



McGRAW-HILL
ENCYCLOPEDIA OF
SCIENCE &
TECHNOLOGY

www.MHEST.com

11 **METE-NIT**



Meteor — Myzostomida

Meteor

The luminous streak lasting seconds or fractions of a second and seen at night when a solid, natural body plunges into the Earth's (or another planet's) atmosphere. The entering object is called a meteoroid and, if any of it survives atmospheric passage, the remainder is called a meteorite. Cosmic dust particles (with masses of micrograms) entering the atmosphere and leaving very brief, faint trails are called micrometeors, with the surviving pieces known as micrometeorites. If the apparent brightness of a meteor exceeds that of the planet Venus as seen from Earth, it is called a fireball; and when a bright meteor is seen to explode, it is called a bolide. *See* METEORITE; MICROMETEORITE.

Visual observation. Under normal, clear atmospheric conditions and dark skies (no moonlight or artificial lights), an observer will see an average of five meteors per hour. The spatial distribution of meteoroid orbits relative to the Sun, and the circumstances of their intersections with the moving Earth are responsible for pronounced variations in meteor rates.

As the Earth moves in its orbit, its velocity points toward that part of the sky (sometimes called the apex of meteor velocities) which is visible from local midnight through morning to local noon, and points away from that part of the sky which is visible from local noon through evening to local midnight. On average, an observer sees more meteors during the early morning hours as the Earth sweeps up objects in its path than in the early evening hours when meteor-producing objects must catch up with the Earth.

Physical characteristics. The heights of appearance and disappearance of a meteor depend on meteoroid initial velocity, angle of entry with respect to the vertical, initial meteoroid mass, and meteoroid material strength. The average meteor seen by the unaided eye starts with a meteoroid velocity of 18 mi/s (30 km/s) and leaves a luminous trail from 67 to

50 mi (110 to 80 km) high. The fainter is the meteor (the smaller the meteoroid mass), the shorter is the meteor length. The faster the meteoroid travels just before hitting the atmosphere, the higher in the atmosphere the meteor trail occurs.

In the end, most, if not all, of the meteoroid material is vaporized, leaving a deposit of metallic atoms (predominantly sodium, calcium, silicon, and iron) in the upper atmosphere. This deposition is an important mechanism in the wind-shear formation of certain types of highly ionized, radio-reflecting, upper-atmospheric phenomena called the sporadic E layers. *See* IONOSPHERE.

The meteor trails themselves are rapidly expanding columns of atoms, ions, and electrons dislodged from the meteoroid by collisions with air molecules, and can be excited to temperatures of several thousand degrees Celsius. For a time after trail formation, the free electrons are dense enough to reflect radio waves in the very high frequency range, and therefore can be used to transmit radio messages that are brief (0.1–15 s) but high in information content (on account of their large bandwidth) for up to 1300 mi (2200 km). Since meteor-reflected signals are not as subject to ionospheric and other disturbing influences as are other means of radio communications, there has been interest in using meteors for certain military and commercial purposes. *See* RADIO-WAVE PROPAGATION.

Under the right circumstances, particularly with high-power ultrahigh-frequency (UHF) radars, the ionization right around and moving with the meteoroid itself is seen. This is known as the head echo, and a determination of its velocity is the most accurate way to determine radar meteor speeds. If the radar beam is very narrow, such as that at the Arecibo National Observatory, the radiant and the orbit around the Sun can also be accurately determined. *See* RADAR; RADAR ASTRONOMY.

Some meteors show relatively long-lasting glows along their paths. The glows are called meteor trains.

Novice observers often confuse this glow with the brief meteor wake that is sometimes seen just behind the luminous gases surrounding the meteoroid itself. These trains are subject to winds in the upper atmosphere, and for many years their apparent motions (twisting and drifting) were the only probes of wind velocities in the mesosphere, or middle atmosphere (42–72 mi or 70–120 km high). The motions of meteor trains and trails detected through the Doppler effect on reflected radar signals have been extensively studied, and global mesospheric wind patterns have been derived. *See* DOPPLER EFFECT; MESOSPHERE.

