice sheets. Conceivably, people reached Alaska only shortly after they occupied Siberia (35,000–30,000 years B.P.), but this remains to be shown. There is a heated debate about when people reached the North American continent and regions farther south. The oldest universally accepted evidence for human presence is only about 15,000 years old. See PALEOINDIAN.

Australia was also probably first colonized in Upper Paleolithic times, 40,000–35,000 years B.P., or possibly by 60,000 years ago. In this case, there was no land bridge to the Asian mainland, and some knowledge of boats is implied. In both the Americas and Australia, the arrival of people appears to have been followed shortly by a wave of animal extinctions. If, as some specialists believe, people were largely responsible, then Upper Paleolithic times witnessed not only important innovations but also the first humanly caused ecological catastrophes.

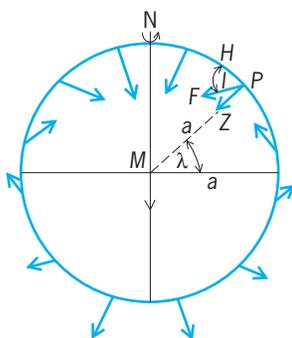
Richard G. Klein; Anne Pike-Tay

Bibliography. O. Bar-Yosef, The Upper Paleolithic Revolution, *Annu. Rev. Anthropol.*, 31:363–393, 2002; R. G. Klein, *The Human Career: Human Biological and Cultural Origins*, 2d ed., University of Chicago Press, 1999; D. Lewis-Williams, *The Mind in the Cave: Consciousness and the Origins of Art*, Thames & Hudson, London, 2002; A. Pike-Tay and R. Cosgrove, From reindeer to wallaby: Recovering patterns of seasonality, mobility, and prey selection in the Paleolithic Old World, *J. Archaeol. Method Theory*, 9(2):101–146, 2002; C. Stringer, R. Barton, and J. Finlayson (eds.), *Neandertals on the Edge*, Oxbow Books, Oxford, 2000; I. Tattersall and J. Schwartz, *Extinct Humans*, Westview Press, 2001.

## Paleomagnetism

The study of the direction and intensity of the Earth's magnetic field through geologic time. Paleomagnetism is an important tool in measuring the past movements of the Earth's tectonic plates. By studying ancient magnetic field directions recorded in rocks, scientists learn how the plates moved relative to the Earth's spin axis and relative to one another. The calibrated history of geomagnetic polarity reversals provides a basis for the temporal correlation of rocks on a local to global geographic scale, called magnetostratigraphy. See PLATE TECTONICS; ROCK MAGNETISM; STRATIGRAPHY.

**Magnetic field behavior.** At all points on the Earth's surface, the geomagnetic field is represented by a vector of specific length and direction. Two features of the geomagnetic field are particularly useful. First, when averaged over  $10^4$  to  $10^5$  years, the field can be sufficiently represented by a dipole magnet located at the Earth's center with its north-south axis aligned with the rotation axis (Fig. 1). In this model, the time-averaged magnetization vector from a rock sequence magnetized on the Equator would be inclined at  $0^\circ$  with respect to local horizontal and would point toward the north geographic pole. If the same rock sequence were magnetized at the north pole, the time-averaged magnetization vector would be inclined vertically downward. For all points in between, the relationship between latitude and inclination is  $I = 2 \tan \lambda$ , where  $I$  is inclination and  $\lambda$  is latitude. Thus, if the magnetization vectors of  $10^8$ -year-old rocks now at  $30^\circ\text{N}$  have a mean inclination of  $0^\circ$ , these rocks (and probably the tectonic plate upon which they lie) have moved  $30^\circ$  northward since the magnetization was acquired. Longitudinal motion is undetectable. For example, rocks formed anywhere on the Equator will acquire a magnetization with  $0^\circ$  inclination, and subsequent motion of the rocks along the Equator results in no latitudinal change.



Key:

$P$ = arbitrary point on the Earth's surface	$H$ = horizontal component
$a$ = radius of the Earth	$M$ = magnetic moment of the dipole
$\lambda$ = latitude	$F$ = total field
$Z$ = vertical component	$I$ = inclination of the total field

Fig. 1. Field of an axial geocentric dipole. (After M. W. McElhinny, *Palaeomagnetism and Plate Tectonics*, Cambridge University Press, 1973)

On time scales shorter than about  $10^5$  years, the Earth's field at a given point exhibits considerable change in both direction and intensity. These variations, termed secular variations, can be useful for detailed stratigraphic correlation over a limited geographical range and time interval.

A second useful feature of the geomagnetic field is its period polarity reversals. The time-averaged axis of the central dipole remains parallel to the rotational axis before and after the reversal, but the north-seeking pole becomes the south-seeking pole and vice versa. Polarity changes from normal (N) to reverse (R) occur over 3000–5000 years and are effectively an instantaneous geological markers for stratigraphic correlation. Complete N-R-N or R-N-R sequences occur on time scales on the order of  $10^5$  to  $10^7$  years, making the geomagnetic polarity time scale a very powerful chronological tool.

**Magnetization of rocks.** Many rocks acquire remanent magnetizations when they form. These magnetizations are usually parallel to the direction of the ambient magnetic field at the time. Magnetic minerals in igneous rocks, such as basalt or gabbro, acquire a thermoremanent magnetization (TRM) as they cool below the Curie point  $T_C$  of the mineral [ $T_C$  for magnetite,  $\text{Fe}_3\text{O}_4$ , is  $1076^\circ\text{F}$  ( $580^\circ\text{C}$ ) and for hematite,  $\text{Fe}_2\text{O}_3$ ,  $1238^\circ\text{F}$  ( $670^\circ\text{C}$ )]. This equilibrium TRM is later "frozen" in the rock at the blocking temperature  $T_B$ , whereupon the magnetization vector is no longer in equilibrium with the Earth's magnetic field and can theoretically remain stable for periods longer than  $10^{11}$  years. Blocking temperatures for many magnetite-bearing rocks range from  $930$  to  $1080^\circ\text{F}$  ( $500$  to  $580^\circ\text{C}$ ). If a preexisting rock is later heated to temperatures near or exceeding  $T_B$ , it may acquire a new magnetization upon cooling. See CURIE TEMPERATURE; IGNEOUS ROCKS.

When sediment accumulates in a basin, magnetic grains sink through the water column and accumulate at the sediment-water interface together with other sedimentary grains. There, they may become aligned with the ambient magnetic field, and a detrital remanent magnetization (DRM) is the result. As sediment is transformed into sedimentary rock by diagenesis and lithification, a previously acquired DRM can be modified by compaction and dewatering, and an additional chemical remanent magnetization (CRM) may be acquired by the rock. CRM is acquired by the formation of new magnetic minerals as they grow through a critical grain size (about 0.2 micrometer for magnetite,  $\text{Fe}_3\text{O}_4$ ). Chemical remanent magnetization may also be acquired at higher temperatures during alteration of plutonic, volcanic, and metamorphic rocks. See SEDIMENTARY ROCKS.

An original TRM, DRM, or CRM acquired by a rock when it formed may be partially or completely overprinted by a younger TRM or CRM in a subsequent geologic event such as metamorphism or hydrothermal alteration. Rocks may also be overprinted by viscous remanent magnetization (VRM), the statistical thermal realignment of the magnetization of smaller grains parallel to a younger field direction. At the surface, lightning strikes create large electric

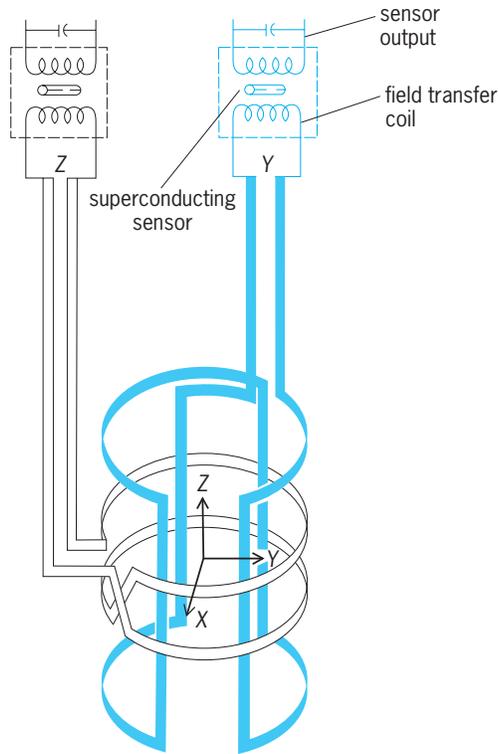


Fig. 2. Measuring head of the SQUID magnetometer (two-axis system).

currents that can impart a spurious isothermal remanent magnetization (IRM). Various demagnetization techniques are employed to selectively remove these undesired magnetizations, isolating the older magnetizations residing in the rock.

**Collection and measurement of samples.** A sequence of carefully oriented samples is collected, spanning a time interval long enough to average geomagnetic secular variation (for a paleomagnetic study) or spanning a stratigraphic interval (for a magnetostratigraphic study). The samples are cut into small cylinders whose magnetization is measured by using a sensitive magnetometer. Modern instruments use fluxgate or SQUID (Superconducting Quantum Interference Device) detectors with very high sensitivity and low noise (Fig. 2). See SQUID.

Resolution of the sample's magnetization vectors is accomplished using partial demagnetization techniques. In alternating field demagnetization, the sample is subjected to a sinusoidal alternating magnetic field of smoothly decreasing amplitude within a region of near-zero magnetic field. This procedure progressively randomizes the magnetic grains whose magnetic coercive force is less than or equal to the peak alternating field intensity. These treatments are useful in removing IRM and in resolving one generation of TRM from another. In thermal demagnetization, the sample is heated and then cooled within a region of near-zero magnetic field. By heating the rock to progressively higher temperatures, magnetizations with differing blocking temperatures can be selectively removed. Thermal demagnetization is useful for isolating multiple TRM vector directions

and for removing recent VRM. Chemical demagnetization is done by immersing rock samples in concentrated hydrochloric acid for periods of days to weeks. The acid progressively dissolves magnetic grains from the outside in, generally the reverse order in which the CRM was acquired. Chemical demagnetization is useful in separating multiple CRM components, and CRM and DRM in sedimentary rocks. Low-temperature treatments by immersion in a cryofluid such as liquid nitrogen are occasionally applied; this procedure tends to remove the magnetization of large, multidomain grains.

**Interpretation.** The end product of the laboratory experiments is a suite of magnetization vector directions defined by the inclination  $I$ , the angle that the magnetization vector makes with the horizontal, and the declination  $D$ , the angle that the projection of the magnetization vector upon a horizontal plane makes with true north. Provided that the samples have recorded the geomagnetic field for  $10^4$ – $10^5$  years, their collective directions will average geomagnetic secular variation, and representative mean  $D$  and  $I$  values and an associated uncertainty in direction may be calculated by using statistical techniques. The mean declination and inclination, together with the inclination-latitude relationship, mentioned previously, and some elementary spherical trigonometry, allow the calculation of a representative paleomagnetic pole from the samples. By connecting paleomagnetic poles of different ages in an ordered time sequence, an apparent polar wander path (APWP) may be constructed for a tectonic plate (Fig. 3). The APWP specifies the displacement

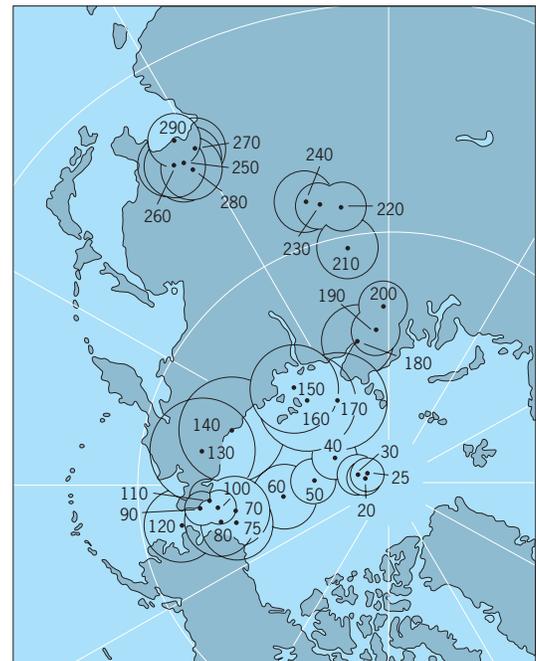


Fig. 3. Apparent polar wander path for North America for Late Carboniferous to recent time. The numbers represent time in millions of years before present. Circles represent one standard error about the mean. (After E. Irving, *Paleopoles and paleolatitudes of North America and speculations about displaced terranes*, *Can. J. Earth Sci.*, 16:669–694, 1979)

history of a plate or continent with respect to the spin axis, and can be directly compared with AP-WPs from other plates to determine whether relative movements have occurred. For times before about  $2.5 \times 10^8$  years before present, APWP information is the only direct method of obtaining relative motion information for the continents. Differing AP-WPs from North America and Europe demonstrated that sea-floor spreading had occurred in the northern Atlantic, hastening the acceptance of continental drift by the earth science community and presaging the development of plate tectonics.

The end product of a magnetostratigraphic study is an ordered set of normal and reversed magnetizations from a rock sequence. By comparing N-to-R and R-to-N transitions in the rock sequence with a reference polarity time scale, and by incorporating other geologic information, magnetostratigraphy makes it possible to correlate rock sequences over intercontinental distances. A reliable polarity time scale has been constructed for about the past  $1.5 \times 10^8$  years. Together with the apparent polar wander path records from continents, this time scale has proved invaluable in developing plate tectonic theory to its present advanced state. While extension of magnetostratigraphy to earlier times is theoretically possible, the technique is proportionally less accurate because uncertainties in the radiometric calibration of the reversal time scale can exceed the length of at least some of the polarity intervals themselves. See GEOMAGNETISM; GEOPHYSICAL EXPLORATION.

Michael McWilliams

**Bibliography.** R. F. Butler, *Paleomagnetism: Magnetic Domains to Geologic Terranes*, 1992; M. W. McElhinny, and P. L. McFadden, *Paleomagnetism: Continents and Oceans*, 2000; R. T. Merrill, M. W. McElhinny, and P. L. McFadden, *The Magnetic Field of the Earth: Paleomagnetism, the Core and the Deep Mantle*, 1996; N. D. Opdyke and J. E. T. Channell, *Magnetic Stratigraphy*, 1996; L. Tauxe, *Paleomagnetic Principles and Practice*, 1998.

## Paleontology

The study of animal history as recorded by fossil remains.

The fossil record includes a very diverse class of objects ranging from molds of microscopic bacteria in rocks more than  $3 \times 10^9$  years old to unaltered bones of fossil humans in ice-age gravel beds formed only a few thousand years ago. Quality of preservation ranges from the occasional occurrence of soft parts (skin and feathers, for example) to barely decipherable impressions made by shells in soft mud that later hardened to rock. See FOSSIL; MICROPALAEONTOLOGY.

The most common fossils are hard parts of various animal groups. Thus the fossil record is not an accurate account of the complete spectrum of ancient life but is biased in overrepresenting those forms with shells or skeletons. Fossilized worms are extremely

rare, but it is not valid to make the supposition that worms were any less common in the geologic past than they are now. See EDIACARAN BIOTA.

The data of paleontology consist not only of the parts of organisms but also of records of their activities: tracks, trails, and burrows. Dinosaur footprints, for example, are very common in the Connecticut Valley. Even chemical compounds formed only by organisms can, if extracted from ancient rocks, be considered as part of the fossil record. Artifacts made by people, however, are not termed fossils, for these constitute the data of the related science of archeology, the study of human civilizations. See ARCHEOLOGY; PALEOBIOCHEMISTRY.

Paleontology lies on the boundary between two disciplines, biology and geology. Various scientists have tried to place it more firmly in one camp than in the other, but such designations emphasize only a small area of the science's domain and do not acknowledge its entire range. If all types of paleontological research are regarded as equal in importance, then the geological and biological aspects must be granted equal weight. See BIOLOGY; GEOLOGY.

### Geological Aspects

A major task of any historical science, such as geology, is to arrange events in a time sequence and to describe them as fully as possible. Geology, the study of Earth history, did not become a modern science until the nineteenth century, when a worldwide time scale based on fossils was established; Earth history could not be deciphered until events that occurred in different places were related to one another by their position in a standard time sequence (Fig. 1). The data provided by fossils are used to accomplish these tasks in the following ways.

**Chronology.** Many geologists of the eighteenth century tried to use rock type as a criterion for judging the relative age of rocks, saying that heavy rocks such as granites were older than lighter sandstones and limestones. But granites formed throughout Earth history and Cenozoic granites cannot be distinguished from Cambrian granites by mineral content. A species of organisms, however, is unique; it lives for only a short time before becoming extinct or evolving into something else, and once gone, it never reappears. Thus fossils can be used to correlate rocks; for if species A is found in both a Nevada limestone and a Colorado sandstone, it may be concluded that the rocks are approximately the same age (Fig. 2). Fossils of certain species, or index fossils, are ideally suited for correlation because they were abundantly distributed over a large geographic area, yet persisted for a very short time before becoming extinct. See INDEX FOSSIL; STRATIGRAPHY.

Of course, fossils only tell that a rock is older or younger than another; they do not give absolute age. The decay of radioactive minerals may provide an age in years, but this method is expensive and time-consuming, and cannot always be applied since most rocks lack suitable radioactive minerals. Correlation by fossils remains the standard method for comparing ages of events in different areas.

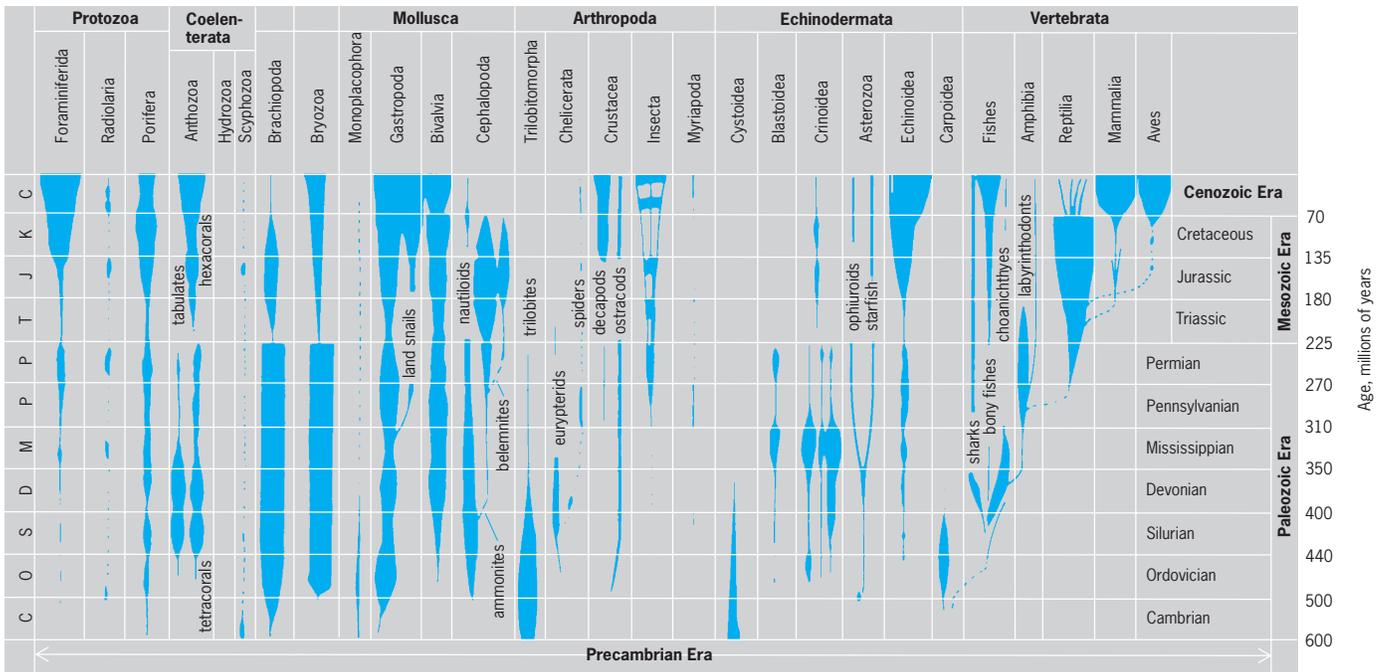


Fig. 1. Geologic distribution of animal life since the beginning of the Cambrian Period.

**Determining ancient Earth's appearance.** The physical appearance and climate of the Earth during a given period of the geologic past can be described from compilation and analysis of the data which is obtained through studies of the habitats of extant fauna, the geographic distribution of fossils, and the climatic preferences of ancient forms of life.

*Local conditions.* Life habits of fossils can be inferred by comparing them with modern organisms to which they are closely related. For example, all modern sea urchins live in marine environments. Since the Cretaceous limestones of southern England contain numerous fossil sea urchins, the conclusion is that this area was covered by ocean waters when these rocks were deposited. The study of fossils in relation to their physical and biological environment is called paleoecology. See PALEOECOLOGY.

*Paleogeography.* Fossils are indispensable guides for determining the positions of continents and seas in former times. The formation of the Isthmus of Panama can be dated, for example, by studying the distributions of marine and terrestrial fossils. Before North and South America were connected by this land bridge, Atlantic and Pacific marine faunas were very similar but the mammals of the two continents were completely different. However, in South American rocks deposited after the Isthmus was formed, there are fossils of North American mammals which had migrated over the newly formed land. Likewise, Atlantic and Pacific marine faunas, isolated from each other by the rise of the Isthmus, began to evolve in different directions; this increasing difference can be traced in the fossils of successively younger rocks. Fossils also help the scientist to decide whether continental drift has occurred. If South America and Africa were once united, their faunas should be similar during that time. As they drifted apart, there should be stronger and stronger faunal differences. See PALEO GEOGRAPHY.

*Paleoclimatology.* The natural occurrence of polar bears always implies a cold climate. In the same way, it is possible to learn about ancient temperatures if the climatic preferences of fossilized organisms are known. There is, for example, a one-celled animal which tends to coil its shell one way when the water temperature is cold and the other way when it is warmer.

During the last million years, the oceans have been successively cooled and warmed as giant glaciers of the ice ages grew and melted. The most recent warming of the oceans (melting of the last continental ice sheet) can be dated by finding the age of the sediments in which the change of the fossil's coiling direction is noted. See GEOLOGIC THERMOMETRY.

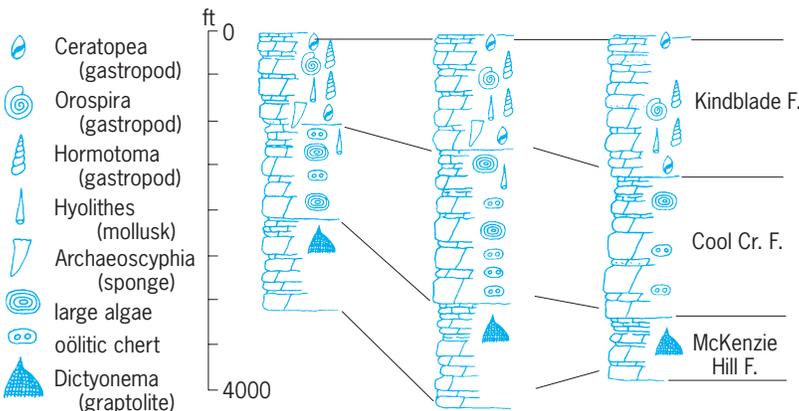


Fig. 2. Correlation of rocks by means of fossils. Although these three rock columns of Ordovician rocks in southern Oklahoma are all very similar in lithology, they can be divided and correlated on the basis of fossils. 1 ft = 0.3 m.

### Biological Aspects

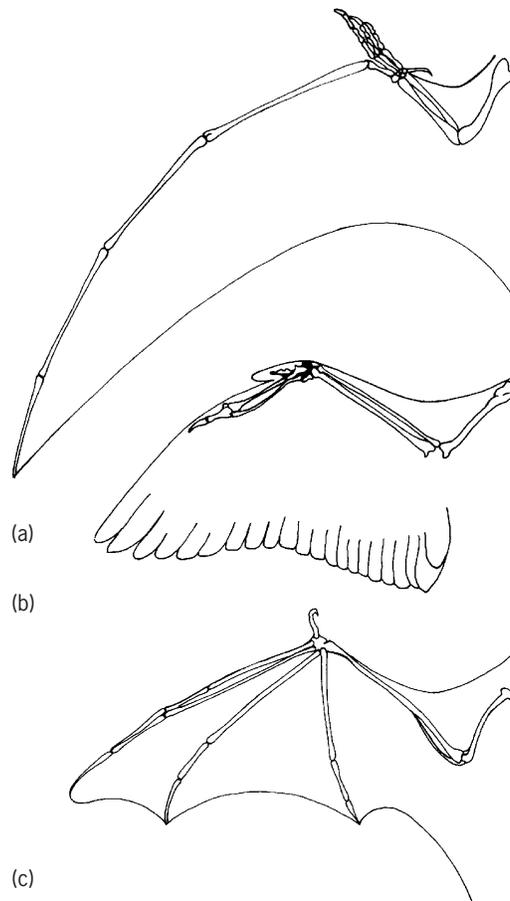
Paleontology as a branch of the life sciences may be called the biology of fossils. Any technique applied by biologists to the study of living organisms is theoretically relevant to fossils, although many cannot be applied in practice because of the limitations of paleontological evidence. The rarity of preserved soft parts, for example, severely limits the possibilities for a physiology of fossil animals. See FOSSIL.

The most fundamental fact of paleontology is that organisms have changed throughout earth history and that each geological period has had its characteristic forms of life. Before Charles Darwin won acceptance for his theory of evolution, various explanations were offered for the differences between ancient and modern life. Some claimed that fossils were not organic remains, but manifestations of a plastic force in rocks; others attributed fossils to the devastation of Noah's flood; still others envisaged a whole series of catastrophes that destroyed life and led to the creation of an entirely new set of organisms. When Darwin published his theory of evolution in 1859, paleontological evidence for evolution was very meager but it began to accumulate quickly. The "feathered reptile" *Archaeopteryx*, an almost perfect link between reptiles and birds, was discovered early in the 1860s and the first fossil humans were found a few years later. In the light of this evidence and the purview of a more scientific culture, evolution was accepted as the only reasonable explanation for change in the history of life. To a paleontologist, life's outstanding attribute is its evolution.

**Evolutionary process and life history.** An evolutionist has two major interests: first, to know how the process of evolution works; this is accomplished by studying the genetics and population structure of modern organisms; second, to reconstruct the events produced by this process, that is, to trace the history of life. This is the paleontologist's exclusive domain. Any modern animal group is merely a stage, frozen at one moment in time, of a dynamic, evolving lineage. Fossils give the only direct evidence of previous stages in these lineages. Horses and rhinoceroses, for example, are very different animals today, but the fossil history of both groups is traced to a single ancestral species that lived early in the Cenozoic Era. From such evidence, a tree of life can be constructed whereby the relationships among organisms can be understood.

**Life properties.** Evolution is responsible for life's outstanding properties: its diversity and its adaptation to environment.

**Diversity.** A paleontologist studies diversity by classifying organisms into basic units known as species. The number and geographic distribution of species in a more inclusive taxonomic group, such as vertebrates, can then be tabulated for different periods of geologic history. When this is done for all major groups, very definite patterns emerge. At the close of the Permian and Cretaceous periods, for example, extensive extinction of species occurred in many major groups at about the same time. Although the



**Fig. 3. Adaptation.** Three groups of vertebrates independently developed wings for flight. (a) Pterodactyl (extinct reptilian pterosaurs). (b) Bird. (c) Bat.

cause for these mass extinctions is unknown (theories range from large-scale shifts of sea level to bursts of cosmic radiation), they are major events in the history of life.

**Adaptation.** The term adaptation refers to the way organic form is fitted to function (Fig. 3). Given the raw materials of life, an engineer could scarcely design a better flying machine than a bird, yet the bird evolved naturally from reptilian ancestors which did not fly. Paleontologists study the development of new adaptations by a variety of techniques. For example, consider the increase in brain size between apelike ancestors and modern humans. To study this a paleontologist employs the following: (1) studying the rocks in which fossil bones are found to look for signs of any environmental change which might have made increased brain size advantageous; (2) determining the advantages of a larger brain if possible; for example, it may, among other things, have conferred upon ancient humans the ability to learn to control fire for warmth and cooking (charred bones of edible animals in caves inhabited by ancient humans could be used as evidence for such an assertion); and (3) applying mathematical techniques to obtain a precise measurement of the rate of evolution. See EXTINCTION (BIOLOGY); SPECIATION.

### Capsule History of Life

Although life was not present on Earth when it first formed about  $4.5 \times 10^9$  years ago, the raw materials necessary for its natural development were available. Electrical discharges and ultraviolet radiation induced the components of the original atmosphere to form complex organic compounds which later combined to form the prototype of a living cell. The oldest recognizable organisms so far discovered are bacterial-like cells from Australian cherts dated at  $3.5 \times 10^9$  years old. By the time the Gunflint cherts of Canada were deposited  $2 \times 10^9$  years ago, complex algae had evolved. Except for a few forms found in rocks just slightly older than  $6 \times 10^8$  years, invertebrate animals are not found until the base of the Cambrian Period,  $6 \times 10^8$  years ago. It is not known why so many groups of invertebrates made their first appearance as fossils in rocks at the same age; perhaps the base of the Cambrian marks a time during which animals first developed hard parts. The first vertebrate fossils, primitive fishes, occur in Lower Ordovician rocks. Vertebrates invaded the land with the evolution of amphibians in the Late Devonian, about  $3.5 \times 10^8$  years ago. Reptiles evolved soon afterward in Mississippian times; the first true mammals are Jurassic in age. The human is a newcomer, a product of the last few million years. Humans are merely a single species, a natural product of an unplanned evolution; yet from a geological perspective, they have altered the Earth far more than it had changed in any comparable time since life evolved. See ANIMAL EVOLUTION; PALEOBOTANY. Stephen J. Gould

Bibliography. E. H. Colbert, *Evolution of the Vertebrates*, 4th ed., 1991; R. Cowan, *History of Life*, 1991; A. Hallam (ed.), *Patterns of Evolution: As Illustrated by the Fossil Record*, 1977; L. B. Halstead, *Search for the Past*, 1983; M. McKinney, *Evolution of Life*, 1993; R. C. Moore, C. G. Lalicker, and A. G. Fischer, *Invertebrate Fossils*, 1952.

## Paleopathology

The study of ancient diseases and their origins. Paleopathology is especially important in the understanding of the origins, prevalence, and spread of infectious diseases, including how humans have contributed to the spread of disease and how they can overcome it. See EPIDEMIOLOGY; INFECTIOUS DISEASE.

Hypothesis testing of populations has contributed to the field of paleopathology, as has application of macroscopic (visual) examination, routine x-ray, computerized tomography (CAT) scans, magnetic resonance imaging (MRI), electron microscopy, and immunologic, chemical, and mass spectrophotometry techniques to skeletons, soft tissue, and even scat (animal droppings).

### Basic Principles and Methods

The scientific method in paleopathology is based upon comparison of archeologic or paleontologic

findings with individuals documented to have the disease. To this end the following basic tenets are observed: (1) Tissue must be adequately preserved to allow recognition of disease and distinguish possible pseudopathology or postdeath artifact. (2) The manifestations of a disease must be sufficiently stable across generations to allow comparison of ancient with modern disease. (3) Analysis of entire skeletons is more accurate than analysis of isolated bones. (4) Analysis of afflicted populations (paleoepidemiology) is more accurate than analysis of isolated skeletons.

The range of diagnostic methods used in paleopathology is extensive. Skeletal remains are visually examined to identify occurrence and nature of alterations, mapping their skeletal distribution. Internal structure can then be assessed, preferably by a nondestructive technique. Even fossils are not simply casts of external surfaces, but have a visualizable internal structure.

Microscopes can look at the surface of tissue or at cut sections. Traditional microscopy often requires tissue sections thin enough to allow light to pass through the specimen; another technique bounces the light off the specimen, so destructive thin sections are not required. Preservation of microscopic details has long been recognized in dinosaurs, as well as in the preserved muscles of fish as old as 300 million years. See X-RAY MICROSCOPE.

Molecules can be identified in tissue using antibodies or even DNA amplification and analysis, and chemicals can be identified by x-ray diffraction. Molecular preservation extends even to the three-dimensional structure of molecules, such as have been identified in 10,000-year-old collagen from birds and mammals and in blood from *Tyrannosaurus*. See ANTIBODY; X-RAY DIFFRACTION.

Rudolf Virchow and Charles Lester Leonard x-rayed mummies as long ago as 1897. As x-ray technique has been refined, computerized tomography and magnetic resonance imaging have also found application. Three-dimensional reformatting has allowed examination of internal surfaces of fossilized skeletons. Computerized tomography scanning with "dissection" of the resulting image and three-dimensional reconstruction can illustrate internal structures, and a plastic model can even be generated. See COMPUTERIZED TOMOGRAPHY; MAGNETIC RESONANCE.

Mummies provide an additional source of information. Rehydration of mummy tissue allows standard soft tissue histology, providing information often transcending that available through study of bones. Anthropologic study of artifacts such as daggers that sometimes accompany mummies and skeletons has also contributed to the understanding of ancient lifestyles and the diseases which impacted them. See HISTOLOGY.

### Diseases of the Past

Following is a review of some of the diseases recognized in living individuals that are believed to have existed as well in ancient life forms.

**Dental.** Dental caries are an ancient phenomenon, with abscesses present as far back as Devonian lungfish. However, such appear to have been rare occurrences until the introduction of agriculture in human populations. *See* DENTAL CARIES.

**Trauma.** The oldest example of a broken limb appears to be the radius of the sail-backed Permian lizard, *Dimetrodon*. Fractures are noted only rarely in flying reptiles (for example, *Pteranodon*), suggesting that they were rare or incompatible with survival.

Evidence of injuries among plant-eating dinosaurs is rare. Analysis of 30,000 elements from horned dinosaurs revealed a frequency in the range of 0.025–1.0%. Most were rib fractures, apparently related to side-butting territorial behavior. The most common injuries in duck-billed dinosaurs was to the tall projections off the back of vertebrae, the neural spines, perhaps related to mating activities. Large meat-eating dinosaurs (for example, *Tyrannosaurus*) seemed to have a much higher frequency of fracture.

Stress fractures related to activity, as have been reported in humans and greyhounds, have also been noted in the toes of horned dinosaurs.

The Neandertal Shanidar I had multiple fractures. The frequency of such fractures has been used to understand ancient “lifestyle.” *See* NEANDERTALS.

**Vascular phenomena.** Death, or necrosis, of bone occurs when its blood supply is compromised. Such has been described in Cretaceous marine lizards and turtles, although its occurrence in marine turtles seems to have diminished in frequency in the early Eocene and nearly disappeared subsequent to the Oligocene. *See* GEOLOGIC TIME SCALE.

Bone necrosis in the mosasaurs, a family of marine lizards, was universal in the deep and repetitive diving genera, yet absent in those whose members remained close to the water surface.

**Infections.** Suppurative (pus-producing) infections of the postcranial skeleton are recognized by irregular bone destruction and associated draining sinuses. Criteria have also been developed to recognize granulomatous infections such as fungus, tuberculosis, and leprosy, and even for treponemal disease (for example, syphilis). *See* INFECTION.

**Tuberculosis.** Tuberculosis accompanied by vertebral collapse and fusion, preserving the posterior joints, has been recognized on both sides of the Atlantic dating up to 6000 years before present (BP). Its diagnosis in mummies has been validated by recognition of the causative organism on microscopic examination, as well as by detection of *Mycobacterium tuberculosis* DNA. However, as bone involvement is relatively rare in tuberculosis and as vertebral collapse and fusion represent only a small portion of tubercular bone disease, only a small proportion of individuals afflicted can be recognized in this way. *See* TUBERCULOSIS.

**Leprosy.** Leprosy, or Hansen’s disease, is an infectious disease primarily affecting the skin and peripheral nerves. It is diagnosed on the basis of changes to the tips of fingers and toes, as well as fractures and other defects in the joints. Leprosy has been quite controversial in history, due to imprecise descrip-



**Fig. 1.** Two views of tibia (lower leg) affected with bejel (nonvenereal syphilis) from nineteenth-century Israel. (a) Anterior bowing of tibia with periosteal reaction on surface of both tibia and fibula. (b) Anterior and lateral view of tibia with bony enlargement and draining sinuses.

tions in biblical and ancient manuscripts. V. Moller-Christensen reported over 300 cases from Medieval hospital cemeteries in Denmark and Sweden. He suggested that 20% of individuals buried in those countries from 800 to 500 years BP had leprosy, but noted it only in isolated cases elsewhere (two dated at 1400 years BP from Egypt, one from France dated 1400 years BP, and 5 from the British Isles dated 1400 years BP). *See* LEPROSY.

**Treponematoses.** The treponematoses (caused by the spirochete, or treponeme, type of bacteria) include three clinically distinct disorders known to afflict bone: syphilis, yaws, and bejel (nonvenereal syphilis) [Fig. 1]. All three diseases cause skin rashes and bone damage. Syphilis is generally transmitted by sexual contact, while yaws and bejel are contracted by skin contact or sharing food utensils. While syphilis can affect the fetus, yaws and bejel are generally contracted in the first decade of life. Bone changes of syphilis are rare in children (less than 5%), in contrast to yaws and bejel, in which 10–20% of children have bone changes. Syphilis can produce heart, aorta, and nervous system damage, while these systems are generally spared in yaws and bejel.

Syphilis has been of such interest that three theories have been proposed for its occurrence: (1) Old World origin (pre-Columbian hypothesis); (2) New World origin (Columbian hypothesis); and (3) independent New and Old World origin.

The pre-Columbian hypothesis suggests that the epidemic of syphilis in late fifteenth century–early sixteenth century Europe simply represented new diagnostic ability to distinguish syphilis from leprosy. E. H. Hudson suggested an Old World origin of treponemal disease, in the form of yaws, that was subsequently transmitted to the New World.

Patterns reproducible for syphilis have been identified in Michigan and West Virginia (dated at 600 years BP), in Florida (700 years BP), in Ecuador

(800 years BP), in Wisconsin (1000 years BP), and in New Mexico (1500 years BP), but in no pre-Columbian European, African, or Asian sites examined. The worldwide presence of yaws, on the other hand, the geographic separation of syphilis and yaws in the New World, and the chronologic progression to recognizable syphilis (as a regional phenomenon) suggest that syphilis derived from yaws.

The history of treponemal disease in the New World can be definitively traced back almost 8000 years. It has been suggested that yaws migrated to the New World with the first humans to enter North America from Asia.

The first recognized case of treponemal disease appears to have been a case of yaws dated at 1.6 million years BP, suggesting that the onset of treponemal disease coincided with the origins of humans in Africa. See SYPHILIS; YAWS.

**Parasitic diseases.** Parasites, the subjects of the specialized field of paleoparasitology, are well represented in the archeologic record. A fish tapeworm dated at 6000 years BP has been identified in Peru and Chile; dog tapeworm in Medieval England and North Dakota (700 years BP); flukes in coprolytes (fossilized feces) from the southwestern United States (700 years BP) and Medieval Germany; roundworms in South American mummies (4000 years BP) and in the southwestern United States (1500 years BP); hookworms in Peru (1100 years BP), Brazil (5000 years BP), and Tennessee (2200 years BP); whipworm in Arizona (1000 years BP), from pre-Columbian Peru, Ecuador, and Brazil, and Japan (1000 years BP); schistosomiasis in the Sudan (2000 years BP); trichinella in mummies from the Aleutians (500 years BP), Viking-era Greenland, and Medieval Denmark and Switzerland; and Chagas' disease in Chile (4000 years BP). See PARASITOLOGY.

**Osteoarthritis.** Osteoarthritis is a disorder in which bone spurs occur at joints. The underlying bone becomes denser and the space between joints narrows. A type of osteoarthritis characterized as trauma-related degenerative disease has been noted in Neandertals, but osteoarthritis is more frequently a phenomenon of aging. Osteoarthritis is commonly observed in contemporary humans, but only within the past 50 years has it been routinely separated from other forms of arthritis. Thus, clinical records do not allow estimates of its frequency earlier than the most recent past.

As the presence of bone spurs defines the disease, historic occurrences of osteoarthritis are easily discernible by examination of skeletons from cemeteries or mass burials. As joint space-narrowing defines severity, occurrences are harder to measure. If the disease is so severe as to eliminate joint cartilage, bone rubs on bone. This polishes the bone, a condition that can be easily recognized in skeletons. As bone polishing occurs also in other diseases, however, only that associated with spurs indicates the most severe form of osteoarthritis.

Osteoarthritis actually appears to be a relatively new disease. Among dinosaurs it is extremely rare,

documented only in *Iguanodon*, affecting the ankles of 2 of 39 individuals examined. This contrasts with the frequent occurrence of osteoarthritis in the flying reptiles, pterosaurs. Osteoarthritis in Pleistocene mammals is quite rare, similar to the frequency observed in contemporary wild caught animals.

The paucity of osteoarthritis in large dinosaurs provides evidence that weight is not the major consideration in its development. Dinosaurs had highly constrained joints. They were not capable of the rotation inherent in the human knee. Thus, they were protected. This finding affirms that it is the stability of the joint, not the weight placed upon it, which determines if osteoarthritis will occur. See ARTHRITIS.

**Rheumatoid arthritis.** Rheumatoid is an inflammatory arthritis that afflicts most of the joints of the body symmetrically but spares the vertebrae (except neck) and sacroiliac (lower back) joints. It originated in the New World. No valid cases earlier than 1785 have been found in the Old World. Rheumatoid arthritis has been found in numerous pre-Columbian New World populations, dating back 6500 years, and the character of this disease has remained unchanged over the time span. Originally found in the Green River region of west-central Kentucky and the west branch of the Tennessee River in northwest Alabama and Tennessee, it remained there for 5000 years, spread into Ohio approximately 1000 years ago, and became more disseminated 200–300 years ago. No valid animal example of rheumatoid arthritis has been identified to date.

The initial geographic localization of the disease, with its subsequent spread in the New World and finally its penetration into the Old World, is highly suggestive of a vector-transmitted disease. Identification of such an agent may allow prevention and possibly specific treatment of the disease in the future.

**Spondyloarthropathy.** Another form of erosive arthritis commonly produces a more limited peripheral arthritis, a different variety of joint erosion, and joint fusion, and at times affects the vertebrae and sacroiliac joints (Fig. 2). There are a number of varieties of disease in this category. Ankylosing spondylitis, psoriatic arthritis, and reactive arthritis are the

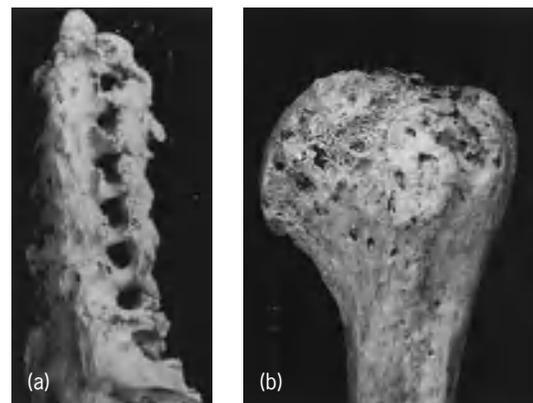


Fig. 2. Spondyloarthropathy. (a) Cervical spine (neck vertebra) fusion and (b) shoulder (upper arm) erosions affecting both margins and joint surface in individuals from early-nineteenth-century Rochester Poorhouse.

most prevalent. While frequency varies in world populations, the nature and skeletal distribution are quite similar through time and even across species lines. Its frequency in afflicted human populations runs 0.5–8%, seemingly dependent upon sanitary conditions, although some Amerindian groups seem to have a genetic predisposition. The oldest human case reported in the New World so far was dated at 5000–8000 years BP. *See* JOINT DISORDERS.

Spondyloarthropathy is believed to occur in 25% of contemporary bears, apes, and baboons, whose antecedents were largely spared. While the disease was undetected in Miocene apes, the frequency of spondyloarthropathy increased to 20–28% of contemporary gorillas and chimpanzees. While it was undetected in Miocene baboons, its frequency increased from 4% in the early part of the twentieth century to 10% at midcentury and 30% in the 1980s. Spondyloarthropathy truly has become epidemic in both free-ranging and colony-raised animals, and now appears pandemic in mammalian orders including carnivores such as cats and dogs, hoofed animals, and even marsupials such as kangaroos.

**Metabolic diseases.** Metabolic diseases involve defects of the metabolic pathway that result in an imbalance in nucleic acids, proteins, lipids, or carbohydrates.

*Gout.* Gout, first reported by Hippocrates, is a disorder in which urate crystals accumulate as space-occupying masses in the blood. Gout has only rarely been reported in ancient humans, but it has been reported in monitor lizards, turtles, alligators and crocodiles, birds, and most recently in *Tyrannosaurus rex*. *See* GOUT.

*Osteoporosis.* Osteoporosis is a disease in which bones become brittle and likely to break, usually due to a loss of bone density. Unfortunately, modern techniques for measurement of bone density are of limited value for ancient skeletons. Variable effects of diagenesis (the chemical and physical changes undergone by buried bones with increasing time, pressure, and temperature) generally preclude direct application of density-measuring techniques. One indirect technique uses x-rays to measure the cortical thickness of hand bones. This seems an effective technique, although it has yet to be applied in routine population studies. *See* OSTEOPOROSIS.

*Osteomalacia and rickets.* These diseases, which cause bowing of bones with widened growth plates, are uncommon in the paleontologic record. A case of rickets from the Upper Paleolithic (1200 BCE) of Italy has been reported, as have cases in children from Bahrain 2000 years BP and from sixth-century Negev. An adult case in Austria 1600 years BP has also been reported. *See* SKELETAL SYSTEM DISORDERS.

**Anemia.** Examination of skeletons may allow recognition of certain forms of congenital anemias (lack of hemoglobin or red blood cells), as well as suggesting those due to the blood loss and other effects of parasitism. A possible case of thalassemia (related to abnormal hemoglobin) in thirteenth-century Greece was reported, as well as cases from late

Bronze age Greece and a Khok Phanom Di site 4000 years BP in central Thailand. A DNA abnormality associated with anemia was demonstrated in a child from the early Ottoman period, and possible sickled cells were reported in skeletons dated 150–330 years BP. *See* ANEMIA.

There has been a misconception that localized overgrowth of skull bone marrow, referred to as porotic hyperostosis, can occur as a complication of primary iron-deficiency anemia. Actually, blood loss can produce both iron-deficiency anemia and porotic hyperostosis. The challenge of determining which came first is always problematic when trying to derive relationships from observed associations.

**Congenital disorders.** Disorders of growth, duplication, or hyper- or hypodevelopment of skeletal elements have been noted as isolated phenomena in both contemporary and ancient animals. Coalitions of foot bones, for example, were noted in the Cretaceous marine lizard *Platecarpus*. Another intriguing isolated phenomenon is the stone baby. A small percentage of fetuses develop outside the uterus. Most die quickly and the body reabsorbs them. Sometimes, the body does not reabsorb the fetus. In those cases, the remains become somewhat mummified. Skeletonized and bound by thickened, calcified membranes, such a fetus—dated at 3100 years BP—was clearly identified in Kerr County, Texas, in 1991.

Spina bifida, failure of the posterior vertebral elements to fully fuse and surround the nerves passing through the sacral canal, has attracted much anthropologic interest. Neural tube defects including spina bifida have been reported from sites in Florida (8000 years BP), Mali (7000 years BP), and northeastern Morocco (10,500–12,070 years BP). *See* CONGENITAL ANOMALIES.

**Tumors.** One of the simplest forms of neoplasia (excessive tissue production) is an overgrowth of bone called an exostosis. One of the earliest hominid examples is occurrence of an exostosis in the femur of *Homo erectus* from Java. Exostosis has also been found on the scapula of the dinosaur *Allosaurus* and the mandible of *Triceratops*. Multiple hereditary exostoses was recognized in a human 500 years BP from southern Ontario.

A more complicated form of neoplasia actually develops a cartilaginous cap, similar to that seen in joints, but without an opposing bone. Such osteochondromas have been reported from Egypt 3500–2500 years BP and Hawaii 500 years BP. It has also been identified in 64% of Oligocene canids (doglike creatures).

Another form of exostosis affects the external auditory canal. An association with cold water is suggested, possibly related to diving activities. Such cases are of great antiquity, reported in Pleistocene human specimens Shanidar I, la Chapelle aux Saints, and skull X from Zhoukoudien, China, as well as in humans of the Neolithic of southwest Asia, Africa, and North and South America. *See* FOSSIL HUMANS.

Isolated reports include benign bone tumors in humans from England 2000 years BP and in ancient marine lizards, and multiple vertebral hemangioma

(vascular tumors) from humans in Egypt 1000 years BP. Hemangiomas have also been recognized in dinosaurs, including a specimen collected from Emery County, Utah. The fragmentary nature of the 10-pound specimen precluded identification of species or even definite bone element identification, but the general morphology suggested a classic hemangioma.

E. Strouhal reviewed malignant tumors (predominantly osteosarcomas) in the Old World, reporting 10 isolated cases from Egypt, Nubia, and Europe, the oldest being Neolithic. He identified osteolytic metastases in 30 individuals from Egypt, Nubia, Europe, and Asia, dating back to the Early Dynastic of Abu Simbel, Nubia. While the blood malignancy multiple myeloma has been suggested, none of these cases fulfill current criteria for this disease. Multiple myeloma still awaits documentation in the ancient record. See TUMOR.