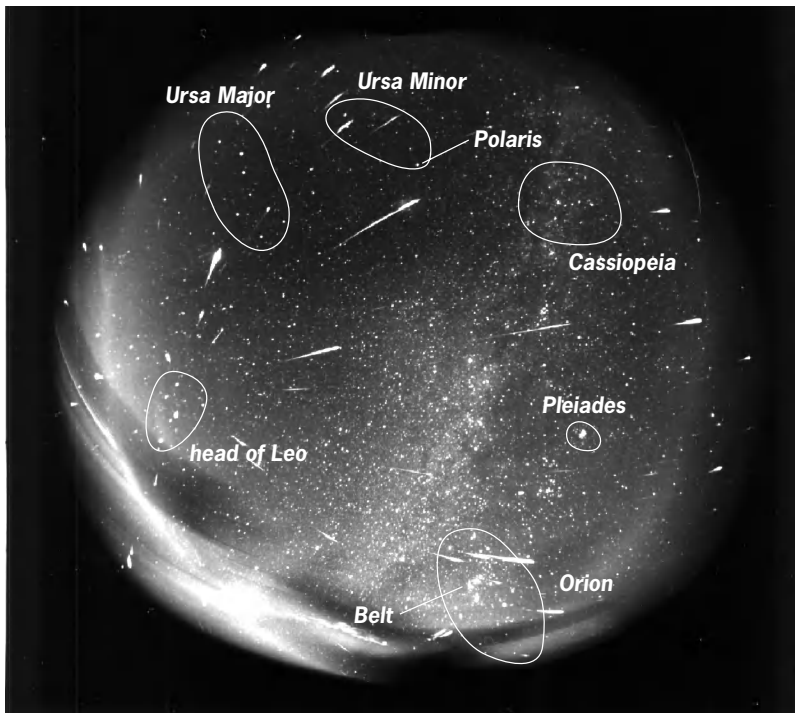
Meteoroid orbits and velocities. The Earth moves around the Sun with an average speed of 18 mi/s (30 km/s). According to the laws of celestial mechanics, if a meteoroid comes from beyond the solar system, its velocity at the Earth's distance from the Sun must be greater than 26 mi/s (42 km/s). If such a meteoroid hits the Earth head-on, indications of preatmospheric speeds in excess of 45 mi/s (72 km/s) would be observed. The fact that the vast majority of observed meteoroids have orbits with Earth-approaching velocities of less than 45 mi/s indicates that most of these are comet and asteroid fragments, and are therefore long-term members of the solar system. Early velocity measurements were quite crude, and the large errors led to the belief that a sizable fraction of meteoroids actually had speeds great enough to escape the solar system or come from the outside. Once these errors were reduced, the number of hypervelocity orbits decreased dramatically, with doubts expressed that any meteoroids could come from beyond the solar system.

However, in the 1980s and 1990s, a combination of spacecraft and high-power radar observations indicated that hypervelocity micrometeoroids do indeed exist with seeming interstellar dust connections. An even more curious connection is the discovery of similarly sized particles with peculiar isotopic abundances within meteorites of presumed asteroidal origin. These micrometeorites appear to predate the solar system with an apparent origin in early supernovae, novae, or supergiant stars. *See* COSMOCHEMISTRY; INTERSTELLAR MATTER.

The slowest velocity of Earth-meteoroid encounter occurs when the meteoroid has to catch up with the Earth. The gravitational attraction of the Earth keeps such an encounter from producing a zero atmospheric velocity. The velocity that an object will achieve by falling toward the Earth from an infinite distance and in the absence of the Sun's gravitational attraction is 7 mi/s (11 km/s). High-power ultrahigh-frequency radars detect a considerable number of micrometeors with velocities well below the Earth-attraction limit. These are either artificial satellite debris or natural debris captured by the Earth from a large, co-moving interplanetary dust zone that has been discovered in the Earth's orbit. *See* CELESTIAL MECHANICS; ESCAPE VELOCITY; INTERPLANETARY MATTER; ORBITAL MOTION.

Radiants and showers. A combination of the meteoroid's and Earth's velocities of travel around the Sun makes the meteor itself seem to originate from a specific direction in the sky called the radiant. If there are numerous meteoroids in nearly the same orbit (sometimes incorrectly called meteor streams), the Earth sweeps them up at specific times of the year and a so-called meteor shower is observed. Meteor showers are named after the constellation or single star in the sky from which they appear to radiate. While shower meteoroids are really moving nearly parallel through space and result in nearly parallel meteor trails, the effects of perspective make the meteors appear to diverge from the radiant (*see illustration*). If the meteor shower is particularly long-lasting, the radiant will appear to drift slowly in position from night to night as the relative directions of the Earth-meteoroid velocities change. Radiants are not geometric points, but areas in the sky that can be several degrees in diameter. Hence, quoted radiant positions are averages over such areas and often differ somewhat from one compilation to another. Meteors that cannot be shown to be associated with a known shower are termed sporadic meteors.

While all meteor showers show both day-to-day and year-to-year variations, some are more reliable than others, and these are called annual or major showers (**Table 1**). There are over 100 minor showers, some of which are of special interest (**Table 2**). Many minor showers have hourly rates so low that they are often not noticed except by experienced observers or when there is an unusual burst of activity. Some showers have been discovered by radio or radar methods to occur only during the hours of daylight, with no visible counterpart seen at night. Most of the radiants listed refer to geocentric radiants where the effects of the Earth's gravitational



Leonid meteor shower observed at the Modra Astronomical Observatory in Slovakia through a fisheye lens. The exposure on November 16–17, 1998, lasted 4 h. About 150 bolides brighter than magnitude -2 can be seen; the brightest fireball is about magnitude -8 . Zenithal hourly rate of the shower estimated from visual observations was about 400. Positions of some well-known constellations and features are indicated. Meteors appear to radiate from the head of Leo. (*Modra Astronomical Observatory*)

TABLE 1. Major meteor showers

| Shower | Duration | Approx. date of maximum | Approx. radiant coordinates, degrees* | | Meteoroid orbital speed | | Strength† | Suggested parent body | Notes |
|----------------|------------------|-------------------------|---------------------------------------|-------------|-------------------------|------|-----------|---|------------------------------|
| | | | Right ascension | Declination | mi/s | km/s | | | |
| Quadrantids | Jan. 1–6 | Jan. 3 | 230 | +49 | 27 | 42 | M | 2003 EH 1 = extinct (?) nucleus of comet C1490 Y1 (?) | Sharp maximum |
| Lyrids | Apr. 20–23 | Apr. 22 | 271 | +34 | 30 | 48 | M-W | 1861 I | Good in 1982 |
| π Puppids | Apr. 16–25 | Apr. 23 | 110 | -45 | — | — | W-M | Grigg-Skjellerup | Highly variable |
| η Aquarids | Apr. 21–May 12 | May 4 | 336 | -2 | 40 | 64 | M | Halley | Second peak May 6 |
| Arietids | May 29–June 19 | June 7 | 44 | +23 | 24 | 39 | S | — | Daytime shower |
| ζ Perseids | June 1–17 | June 7 | 62 | +23 | 18 | 29 | S | — | Daytime shower |
| β Taurids | June 24–July 6 | June 29 | 86 | +19 | 20 | 32 | S | Encke | Daytime shower |
| S. δ Aquarids | July 21–Aug. 25 | July 30 | 333 | -16 | 27 | 43 | M | — | Primary radiant |
| S. ο Aquarids | July 15–Aug. 25 | Aug. 6 | 333 | -15 | 19 | 31 | W | — | Primary radiant |
| Perseids | July 23–Aug. 23 | Aug. 12 | 46 | +57 | 37 | 60 | S | 1862 III | Best-known shower |
| Orionids | Oct. 2–Nov. 7 | Oct. 21 | 94 | +16 | 41 | 66 | M | Halley | Trains common |
| S. Taurids | Sept. 15–Nov. 26 | Nov. 3 | 50 | +14 | 17 | 27 | M | Encke | Known fireball producer |
| Leonids | Nov. 14–20 | Nov. 17 | 152 | +22 | 45 | 72 | W-S | 1866 I | Many peaks seen 1996 to 2003 |
| Puppids-Velids | Nov. 27–Jan. | Dec. 9 | 135 | -48 | — | — | M | 2102 Tantalus? | Many radiants in region |
| Geminids | Dec. 4–16 | Dec. 13 | 112 | +32 | 23 | 36 | S | 3200 Phaethon | Many bright meteors |