Bruce M. Rothschild

**Bibliography.** J. E. Buikstra and D. H. Ubelaker, Standards for Data Collection from Human Skeletal Remains, *Arkansas Archeol. Surv. Res. Ser.*, no. 44, 1994; M. G. Fiori and M. G. Nunzi, The earliest documented application of X-rays to examination of mummified remains and archaeological materials, *J. Roy. Soc. Med.*, 88:67-69, 1995; C. Greenblatt, *Digging for Pathogens: Ancient Emerging Diseases—Their Evolutionary, Anthropological and Archaeological Context*, Balaban Publishers, Rehovot, Israel, 1998; B. M. Rothschild et al., First European exposure to syphilis: The Dominican Republic at time of Columbian contact, *Clin. Infect. Dis.*, 31:936-941, 2000; B. M. Rothschild et al., Geographic distribution of rheumatoid arthritis in ancient North America: Implications for pathogenesis, *Semin. Arthritis Rheum.*, 22:181-187 1992; B. M. Rothschild and L. Martin, *Paleopathology: Disease in the Fossil Record*, CRC Press, London, 1993; B. M. Rothschild and C. Rothschild, Treponemal disease revisited: Skeletal discriminators for yaws, bejel, and venereal syphilis, *Clin. Infect. Dis.*, 20:1402-1408, 1995; B. M. Rothschild and R. J. Woods, Spondyloarthropathy in the Old World, *Semin. Arthritis Rheum.*, 21:306-316, 1992; E. Strouhal, Malignant tumors in the Old World, *Paleopath. Newsl.*, 85:1-6, 1994; P. L. Thillaud and P. Charon, *Lesions Osteoarcheologiques: Recueil et Identification*, Kronos, Sceaux, France, 1994; R. A. Tyson, *Human Paleopathology and Related Subjects: An International Bibliography*, San Diego Museum of Man, 1997.

## Paleoseismology

The study of geological evidence for past earthquakes. This is a scientific discipline that has contributed greatly to modern understanding of the nature of earthquakes. The patterns of earthquakes, in both space and time, evolve over centuries and millennia and cannot be discovered by modern instruments. Knowledge of these patterns is important for understanding the physics of earthquakes and for forecasting future destructive earthquakes.

In certain natural environments, the features related to ancient earthquakes are preserved in the landforms and superficial layers of the Earth's surface. Geologists use this paleoseismological evidence to extend the short historical and instrumental record of earthquakes into ancient centuries and millennia. Paleoseismological studies have been extended into the ancient past; they have clarified the earthquake record of many parts of the world, including the midcontinent and east coast of the United States, northern Africa, southern Europe, China, Japan, Indonesia, and New Zealand.

**Sedimentary record of ancient earthquakes.** In certain geological settings, the disruptions produced during large earthquakes become preserved in geological strata. These sediments, therefore, provide a record of the earthquake. In this respect, the sediments are like seismograms. Geologists study the disruptions of these sediments to gain a better understanding of the earthquake. The San Andreas Fault in southern California provides one of the clearest examples.

Most Californians live within 100 km (60 mi) of the San Andreas Fault. Along the fault, coastal California, including Los Angeles, Monterey, and San Diego, has slipped northwestward over 300 km (180 mi), relative to the rest of North America. Despite this large horizontal translation over millions of years, no more than a few meters has occurred along the fault in the two centuries of historical record. These small movements occurred primarily during two great earthquakes—the 1906 San Francisco earthquake in northern California and the 1857 Fort Tejon earthquake in the southern part of the state. See FAULT AND FAULT STRUCTURES.

With no more than the short historical and instrumental records, scientists had no clear understanding of the frequency of large, destructive earthquakes produced by the fault. Fortunately, paleoseismological studies in southern California have extended the known record of earthquakes back nearly two millennia. Northeast of Los Angeles, a record of 10 large earthquakes is preserved in peaty and sandy stream and marsh deposits. Faulted sediments record evidence of the great earthquake that occurred around 1480. Overlying layers were deposited after this strand of the fault broke, and thus are undisturbed.

**Dating ancient earthquakes.** In addition to determining the amount and type of motion along a fault in an ancient earthquake, it is necessary to determine the date. Radiocarbon dating is the most common of many methods. Precise dating of earthquakes along the San Andreas Fault has shown that earthquakes along one part of the fault have struck southern California about every 130 years on average. In the past, 10 large earthquakes occurred along the San Andreas Fault northeast of Los Angeles. The 1857 earthquake is known from historical records. The 1812 earthquake is a historical event, but was not known to have been generated by the San Andreas Fault until the completion of paleoseismologic studies of disturbed trees along the fault. The other quakes have been dated by the radiocarbon method. It is

curious that the intervals between the earthquakes have ranged between 44 and about 330 years. See RADIOCARBON DATING.

The reason for the large variation in the dates of the earthquakes is the subject of much debate. Such large natural variations make forecasts of future earthquakes uncertain. Knowledge of the average period between large earthquakes, however, has given Californians a more informed basis for earthquake preparedness. Based upon paleoseismic data from several localities along the fault, the southern 200 km (120 mi) of the fault, near Palm Springs and San Bernardino, are considered the most likely to produce the next great earthquake in California. See SEISMIC RISK.

**Geomorphic record.** The world's greatest earthquakes occur at subduction zones, where great slabs of oceanic lithosphere are sinking diagonally into the Earth's interior. One of the largest earthquakes of the twentieth century, of magnitude 9.2, was generated on the subduction zone of southern Alaska in 1964, when a huge slab of the Pacific Plate lurched suddenly downward about 30 m (100 ft) beneath Alaska. Southern coastal communities were devastated by the long and heavy shaking, by the seismic sea waves (tsunamis) that swept over them, and by permanent submergence of the land. No earthquake of this size had struck this portion of Alaska in the entire two centuries of recorded history. See PLATE TECTONICS; SUBDUCTION ZONES; TSUNAMI.

Studies of the effects of this devastating earthquake revealed paleoseismological evidence of previous great earthquakes in this region. There was clear evidence of earlier uplifts on a small island, far out to sea, that had risen 3.4 m (11 ft) in 1964. The flat-topped island exhibits a set of six concentric ancient shorelines, each a few meters higher than its seaward neighbor; these appear as small dark cliffs. Radiocarbon dating of the ancient shorelines revealed that the top of the island, now 50 m (165 ft) above sea level, first rose above the sea about 5000 years ago. In the past 5000 years, uplift occurred episodically, during six ancient earthquakes, 500 to 1500 years apart. The shoreline of early 1964, now also dry and 3.4 m (11 ft) high, forms a ring just inland from the modern shoreline.

Other than a few moderate earthquakes, the Cascadian subduction zone of the Pacific Northwest of the United States has been quiet during the period of historical record. Older building codes in Oregon reflect a historical lack of concern about earthquakes there. Along the coasts of Oregon and Washington, however, there are clear records of sudden and recurring submergence, a paleoseismic record that has been related to great earthquakes on the Cascadian subduction zone. In many coastal estuaries, geologists have found layers of peat intercalated with silt beds. The peats are composed of plants known to grow at certain elevations in the estuaries. The silts, carried to the sea by nearby rivers, are known to settle upon the floor of the estuary in deeper water. Yet, the paleoseismologists have found that the peats now reside well below the level at which they grew

and were buried by the silts. Radiocarbon dating of the peats indicates that sudden submergence of the estuaries has occurred several times in the past few thousand years, most recently about 300 and about 1200 years ago. In some localities, geologists have found sand atop the peats, carried into the estuaries from the sea by tsunamis. Since the discovery of paleoseismic evidence for great earthquakes in the Pacific Northwest, much more civic attention has been focused upon seismic building codes and earthquake preparedness there.

The geological preservation of ancient earthquakes also has enabled scientists to compare modern earthquakes with those of the ancient past. In 1983, for example, a sparsely populated region of Idaho was struck by a magnitude-7.3 earthquake. Investigations after the earthquake revealed a fresh, 30-km-long (18-mi) fault scarp running along the western base of the lofty Lost River Range. During the earthquake, Borah Peak, which crowns the range, had jumped 2 m (7 ft) skyward. Inspection of the fresh escarpment produced during the earthquake revealed that it is surmounted by a more subdued, vegetated escarpment of nearly identical length and height. Excavations across this ancient fault scarp showed that it had formed during an event very similar to the earthquake of 1983 but about 5000 years earlier. This is one of several examples of what paleoseismologists call a characteristic earthquake. It appears, from such examples, that some earthquakes are nearly identical repetitions of their predecessors. If nature were always as regimented as this, the prediction of earthquakes would be far simpler. Unfortunately, many examples of irregular behavior of faults also exist. See EARTHQUAKE; SEISMOLOGY.

Kerry Sieh

**Bibliography.** U.S. Geological Survey, *Earthquakes and Volcanos*, bimonthly; R. E. Wallace (ed.), *The San Andreas Fault System, California*, USGS Prof. Pap. 1515, 1990; R. Wesson and R. E. Wallace, Predicting the next great earthquake in California, *Sci. Amer.*, 252:35-43, 1985.

## Paleosol

A soil of the past, that is, a fossil soil. Paleosols are most easily recognized when they are buried by sediments. They also include surface profiles that are thought to have formed under very different conditions from those now prevailing, such as the deeply weathered tropical soils of Tertiary geological age that are widely exposed in desert regions of Africa and Australia. Such profiles are generally known as relict paleosols. Those that can be shown to have been buried and then uncovered by erosion are known as exhumed paleosols. The main problem in defining the term paleosol comes not so much from complications such as these arising from its fossil nature, but from defining what is meant by soil, a term that has very different meanings for agronomists, engineers, geologists, and soil scientists. Considering research on soils of Antarctica and Mars and

on paleosols in a variety of rocks ranging back to  $3.5 \times 10^9$  years old, soil can be considered distinct from sediment in that it forms in place, but soil need not necessarily include traces of life. At its most general level, soil is material forming the surface of a planet or similar body and altered in place from its parent material by physical, chemical, or biological processes.

Soils buried by till, colluvium, and flood deposits are commonly encountered during roadwork or excavations for foundations. The ways in which such soils can overlap or be eroded are complex (Fig. 1). Weakly developed soils may retain features of their parent material, such as relict sedimentary structures. They also may contain fragments or horizons of preexisting soils (pedorelicts). Especially distinctive soil materials, such as laterites, that have been eroded and redeposited are known as pedoliths. Despite these complications, many paleosols are unmis-

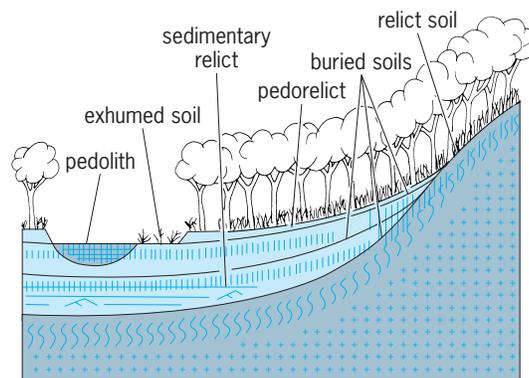


Fig. 1. Diagrammatic landscape profile illustrating key concepts and terms in the study of paleosols.



Fig. 2. Triassic paleosol from the sea cliffs at Long Reef, near Sydney, in southeastern Australia, showing conspicuous drab-haloed root traces (near 1-ft or 0.3-m scale) and the sharply truncated top and gradational lower horizons characteristic of paleosols.



Fig. 3. Middle Miocene paleosols and a similar modern soil in the quarry at Forth Ternan National Monument, southwestern Kenya, showing finely structured surface horizons (crumb peds) that have developed during soil formation at the expense of relict bedding (at level of hammer).

takable with their prominent root traces, gradational soil horizons, and soil structures, such as peds and cutans (Figs. 2 and 3). Buried land surfaces of numerous laterally contiguous paleosols are known as geosols and have been widely used as stratigraphic marker horizons for the mapping and interpretation of Quaternary deposits. See LATERITE; QUATERNARY; STRATIGRAPHY.

Paleosols also are found at major geological unconformities in the rock record, including the unconformity at the land surface in continental regions that have been exposed to weathering for many millions of years. Some of these relict deep-weathering profiles dating back to the Cretaceous in Australia, Africa, and Brazil are up to 100 m (330 ft) thick. Their bauxitic and lateritic horizons are valued as a source of clay for china, aggregate for construction material, and ores of aluminum and iron. A long fossil record of deeply weathered paleosols at major geological unconformities as old as  $3\text{--}5 \times 10^9$  years has been subject to exploration as evidence for prehistoric changes in the atmosphere and life on land. See BAUXITE; CRETACEOUS.

Paleosols are especially abundant in volcanic, alluvial, and eolian sedimentary sequences. Along with the fossils, sedimentary structures, and volcanic rocks found in such deposits, paleosols provide an additional line of evidence for ancient environments during times between eruptions and depositional events. See PALEOCLIMATOLOGY; SEDIMENTOLOGY.

Paleosols can be characterized by their root traces, horizons, and soil structures visible in the field, supported by laboratory studies of their chemical and

mineralogical composition and their microscopic fabrics as seen in petrographic thin sections. Many ancient paleosols have been altered during burial by physical compaction, by local chemical reduction producing green-gray areas around buried organic matter, by reddening from dehydration of ferric oxyhydrate minerals, and by illitization of clays from the dissolution during burial of microcline and other potash-bearing minerals. Such alteration after burial can limit the reconstruction of a paleosol, its identification in a soil classification, and interpretation of soil-forming factors during its formation. Nevertheless, many features of paleosols, such as horizons rich in clay skins or in calcareous nodules, are surprisingly robust in the face of alteration during deep burial and metamorphism. *See* PETROFABRIC ANALYSIS; SOIL.

Gregory J. Retallack

Bibliography. P. W. Birkeland, *Soils and Geomorphology*, 1999; G. J. Retallack, *A Color Guide to Paleosols*, 1997.

## Paleozoic

A major division of time in geologic history, extending from about 540 to 250 million years ago (Ma). It is the earliest era in which significant numbers of shelly fossils are found, and Paleozoic strata were among the first to be studied in detail for their biostratigraphic significance. Western Europe, especially the British Isles, was the cradle of historical geology. Early work with rock strata and their fossils was strictly practical; the relative ages of rock units were essential for correlating scattered outcrops to search for natural resources—particularly coal—in the early part of the nineteenth century.

During its first four decades, natural groupings of strata were studied and named for easy reference. Thus the several subdivisions of the Paleozoic, ultimately the six standard systems, were established. The original basis for establishing sequence was superposition. The operational stratigraphic hypothesis is that, in most instances, the strata at the bottom of a sequence are the oldest and the overlying beds are progressively younger. Thus, the basal system of the Paleozoic, in which primitive shelly fossils are found, is the Cambrian. As younger and younger layers were studied, their fossils collected, and the biological affinities suggested, the concept of evolution from simpler to more complex life forms took shape in the minds of the paleontologists and geologists who were studying the rocks. This process did not take place in an orderly way, from oldest to youngest strata, but rather as a consequence of fulfilling a need of the moment, whether to complete a geologic map or to solve a problem of stratigraphic correlation. Consequently, the first Paleozoic system to be named and studied in some detail was the Carboniferous—the great “coal-bearing” sequence—given that name by W. D. Conybeare and W. Phillips in 1822. These strata were to provide the world’s major energy resources during the next century and

CENOZOIC	QUATERNARY	
	TERTIARY	
MESOZOIC	CRETACEOUS	
	JURASSIC	
	TRIASSIC	
PALEOZOIC	PERMIAN	
	CARBONIFEROUS	Pennsylvanian
		Mississippian
	DEVONIAN	
	SILURIAN	
	ORDOVICIAN	
	CAMBRIAN	
PRECAMBRIAN		

a half. Most of the Northern Hemisphere’s coal fields, and much of its oil and gas as well, were produced from Carboniferous rocks. *See* SUPERPOSITION PRINCIPLE.

In the 1830s and 1840s two British geologists, R. Murchison and A. Sedgwick, studied and named the natural groupings of rock strata in the British Isles. Sedgwick named the Cambrian System in 1835, for a sequence of strata that overlies the Primordial (Precambrian) rocks in northwest Wales. Four years later, Murchison gave the name Silurian to the early Paleozoic rocks found in the Welsh borderland. However, there was an almost complete overlap of the Cambrian by Murchison’s Silurian. It was not until 1879, when C. Lapworth named the Ordovician System for rocks intermediate between the Cambrian and the “upper” Silurian, that the three early Paleozoic systems were sorted out in the correct order. In the meantime, Murchison and Sedgwick managed to agree on the rocks above the Silurian and, in 1839, they named the Devonian System for rocks exposed in Devonshire, England. The final Paleozoic system, the Permian, was named by Murchison in 1841, after an expedition to Russia, where he recognized the youngest Paleozoic fossil assemblages in the carbonate rocks exposed in the province of Perm. *See* PRECAMBRIAN.

**Subdivisions.** The Paleozoic Era is divided into six systems; from oldest to youngest they are Cambrian, Ordovician, Silurian, Devonian, Carboniferous, and Permian. The Carboniferous is subdivided into two subsystems, the Mississippian and the Pennsylvanian which, in North America, are considered systems by many geologists. The Silurian and Devonian systems are closer to international standardization than others; all the series and stage names and lower

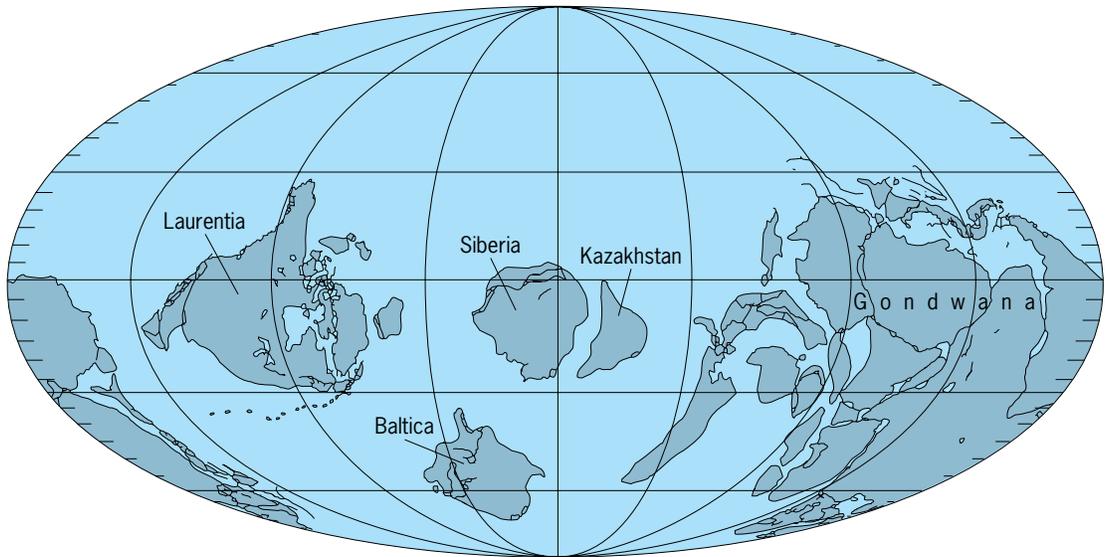


Fig. 1. Paleogeography of the Cambro-Ordovician (Tremadoc) showing most of the northern plates spread east-west in the equatorial regions. (After W. S. McKerrow and C. R. Scotese, eds., *Paleozoic Palaeogeography and Biogeography*, Geol. Soc. Mem. 12, Geological Society, London, 1990)

boundaries have been agreed upon, and most have been accepted. Despite continuing revisions, the major subdivisions of the geologic time scale have been relatively stable for nearly a century. See CAMBRIAN; CARBONIFEROUS; DEVONIAN; ORDOVICIAN; PERMIAN; SILURIAN.

**Paleotectonics.** The Paleozoic oceans, just as those today, surrounded a series of landmasses that formed

the cores of ancient plates, always in motion as are their modern counterparts. Sediments were supplied to the seas through a network of river drainage systems and distributed in the oceans, by currents and gravity, very like today. Clastic sediments were supplied by the mountainous regions that were uplifted and eroded in cyclic patterns as the major plates collided and parted; and subduction at the

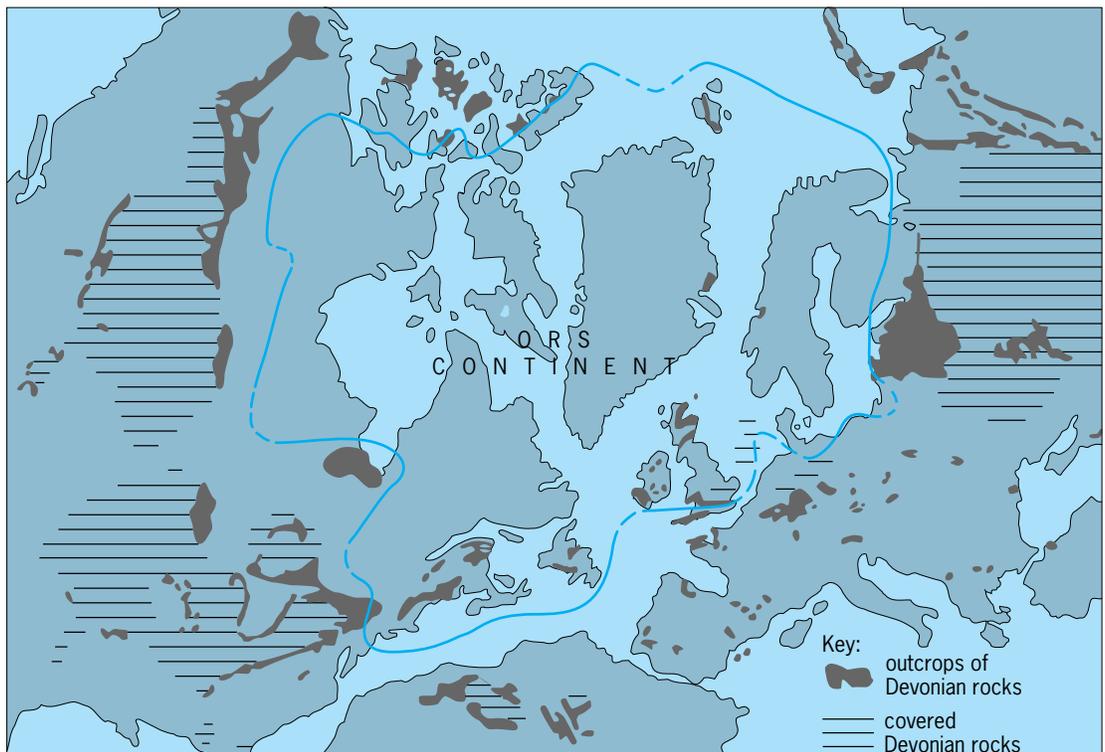


Fig. 2. Paleogeography of Old Red Sandstone continent (outlined in color) in Late Devonian. (After R. Goldring and F. Langenstrassen, *Open shelf and near-shore clastic facies in the Devonian*, in M. R. House, C. T. Scrutton, and M. G. Bassett, eds., *The Devonian System*, Spec. Pap. Palaeont. 23, Palaeontological Association, London, 1979)

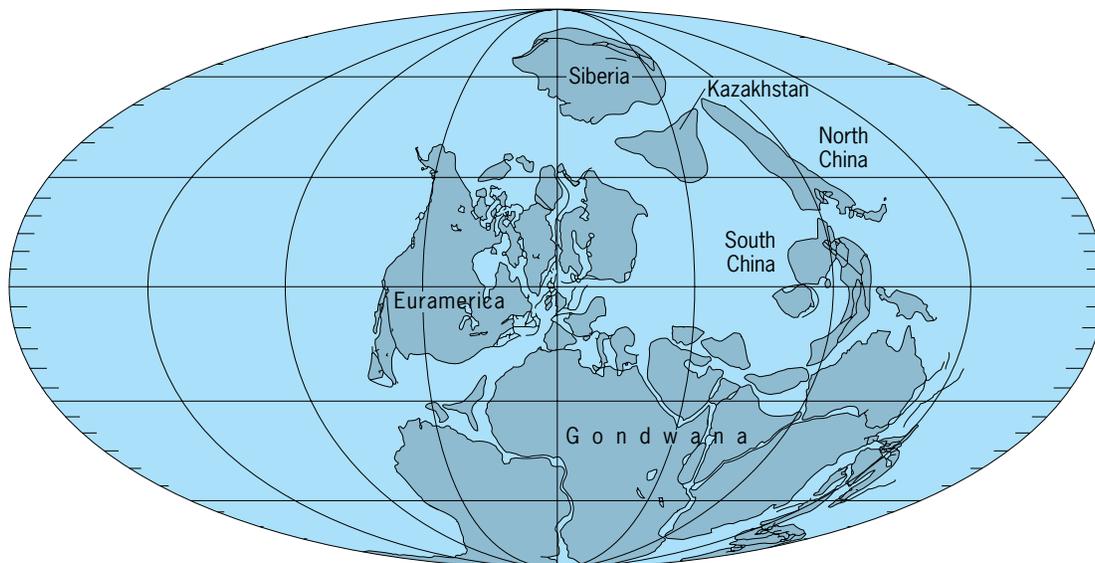


Fig. 3. Paleogeography of the Early Carboniferous (Viséan), showing the Euro-American megaplate centered in equatorial region. (After W. S. McKerrow and C. R. Scotese, eds., *Paleozoic Palaeogeography and Biogeography*, *Geol. Soc. Mem. 12*, Geological Society, London, 1990)

leading edges of some plates produced volcanic highlands. The plate tectonic theory provides a template for sorting out the periods of mountain building during the Paleozoic. Like the discovery of the stratigraphic systems, periods of orogeny, with their concurrent volcanic and intrusive igneous activities, were revealed by field studies. Tectonic effects (folding and faulting) were analyzed by geologic mapping, as were crosscutting igneous relations and unconformities in the sedimentary sequence. Regional orogenic terranes were named and the general time sequence assigned; these were sharpened as the use of isotopic age analyses of the igneous components became possible in the twentieth century.