* When the stream velocity is unknown or poorly known, the radiant coordinates are those of the apparent radiant rather than the geocentric radiant.

† Estimate of relative meteor hourly rate for visual observers: S = strong (sometimes above 30 per hour at peak); M = moderate (10 to 30 per hour at peak); W = weak (5 to 10 per hour at peak).

attraction have been removed. Radiant positions that have not been corrected for the Earth's attraction (usually because the meteoroid velocities are unknown) are called apparent radiants.

Stationary meteors. Occasionally, a meteor is seen coming directly toward an observer and shows a bright point of light rather than a trail. Such me-

teors are called, somewhat inaccurately, stationary meteors. The positions of these head-on meteors define the radiant precisely and are quite important to observe, particularly during showers. However, short glints of light from Earth-orbiting spacecraft and space debris are sometimes mistaken for stationary meteors. It has been suggested that a certain

TABLE 2. Minor meteor showers*

| Shower | Duration | Approx. date of maximum | Approx. radiant coordinates, degrees† | | Meteoroid orbital speed | | Suggested parent body | Notes |
|-------------------|------------------|-------------------------|---------------------------------------|-------------|-------------------------|------|-----------------------|----------------------------|
| | | | Right ascension | Declination | mi/s | km/s | | |
| Coma Berenicids | Dec. 12–Jan. 23 | Jan. 17 | 186 | +20 | 40 | 65 | 1913 I | Uncertain radiant position |
| α Centaurids | Jan. 28–Feb. 23 | Feb. 8 | 209 | -59 | — | — | — | Colors in bright meteors |
| δ Leonids | Feb. 5–Mar. 19 | Feb. 26 | 159 | +19 | 14 | 23 | — | Slow, bright meteors |
| Virginids | Feb. 3–Apr. 15 | Mar. 13? | 186 | +00 | 21 | 35 | — | Other radiants in region |
| δ Normids | Feb. 25–Mar. 22 | Mar. 14 | 245 | -49 | — | — | — | Sharp maximum |
| δ Pavonids | Mar. 11–Apr. 16 | Apr. 6 | 305 | -63 | — | — | Grigg-Mellish | Rich in bright meteors |
| σ Leonids | Mar. 21–May 13 | Apr. 17 | 195 | -05 | 12 | 20 | — | Slow, bright meteors |
| α Scorpids | Apr. 11–May 12 | May 3 | 240 | -22 | 21 | 35 | — | Other radiants in region |
| τ Herculids | May 19–June 14 | June 3 | 228 | +39 | 9 | 15 | — | Very slow meteors |
| Ophiuchids | May 19–July | June 10 | 270 | -23 | — | — | — | One of many in region |
| Corvids | June 25–30 | June 26 | 192 | -19 | 6 | 11 | — | Very low speed |
| June Draconids | June 5–July 19? | June 28 | 219 | +49 | 8 | 14 | Pons-Winnecke | Maximum only 1916 |
| Capricornids | July–Aug. | July 8 | 311 | -15 | — | — | — | May be multiple |
| Piscis-Australids | July 15–Aug. 20 | July 31 | 340 | -30 | — | — | — | Poorly known |
| α Capricornids | July 15–Aug. 25 | Aug. 2 | 307 | -10 | 14 | 23 | 1948 XII (1948n) | Bright meteors |
| N. α Aquarids | July 14–Aug. 25 | Aug. 12 | 327 | -06 | 26 | 42 | — | Secondary radiant |
| κ Cygnids | Aug. 9–Oct. 6 | Aug. 18 | 286 | +59 | 15 | 25 | — | Bursts of activity |
| N. Aquarids | July 15–Sept. 20 | Aug. 20 | 327 | -06 | 19 | 31 | — | Secondary radiant |
| S. Piscids | Aug. 31–Nov. 2 | Sept. 20 | 6 | +00 | 16 | 26 | — | Primary radiant |
| Andromedids | Sept. 25–Nov. 12 | Oct. 3 | 20 | +34 | 11 | 18 | — | "Annual" version |
| October Draconids | Oct. 10 | Oct. 10 | 262 | +54 | 14 | 23 | Giacobini-Zinner | Can be spectacular |
| N. Piscids | Sept. 25–Oct. 19 | Oct. 12 | 26 | +14 | 18 | 29 | — | Secondary radiant |
| Leo Minorids | Oct. 22–24 | Oct. 24 | 162 | +37 | 38 | 62 | 1739 | Probable comet association |
| μ Pegasids | Oct. 29–Nov. 12 | Nov. 12 | 335 | +21 | 7 | 11 | 1819 IV | Probable comet association |
| Andromedids | Nov. 25 | Nov. 25 | 25 | +44 | 10 | 17 | Biela | Once only, 1885 |
| Phoenicids | Dec. 5 | Dec. 5 | 15 | -50 | — | — | — | Once only, 1956 |
| Ursids | Dec. 17–24 | Dec. 22 | 217 | +76 | 20 | 33 | — | Good in 1986 |

* Peak strength for visual observers usually less than 5 per hour. Only showers of special interest are listed here.

† When the stream velocity is unknown or poorly known, the radiant coordinates are those of the apparent radiant rather than the geocentric radiant.

number of visible light pulses from cosmic x-ray and gamma-ray sources are also mistaken for stationary meteors. Such cosmic-ray events are, however, far too brief to be seen with conventional meteor equipment or by the unaided eye. *See* GAMMA-RAY ASTRONOMY; X-RAY ASTRONOMY.

Bright meteors. Extremely bright meteors rivaling even the full moon (often bolides with associated sonic phenomena) are usually not associated with the major showers. Instead, they have meteoroid orbits that are more characteristic of those minor planets and short-period comets that are in highly eccentric orbits in the inner solar system. If the observed luminous-trail end points of these events are less than 12 mi (20 km) high in the atmosphere, there is a good chance that recognizable fragments (meteorites) will fall to Earth and perhaps be recovered. Most fireballs that drop meteorites occur in the afternoon or early evening, when velocities are low. For example, both a meteorite fall on October 9, 1992, which deposited a stone in Peekskill, New York, and one on June 14, 1994, which scattered numerous fragments over Quebec province, east of Montreal, appeared in the early evening. Thus most types of meteorites are believed to be samples of minor planets and possibly certain short-period comets. This connection with minor planets is verified by the fact that the reflecting properties of many minor planets resemble reflections from known meteoritic materials.

It is usually very difficult to detect even the brightest meteors in the atmospheres of other planets, although there have been several instances of spacecraft detection. An exception occurred July 16–22, 1994, when over 20 fragments of Comet Shoemaker-Levy 9 crashed into Jupiter in one of the most spectacular astronomical events ever observed. Even though the impacts happened only on the night side of Jupiter, several of their flashes were seen by reflection from the Jovian satellites. Several spacecraft were able, because of their offset position from the Earth, to view a larger portion of the Jovian night side than was visible from Earth and hence saw the impacts directly. Numerous observatories with large telescopes, including the *Hubble Space Telescope*, watched as hot plumes of material rotated into view at Jupiter's edge. Most impacts left mysterious dark clouds in Jupiter's atmosphere, several of which were visible for months even in small telescopes. Similar impacts on the Earth are thought to be responsible for the extinction of the dinosaurs 65 million years ago, and several other bioextinctions. *See* ASTEROID; COMET; JUPITER.