Because Alpine and Appalachian mountain chains were among the first studied in detail, orogenies were first named there. In eastern North America, mountain-building effects during the early Paleozoic were ascribed to the Taconic orogeny (Middle and Late Ordovician); middle Paleozoic events were assigned to the Acadian orogeny (Middle and Late Devonian); and late Paleozoic movements were called Appalachian (more accurately Alleghenian) for Permian and, perhaps, Triassic events. Similar, but not precisely correlative, orogenic episodes in western Europe are ascribed to the early Paleozoic Caledonian and the late Paleozoic Variscan (or Hercynian) orogenies. This regionalization, overlapping of

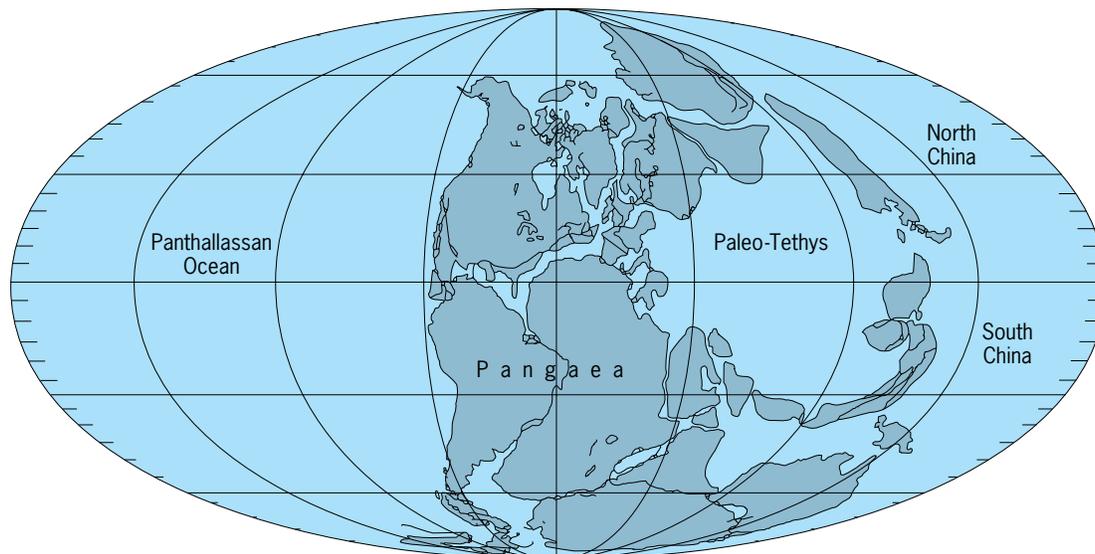


Fig. 4. Paleogeography at the end of the Permian showing the consolidation into Pangaea preparatory to the onset of the Mesozoic tectonic cycle. (After W. S. McKerrow and C. R. Scotese, eds., *Paleozoic Palaeogeography and Biogeography*, *Geol. Soc. Mem. 12*, Geological Society, London, 1990)

timing of events and lack of correlation of intrusive phases, tectonics, and sedimentation cycles emphasize the universality of the ever-moving plates as the global mechanism responsible for all tectonic events. *See* DATING METHODS; ISOTOPE; OROGENY; PLATE TECTONICS; UNCONFORMITY.

**Lithofacies.** The major changes in lithofacies during the Paleozoic were also effected by biotic evolution through the era. Limestone facies became more abundant and more diversified in the shallow warm seas as calcium-fixing organisms became more diverse and more widespread. Sediment input from the land was modified as plants moved from the

seas to the low coastal plains and, eventually, to the higher ground during the Devonian. Primitive vertebrates evolved during the Cambro-Ordovician, but true fishes and sharks did not flourish until the Devonian. Amphibians invaded the land during the Late Devonian and early Carboniferous at about the same time that major forests began to populate the terrestrial realm. These changes produced an entirely new suite of nonmarine facies related to coal formation, and the Carboniferous was a time of formation of major coal basins on all continental plates.

Climate continually influenced depositional patterns and lithofacies both on land and in the seas. In the major carbonate basins and platforms, particularly from the Late Ordovician onward, cyclic climatic changes resulted in changes from calcitic to magnesian carbonates and, ultimately, to various saline deposits as the basins dried up. There were great salt deposits in several systems, but spectacular thicknesses developed in many basins during the Silurian and Permian. Major cycles of cold and warm climates were overlaid on depositional and evolutionary patterns, producing periods of continental glaciation when large amounts of the Earth's water were tied up in ice during the Late Ordovician, the Late Devonian, and the late Permian. During the earliest and latest of these periods, icesheets were concentrated in the Southern Hemisphere on a single large Paleozoic continental mass—Gondwana. *See* DEPOSITIONAL SYSTEMS AND ENVIRONMENTS; FACIES (GEOLOGY); PALEOCLIMATOLOGY.

**Paleogeography.** Paleogeographic changes naturally followed the shifting of plates on a megacyclic scale during the Paleozoic. In general, the Paleozoic featured a single southern landmass (Gondwana) for most of the era. This megaplate moved relatively sedately northward during this entire time interval (540–250 Ma) and always contained the magnetic and geographic south poles. Consequently, many of the facies and biologic provinces in the Gondwanan region were influenced by the cooler marine realms and continental and mountain glaciers in nearly every Paleozoic period. Most of the tectonic action that produced major periods of collision, mountain building, carbonate platform building, back-arc fringing troughs with their distinctive faunas and lithofacies, and formation of coal basins and evaporites took place in the Northern Hemisphere. These pulsations produced combinations of Laurentian (North American), Euro-Baltic, Uralian, Siberian, and Chinese plates at various times during the Paleozoic; and these combined units, in turn, moved slowly across the latitudes, producing climatic change; lithofacies changed in response to both the climate and the plate tectonics.

Representative geographies that show the range of change have been deduced (Figs. 1–4). The map for the Cambro-Ordovician portrays the general early Paleozoic patterns (Fig. 1); these hold for the entire span of time from the Early Cambrian (about 540 Ma), through the Cambrian, Ordovician, and Silurian, into the early Devonian (about 400 Ma). There were consolidations in the Northern Hemisphere in

System	Selected fossil groups									
Neogene										
Paleogene										
Cretaceous										
Jurassic		conodonts							silicoflagellates	diatoms
Triassic			goniatites	brachiopods				fusulinids		
Permian				ammonites				endothyrids	rotalids	calcareous nanoplankton
Pennsylvanian										dimoflagellates
Mississippian	trilobites	graptolites	tentaculifids							non-cephalopod mollusks
Devonian		nautiloids			chitinozoans					
Silurian					spore-pollen	radiolarians				
Ordovician										
Cambrian										

Fig. 5. Major fossil groups used for detailed biostratigraphy of the Paleozoic and younger strata. These are not total ranges of all groups. (After J. T. Dutro, R. V. Dietrich, and R. M. Foose, eds., *AGI Data Sheets, 3d ed.*, American Geological Institution, 1989)

the Devonian, leading to a northern landmass—the so-called Old Red Continent (Fig. 2)—the forerunner of the Euro-American megaplate of the Carboniferous (Fig. 3), and culminating at the end of the Permian in the Pangaeon continental mass (Fig. 4). This, in turn, set the stage for the breakup of Pangaea during the subsequent Mesozoic tectonic megacycle. See PALEO GEOGRAPHY.

**Biostratigraphy and biostratigraphy.** The complexities of evolution from relatively simple forms at the beginning of the era to more advanced faunas and floras at the beginning of the Mesozoic produced a web of distributions in both time and space during the Paleozoic. In general terms, there were fewer and simpler life forms in the Cambrian—often termed the Age of Trilobites. All groups of invertebrates and plants became more numerous through geologic time. For example, 7 major invertebrate animal groups at the beginning of the Cambrian doubled to 14 by the end of the period, 20 by the end of the Ordovician, 23 at the end of the Devonian, and 25 at the end of the Paleozoic. The pattern for plant diversification, although starting later, is similar. Three simple plant groups became 5 by the end of the Silurian, 7 at the end of the Devonian, and 13 at the end of the Paleozoic. The vertebrates also diversified very slowly. From one or two groups in the Cambro-Ordovician (conodonts are now considered primitive vertebrates), the number of major kinds rose to 6 at the end of the Devonian and 8 at the end of the Paleozoic.

Biostratigraphic usefulness of fossils varies widely. Certain groups have been shown empirically to be more useful than others, and the abundance and diversity within these groups change from system to system during the Paleozoic. Groups with wide dispersal, occurrences in several facies, and rapid rates of evolution have proved most useful (Fig. 5). In the Paleozoic, trilobites are most valuable in the Cambrian and Ordovician; conodonts are more widely studied and are providing detailed biochronologic control for many system, stage, and zonal boundaries. Graptolites are indispensable in the deeper-water facies of the Ordovician through Early Devonian; goniatite cephalopods provide standards in the Devonian through the Permian; and fusulinids have long been essential for detailed work in the Carboniferous and Permian. Of course, all groups are useful for other kinds of paleobiologic research. Paleoenvironmental, paleoecological, and paleobiogeographic reconstructions use all appropriate biologic, chemical, and physical data in developing models of ancient Paleozoic worlds. See BIOGEOGRAPHY; CEPHALOPODA; CONODONT; FUSULINACEA; GEOLOGIC TIME SCALE; GRAPTOLITHINA; INDEX FOSSIL; PALEOECOLOGY; STRATIGRAPHY; TRILOBITA.

J. Thomas Dutro, Jr.

**Bibliography.** A. F. Embry, B. Beauchamp, and D. J. Glass (eds.), *Pangea: Global Environments and Resources*, Canadian Society Petroleum Geologists Memoir 17, 1994; F. M. Gradstein, J. G. Ogg, A. G. Smith (eds.), *A Geological Time Scale 2004*, 2005; M. R. House, C. T. Scrutton, and M. G. Bassett (eds.),

*The Devonian System*, Spec. Pap. Palaeont. 23, Palaeontological Association, London, 1979; E. G. Kauffman and J. E. Hazel (eds.), *Concepts and Methods of Biostratigraphy*, 1977; W. S. McKerrow and C. R. Scotese (eds.), *Palaeozoic Palaeogeography and Biogeography*, Geol. Soc. Mem. 12, Geological Society, London, 1990; R. C. Moore et al., *Treatise on Invertebrate Paleontology*, Part A, 1979; G. C. Young and J. R. Laurie (eds.), *An Australian Phanerozoic Time Scale*, Oxford University Press, 1996.

## Palladium

A chemical element, Pd, atomic number 46, and atomic weight 106.4. A transition metal, palladium occurs in combination with platinum (Pt) and is the second most abundant platinum-group metal, accounting for 38% of the reserves of these metals. See PERIODIC TABLE; PLATINUM.

1																	18
H																	He
3	4											5	6	7	8	9	10
Li	Be											B	C	N	O	F	Ne
11	12											13	14	15	16	17	18
Na	Mg	3	4	5	6	7	8	9	10	11	12	Al	Si	P	S	Cl	Ar
19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36
K	Ca	Sc	Ti	V	Cr	Mn	Fe	Co	Ni	Cu	Zn	Ga	Ge	As	Se	Br	Kr
37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54
Rb	Sr	Y	Zr	Nb	Mo	Tc	Ru	Rh	Pd	Ag	Cd	In	Sn	Sb	Te	I	Xe
55	56	71	72	73	74	75	76	77	78	79	80	81	82	83	84	85	86
Cs	Ba	Lu	Hf	Ta	W	Re	Os	Ir	Pt	Au	Hg	Tl	Pb	Bi	Po	At	Rn
87	88	103	104	105	106	107	108	109	110	111	112	113					
Rf	Ra	Lr	Rf	Db	Sg	Bh	Hs	Mt	Ds	Rg							
lanthanide series		57	58	59	60	61	62	63	64	65	66	67	68	69	70		
		La	Ce	Pr	Nd	Pm	Sm	Eu	Gd	Tb	Dy	Ho	Er	Tm	Yb		
actinide series		89	90	91	92	93	94	95	96	97	98	99	100	101	102		
		Ac	Th	Pa	U	Np	Pu	Am	Cm	Bk	Cf	Es	Fm	Md	No		

Palladium is soft and ductile and can be fabricated into wire and sheet. The metal forms ductile alloys with a broad range of elements. Palladium is not tarnished by dry or moist air at ordinary temperatures. At temperatures from 350 to 790°C (660 to 1450°F) a thin protective oxide forms in air, but at temperatures from 790°C (1450°F) this film decomposes by oxygen loss, leaving the bright metal. In the presence of industrial sulfur-containing gases a slight brownish tarnish develops; however, alloying palladium with small amounts of iridium or rhodium prevents this action. Important physical properties of palladium are given in the table. See ALLOY; METAL.

At room temperature, palladium is resistant to nonoxidizing acids such as sulfuric acid, hydrochloric acid, hydrofluoric acid, and acetic acid. The metal is attacked by nitric acid, and a mixture of nitric acid and hydrochloric acid is a solvent for the metal. Palladium is also attacked by moist chlorine (Cl) and bromine (Br). See ELECTROPLATING OF METALS; NON-STOICHIOMETRIC COMPOUNDS.

The major applications of palladium are in the electronics industry, where it is used as an alloy with silver for electrical contacts or in pastes in miniature solid-state devices and in integrated circuits. Palladium is widely used in dentistry as a substitute for

Physical properties of palladium	
Property	Value
Atomic weight	106.4
Naturally occurring isotopes (percent abundance)	102 (0.96)
	104 (10.97)
	105 (22.23)
	106 (27.33)
	108 (26.71)
	110 (11.81)
Crystal structure	Face-centered cubic
Thermal neutron capture cross section, barns	8.0
Density at 25°C (77°F), g/cm <sup>3</sup>	12.01
Melting point, °C (°F)	1554 (2829)
Boiling point, °C (°F)	2900 (5300)
Specific heat at 0°C (32°F), cal/g	0.0584
Thermal conductivity, (cal·cm)/(cm <sup>2</sup> ·s·°C)	0.18
Linear coefficient of thermal expansion, (μin./in.)/°C	11.6
Electrical resistivity at 0°C (32°F), μΩ·cm	9.93
Young's modulus, lb/in. <sup>2</sup> , static, at 20°C (68°F)	16.7 × 10 <sup>6</sup>
Atomic radius in metal, nm	0.1375
Ionization potential, eV	8.33
Binding energy, eV	3.91
Pauling electronegativity	2.2
Oxidation potential, V	-0.92

gold. Other consumer applications are in automobile exhaust catalysts and jewelry. *See* INTEGRATED CIRCUITS.

Palladium supported on carbon or alumina is used as a catalyst for hydrogenation and dehydrogenation in both liquid- and gas-phase reactions. Palladium finds widespread use in catalysis because it is frequently very active under ambient conditions, and it can yield very high selectivities. Palladium catalyzes the reaction of hydrogen with oxygen to give water. Palladium also catalyzes isomerization and fragmentation reactions. *See* CATALYSIS.

Halides of divalent palladium can be used as homogeneous catalysts for the oxidation of olefins (Wacker process). This requires water for the oxygen transfer step, and a copper salt to reoxidize the palladium back to its divalent state to complete the catalytic cycle. *See* HOMOGENEOUS CATALYSIS; TRANSITION ELEMENTS.

D. Max Roundhill

Bibliography. G. W. Gribble and J. J. Li, *Palladium in Heterocyclic Chemistry*, Pergamon, 2000; F. R. Hartley, *Chemistry of Platinum and Palladium*, 1973; J. Tsuji (ed.), *Palladium in Organic Synthesis (Topics in Organometallic Chemistry)*, Springer, 2005; J. Tsuji, *Palladium Reagents and Catalysts: New Perspectives for the 21st Century*, Wiley, 2d ed., 2004.

## Palpigradi

An order of rare arachnids comprising 21 known species from tropical and warm temperate regions. American species occur in Texas and California. All are minute, whitish, eyeless animals, varying from

0.03 to 0.11 in. (0.7 to 3 mm) in length, that live under stones, in caves, and in other moist, dark places. The elongate body terminates in a slender multisegmented flagellum set with setae. In a curious reversal of function, the pedipalps, the second pair of head appendages, serve as walking legs. The first pair of true legs, longer than the others and set with sensory setae, has been converted to tactile appendages which are vibrated constantly to test the substratum. *See* ARACHNIDA. Willis J. Gertsch

## Palynology

The study of pollen grains and spores, both extant and extinct, as well as other organic microfossils. Although the origin of the discipline dates back to the seventeenth century, when modern pollen was first examined microscopically, the term palynology was not coined until 1944.

The term palynology is used by both geologists and biologists. Consequently, the educational background of professional palynologists may be either geologically or biologically based. Considerable overlap exists between some areas of the fields, however, and many palynologists have interdisciplinary training in both the earth and life sciences. Palynologists use a range of sophisticated methodologies and instruments in studying both paleopalynological and neopalynological problems, but the utilization of modern microscopy is fundamental in both subdisciplines.

Palynologists study microscopic bodies generally known as palynomorphs. These include an array of organic structures, each consisting of a highly resistant wall component. Examples include acritarchs and chitinozoans (microfossils with unknown affinities), foraminiferans (protists), scolecodonts (tooth and mouth parts of marine annelid worms), fungal spores, dinoflagellates, algal spores, and spores and pollen grains of land plants. This discussion will focus on the palynomorphs produced by land plants, beginning with a general description of pollen grains and spores and then providing an overview of the primary areas of investigation within neo- and paleopalynological subdisciplines. *See* MICROPALYNOLOGY.

### Pollen Grains and Spores: An Overview

Spores and pollen grains are reproductive structures and play a paramount role in the life history of land plants. The sporophyte generation of nonseed-bearing plants (ferns, for example) produces single-celled spores that ultimately germinate to grow into the haploid gametophyte generation. Homosporous species produce a single type of spore, whereas heterosporous species produce two spore types. Microspores germinate and grow into "male" sperm-producing microgametophytes, and megaspores develop into "female" egg-producing megagametophytes. The gametophytes of most nonseed plants are multicellular and proliferate outside the spore wall during development. All seed-bearing plants

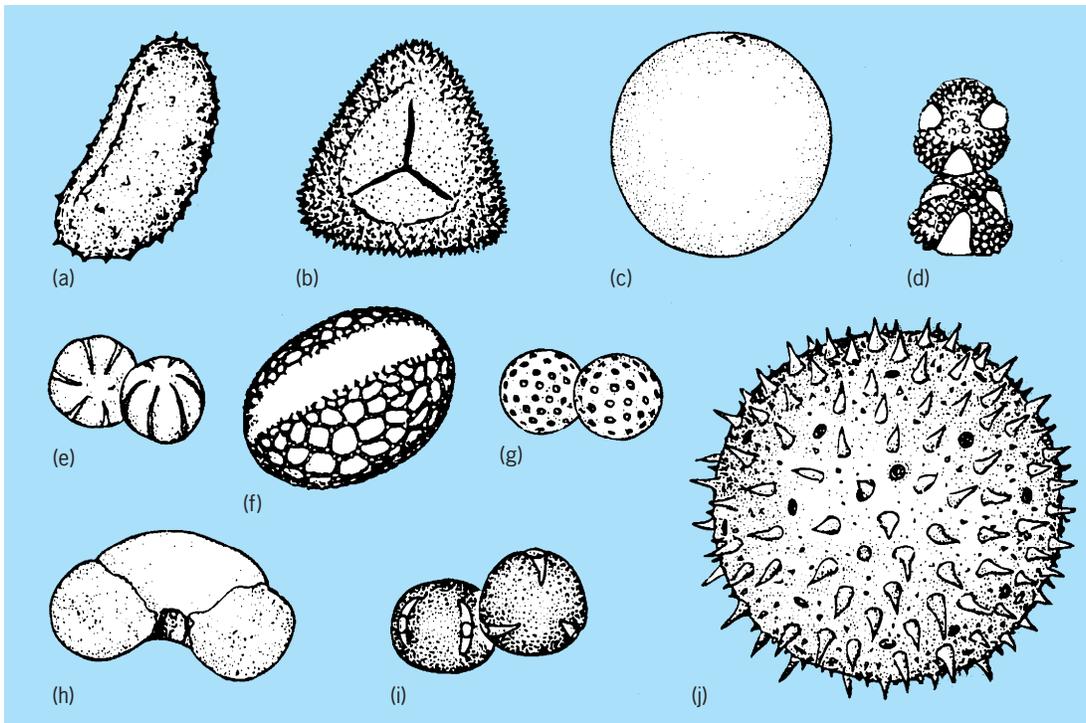


Fig. 1. Pollen and spore morphology, (a, b) Spores. (c–j) Pollen grains. (After Y. Iwanami, T. Sasakuma, and Y. Yamada, *Pollen: Illustrations and Scanning Electron Micrographs*, Kodansha and Springer, 1988)

(gymnosperms and angiosperms) are heterosporous, and their pollen represents the microgametophyte generation. Pollen grains consist of just three to a few cells, and these remain within the microspore wall, where they originally developed. See REPRODUCTION (PLANT).

Spores and pollen grains are formed in multiples of fours following meiotic divisions. During development, the four are united into a tetrad that, in most plants, subsequently dissociates into the four individual propagules. In nonseed plants, each spore commonly bears a mark on its proximal surface indicating where it made contact with the others at the center of the tetrad. In most spores this external mark is either straight or Y-shaped (Fig. 1), and it is typically characterized by a suture that spans the spore wall and is the site through which germination occurs. In contrast, most pollen grains lack sutures and germinate through thin areas in the wall, or apertures. Apertures are typically located in either a distal or an equatorial position. Common aperture types include elongate furrows, pores, and furrows with a central pore. Aperture type, number, and position are important systematic characters by which fossil and modern taxa can be compared. Other descriptively and systematically relevant characters include size, shape, presence and structure of air bladders, surface ornamentation, and wall ultrastructure. See POLLEN.

The wall of spores and pollen grains is known collectively as the sporoderm (or “skin of the spore”) and actually consists of two distinct walls (Fig. 2). The inner wall, or intine, is primarily composed of

cellulose and pectin; as such, it is similar to most other plant cell walls. The outer wall, or exine, is principally composed of sporopollenin, a chemically enigmatic macromolecule that is resistant to biological decay and geological degradation. The exine is further characterized by several ultrastructural layers and an array of sculptural elements. It is the very presence of the exine that allows for the spectacular preservation of pollen and spores in the fossil record.

### Neopalynology

This discussion focuses on several subdisciplines of neopalynology, including taxonomy, genetics, and evolution; development, functional morphology, and pollination; aeropalynology; and melissopalynology.

**Taxonomy, genetics, and evolution.** Taxonomy and systematics are concerned with classifying organisms into hierarchical ranks that reflect evolutionary, or phylogenetic, relationships. Pollen and spore morphology is important systematically, with particular features characteristic of different taxonomic ranks. For example, distinguishing characters may include aperture type for a family, different ornamentation patterns for its subordinate genera, and variation in exine ultrastructure for its congeneric species. Palynological characters are especially useful systematically when evaluated in conjunction with other characters (for example, plant morphological and molecular characters). Cladistics is one technique that has employed such an integrated approach. Cladistic analyses are based on numerical algorithms

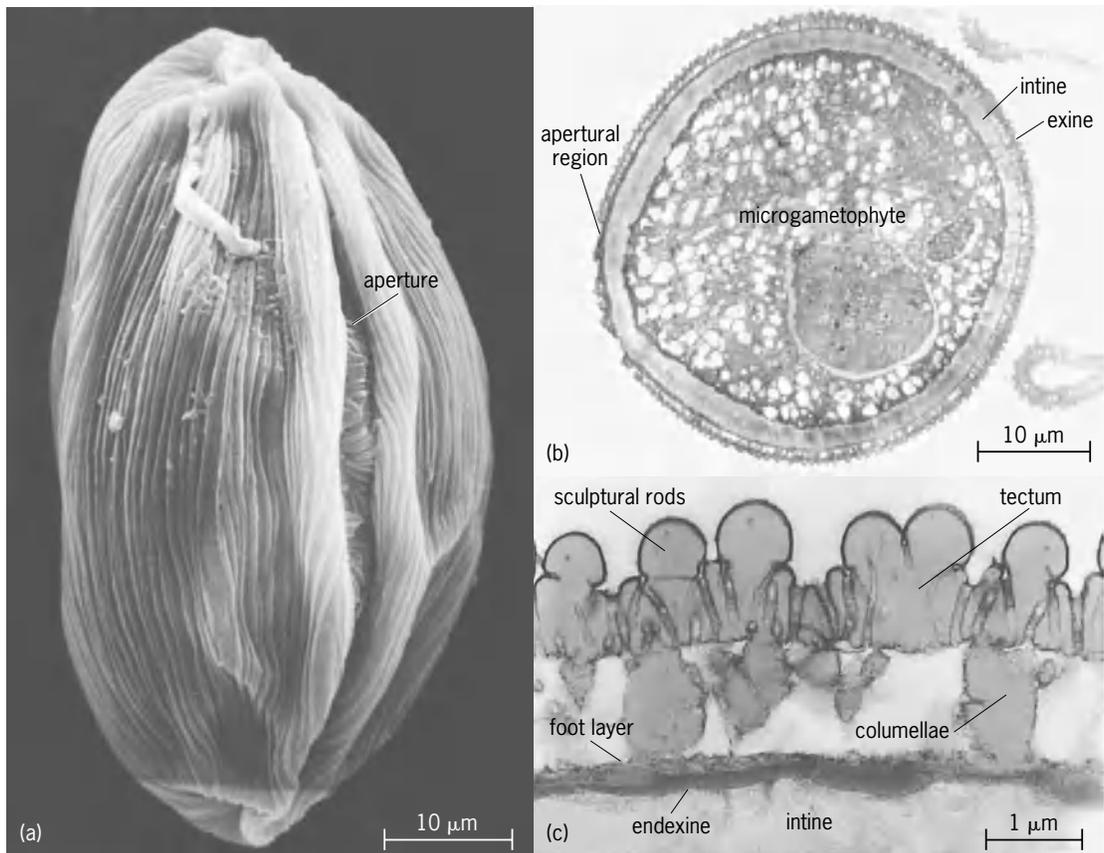


Fig. 2. Pollen morphology and sporoderm ultrastructure of *Cabomba caroliniana* (Cabombaceae), a modern water lily and primitive flowering plant. (a) Distal view of grain. (b) Cross section through the entire grain. (c) Cross section through the sporoderm.

that produce trees, or cladograms, demonstrating phylogenetic lineages among the organisms examined. See PLANT EVOLUTION; PLANT TAXONOMY.

Assessing pollen flow is another approach used to study evolutionary questions. Because pollen is the sperm-producing generation, the patterns and rates of pollen transfer are important factors in determining the spread of genes throughout a population. Competition can occur among reproductive organs, and pollen flow may be influenced by the correlation of pollen characters with those of reproductive organs. For example, the ovulate, or female, cones of some gymnosperms, such as pines, are aerodynamically adapted for entraining airborne pollen grains that themselves have particular aerodynamic characters, such as extended air bladders. Furthermore, competition may exist among individual pollen grains. Following pollination in some flowering plants, intraspecific variation can result in the grains of a more reproductively fit plant producing faster-growing pollen tubes than others of the same species. See POPULATION GENETICS.

#### Development, functional morphology, and pollination.

In most seed plants, a layer of callose, a carbohydrate, surrounds the entire tetrad and separates each immature pollen grain during development. Formation of the pollen wall and positioning of the aperture begin while each grain is encased within the callose layer. Both the internal ultrastructure and

the sculptural surface ornamentation of the outer pollen wall, the exine, are dependent upon the depositional pattern of the chemical that makes up the exine, sporopollenin. Sporopollenin is primarily derived from the developing pollen grain, but can also be released from a specialized layer of cells known as the tapetum, which surrounds the developing pollen grains. The inner pollen wall, or intine, is synthesized last.

When the tapetum breaks down, or undergoes programmed cell death (apoptosis), it also releases several proteins, lipids, and other substances that become isolated within the spaces and on the surface of the developing pollen wall. In flowering plants, many tapetum-derived proteins function as recognition molecules in pollination systems and are important in determining the extent of compatibility of a particular pollen grain on a floral stigma. Other pollen-derived proteins become stored within the intine and may also be involved in compatibility-incompatibility reactions. Several tapetal lipids also play important roles in pollination. Pollenkitt, for example, functions in pollen adhesion and acts as a visual and olfactory attractant. Additionally, pollen morphology and exine architecture may be correlated with pollination systems. For example, the pollen of wind-pollinated plants is typically smooth, has a thin exine, lacks pollenkit, and may have air bladders, whereas that of animal-pollinated plants is

commonly highly ornamented and bears pollenkitt. See FERTILIZATION.

**Aeropalynology.** Aeropalynology is the study of pollen grains and spores that are dispersed into the atmosphere. Wind-pollinated plants typically produce copious amounts of pollen, thereby enhancing successful pollinations. The abundance of airborne pollen commonly causes allergic reactions in a large proportion of the human population. Pollinosis, allergen rhinitis or hay fever, is elicited when allergen-containing pollen makes contact with the mucous membranes lining the nose, trachea, or bronchi, and the cornea of the eye. Allergens leach out of the pollen and bind to immunoglobulin E antibodies. The antibodies are linked to mast cells that release histamine and other inflammatory chemicals, producing allergy symptoms. Ironically, the allergens that induce pollinosis include many of the same compatibility-incompatibility, recognition proteins involved in pollination.

Knowledge of the temporal, seasonal, and environmental aspects of pollen dispersal is also important in understanding and avoiding hay fever. Flowering time and season vary widely for different plants, and the release of airborne pollen is typically inhibited by high humidity or rain. To monitor risks of pollinosis, the diversity and quantity of various pollen types are assessed by filtering the air throughout the year. See ALLERGY; ANTIBODY.

**Melissopalynology.** Honeybees are the primary pollinators of many flowering plants. Honeybees, and other bees, visit flowers to collect nectar and large quantities of pollen (pollen loads), both of which are used as food sources for developing larvae. Melissopalynologists analyze bee pollen loads and the pollen component within honeys. Although bees primarily produce honey from nectar, 1 mL of honey may contain more than 20,000 pollen grains. The foraging behavior of bees can be determined by microscopically examining their pollen loads and taxonomically identifying the pollen constituents. Honey purity can also be assessed by examining the diversity of pollen grains found within the honey.

### Paleopalynology

The main fields of study within paleopalynology are discussed below. The areas addressed involve paleobotany; past vegetation and climate reconstruction; geochronology and biostratigraphy; and petroleum and natural gas exploration.

**Paleobotany.** Fossil pollen and spores typically consist of only fossilization-resistant exine layers. However, these propagules did at one time contain both gametophytic cells and pectocellulosic intines, and functioned in a similar way to that of their extant counterparts. Fossil pollen and spores can be distinguished into two categories based on the general way in which the palynomorphs are preserved. *Sporae dispersae* grains are those occurring within sediments in a dispersed condition; in most cases, information about the parent plants is unavailable. Investigation of dispersed grains is especially impor-

tant in the fields addressed below. *In situ* grains occur within intact, megafossil reproductive organs (like flower anthers); as such, morphological data from the parent plant are available and provide for better systematic evaluation. Studies of *in situ* pollen or spores also afford the opportunity to evaluate and interpret fossils in a biological context. For example, developmental information can be inferred by examining pollen-containing organs preserved in various ontogenetic stages, and ancient pollination events, such as pollen germination and pollen tube growth, can be assessed when grains are recovered on receptive structures. These types of data allow paleobotanists to better understand and reconstruct the complete life history of fossil plants.

**Past vegetation and climate reconstruction.** A significant focus of palynology involves reconstructing the Earth's vegetational history since the last major glaciation event, within the past 10,000 years, or during the Holocene Epoch. This area of postglacial palynology is known as pollen analysis and primarily includes the study of palynomorphs from lake sediments and peat deposits. Sediments are obtained by several methods (mostly core sampling), and palynomorph diversity, distribution, and abundance are plotted on pollen profiles. Pollen analysis can provide historical information regarding both individual taxa and larger plant communities, including data about vegetational succession. Such analyses must consider all possible sources of palynomorphs and take precaution during sample preparation to avoid contamination with extant pollen because many modern taxa may have also existed in the Holocene. However, because of the excellent preservation potential of key palynological characters, such as those described above, fossil pollen grains yield a high degree of taxonomic resolution.

Because many plants inhabit areas exhibiting particular environmental regimes and have limited geographic distributions, palynological analyses contribute to an understanding of paleoclimatic conditions. For example, a palynoflora may be indicative of source vegetation occupying a polar, temperate, subtropical, or tropical habitat. Therefore, palynological information can also be used in conjunction with other megafossil indicators of climate, such as tree ring data. See POSTGLACIAL VEGETATION AND CLIMATE.

**Geochronology and biostratigraphy.** Palynological analyses play a significant role in age determinations of rocks, or geochronology. Dating the geological ages of palynomorph-bearing rocks is dependent upon knowledge of the stratigraphic ranges of extinct plant groups. Because different plant groups are known to have restricted geological time ranges, the pollen and spores produced by their plants are characteristic of particular ages and may serve as index fossils. Comparisons may be made against well-established reference palynomorphs, or index fossils, and palynofloras. Palynological dating techniques are especially useful when correlated with ages of rocks that have been radiometrically dated. See FOSSIL; INDEX FOSSIL.

Comparisons of palynomorphs within a given rock section from one site with those of units from other localities are important in documenting stratigraphic similarities among the rock sections, even if the sections exhibit different thicknesses and lithologies. When the occurrence, diversity, and abundance of fossils (palynomorphs, megafossils, or both) are used to correlate geographically separated rock sequences, this is known as biostratigraphy. Historically, biostratigraphic correlation has provided supporting evidence for continental drift theory. For example, some present-day continents, such as Antarctica, Africa, and India, have distinguishing index fossils, of various ages, that are present on these continents and absent from others. These intercontinental correlations are supportive of the previous existence of the single Southern Hemisphere landmass known as Gondwana. *See* PALEOGEOGRAPHY.

**Petroleum and natural gas exploration.** Economically, palynological biostratigraphy is an important technique used in the exploration for natural gas and petroleum. Biostratigraphic correlations in this context are conducted on a smaller scale, typically within an existing oil field. Besides identifying the locality, it is critical to determine the appropriate level at which to drill. For this endeavor, the palynologist is not necessarily interested in references of depth, but in important palynological indicators of known oil and gas production levels. In addition to the identification of key index fossils, a color evaluation is relevant. Following standard preparations, palynomorphs exhibit a range of colors that indicate their degree of geothermal alteration. Certain palynomorph colors are characteristic of rocks with either oil or gas reservoirs. *See* PALEOBOTANY; STRATIGRAPHY.

Jeffrey M. Osborn

**Bibliography.** K. Faegri, P. Kaland, and K. Krzywinski, *Textbook of Pollen Analysis*, 4th ed., Wiley, Chichester, 1989; M. M. Harley, C. M. Morton, and S. Blackmore (eds.), *Pollen and Spores: Morphology and Biology*, Royal Botanic Gardens, Kew, 2000; J. Jansonius and D. C. McGregor (eds.), *Palynology: Principles and Applications*, vols. 1-3, American Association of Stratigraphic Palynologists, Salt Lake City, 1996; R. O. Kapp, O. K. Davis, and J. E. King, *Ronald O. Kapp's Pollen and Spores*, 2d ed., American Association of Stratigraphic Palynologists, College Station, 2000; S. Nilsson and J. Pragłowski (eds.), *Erdtman's Handbook of Palynology*, 2d ed., Munksgaard, Copenhagen, 1992; A. Traverse, *Paleopalynology*, Unwin Hyman, Boston, 1988.

## Pancreas

A composite gland in most vertebrates, containing both exocrine cells—which produce and secrete enzymes involved in digestion—and endocrine cells, arranged in separate islets which elaborate at least two distinct hormones, insulin and glucagon, both of which play a role in the regulation of metabolism, and particularly of carbohydrate metabolism. This article discusses the anatomy, histology, embryology,

physiology, and biochemistry of the vertebrate pancreas. *See* CARBOHYDRATE METABOLISM.

### Anatomy

The pancreas is a more or less developed gland connected with the duodenum. It can be considered as an organ which is characteristic of vertebrates.

**Chordates and lower vertebrates.** In *Branchiostoma* (*Amphioxus*) a pancreatic anlage is found in young stages as a thickening of the gut caudal to the liver. The pancreas of cyclostomes, arising from the gut epithelium or from the liver duct, seems to be purely endocrine; it degenerates in later stages of development.

A true pancreas is found in selachians, with an exocrine portion opening into the intestine and an endocrine portion represented by cellular thickenings of the walls of the ducts.

**Higher vertebrates.** The ganoids (palaeopterygian fishes) show a diffuse pancreas—its principal mass lying between the gut and the liver—in which typical islets of Langerhans are observed. The pancreas of teleosts is either of the massive or dispersed type. Many species, such as the pike, show enormous islets of Langerhans, 10 × 5 mm, from which J. McLeod (1922) extracted insulin. The existence of a pancreas in dipneusts, such as *Protopterus*, is doubtful.

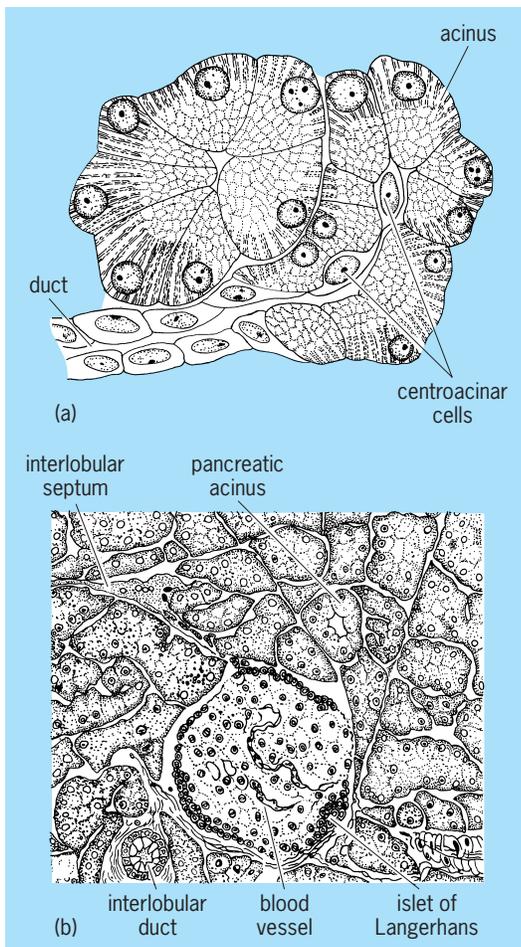
The compact pancreas of the amphibians is located in the gastrohepatic omentum and extends toward the hilus of the liver and along the branches of the portal vein. It develops from three anlagen, one dorsal and two ventral, the evolution of which varies from one species to another. The dorsal anlage would be the only source of endocrine islands. The pancreas of reptiles is very similar to that of amphibians; the number of excretory ducts varies from one to three.

In birds, the massive pancreas always lies in the duodenal loop. It develops from many dorsal and two ventral thickenings of the duodenal epithelium; one (sometimes two) excretory duct persists. The median portion of the dorsal anlage develops into a single mass which subdivides into typical islets of Langerhans. A complete ring of pancreatic tissue surrounds the portal vein.

The pancreas of mammals shows the same variations as in the fishes. The extremes are the unique, massive pancreas of humans, and the richly branched organ of the rabbit. Usually, the main duct, the duct of Wirsung, opens into the duodenum very close to the hepatic duct. Many rodents have this opening of the pancreatic duct as far as 40 cm (15.7 in.) from the hepatic duct. In humans, the pancreas weighs about 70 g (7.5 oz). It can be divided into head, body, and tail. A portion called the uncinata process is more or less completely separated from the head. Accessory pancreases are frequently found anywhere along the small intestine, in the wall of the stomach, and in Meckel's diverticulum. *See* DIGESTIVE SYSTEM.

### Histology

The pancreatic parenchyma is formed by two elements; one is the exocrine tissue of which the



**Fig. 1. Histology of the pancreas. (a)** Centroacinar cells in the pancreas of a guinea pig (after J. F. Nonidez and W. F. Windle, *Textbook of Histology*, 2d ed., McGraw-Hill 1953). **(b)** A section through the pancreas of the rat (after C. D. Turner, *General Endocrinology*, 2d ed., Saunders, 1955).

secretion empties into the pancreatic ducts and ultimately into the duodenum; the latter is the endocrine islands whose secretions enter the blood vessels.

**Exocrine pancreas.** The acini are the part of the exocrine pancreas that produces the enzymes of the pancreatic juice (**Fig. 1**). The acinar cells are pyramidal in shape and all are of the serous type. The apical pole of an acinar cell is filled with granules whose number varies according to the cell activity: they are particularly numerous in fasting animals, and few after a meal or after injection of pilocarpine. Since the granules contain the enzyme precursors, they are called zymogen granules. The basal pole of an acinar cell is intensely basophilic and contains a material formerly called ergastoplasm, which consists of staggered cisternae of rough endoplasmic reticulum. The well-developed Golgi apparatus is located in the supranuclear region of the cell.

All cells converge toward a central lumen, which is rather narrow when the organ is at rest and becomes dilated during secretion. On their lateral faces, the cells are separated by narrow spaces, the secretion capillaries, which are connected with the central lumen.

The digestive enzymes of the pancreatic juice are peptidases, lipases, esterases, amylase, and nucleases. Their synthesis starts in the basal cytoplasm, inside the endoplasmic reticulum through which they are transported to the Golgi system, where they are segregated in Golgi vesicles and later concentrated into typical granules. The secretory product flows out of the cell through a fissure created by the fusion of the granule-limiting membrane with the plasma membrane. All enzymes, except one lipase and the amylase, are inactive when secreted and become active only in the duodenal lumen. Their production is stimulated by various hormones originating from the gastric and duodenal mucosae.

The lumen of each acinus is continuous with the lumen of a small canal that is limited by so-called centroacinar cells. These cells are pale when stained as compared to the acinar cells because of the scarcity of their organelles. Where the ducts begin, their wall may be made partly with centroacinar cells and partly with acinar cells themselves.

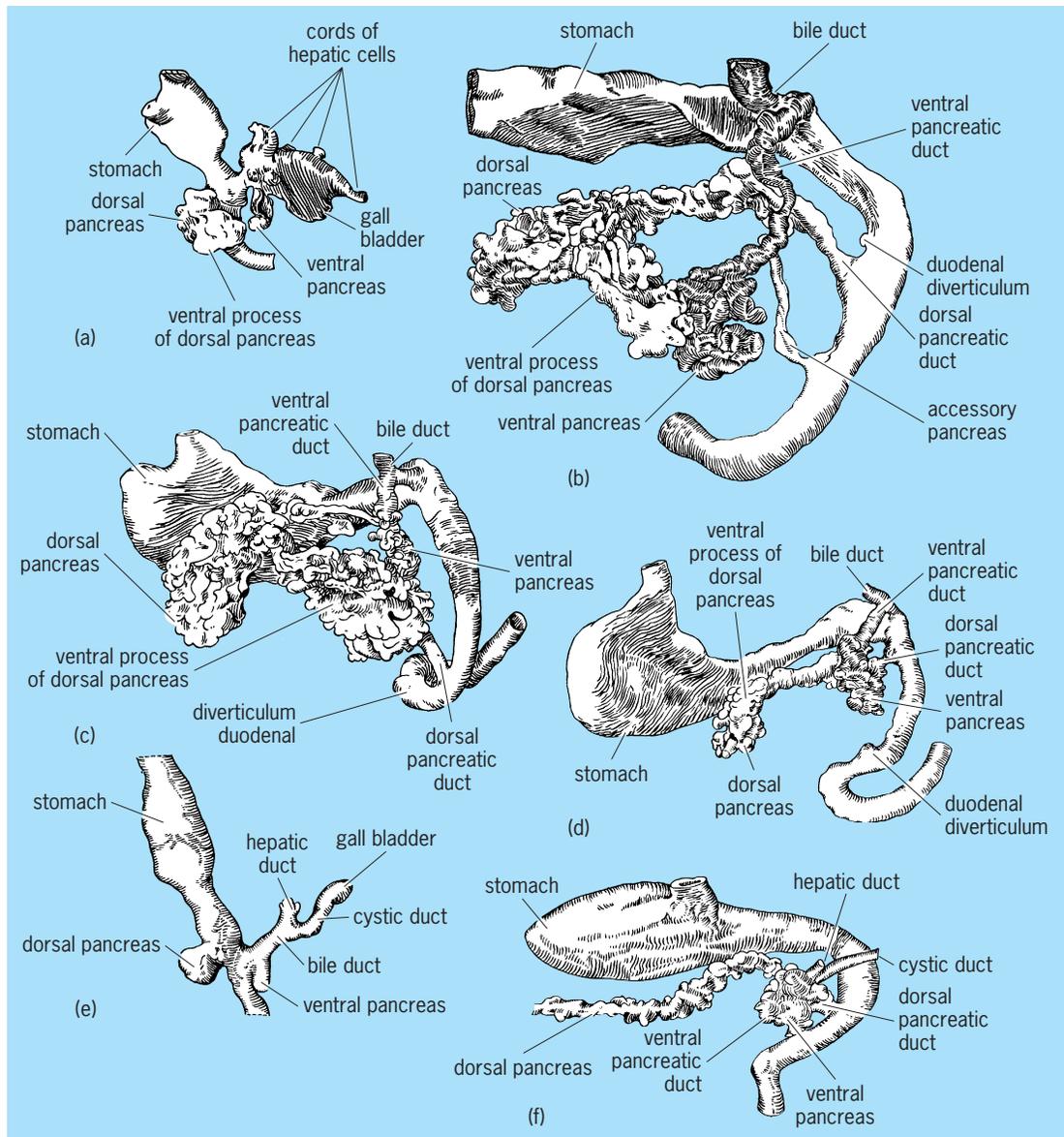
The distal part of the duct system drains proximally into the intralobular or intercalated ducts, which are lined by a low cuboidal epithelium and join together to form the interlobular ducts. These ducts are surrounded by connective tissue. Their wall structure varies with lumen diameter: when the diameter increases, the cuboidal epithelium that lines the ducts becomes cylindrical; goblet cells and isolated hormonal cells are interspersed between the other epithelial cells; some small mucous glands are annexed to the epithelium.

All interlobular ducts join the two main pancreatic ducts. The major pancreatic duct, the duct of Wirsung, runs the entire length of the pancreas, receiving numerous branches, and so gradually increases in size throughout its course. The accessory duct of Santorini lies cranial to the duct of Wirsung and only in the head of the pancreas. In both main ducts there is a layer of dense connective tissue containing elastic fibers that surrounds the stratified columnar epithelium.

**Endocrine pancreas.** The endocrine portion of the pancreas consists of cellular masses called the islets (or islands) of Langerhans scattered throughout the exocrine portion. In the adult human their number is estimated to range from 200,000 to 2,300,000 and their diameter varies from 30 to 300 micrometers. The islets are demarcated from the surrounding tissue by an irregular, thin layer of connective tissue.

The islet cells are arranged into cords and, in routine preparations, are paler when stained than the acinous cells. The two main cell types are the alpha or A cells and the beta or B cells, which exist usually in the proportion of 1:4. Another cell type, less abundant (about 5%), is the delta or D cell. The localization of these three cell types varies from species to species. In humans, B cells are usually central and surrounded by A and D cells.

The A cell contains acidophilic granules that are insoluble in alcohol. In electron micrographs these granules, whose diameter ranges from 190 to 310 nanometers, are spherical and very dense. They are



**Fig. 2. Pancreas models in vertebrate embryos. (a) 5.5-mm pig embryo. (b) 20-mm pig embryo. (c) 11-mm rabbit embryo. (d) 10.7-mm cat embryo. (e) 7.5-mm human embryo. (f) 13.6-mm human embryo. (After F. W. Thyng, *Models of the pancreas in embryos of the pig, rabbit, cat and man*, *Amer. J. Anat.*, 7:488-503, 1907)**

limited by a membrane which is separated from the granule by a narrow clear zone. The core of the granule consists probably of glucagon or of glicentin (GLI 1), the prehormonal form of glucagon. The paler peripheral part of the granule should contain a glicentine-related pancreatic peptide (GRPP). The contents of granules are excreted by exocytosis, and this secretion is stimulated by the lowering of the glucose blood level.

The B cell contains basophilic granules that are soluble in alcohol. In electron micrographs these granules, whose diameter ranges from 225 to 375 nm, are composed from a dense core surrounded by a large very clear halo. It has been shown by immunocytochemical methods that the central part contains insulin and that the peripheral part contains the connection peptide which links the two chains of

insulin. The crystals found in the central core are formed by insulin with zinc.