Origins of shower meteors. A number of meteor showers have been observed to be in orbits that are similar to those traveled by known comets. Thus an association between shower meteors and comets has gradually become a firmly entrenched concept (Tables 1 and 2). There are numerous theoretical scenarios where vaporization of the more volatile cometary ices ejects small solid particles from the surface of the nucleus. A fair proportion of these fragments, particularly the smaller dust-sized ones, escape and take up their own orbits as meteoroids.

Cometary nuclei have been known to split into two or more pieces and, when this occurs, it is likely that particles larger than dust size are released as well.

Gravitational attractions of the major planets and the disturbing effect of solar radiation pressure on individual particles tend to spread meteoroids out from the parent object position. Thus "young" showers are those that last only briefly (some as short as an hour or less), while "old" showers may show a few meteors per night but last a month or more. Many showers have a nonuniform structure along their orbits with highest meteoroid densities near the parent body. Lacking orbital synchronism with the Earth's position, these showers do not have an annual appearance at a reliable level. Instead, they show a tendency for strong showings to be separated by intervals roughly equal to the meteoroid orbital periods. The concentration of particles in these orbiting clumps can be relatively high, giving rise to brief deluges called meteor storms, where equivalent rates of thousands of meteors per hour have been noted for times that are at most an hour or so long. These numbers, however, give a false impression of the actual number density of meteoroids in space. With relative velocities on the order of tens of miles per second and a collecting area for each observer of a few hundreds of miles in diameter, the average separation between individual meteoroids is still a few thousand miles. Away from these maxima, however, the meteoroid number densities and hence the observed meteor hourly rates are quite low.

There are some instances where the Earth crosses the meteoroid stream twice per year, giving rise to two separate meteor showers. For example, Comet Halley gives rise to the May Aquarids and the October Orionids, while Comet Encke gives rise to the June Taurids and the November Taurids. *See* HALLEY'S COMET.

While the parent comet idea nicely explains many features of meteor showers, there are problems with this simple picture. First, certain minor planets resemble what might be termed extinct comet nuclei. Some of these have been observed at times with faint atmospheres, a main characteristic of comets. These identifications have been strengthened by the spacecraft observation of the properties of the nucleus of Comet Halley. Second, a perfectly respectable minor planet, 3200 Phaethon, which was discovered by the *Infrared Astronomical Satellite (IRAS)*, has an orbit nearly the same as the prominent Geminid annual shower visible in December. A minor planet, 2003 EH 1 (designated for now as a minor planet, but possibly the extinct nucleus of an ancient comet), has been suggested as the parent of the Quadrantid stream (Table 1). Evidence has also been found for several other minor planet connections with a small number of minor meteor showers. There is convincing evidence that a major source of micrometeoroids with orbits that go out through the asteroid belt is a relatively small number of asteroidal parent objects. There is also accumulating evidence that much of what is termed the sporadic meteor background is really a superposition of millions of ancient meteoroid

streams of various origins—asteroidal, cometary, outer solar system, and possibly interstellar. The meteoroid parent body question is apparently more complicated than previously thought.

The mechanisms that convert cometary or minor-planet fragments into meteoroid streams are not well understood either. The simplistic picture assuming that such fragments are in symmetrical swarms about the parent object are, at best, crude approximations. Thus the prediction of intense meteor storms must be treated with some suspicion. In 1992, when periodic comet Swift-Tuttle, the assumed parent object of the Perseid shower, was again sighted coming in toward the Sun, some astronomers predicted a brief but spectacular Perseid meteor storm for August 1993. In fact, it was nowhere near “storm” level.