The D cell contains granules considerably less dense than the A or B granules and totally devoid of a clear ring. The D cells were first considered as altered A cells, but it is now clearly demonstrated that they produce somatostatin. It has been speculated that the secretion of somatostatin is necessary for the local regulation of the other pancreatic endocrine secretions. The long slender processes of the D cells which extend between the other cells and are rich in granules could be the morphological support of this paracrine regulation. D cells are most frequent in newborns.

Beside the A, B, and D cells, six other cell types have been identified mostly by immunocytochemical methods. These produce various substances such

as the pancreatic polypeptide, gastrin, secretin, and bombesin; their physiological significance in the islets of Langerhans still remains to be explained.  
S. Haumont

### Embryology

Ontogenetically, the organ is of entodermal origin. It arises initially from a series of outpocketings of the embryonic digestive tract, which eventually fuse, to various extents in various species, to form a single organ (Fig. 2). The adult gland in mammals results from the fusion of two such evaginations, one dorsal, opposite the hepatic diverticulum, and one ventrolateral, at the base of the biliary duct. In forms such as reptiles, anuran amphibians, and birds, two bilateral ventral evaginations are formed in addition to the single dorsal primordium. The elasmobranch pancreas is derived solely from a dorsal evagination, the ventral components never arising.

**Primordia.** The pancreatic diverticula make their appearance early in development, the dorsal element preceding that of the ventral component or components. In the chick, the pancreatic primordia appear between the third and fourth days of incubation, the dorsal primordia arising about the third day and the ventral on the fourth day. The dorsal component in the pig appears at the 4-mm stage, followed by the formation of the ventral diverticulum at the 5-mm stage. The pancreatic primordia of amphibians appear between the 8- and 10-mm stages, depending on the species.

**Pancreatic ducts and tabules.** The base of the evaginations eventually develop into the pancreatic ducts, Santorini's duct from the dorsal rudiment, Wirsung's duct from the ventral (Fig. 3). Santorini's duct, when it persists, opens directly into the duodenum, while the duct of Wirsung joins the common bile duct. Although both ducts may be retained throughout adult life in some species, such as the chick, horse, and dog, one or the other duct usually disappears, leaving a single pancreatic duct. Therefore, Wirsung's duct is the adult pancreatic duct in humans, sheep, ganoid fish, teleost fish, and the frog; on the other

hand, the duct of Santorini serves the adult elasmobranch, pig, and ox.

The original pancreatic primordia proliferate into masses of undifferentiated cells which then grow into solid cords of cells. These cords eventually are transformed into primitive pancreatic tubules. By a process of budding, the exocrine acini are formed from the primitive tubules. In addition, masses of loosely connected cells, centroacinar cells, are derived from the primitive tubules, and form an anastomosing network between the primary tubule and acinar system.

**Endocrine elements.** The endocrine cells of the pancreas are grouped in aggregates or islets of varying size which have no connection to the duct system. These aggregates, the islets of Langerhans, are highly vascularized and distributed throughout the organ. The distribution of islets is not uniform, and it has been suggested that this unequal distribution is a reflection of the fact that most of the endocrine cells are derived from the dorsal pancreas. This contention is best supported by the evidence presented from studies of certain urodele species in which fusion of the pancreatic diverticula occurs late and is restricted to the formation of a narrow isthmus of tissue between the pancreatic lobes. The existence of a separate giant islet of endocrine cells in certain teleosts also suggests the predominant role of one diverticulum in the development of the endocrine component.

**Islet morphogenesis.** The early endocrine elements of the pancreas arise from the cells of the diverticulum itself and from the primitive tubules. These cells form the primary islets of the pancreas, sometimes referred to as the islets of Laguesse. Considerable controversy has been waged over the eventual fate of the primary islets. The alternative views are (1) that the primary islets degenerate and are replaced by the definitive islet tissue, and (2) that the early islets persist into adult life. A second generation of islet cells occurs in the developing pancreas, although many investigators do not make a sharp distinction between primary and secondary islets. These secondary islets are derived largely from centroacinar cells and from the cells of the developing duct system. It has also been suggested that differentiated acinar cells can, by a process of dedifferentiation, give rise to the secondary islets. This contention, however, has been hotly disputed. Transitional stages in the formation of these islets of endocrine cells from ductules have been described in the embryonic organ. In adult organs of certain primitive species, such as the Elasmobranchii, these transitional stages may persist throughout adult life. For example, in some species of elasmobranchs, the islet cells are arranged in sheets next to the duct cells; in others, the islets may be connected to a ductule by a solid cord of cells. Phylogenetically, a progression of pancreatic types exists which resembles the development stages occurring in any one of the higher vertebrates.

**Lower chordates.** Scattered cells resembling pancreatic cells have been described in the digestive tube of the lancet *Branchiostoma*. In the lamprey

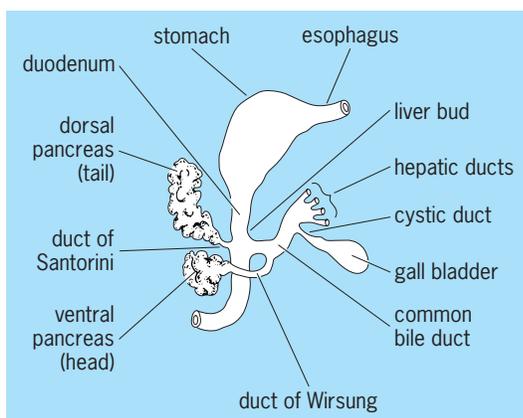


Fig. 3. Embryonic development of liver and gallbladder and of dorsal and ventral pancreatic buds before fusion of the latter. (After G. C. Kent, *Comparative Anatomy of the Vertebrates*, McGraw-Hill, 1954)

*Petromyzon*, the pancreas remains embedded in the wall of the intestine, having lost, secondarily, its pancreatic duct. The suggestion has been made that in the Cyclostomata, like the lamprey, the pancreas is primarily endocrine in function. Certain anatomical and histological features of the development of the digestive tract of the ammocoete larvae of the lamprey have led several investigators to speculate that this organism represents a transitional condition in the evolution of the vertebrate pancreas. At the junction of the esophagus and intestine, a mass of darkly staining cells forms from follicles budding off from the intestinal wall. These follicles lose their connection with the intestinal epithelium and, based on the histological features of their component cells, have been identified with the islet tissue of the higher vertebrates. In the anterior portion of the ammocoete intestine, two bilaterally arranged areas containing cells with a distinctly granular cytoplasm have been described, and it has been suggested that they are homologs of the exocrine portion of the vertebrate pancreas. The granules of these cells exhibit staining properties similar to the zymogen granules of the acinar cells of the pancreas of higher vertebrates. The production of a proteolytic enzyme of the tryptic type, thus similar to one of the digestive enzymes of the pancreas, has been localized in the anterior portion of the intestine. The hypothesis is further strengthened by the observation that in the larva of an Australian lamprey, *Geotria australis*, the intestine grows anteriorly as a pair of blind pouches. These pouches are lined with the granular cell type. It is difficult to escape the temptation to speculate that the ammocoete larva represents an evolutionary stage in which the endocrine pancreas has already separated from its original location and that the zymogenlike cells are preparing to follow suit, namely, the condition in the larva of *G. australis*. The next phylogenetic advance is observed in the elasmobranch. In these forms, a single dorsal pancreatic primordium and, consequently, the single duct of Santorini develop. Thus, phylogenetically as well as ontogenetically, the dorsal pancreas arises first. The first orders in which a pancreas similar to the mammalian organ appears are the ganoids and teleosts.

Another observation bespeaking the embryonic origin of the pancreas from the epithelium of the primitive digestive tract is the demonstration that cells with staining properties similar to those of the alpha cells occur in the pyloric portion of the bovine stomach. Extracts of this region give a positive test for the hormone glucagon.

**Cell types.** Within an islet of Langerhans, at least three distinct granular cell types can be distinguished on the basis of tinctorial differences among the granules demonstrable with polychrome stains. Acidophilic alpha cells are concerned with the synthesis of the hormone glucagon, and the basophilic beta cell is the site of insulin synthesis. The third cell type, the delta cell, may be a transitional form. In addition, a fourth, agranular, cell type may be present in some species. Although, in most species, each islet contains all of the representative cell types, it has been

reported that, in the domestic fowl, the individual islets may be predominantly alpha-cell or beta-cell islets. Considerable variation exists as to the time and order of appearance during development of the cell types among the various species. In the rat, the eta cell can first be discerned in the 18 $\frac{1}{2}$ -day fetus; alpha cells cannot be demonstrated until 2 days postpartum. The alpha cell, on the other hand, is the first islet cell type recognized in the chick pancreas, appearing on the eighth day of incubation. On the twelfth day, concomitantly with the appearance of large numbers of degenerating cells, beta cells can be distinguished. As yet, no definitive evidence is available which would permit a correlation between the detection of a particular cell type and the production of the hormonal product of that cell type. However, on the basis of two types of observations, it can be safely assumed that pancreatic hormones are elaborated during embryonic life. First, the effects of pancreatic insufficiency or ablation are temporarily alleviated in the pregnant female. Second, insulin and glucagon have been isolated from the pancreata of fetal cattle and swine, the embryonic organ yielding more hormone per unit weight than adult glands.

Irwin R. Konigsberg

### Exocrine Physiology

The exocrine function of the pancreas is subserved by acinar and duct cells which form pancreatic juice. Acinar cells make up about 90% and duct cells only about 4% of pancreatic tissue, the rest being endocrine tissue. While secretion of electrolytes and water originates from both acini and excretory ducts, enzyme secretion occurs only in acinar cells. Enzymes are synthesized in the ribosomes attached to the surface of the endoplasmic reticulum, are then transported across the boundary of the endoplasmic reticulum, and are formed into granules within the intracisternal spaces of the reticulum. The granules move toward the Golgi apparatus and become invested with a membrane to become mature zymogen granules which are stored in the apical region of the acinar cell. On stimulation they fuse with the luminal plasma membrane and release their contents into the lumen. The zymogen granule membranes are retrieved and reutilized after degradation into proteins by lysosomes and resynthesis in the endoplasmic reticulum. The juice passes via the duct system to be collected by the main duct which opens into the duodenum, where digestion of proteins, fat, and carbohydrates takes place.

**Stimulation of secretion.** In humans some 500–800 ml of juice is secreted daily. Secretion is controlled by the vagus nerve and by two main peptide hormones; secretin and cholecystokinin-pancreozymin. While secretin evokes a plasmaitonic fluid rich in bicarbonate from duct cells, both acetylcholine and cholecystokinin-pancreozymin stimulate enzyme secretion and in some animal species, such as rat, they also stimulate an isotonic NaCl fluid from acinar cells. With increasing flow rate as stimulated by secretin, bicarbonate concentration increases up to ~120 meq/liter and chloride

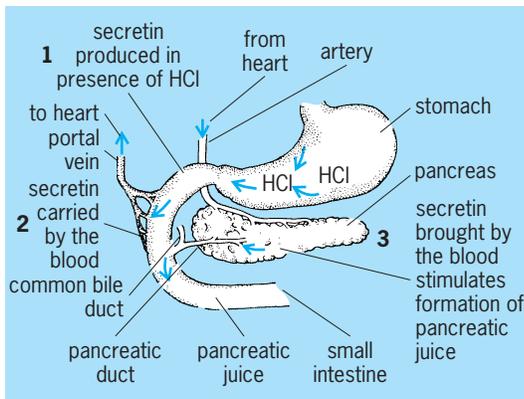


Fig. 4. The path and action of secretin in stimulating production of pancreatic juice. (After T. I. Storer and R. L. Usinger, *General Zoology*, 3d ed., McGraw-Hill, 1958)

concentration decreases, while sodium and potassium concentrations remain constant.

Secretin also evokes enzyme secretion from acinar cells in a variety of animal species. The intracellular mechanism of secretin-stimulated enzyme secretion, however, is different as compared to that of acetylcholine or cholecystokinin-pancreozymin-evoked enzyme secretion.

**Intracellular messengers.** Stimulation of enzyme secretion by secretagogues, such as acetylcholine or cholecystokinin-pancreozymin from pancreatic acinar cells, is mediated by rise of the cytosolic free  $\text{Ca}^{2+}$  concentration. This increase in  $\text{Ca}^{2+}$  is partly due to secretagogue-induced release of  $\text{Ca}^{2+}$  from intracellular stores and partly due to increased  $\text{Ca}^{2+}$  influx into the cell.

Enzyme secretion from pancreatic acinar cells is stimulated by secretin and mediated by increase in intracellular adenosine-3',5'-cyclic monophosphate (cAMP). Fluxes of  $\text{Ca}^{2+}$  are not changed by secretin; rather, secretin potentiates the response of cholecystokinin-pancreozymin or of acetylcholine to stimulate enzyme secretion. Secretagogues which induce stimulation by mediation of  $\text{Ca}^{2+}$  also elicit a variety of other intracellular signals, such as increase in guanosine 3',5'-cyclic monophosphate or of metabolites related to the hydrolysis of phospholipids, such as diacylglycerol, phosphatidic acid, arachidonic acid, and prostaglandins. For some, such as prostaglandins and diacylglycerol, participation in stimulus-secretion coupling seems to be likely, whereas for others, such as cyclic guanine monophosphate, a direct role in enzyme secretion cannot be assumed.

**Regulation of secretion.** Within a few minutes of taking food, pancreatic juice is secreted. This rapid response occurs reflexively via the vagus nerve and only occurs if the vagi are intact. Conversely, stimulation of the vagus provokes enzyme-rich juice.

The hormones secretin and pancreozymin are produced by the duodenal and jejunal mucosa. The acidity of the duodenal contents as well as fat and products of fat digestion are the most effective stimulants for secretin release, whereas fat, peptones, carbohydrates, and amino acids are most effective

in liberating cholecystokinin-pancreozymin. On liberation into the portal venous blood, these hormones gain access to the systemic circulation and reach the pancreatic tissue via the arterial blood (Fig. 4).

Irene Schulz

### Endocrine Physiology

In 1889 J. von Mehring and O. Minkowsky were the first to remove completely the pancreas of a dog, thereby causing true diabetes. This experiment is easily performed in most animals with the same results. In the rabbit, however, the complete removal of the gland is difficult, because of its ramified constitution. In birds, pancreatectomy is followed only by a very transient hyperglycemia. F. Banting and C. Best (1922) prepared pancreatic extracts which were able to prevent the lethal effects of pancreatectomy. The same effect was obtained with extracts from pancreas in which, after ligation of the duct of Wirsung, the exocrine portion of the gland had disappeared. The insular origin of the active factor, called insulin by De Meyer in 1909, was proved. A considerable amount of histological, embryological, experimental, and pathological material permits the differentiation of the roles of the two main cellular constituents of the islets of Langerhans, namely, the alpha and the beta cells. See ENDOCRINOLOGY; GLAND.

**Effect of alloxan.** The beta cells secrete insulin, that is, the hypoglycemic factor. The injection of alloxan, according to J. Dunn and N. McLetchie (1943), causes a selective destruction of the beta cells and the appearance of the various symptoms of diabetes. The mechanism of the diabetogenic action of alloxan is still a matter of discussion. Repeated injections of the extract of the anterior lobe of the pituitary gland cause alterations of the beta cells and diabetes; the severity of this diabetes is proportional to the degree of degranulation of the cells. The histological study of the pancreases of individuals who have suffered from diabetes frequently shows marked alterations of the beta elements. Certain types of pancreatic adenomas found in the human are essentially constituted by beta cells, and they are associated with hyperinsulinism and hypoglycemia. As further arguments in favor of the beta-cell origin of insulin, the damaging effect of dithizone on the dog's beta cells, associated with diabetes and with the disappearance of zinc from the cells, may be mentioned. Dialuric acid has identical effects, followed by regeneration of beta cells from acinous tissue. Various substances that have been studied by Kadota and Midokawa cause alterations of which the specificity is not evident. See DIABETES; INSULIN.

**Effect of HGF.** Intravenous injections of insulin cause a transitory hyperglycemia, soon followed by the typical hypoglycemia. This was attributed to the existence in the injected insulin or pancreatic extract of another substance, called the hyperglycemic factor or glucagon. This factor has been purified, and its principal effect is to cause hyperglycemia through a process of glycogenolysis. Thus, it is known as the hyperglycemic-glycogenolytic factor or HGF. Besides

the upper two-thirds of the gastric mucosa of dogs and rabbits, the pancreas is the only organ yielding HGF. The insular origin of HGF is shown by the strong correlation between glucagon content and density of insular tissue, the highest content being found in fetal pancreas and the tail part of the adult pancreas. Atrophy of the acinous tissue after ligation of the ducts and degeneration of the beta cells following injection of alloxane do not affect the glucagon content. *See* ENDOCRINE MECHANISMS.

Ernest L. Van Campenhout

### Biochemistry

The alpha and beta cells in the islets of Langerhans are the sources of two hormones, insulin from the beta, and glucagon, also known as the hyperglycemic factor, from the alpha cells. The former is a hormone which influences carbohydrate metabolism, enabling the organism to utilize sugar. The latter accelerates the conversion of liver glycogen, the form in which carbohydrate is stored in liver and muscles until needed by the body, into glucose, the principal sugar used by the body to meet its energy requirements. Thus, glucagon elevates the blood sugar level, and its effects are the opposite of those of insulin, so that the two hormones together maintain the sugar metabolism of the body in balance. When the level of sugar in the blood becomes too low, the secretion of glucagon is stimulated.

**Insulin.** In the normal organism, the blood sugar level remains relatively constant, no matter what the intake or output. When the amount of sugar in the blood is raised, the islets of Langerhans in the pancreas increase the secretion of insulin, which then facilitates the oxidation of sugar as a source of energy. Although the correlation between pancreatic activity and diabetes was made as early as 1889 by J. von Mehring and O. Minkowsky, it was not until 1922 that insulin was discovered, by F. G. Banting, C. H. Best, J. J. R. Macleod, and J. B. Collip, as a life-saving treatment for diabetes mellitus. The hormone was prepared in crystalline form in 1926 by J. J. Abel and coworkers. The effect of insulin in diabetic patients was found to be remarkable; as little as 1 or 2 mg/day administered to an individual unable to manufacture insulin was sufficient to restore his whole sugar metabolism to normal. However, the exact mechanism whereby insulin exerts this action is still a subject of debate. Some investigators believe that the hormone acts on the permeability of the cell membrane, thus permitting the entrance of sugar into cells. Others believe that the enzyme hexokinase, which catalyzes the first reaction in the synthesis of glycogen from glucose, is controlled by insulin. Sugar metabolism is controlled by other factors besides insulin. For example, in 1924 A. B. Houssay demonstrated that the anterior pituitary plays an important role in the etiology of the diabetic state of an organism. *See* DIABETES.

The biological potency of insulin preparations is given in terms of international units (IU), defined as the hormonal activity of 0.125 mg of an interna-

tional standard preparation. It is generally assayed by injecting the hormone subcutaneously into fasting rabbits and then determining the rate and extent of the lowering of the blood sugar level. Pure insulin possesses an activity equivalent to 24 IU/mg. Insulin loses its physiological activity when treated with alkali to pH 10 or with proteolytic enzymes such as chymotrypsin.

Insulin is a polypeptide with a molecular weight of approximately 6000. It is generally prepared from beef pancreas, and this bovine insulin contains 48 amino acids in two peptide chains (A chain and B chain) joined by sulfur atoms or —S—S—bridges. The sequence of amino acids constituting the primary structure of bovine insulin was described by F. Sanger and coworkers in 1954. In 1972 a three-dimensional (tertiary) structure of insulin was proposed by D. Hodgkin and coworkers; it was derived from the atomic arrangement found by x-ray analysis of rhombohedral crystals of the hormone.

Insulin preparations from other species, such as pig, sheep, horse, and whale, have also been examined by Sanger and coworkers and have been found similar in structure to bovine insulin, except for the amino acid sequence comprising positions 8, 9, and 10 in the A chain of the molecule. The biological potency of insulin has not been found to differ from species to species, nor has it been possible to differentiate the various preparations immunologically. Insulin was isolated from a primitive vertebrate, the Atlantic hagfish (*Myxine glutinosa*) by D. F. Steiner and coworkers in 1975. The complete amino acid sequence of the fish hormone was determined, and it was found that hagfish insulin is in many ways structurally similar to other vertebrate insulins. Apparently, the structure of the insulin molecule has been effectively conserved during the evolution of the vertebrates. The chemical synthesis of insulin was achieved by three groups of investigators in 1963–1965 and confirmed its chemical structure as proposed by Sanger.

In 1974 biologically active, crystallized human insulin was successfully synthesized in a laboratory.

The difference between human and pig insulin is located at the residue position 30 in the B chain: Thr in humans is replaced by Ala in pigs. The conversion of pig insulin to human insulin has been accomplished by chemical means.

While studying the biosynthesis of insulin in 1967, Steiner and P. E. Oyer demonstrated that the hormone itself is formed by the cleavage of the smaller insulin molecule from a larger parent molecule called proinsulin in which the polypeptide chains have been sequentially synthesized. In 1968 Steiner and coworkers and R. E. Chance and coworkers independently isolated and characterized bovine and porcine proinsulins. Chance and his group also determined the complete amino acid sequence of porcine proinsulin (**Fig. 5**).

The precursor forms of rat insulin were identified and characterized by D. F. Stein, W. Gilbert, and their collaborators in 1976–1978. The

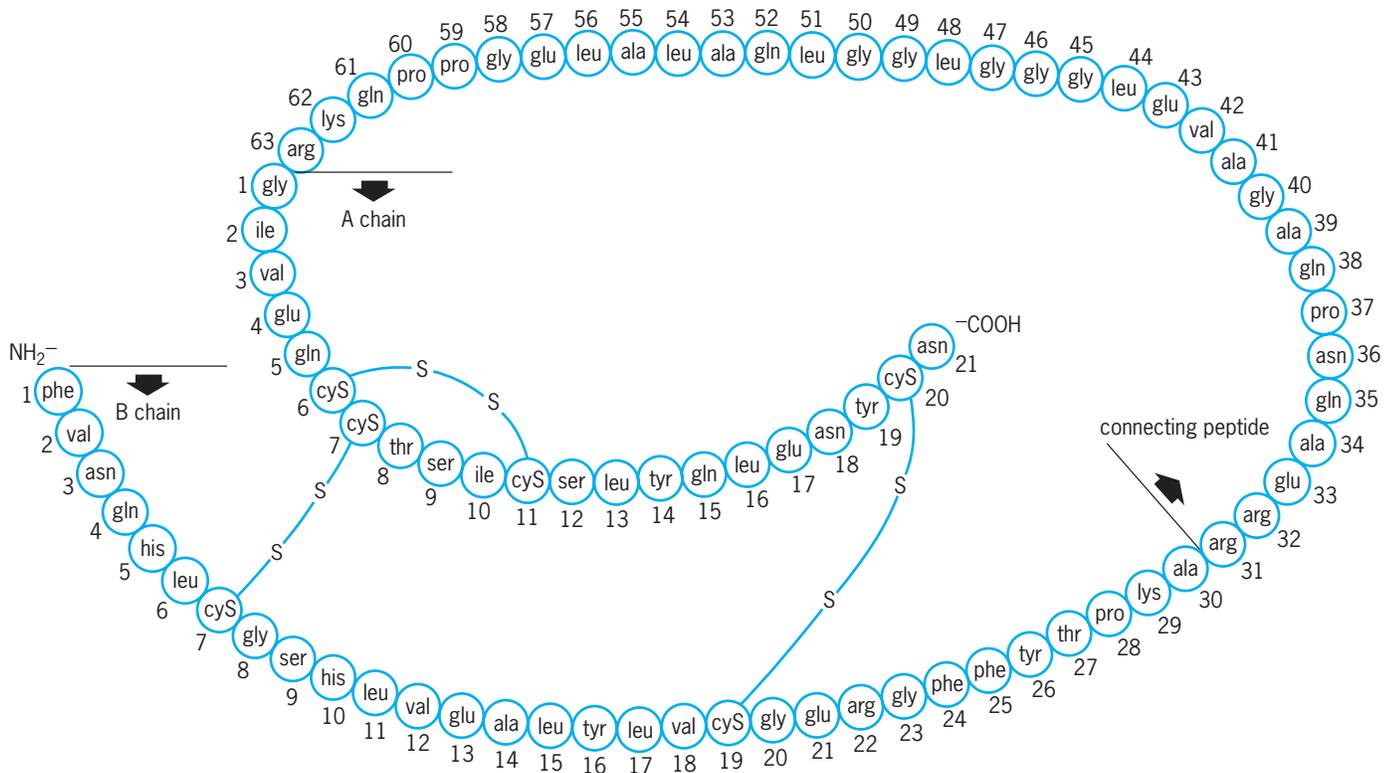


Fig. 5. Amino acid sequence of porcine proinsulin.

recombinant DNA-cloning synthesis of human insulin chains was achieved by D. V. Goeddel and coworkers in 1979.

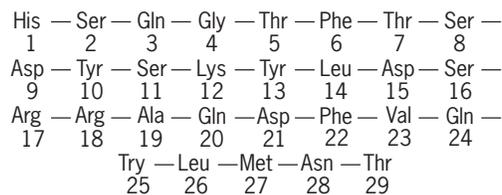
The first structurally abnormal mutant insulin found in a diabetic patient was suggested in 1980 to be [Leu<sup>B24</sup>]-insulin on the basis of its antagonistic effect. This was subsequently shown to be [Leu<sup>B25</sup>]-insulin in 1983. In the same year, the second case of a diabetic patient with an abnormal insulin was demonstrated to be [Ser<sup>B24</sup>]-insulin.

The initial step of insulin action is the insulin binding to its receptors. The binding kinetics of insulin-receptor interactions have been studied extensively and shown to be similar among different cells, with a dissociation constant of about 1 nM. The insulin receptor has been shown to be a glycoprotein consisting of two  $\alpha$  (apparent molecular weight of 125,000) and two  $\beta$  (apparent molecular weight of 94,000) subunits linked by disulfide bonds.

**Glucagon.** The existence of glucagon in pancreatic extracts was first demonstrated by O. Kimball and J. Murlin in 1923. It is frequently present in commercial preparations of insulin as an impurity; the administration of such insulin preparations causes an initial rise in blood sugar, followed by the lowering of blood sugar (hypoglycemia) characteristic of insulin action. See GLUCAGON.

In 1955 A. Staub, O. K. Behrens, and their coworkers obtained glucagon in crystalline form; 2 years later the same group of investigators established the complete structural formula of the hormone as a sin-

gle polypeptide chain of 29 amino acids with the sequence shown below, where His is histidine; Ser, ser-



ine; Gln, glutamine; Gly, glycine; Phe, phenylalanine; Thr, threonine; Asp, aspartic acid; Asn, asparagine; Tyr, tyrosine; Lys, lysine; Leu, leucine; Arg, arginine; Ala, alanine; Val, valine; Try, tryptophan; Met, methionine.

The three-dimensional structure of crystalline glucagon was determined in 1975; total synthesis of glucagon was achieved in 1967 and again in 1975. The precursor forms of angelfish and rat glucagons have been identified and characterized with molecular weights of 12,000 and 18,000, respectively.

**Pancreatectomy.** Within 24 h after the surgical removal of the pancreas, hyperglycemia, or elevation of blood sugar, is detectable, followed by glycosuria, or excess sugar secreted in the urine. Without the pancreas, the organism is unable to store or to oxidize sugar. This disturbance of sugar metabolism is also accompanied, when there is pancreas deficiency, by defective fat metabolism. The organism, in an attempt to compensate for the loss of its primary source of energy, sugar, increases the catabolism or

breakdown of fats. When the stores of carbohydrate in the body are depleted, ketone bodies, acid products which are responsible for the metabolism of fats, increase in both the blood and the urine, with serious consequences. The accumulation of these acid products leads to a depletion from the blood of the alkali reserve that balances them in the normal organism. As a consequence, the organism is unable to remove accumulating carbon dioxide in the blood, resulting in acidosis, or air hunger, which leads to coma and death. These symptoms are known to be associated with diabetes mellitus, which in humans is characterized by loss of weight, hyperglycemia, glycosuria, ketonemia, ketonuria, and polyuria.

Another pancreatic disorder, which has been observed in rare instances of tumors of the islet tissue, is known clinically as hyperinsulinism. The disease, associated with low blood sugar and indistinguishable from the consequences of insulin overdosage, is caused by excessive production of insulin by the tumor. See CARBOHYDRATE METABOLISM; PANCREAS DISORDERS.

Choh Hao Li

Bibliography. R. E. Chance et al., *Science*, 161:165–167, 1968; V. L. W. Go (ed.), *The Pancreas: Biology, Pathology, and Disease*, 2d ed., 1993; D. C. Hodgkin, *J. Endocrinol.*, 63:1–14, 1974; P. G. Kataoyannis, *Science*, 154:1509–1514, 1966; I. Schulz, Messenger role of calcium in function of pancreatic acinar cells, *Amer. J. Physiol.*, 239:G335–G347, 1980; P. Sieber et al., *Helv. Chim. Acta*, 57:2617–2621, 1974; D. F. Steiner et al., *Ann. N. Y. Acad. Sci.*, 343:1–16, 1980; H. Streb et al., Release of  $Ca^{2+}$  from a nonmitochondrial intracellular store in pancreatic acinar cells by inositol-1,4,5-trisphosphate, *Nature*, 306:67–69, 1983; H. S. Tager et al., *Ann. N.Y. Acad. Sci.*, 343:133–147, 1980.

## Pancreas disorders

The pancreas is affected by a variety of congenital and acquired diseases. Because of the dual functional role, the diseases of the exocrine portion of the pancreas will be separated from the endocrine lesions in this discussion.

**Congenital disorders.** The most frequent congenital lesion of the pancreas is more appropriately designated as a developmental abnormality—ectopic or aberrant pancreas. Ectopic pancreas can be found anywhere within the gastrointestinal tract, but is more frequent in the stomach and duodenum. Both acinar and islet tissue may be present in aberrant pancreas. Clinically these tissues have little significance.

**Cystic fibrosis.** Cystic fibrosis (mucoviscidosis) is a systemic disease in which mucus secretion is altered so that a viscid mucus is produced. The disease is inherited as a mendelian recessive. Cystic fibrosis affects all exocrine glands, including the acinar portion of the pancreas. Production of altered mucus leads to dilation of the exocrine ducts (cystic), destruction of acinar tissue, and replacement of the destroyed tissue by fibrous connective tissue (fibrosis). The islets are not affected by this disease. Elevation

in secretion of sodium and chloride in sweat is also common. In addition to affecting the pancreas, the sweat glands, gastrointestinal tract, lungs, liver, and salivary glands are also involved in cystic fibrosis. Destruction of the pancreas can be so extensive that pancreatic insufficiency develops. Careful medical management successfully compensates for pancreatic insufficiency. Involvement of the lungs leads to pneumonia, bronchiectasis, and often death during childhood.

**Pancreatitis.** Acute hemorrhagic pancreatitis is a serious disease of unknown etiology which causes sudden liberation of activated pancreatic enzymes that digest the pancreatic parenchyma. The digestive process leads to dissolution of fat and production of calcium soaps. In addition, rupture of pancreatic vessels occurs with resultant hemorrhage and shock. This disease is associated with biliary tract disease, especially gallstones (cholelithiasis), alcoholism, hyperlipidemia, and hypercalcemia. See ALCOHOLISM; GALLBLADDER DISORDERS.

Chronic pancreatitis, perhaps better designated chronic relapsing pancreatitis, is a condition in which recurrent episodes of pancreatitis occur without the production of symptoms or with the production of mild symptoms. Destruction of the pancreatic tissue, with repair by fibrosis, calcification, and cyst formation, is frequent.

**Diabetes.** Diabetes mellitus is the principal disease associated with the endocrine portion of the pancreas. Two clinical forms of the disease are recognized—insulin-dependent diabetes mellitus and non-insulin-dependent diabetes mellitus. While many factors are involved in the causation of this disease, basically the disease is a result of the failure of the beta cells of the pancreas to produce appropriate kinds and amounts of insulin to meet metabolic needs. Diabetes mellitus is a disease affecting many organs—a systemic disease. Long-term consequences include the early development of lesions in arteries, especially atherosclerosis, the development of lesions in small blood vessels throughout the body (microangiopathy), and abnormalities of the kidneys and nerves.

The insulin-dependent form frequently occurs in the younger age group, hence it is referred to as juvenile-onset diabetes mellitus. Factors involved in the pathogenesis of this form of diabetes mellitus include injury to the islet cells by viral infection or autoimmune processes. Insulin is necessary in the therapy of these individuals. The majority of individuals with diabetes mellitus develop the disease process as adults—maturity-onset (non-insulin-dependent) diabetes mellitus. Many factors are included in the causation of diabetes mellitus in this group. In maturity-onset diabetes mellitus, the body may have difficulty in utilizing the insulin produced by the pancreatic islets. Increased production of another pancreatic hormone, glucagon, may also play a role. Genetic influences are certainly participating in the development of diabetes mellitus, but the exact role of inheritance has yet to be defined. Environmental factors such as obesity, pregnancy, and infections also

contribute to the development of the disease. See DIABETES.

**Tumors.** Tumors of the pancreas can be either benign or malignant. They affect both the endocrine and exocrine portions of the pancreas.

Benign tumors of the exocrine pancreas are extremely rare. Malignant tumors of the exocrine pancreas arise most frequently from the pancreatic ducts. Acinar carcinomas also exist but are very rare. Exocrine pancreatic carcinomas are very malignant tumors. They are clinically silent until well advanced. Cure by surgery or other therapeutic approaches is rare. Carcinoma of the pancreas affects individuals in the 60- to 80-year-old age group and occurs most frequently in the head and least frequently in the tail of the pancreas. Individuals develop jaundice, as obstruction of the common bile duct is frequent in this disease. Involvement of the nerves in and around the pancreas cause severe back pain. The most frequent histologic type of pancreatic carcinomas is a glandular pattern (adenocarcinoma).

Islet cell lesions are quite rare but may be associated with increased hormone production. The tumors can be single or multiple, benign or malignant, and they can form anywhere in the pancreas. Hyperfunction of the islets of Langerhans can result in three distinct clinical syndromes: hyperinsulinism and hypoglycemia, the Zollinger-Ellison syndrome (gastrinoma), and multiple endocrine neoplasia.

Beta-cell tumors (insulinomas) are the most common tumor, with elaboration of insulin and production of hypoglycemia. Most insulinomas are the result of a solitary benign tumor (adenoma). However, multiple adenomas, islet-cell carcinoma, and islet-cell hyperplasia can also cause the syndrome.

The Zollinger-Ellison syndrome is characterized by a pancreatic islet tumor that produces excessive amounts of gastrin, with a resultant hypersecretion of the mucosa of the stomach and ulceration of the upper gastrointestinal tract (peptic ulceration). The majority of gastrinomas are malignant, with metastasis to lymph nodes and liver. Gastrin is not produced by islet cells under normal conditions.

Individuals with islet-cell tumors may have tumors in other endocrine organs—a condition known as multiple endocrine neoplasia syndrome—including the pituitary, thyroid, parathyroid, and adrenal glands. Several different syndromes have been recognized and are characterized by the organs involved. See ONCOLOGY; PANCREAS. H. Thomas Norris

**Bibliography.** G. P. Burns and S. Bank, *Disorders of the Pancreas*, 1989; R. S. Cotran, V. Kumar, and S. L. Robbins (eds.), *Robbins' Pathologic Basis of Disease*, 6th ed., 1999.

## Panda

The family Ailuropodidae contains two species of pandas—the giant panda (*Ailuropoda melanoleuca*) and the lesser or red panda (*Ailurus fulgens*). Until relatively recently, the giant panda had been classified in the family Ursidae with the



**Fig. 1.** Giant panda (*Ailuropoda melanoleuca*) (Copyright © 2001 John White)

bears, and the red panda had been included in the family Procyonidae along with raccoons, ringtails, and coatis.

**Giant panda.** Giant pandas are an endangered species, surviving only in small isolated populations in China. Approximately 1000 giant pandas live in an area of about 29,500 km<sup>2</sup> (10,600 mi<sup>2</sup>). They inhabit a narrow zone of bamboo forest from 1200 to 3300 m (3800 to 10,600 ft) in elevation. Below this zone is cultivated land; above this zone is the treeline. There are currently 12 wildlife reserves comprising about 6000 km<sup>2</sup> (2200 mi<sup>2</sup>) in which the pandas live.

Giant pandas have massive heads and bodies with an unmistakable black and white fur pattern (**Fig. 1**). The body is whitish except for the ears, eye spots, nose, limbs, and shoulders which are black. The limbs are relatively short, and each forepaw possesses a “pseudthumb,” or sixth toe, which is an adaptation for stripping bamboo leaves from the stalk. The dental formula is I 3/3, C 1/1, Pm 4/4, M 2/2 × 2, for a total of 40 teeth. Adults are 150–180 cm (5–6ft) long with a 10–15 cm (4–6 in.) tail. They weigh 75–110 kg (165–242 lb). See DENTITION.

Giant pandas are active throughout the year. They are solitary and have broadly overlapping home ranges. Their primary food consists of the sprouts, stems, and leaves of bamboo, although they occasionally consume bulbs, grasses, insects, and rodents. Breeding occurs from mid-March to mid-May. Following a gestation period of 97 to 163 days, one or two cubs are born in a hollow tree or cave, mainly in August and September. Gestation is lengthened because, like bears and some other carnivores, pandas experience delayed implantation of the fertilized egg in the uterine wall. Newborn pandas are blind, pink, and almost hairless. The eyes open at about 45 days, and cubs begin feeding on bamboo at about 5 months of age. They remain with their mother for about 18 months; thus, females give birth only once every 2 or 3 years from age 4 to at least 20. Reproduction in captivity is poor, although artificial insemination and hand raising of young are increasingly successful. A captive specimen lived to an age of approximately 34 years.

**Red panda.** Red pandas have long, soft, thick rusty to chestnut-brown fur on their dorsal surface (**Fig. 2**). Small dark-colored patches are present



Fig. 2. Red panda (*Ailurus fulgens*). (Copyright © 2004 John White)

beneath each eye, and the muzzle, lips, cheeks, and edges of the ears are white. The backs of the ears, the limbs, and the underparts are dark reddish-brown to black. The claws are sharp and semiretractile, and the feet have hairy soles. The soles of the feet have numerous glands opening through minute pores, whose secretions are used for marking runways and trails. Locomotion is plantigrade (individuals walk on the soles of their feet). The bushy nonprehensile tail has 12 inconspicuous dark brown rings on a reddish background. The head is rounded with a short snout, and the ears are large and pointed. The dental formula is I 3/3, C 1/1, Pm 3/4, M 2/2, for a total of 38 teeth. Adults are 79–110 cm (31–44 in.) in total length, including a 28–48 cm (11–19 in.) tail. They weigh 3–6 kg (8–13 lb).

This species inhabits the mountain forests of China and the southeastern mountainsides of the Himalayas ranging in elevation from 2200 to 4800 m (7000 to 15,000 ft). Preferred habitat consists of giant rhododendron, oak, and bamboo forests. They are generally solitary and primarily nocturnal, although researchers have found them to be active in daylight, especially during the summer. Although red pandas are capable climbers, they do most of their feeding on the ground. The principal food is bamboo, but the diet is supplemented by grasses, acorns, fruits, roots, lichens, insects, eggs, young birds, and small rodents. A “pseudothumb” is present to facilitate the handling of bamboo leaves and stems.

Breeding occurs from January to March. Following a gestation of 90–145 days, one to four (usually two) fully haired, blind cubs are born from mid-June to late July in a nest in a hollow tree or a rock crevice. The eyes open at 17–18 days. They attain their adult coloration, leave the nest, and begin feeding on solid food at about 90 days of age. Sexual maturity is attained at 18 months. They may live up to 14 years.

The specialized diet, low reproductive rate, and low population density make this species highly vulnerable. It is classified as endangered by the IUCN (World Conservation Union) which estimates that fewer than 2500 mature individuals exist. See CARNIVORA; MAMMALIA.

Donald W. Linzey

Bibliography. *Grzimek's Encyclopedia of Mammals*, McGraw-Hill, 1990; D. Lindburg and K.

Baragona, *Giant Pandas: Biology and Conservation*, University of California Press, 2004; D. McDonald (ed.), *The Encyclopedia of Mammals*, Andromeda Oxford Limited, 2001; R. M. Nowak, *Walker's Mammals of the World*, 6th ed., Johns Hopkins University Press, 1999; G. B. Schaller et al., *The Giant Pandas of Wolong*, University of Chicago Press, 1985; G. B. Schaller, *The Last Panda*, University of Chicago Press, 1993.

## Pandanales

An order of monocotyledons, the composition of which only recently has been revealed by deoxyribonucleic acid (DNA) sequence studies of four genes. Included are four families, two of which have been generally viewed as closely related: Pandanaceae (800 species; the screw pine family) and Cyclanthaceae (230 species; the Panama hat family). These two families have often been considered to be related to Arecales (with a single family, Arecaceae, palms), but actually the relationship is distant. In addition, the order includes Stemonaceae (35 species) and Velloziaceae (200 species), which have been thought to be related to Dioscoreaceae (Dioscoreales) and Bromeliaceae (Commelinales) or Hypoxidaceae (Asparagales), respectively. Pandanaceae are often lianas or large herbs from the Old World; Cyclanthaceae are herbs or lianas from the New World tropics; Stemonaceae are herbs or lianas of the Old World (but with one species in the southeastern United States); and Velloziaceae are herbs or small shrubs of Africa and particularly South America (with one genus in southwest China). Cyclanthaceae, Pandanaceae, and Stemonaceae have flower parts in twos or fours, which is unusual among monocotyledons, in which threes are most common.

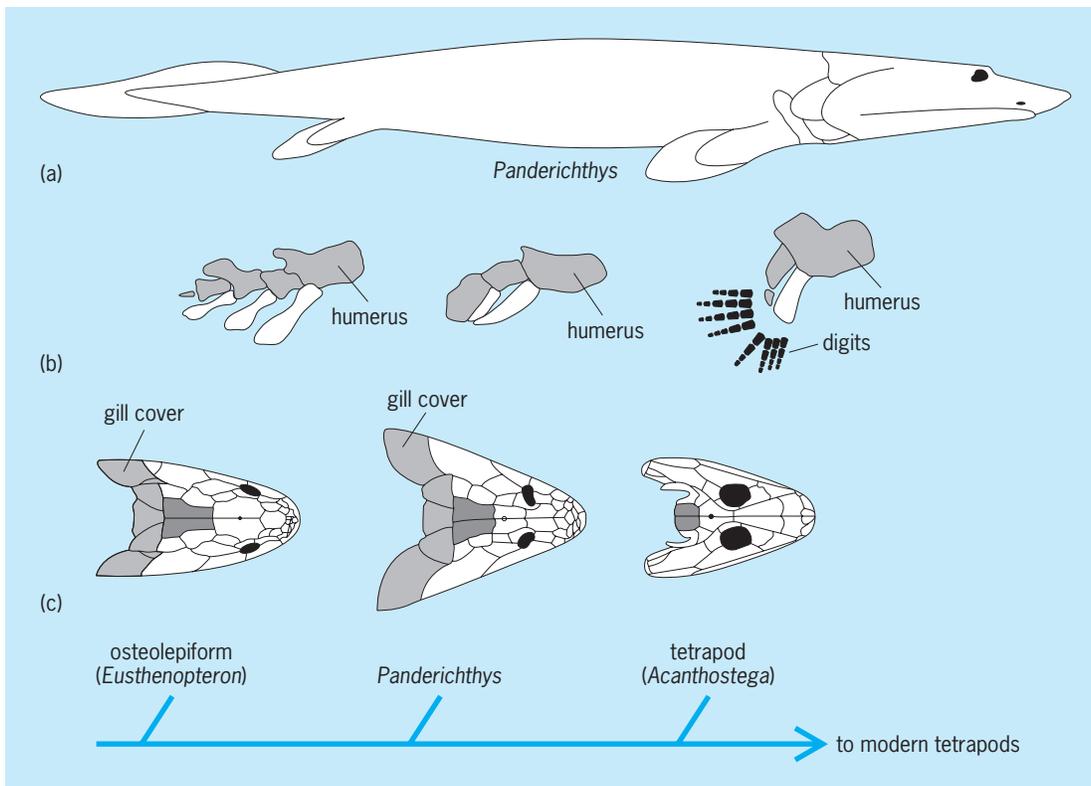
Several species in Pandanales are economically important. The leaves of Pandanaceae are fibrous and used for making rope and roofing, and the fruits are eaten in many areas. Cyclanthaceae leaves are fibrous and have similar uses, including the manufacture of Panama hats. See ARECIDAE; LILIOPSIDA; MAGNOLIOPHYTA; PLANT KINGDOM.

Mark W. Chase

## Panderichthys

A fossil lobe-finned (sarcopterygian) fish from rock strata of the Baltic region dating to the Middle-Late Devonian Period (about 372–368 million years ago), which provides key evidence about the evolutionary transition from fish to land vertebrates (tetrapods). As the most tetrapod-like fish known, it reveals the earliest steps of the transformation that adapted the fish body plan to terrestrial life. See SARCOPTERYGII; TETRAPODA.

**Distribution.** Fragmentary fossils of *Panderichthys* are widespread in Latvia, Estonia, and western Russia, in sandstones and clays that formed in river deltas along a tropical coast. But only one locality—the clay quarry of Lode in Latvia—yields complete



**Panderichthys anatomy.** (a) *Panderichthys rhombolepis*, whole-body reconstruction based on specimens from Lode, Latvia. Total length approximately 1.3 m (4.3 ft). Note the crocodile-like head shape, raised “eyebrows,” and loss of all midline fins except the tail (reprinted by permission from Nature, P. E. Ahlberg and A. R. Milner, *The origin and early diversification of tetrapods*, 368:507–514, © 1994, Macmillan Publishers Ltd.). (b) Pectoral fin/forelimb skeletons of *Panderichthys*, an osteolepiform, and a primitive tetrapod showing proportional changes and the abrupt appearance of digits in the earliest tetrapods. (c) Skulls of *Panderichthys* and a primitive tetrapod in dorsal view showing proportional changes (note the position of the eyes and the size of the shaded bones) and the loss of the gill cover during the fish-tetrapod transition.

bodies that reveal what the animal looked like (illus. a). Its anatomy proves to be a curious mixture of fish and tetrapod traits.

**Anatomy.** *Panderichthys* fits into the tetrapod phylogeny (family tree) above the osteolepiform fishes and immediately below the first tetrapods such as *Ichthyostega* and *Acanthostega*. It shares certain tetrapod-like characteristics with the osteolepiforms, such as a pair of internal nostrils (choanae) on the palate, and paired fin skeletons that contain recognizable equivalents of the major tetrapod limb bones, such as the humerus and femur (illus. b). However, whereas osteolepiforms still had the body form and fins of “normal” open-water fishes, *Panderichthys* looked quite different: it had a crocodile-like head with a long snout and eyes raised on top under distinct bony “eyebrows” (illus. c); a body that is slightly flattened from top to bottom; and a tail carrying only a simple fin fringe (illus. a), rather than the separate dorsal, caudal, and anal fins of osteolepiforms. All these features are shared with early tetrapods.

The internal anatomy reveals additional tetrapod features that are not present in osteolepiforms. For example, the skeleton of the pectoral fin, equivalent to the forelimb of tetrapods (illus. b), and shoulder girdle resembled those of early tetrapods, not just in construction but in shape, suggesting that the range of movement was becoming limblike. The ribs,

which help support the internal organs of the body, were also much larger than in osteolepiforms. However, unlike tetrapods, *Panderichthys* retained well-developed gills, and its paired appendages were still fins with fin rays, not limbs with toes. See ICHTHYOSTEGA; OSTEOLEPIFORMES.

**Mode of life.** These features not only identify *Panderichthys* as a close relative of tetrapods but give hints to its mode of life. Its pectoral fins appear adapted for some degree of “walking,” whether on land or in water, while the crocodile-like body form and raised eyes suggest that it operated in very shallow water and habitually looked out above the surface. The retention of gills and a tail fin indicate that *Panderichthys* had not really become terrestrial, while sharp, pointed teeth show that it was a formidable predator. *Panderichthys* can be envisioned as a specialist shallow-water predator of deltaic environments, which used its fins to crawl through water too shallow to support its body, and perhaps also for short excursions over land from one channel to the next. It is possible that it exploited the intertidal zone, emerging to feed on fishes stranded on mudflats and in tidal pools. See ANIMAL EVOLUTION; DEVONIAN; FOSSIL.

Per E. Ahlberg

**Bibliography.** J. A. Clack, *Gaining Ground*, Indiana University Press, 2002; J. A. Long, *The Rise of Fishes*, John Hopkins University Press, 1996; C. Zimmer, *At Water's Edge*, Free Press, New York, 1998.

### Panel heating and cooling

A system in which the heat-emitting and heat-absorbing means is the surface of the ceiling, floor, or wall panels of the space which is to be environmentally conditioned. The heating or cooling medium may be air, water, or other fluid circulated in air spaces, conduits, or pipes within or attached to the panels. For heating, electric current may flow through resistors in or on the panels. *See* ELECTRIC HEATING.

Warm or cold water is circulated in pipes embedded in concrete floors or ceilings or plaster ceilings or attached to metal ceiling panels. The coefficient of linear expansion of concrete is 0.000095; for steel it is 0.00081, or 15% less than for concrete. For copper it is 0.000112, or 20% more than for concrete, and for aluminum it is 0.000154, or 60% more than for concrete. Since the warmest or coolest water is carried on the inside of the pipes and the heat is transmitted to the concrete, only steel pipe should be used for panel heating and cooling systems, except when metal panels are used.

Cracks are very likely to develop in the concrete or plaster, breaking the bonds between the pipes and the concrete or plaster. The pipes move freely, causing scraping noises. An insulating layer of air is formed between the concrete or plaster, and this markedly reduces the coefficient of conductivity between the liquid heating medium and the active radiant surfaces.

**Heat transfer.** Heat energy is transmitted from a warmer to a cooler mass by conduction, convection, and radiation. Radiant heat rays are emitted from all bodies at temperatures above absolute zero. These rays pass through air without appreciably warming it, but are absorbed by liquid or solid masses and increase their sensible temperature and heat content. *See* HEAT TRANSFER.

The output from heating surfaces comprises both radiation and convection components in varying proportions. In panel heating systems, especially the ceiling type, the radiation component predominates. Heat interchange follows the Stefan-Boltzmann laws of radiation; that is, heat transfer by radiation between surfaces visible to each other varies as the difference between the fourth power of the absolute temperatures of the two surfaces, and is transferred from the surface with the higher temperature to the surface with the lower temperature.

The skin surface temperature of the human body under normal conditions varies from 87 to 95°F (31 to 35°C) and is modified by clothing and rate of metabolism. The presence of radiating surfaces above these temperatures heats the body, whereas those below produce a cooling effect. *See* RADIANT HEATING.

**Cooling.** When a panel system is used for cooling, the dew-point temperature of the ambient air must remain below the surface temperature of the heat-absorbing panels to avoid condensation of moisture on the panels. In regions where the maximum dew-point temperature does not exceed 60°F (16°C), or

possibly 65°F (18°C), as in the Pacific Northwest and the semiarid areas between the Cascade and Rocky mountains, ordinary city water provides radiant comfort cooling. Where higher dew points prevail, it is necessary to dehumidify the ambient air. Panel cooling effectively prevents the disagreeable feeling of cold air blown against the body and minimizes the occurrence of summer colds. *See* DEHUMIDIFIER.

Fuel consumption records show that panel heating systems save 30–50% of the fuel costs of ordinary heating systems. Lower ambient air temperatures produce a comfortable environment, and air temperatures within the room are practically uniform and not considerably higher at the ceiling, as in radiator- and convector-heated interiors. *See* COMFORT HEATING; HOT-WATER HEATING SYSTEM.

Erwin L. Weber; Richard Koral

**Bibliography.** American Society of Heating, Refrigerating, and Air Conditioning Engineers, *ASHRAE Handbook and Product Directory: Systems*, 1992.

### Pantodonta

An extinct order of relatively large placental mammals represented by Paleocene-Eocene fossils from western Europe, North America, and eastern Asia. Pantodonts were an early evolutionary experiment in large-bodied herbivory by primitive placental mammals. They first appeared in Asia during the early Paleocene and disappeared during the middle Eocene, leaving no descendants. With the possible exception of the most primitive pantodonts, all were herbivores, and pantodonts were either the largest or among the largest mammals of their time. The adaptive radiation of pantodonts was diverse and encompassed mammals as different as small [1 kg (2 lb) or less in body mass], arboreal herbivores (Asian *Archaeolambda*), and large [650 kg (1430 lb)], ground-sloth-like, terrestrial herbivores (North American *Barylambda*). *See* EOCENE; PALEOCENE.

Pantodonts are unique among placental mammals in having upper third and fourth premolar tooth crowns in which there are V-shaped crests. The lower premolars and molars had broad, transverse crests, and this feature suggests that pantodonts sheared and pulped plant matter as their primary, or only, source of food. All pantodonts were obligate quadrupeds, with four nearly equal-sized limbs. Their wrists and ankles were generally short and massive, and each hand or foot had five toes that bore small hooves, except in a few pantodont genera that had claws. Pantodont skulls were generally small relative to body size and had long, boxlike snouts, small braincases, and small eyes. Most pantodonts had large tusks and long, heavy tails. When compared to living mammals, many pantodonts would have looked somewhat like bears, pigs, or small hippos. *See* DENTITION.

The most primitive pantodonts are assigned to the family Bemalambdidae. All other pantodonts belong to the suborder Eupantodonta, which diverged early into the superfamilies Pantolambdodontoidea

and Pantolambdaidea. The pantolambodontoids include the Asian family Pantolambodontidae, the North American family Titanoideidae, and the South American genus *Alcidedorbigniya*. The pantolambodonts include the North American families Pantolambidae and Barylambdidae and the more cosmopolitan (known from North America and Eurasia) family Coryphodontidae.

The fossil evidence suggests that pantodonts originated in Asia from an insectivorous ancestry during the earliest Paleocene. Thus, the Chinese early Paleocene bemalambid genera *Bemalambda* and *Hypsilolambda* are the oldest and the most primitive pantodonts. Unlike later pantodonts, their sharp and pointed tooth crowns indicate that they were insectivores or perhaps omnivores.

From their Asian origin, pantodonts rapidly emigrated into North and South America during the early Paleocene. By the late Paleocene, pantolambodontoids were present in Asia (*Harpyodus*, *Pastoralodon*, *Archaeolambda*, *Altlambda*, and *Pantolambodon*), North America (*Titanoides*), and South America (*Alcidedorbigniya*). The Asian (and South American) evolution of pantolambodontoids encompassed small mammals [1 kg (2 lb) or less body mass]. Assuming that the skeleton of Chinese late Paleocene *Archaeolambda tabiensis* is characteristic, Asian pantolambodontoids were lightly built, arboreal herbivores. By contrast, the North American pantolambodontoid *Titanoides* was a large [150 kg (330 lb)], terrestrial, clawed quadruped with saber-like upper canines. This animal may have done some digging to obtain its food.

The diversification of the remaining North American pantodonts began during the early Paleocene with *Pantolambda*, a terrestrial quadrupedal browser. Late Paleocene *Caenolambda* was a somewhat unusual (notably, the bladelike first premolar), though functionally similar close relative. These two genera constitute the family Pantolambidae.

The late Paleocene barylambdid pantodonts were confused early with Asian pantolambodontids because both families converged on some dental features, especially in their small tusks and incisors. However, the skeletons of the pantodonts in these families are very different. *Barylamba* is the best-known barylambdid and was functionally somewhat analogous to the large ground sloths of the Pleistocene. The graviportal pelvis, massive hindlimbs, and heavy tail suggest that *Barylamba* was capable of bipedal browsing. *Leptolambda* shows some of these modifications as well, but the two smaller barylambdids, *Haplolambda* and *Ignatiolambda*, are too poorly known postcranially to allow firm conclusions on their mode of life.

From the standpoint of abundance of fossils, longevity, diversity, and breadth of geographic distribution, *Coryphodon* was the most successful pantodont. It was the largest land mammal of the latest Paleocene–early Eocene, and its fossils come from the western United States, Ellesmere Island in Arctic Canada, western Europe, Mongolia, and China. *Coryphodon* was a terrestrial, graviportal, subdigitigrade quadruped with an adult body weight of 150–300 kg (330–660 lb). Its teeth somewhat resemble those of a living tapir, suggesting that *Coryphodon* was a browser. There is no compelling evidence that it was amphibious other than the abundance of its fossils in some fluvial and lacustrine deposits. The skeleton of *Coryphodon* most closely resembles that of the living pygmy hippopotamus *Hexaprotodon*.

The extinction of *Coryphodon* at the end of the early Eocene marked the disappearance of pantodonts in North America and Europe. But in Asia, other coryphodontid genera (*Eudinoceras*, *Hypercoryphodon*) persisted through the middle Eocene. Their extinction ultimately was due to changes in climate and vegetation that promoted the rise of early tapirs and rhinoceroses. See MAMMALIA.

Spencer G. Lucas

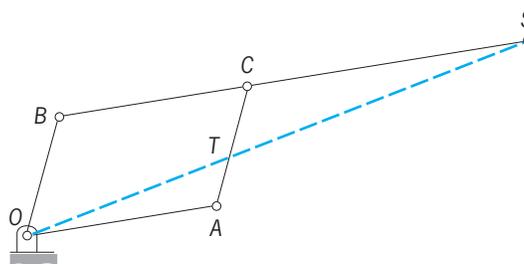
**Bibliography.** M. J. Benton, *Vertebrate Palaeontology*, 1990; R. L. Carroll, *Vertebrate Paleontology and Evolution*, 1988; C. M. Janis, K. M. Scott, and L. L. Jacobs, *Evolution of Tertiary Mammals of North America*, 1998; R. J. G. Savage and M. R. Long, *Mammal Evolution: An Illustrated Guide*, 1986.

## Pantograph

A four-bar parallel linkage, with no links fixed, used as a copying device for generating geometrically similar figures, larger or smaller in size, within the limits of the mechanism. In the **illustration** the curve traced by point *T* will be similar to that generated by point *S*. This similarity results because points *T* and *S* will always lie on the straight line  $\overline{OTS}$ ; triangles  $OBS$  and  $TCS$  are always similar because lengths  $\overline{OB}$ ,  $\overline{BS}$ ,  $\overline{CT}$ , and  $\overline{CS}$  are constant and  $\overline{OB}$  is always parallel to  $\overline{CT}$ . Distance  $\overline{OT}$  always maintains a constant proportion to distance  $\overline{OS}$  because of the similarity of the above triangles. Numerous modifications of the pantograph as a copying device have been made. See FOUR-BAR LINKAGE.

James Watt applied the pantograph as a reducing motion in his beam engine. The pantograph has also served as a reducing motion for an engine indicator: *S* is attached to the engine crosshead, while the indicator cord is attached to *T*. See STRAIGHT-LINE MECHANISM.

A second use of the pantograph geometry is seen in the collapsible parallel linkage used on electric locomotives and rail cars to keep a current-collector bar or wheel in contact with an overhead wire. Two such congruent linkages in planes parallel to the



Similar triangles of a pantograph.

train's motion are affixed securely on the top of the locomotive with joining horizontal members perpendicular to each other. The uppermost member collects the current, and powerful springs thrust the configuration upward with sufficient pressure normally to make low-resistance contact from wire to collector. The inevitable variation of distance from wire to track produces, during high speeds, undesirable dynamic behavior of both wire and collector-pantograph, with oscillations of both that can break the connection between them or cause excessive arcing and wear with attendant loss of transmission. Stitching the wire to a stouter supporting wire can minimize the bounce of both collector and wire and assure more reliable current flow. Douglas P. Adams

Bibliography. N. P. Chironis, *Mechanisms and Mechanical Devices Sourcebook*, 3d ed., 2001.

## Papaverales

An order of flowering plants, division Magnoliophyta (Angiospermae), subclass Magnoliidae of the class Magnoliopsida (dicotyledons). The order consists of only two families: Papaveraceae with some 200 species, and Fumariaceae with about 400 species. Within its subclass, the order is marked by its syncarpous gynoecium, parietal placentation, and only two (seldom three) sepals. Most of the species are herbaceous, and many of them contain isoquinoline alkaloids similar to those in the order Ranunculales. The Papaveraceae, with regular flowers, numerous stamens, and a well-developed latex system, include the poppies (*Papaver* and related genera; see **illus.**), bloodroot (*Sanguinaria*), and celandine (*Chelidonium*). *Papaver somniferum* is the source of opium. The Fumariaceae, with four or six stamens, irregular flowers that usually have some of the petals spurred or saccate, and no latex system, include the bleeding heart (*Dicentra spectabilis*) and some other common ornamentals.

The Papaverales have often been grouped with the Capparales into an expanded order, Rhoeadales, but more recently it has been agreed that the two groups are not closely allied and that the affinities of the Papaverales are with the families of the Ranunculales. See MAGNOLIIDAE; MAGNOLIOPHYTA;



Oriental poppy (*Papaver orientale*) of the family Papaveraceae and the order Papaverales. (John H. Gerard, National Audubon Society)

MAGNOLIOPSIDA; PLANT KINGDOM; POPPY; RANUNCULALES.  
Arthur Cronquist; T. M. Barkley

## Paper

A flexible web or mat of fibers isolated from wood or other plants materials by the operation of pulping. Nonwovens are webs or mats made from synthetic polymers, such as high-strength polyethylene fibers, that substitute for paper in large envelopes and tote bags.

Paper is made with additives to control the process and modify the properties of the final product. The fibers may be whitened by bleaching, and the fibers are prepared for papermaking by the process of refining. Stock preparation involves removal of dirt from the fiber slurry and mixing of various additives to the pulp prior to papermaking. Papermaking is accomplished by applying a dilute slurry of fibers in water to a continuous wire or screen; the rest of the machine removes water from the fiber mat. The steps can be demonstrated by laboratory handsheet making, which is used for process control.

Although paper has numerous specialized uses in products as diverse as cigarettes, capacitors, and counter tops (resin-impregnated laminates), it is principally used in packaging (~50%), printing (~40%), and sanitary (~7%) applications. Paper was manufactured entirely by hand before the development of the continuous paper machine in 1804; this development allowed the United States production of paper to increase by a factor of 10 during the nineteenth century and by another factor of 50 during the twentieth century. In 1960, the global production of paper was 70 million tons (50% by the United States); in 1996 it was 240 million tons (30% by the United States). The annual per capita paper use in the United States is 300 kg (660 lb), and about 45% is recovered for reuse.

Material of basis weight greater than 200 g/m<sup>2</sup> is classified as paperboard, while lighter material is called paper. Production by weight is about equal for these two classes. Paperboard is used in corrugated boxes; corrugated material consists of top and bottom layers of paperboard called linerboard, separated by fluted corrugating paper. Paperboard also includes chipboard (a solid material used in many cold-cereal boxes, shoe boxes, and the backs of paper tablets) and food containers.

Mechanical pulp is used in newsprint, catalog, and other short-lived papers; they are only moderately white, and yellow quickly with age because the lignin is not removed. A mild bleaching treatment (called brightening) with hydrogen peroxide or sodium dithionite (or both) masks some of the color of the lignin without lignin removal. Paper made with mechanical pulp and coated with clay to improve brightness and gloss is used in 70% of magazines and catalogs, and in some enamel grades. Bleached chemical pulps are used in higher grades of printing papers used for xerography, typing paper, tablets, and envelopes; these papers are termed uncoated wood-free (meaning free of mechanical

pulp). Coated wood-free papers are of high to very high grade and are used in applications such as high-quality magazines and annual reports; they are coated with calcium carbonate, clay, or titanium dioxide.

Like wood, paper is a hygroscopic material; that is, it absorbs water from, and also releases water into, the air. It has an equilibrium moisture content of about 7–9% at room temperature and 50% relative humidity. In low humidities, paper is brittle; in high humidities, it has poor strength properties.

**Wood.** Wood, a diverse, variable material, is the source of about 90% of the plant fiber used globally to make paper. Straw, grasses, canes, bast, seed hairs, and reeds are used to make pulp, and in many regards their pulp is similar to wood pulp. Fibers are tubular elements of plants and contain cellulose as the principal constituent. Softwoods (gymnosperms) have fibers that are about 0.12–0.2 in. (3–5 mm) long, while in hardwoods (angiosperms) they are about 0.03–0.06 in. (0.8–1.6 mm) long. In both cases, the length is typically about 100 times the width. Softwoods are used in papers such as linerboard where strength is the principal intent. Hardwoods are used in papers such as tissue and printing to contribute to smoothness. Many papers also include some softwood pulp for strength and some hardwood pulp for smoothness.

Wood consists of three major components: cellulose, hemicellulose, and lignin. The first two are white polysaccharides of high molecular weight that are desirable in paper. Cotton is over 98% cellulose, while wood is about 45%. Hemicellulose, although not water soluble, is similar to starch. The hydroxyl groups of these materials allow fibers to be held together in paper by hydrogen bonding. Adhesives are not required to form paper, but some starch is usually used and has a similar effect to the hemicelluloses in helping the fibers bond together. Lignin makes up about 25–35% of softwoods and 18–25% of hardwoods. It is concentrated between fibers. In mechanical pulping, the lignin is made relatively soft so the fibers can be separated by mechanical action, but the lignin is not removed. In chemical pulping, the lignin is dissolved to separate the fibers. Lignin is a tan material in wood that turns dark brown during chemical pulping. For example, brown paper bags used in grocery stores contain about 10% lignin. Lignin impedes the fiber-to-fiber bonding; therefore, mechanical pulps form much weaker papers than chemical pulps. *See* CELLULOSE; HEMICELLULOSE; LIGNIN.

Extractives are a minor component of wood (1–8%) and comprise relatively small molecules that can be easily removed in the laboratory by using various liquids. Many wood extractives belong to the large class of materials called terpenes. Turpentine (which can be manufactured into pine oil) consists of volatile monoterpenes. This fraction is recovered as the wood chips are heated prior to chemical pulping. Rosin acids and fatty acids are recovered in the kraft recovery process. *See* PINE TERPENE; ROSIN; TALL OIL; TERPENE; WOOD CHEMICALS.

**Pulping.** The fibers of wood or other materials are separated by mechanical action, chemical ac-

tion, or a chemical pretreatment followed by mechanical action; these processes are termed mechanical, chemical, and semichemical pulping, respectively. In fact, there are many variations in these processes. However, thermomechanical pulping and kraft chemical pulping account for over 80% of all pulp.

*Mechanical pulping.* Previously, cotton was the most important source of papermaking fiber because the fibers were already separated from each other, so they did not require pulping, and they have a high cellulose content. Wood bolts were ground against stones in the stone groundwood process, just as flour is made from wheat or corn. This pulp was very weak but was useful as an extender for cotton. As the process improved, the fiber was used in larger and larger amounts. The original grinding stone was sandstone, but it was replaced by synthetic stones embedded with aluminum oxide or silicone carbide.

Stone groundwood pulp has been supplanted by thermomechanical pulp, which has a much higher strength. In this process, wood chips are pulped in refiners with large metal disks that revolve in opposite directions. These disks have bars on them to allow the pulping process to occur. The process is carried out in two steps: a pressurized refiner operating at 110–130°C (230–265°F) softens the lignin and separates the fibers; an atmospheric pressure refining stage refines the pulp. The chips are introduced in the center of the disk, and centrifugal force pushes the pulp outward as it is formed. Energy requirements are 1900–2900 kW-h/ton. The use of a mild chemical pretreatment (using sodium sulfite or hydrogen peroxide under alkaline conditions) prior to thermomechanical pulping is called chemithermomechanical pulping. Nearly 20% of all pulp is made by the thermomechanical pulping process. Hydrogen peroxide is used to brighten mechanical pulp.

*Chemical pulping.* Kraft pulping (invented in 1879) is the most widely used chemical pulping process because kraft pulp is stronger than any other type. The process is also called the sulfate process because sodium sulfate can be used as a make-up chemical (to replace a small amount of chemical that leaves with the pulp), but sulfate is not involved during the actual pulping process. The white liquor used to cook the wood chips is a highly alkaline aqueous solution of sodium hydroxide and sodium sulfide. The wood is cooked at 160–180°C (320–356°F) in sealed digesters at a pressure of about 800 kilopascals for 0.5–2.0 h. Continuous digesters that may be several hundred feet tall, where wood chips continuously enter the top and pulp continuously exits the bottom, are used to produce most kraft pulp, but batch digesters and other designs are used too. About 65% of all pulp is produced by the kraft process. The pulp yield is only 45–55%.

Kraft pulping must be practiced with a chemical recovery process, which is much more complicated than the pulping process. The recovery process allows the sodium hydroxide and sodium sulfide to be reused, burns the materials dissolved from the wood so they are not a disposal problem, and

creates steam to drive the various processes within the mill. The spent liquor (called black liquor) is washed from the brown pulp (the brownstock). The black liquor is concentrated and burned in a boiler; water-filled tubes inside the boiler capture the heat and form steam. The inorganic chemicals are recovered as smelt (sodium carbonate and sodium sulfide) from the bottom of the boiler; the smelt is dissolved in water to form green liquor; the green liquor is treated with calcium hydroxide to convert the sodium carbonate of the green liquor to sodium hydroxide, forming the white liquor; and the calcium carbonate is washed and burned in a lime kiln to regenerate calcium oxide, which forms calcium hydroxide when it is dissolved in water. Tall oil is recovered from the concentrated black liquor by skimming before it is burned. *See* TALL OIL.

Kraft pulp mills have an odor somewhat like rotten eggs. A few parts per billion of mercaptans in the atmosphere cause this smell. Kraft mills have made much progress in recent years in the reduction of air emissions. Before 1930, the acid sulfite process was the main chemical pulping process. Since chemical recovery was not possible, it was modified in the 1950s as a magnesium or sodium process. However, the acid weakens cellulose and results in a pulp of only moderate strength. Therefore, only smaller, older mills use this method. Sulfite pulp can be used as a source of high-purity cellulose for conversion into cellophane, cellulose acetate, rayon, cellulose nitrate smokeless powder, and other materials. *See* CELLOPHANE; MERCAPTAN.

**Bleaching.** Chemical pulp for printing paper has 3–6% lignin, which gives the pulp a brown color. This lignin is removed with bleaching chemicals in four to eight stages. Each stage consists of a pump to mix the bleaching agent with the pulp, a retention tower to allow the chemical to react with the pulp for 30 min to several hours, and a washing unit to remove the solubilized lignin and residual chemicals from the pulp. Usually an oxidizing material is followed by a stage of alkali extraction, since lignin becomes more soluble at high pH. A common bleaching sequence involves elemental oxygen, alkali extraction, chlorine dioxide, alkali extraction, and, finally, chlorine dioxide.

**Refining, stock preparation, and additives.** Chemical pulp fibers are refined prior to papermaking to increase their flexibility and surface area, both of which promote fiber-to-fiber bonding. This is done with a double disk refiner, similar to those used in thermomechanical pulping, except the refining is done at ambient temperature with a finer bar pattern.

Dyes, retention aids, drainage aids, wet-strength adhesives (for paper towels that must hold their strength when wet), biocides, defoamers, and other materials may be added. Fillers are a major additive in printing papers and include clay, precipitated calcium carbonate, and titanium dioxide. These materials add to the brightness of the paper and improve paper opacity, which allows reading one side of the paper without having the print on the other side

showing through. Titanium dioxide is especially useful, but its high expense limits it to thin (bible) papers.

Sizing, that is, the ability to resist rapid water absorption, is important to printing paper; otherwise ink spreads rapidly. Internal sizing has been carried out since the early nineteenth century by adding small amounts of rosin and aluminum sulfate to the dilute slurry of fibers at pH 4.5–5. This makes the paper sufficiently acidic to cause the cellulose to lose its strength over a period of decades. Most high-quality printing papers now use calcium carbonate as filler, which requires the slurry to be at neutral or slightly alkaline pH. Sizing agents that work at high pH are also used, and these papers do not deteriorate appreciably with time.

**Manufacture.** Paper must be made from a dilute slurry of fibers and additives. This slurry travels through screens and cleaners to remove debris prior to papermaking. There is 1 part fiber for every 200 parts water. This low consistency (solids content) is required to make a sheet with good formation, that is, a sheet of even density in all parts of the paper. The purpose of the paper machine is to remove all of this water.

Paper machines are as wide as 33 ft (10 m) and may operate at speeds over 60 mi/h (25 m/s). The dilute slurry of pulp is applied to a continuously moving plastic screen by the headbox. Water drains out by gravity, and then suction devices of increasing strength pull more water out. When the consistency is at 20%, the web is strong enough to be transferred to a felt material. The felt-supported web of fibers travels between three press nips formed where press rolls come together (as in old wringer washing machines) to press additional water from the web into the felt. The consistency at this point is 40–50%. Most of the remaining water is removed by drying the paper against a series of 20–60 steam heated dryer cans that are 6 ft (2 m) in diameter. Surface sizing with a starch or synthetic polymer solution may be accomplished between dryer sections.

The heaviest grades of papers, such as chipboard, are made on multiformer (cylinder) machines that form three to eight layers of fiber mats. These fiber mats are combined prior to pressing and drying. The lightest grades of paper, tissues, cannot withstand numerous felt transfers and are dried on very large Yankee dryers.

The paper may be smoothed against a series of rolls made from metal or rubbery material to impart smoothness or gloss to the paper. The paper may also be coated with a paintlike material to give it high brightness and gloss. In addition, numerous other converting operations may be performed on paper. Christopher J. Biermann

**Bibliography.** C. J. Biermann, *Handbook of Pulp and Papermaking*, 1996; J. E. Kline, *Paper and Paperboard: Manufacturing and Converting Fundamentals*, 2d ed., 1991; E. Sjöström, *Wood Chemistry: Fundamentals and Applications*, 2d ed., 1993; D. C. Smith, *History of Papermaking in the United States: 1691–1969*, 1970.

---

**Paprika**

A type of pepper, *Capsicum annuum*, with nonpungent flesh, grown for its long red fruit. A member of the plant order Polemoniales and of American origin, it is most popular in Hungary and adjacent countries.

Seeds are removed from the mature fruit, and the flesh is dried and ground to prepare the dry condiment commonly referred to as paprika. Production in the United States is limited, with California the only important producing state. *See* PEPPER; SOLANALES.

H. John Carew