Photographic and electronic observations. The strategy of photographic or electronic measurements is to place at least two cameras 10–52 mi (15–85 km) apart over a known baseline, but arranged to examine the same volume of space at a height of about 56 mi (90 km). Each camera has a rotating shutter so that the meteor trail consists of a line of bright dashes (see illustration). Since meteor durations are fairly short, the shutters must spin at speeds of up to 10 times per second to accurately determine the meteor speed. While the early techniques utilized photographic film, ultrasensitive television and other electronic imaging devices have been increasingly used. It has also been customary to place a prism or grating over the front aperture of at least one of the shutter-equipped cameras, so that spectroscopic information can be obtained as well. Television observation of the evolution of meteor trains and wakes has led to a better understanding of upper-atmosphere photochemistry. *See* CAMERA.

Meteor imaging is one of the most difficult areas of astronomical detection, even with ultrafast cameras. Chances for getting an image are better during major showers, because bright meteors are generally more numerous then. Meteor spectroscopy is even more difficult since the light is spread out over areas hundreds of times larger than the meteor trail itself. In spite of these difficulties, several amateur and professional astronomy groups worldwide operate successful bright meteor patrols using photography, but electronic imaging is gaining considerable favor with both professionals and amateurs. *See* ASTRONOMICAL PHOTOGRAPHY; ASTRONOMICAL SPECTROSCOPY.

Radio and radar observations. Radio and radar observations depend on the fact that the initial ion-electron densities in a meteor trail are considerably higher than the average for the ionosphere at an altitude of 56 mi (90 km). For a very high frequency (VHF) or somewhat lower-frequency radar system, the maximum reflected signal occurs when the meteor trail is at right angles to the outgoing wave, with head echoes rarely seen. For a forward-scatter system where the transmitter and receiver are separated, the maximum signal occurs when the meteor trail makes equal angles with the transmitter and receiver lines of sight. As the meteor trail forms, the meteor speed can be inferred from the changing

diffraction pattern that results from reflections by different parts of the ionized trail. At ultrahigh frequencies (UHF), radar reflections from the head-echo predominate. From these, high-accuracy radial velocities are determined directly, using the Doppler effect. Thus, original meteoroid speeds can be estimated and, in certain cases, meteoroid radiants and orbits can be determined without reference to optical images. Radio and radar observations generally require a considerable amount of expertise to obtain scientifically valuable results. *See* DOPPLER EFFECT; RADIO ASTRONOMY.

Nontraditional meteor studies and research. Scientific interest in ground-based meteor studies generally declined in North America for three decades from around 1970, when spacecraft methods of directly investigating interplanetary dust became available. In the late 1990s, however, renewed interest in meteor science was generated in the United States with the impending appearance of the Leonid meteor shower produced by comet 1866 I on another of its periodic (33-year intervals) approaches to the Sun. This was due in part to the perceived danger of meteoroids to spacecraft, and in part to rapid advances in applicable research techniques. Along with the usual Leonid ground-based preparations were the increasing professional use of high-power high-frequency (HF), very-high-frequency, and ultrahigh-frequency radars to observe hypervelocity micrometeors, the use of military spy satellites to observe giant bolides, and the increasing use of large jet aircraft as astronomical and geophysical observing platforms.

The 1990s saw claims of interstellar meteors based on ground-based radar investigations, but these particles appeared to have properties greatly different from those seen by the spacecraft instruments. In 2001, it was possible to show that the properties of interstellar micrometeors detected using the Arecibo UHF radar were actually very similar to those of interstellar dust particles observed by the *Ulysses* and *Galileo* space probes, extrapolated to higher masses. *See* SPACE PROBE.

Leonids in 1996–2003. With the appearance of the Leonid comet, it was expected that the strong meteor storm that happened last in 1966 would again make a brief but spectacular appearance. However, perturbations by the outer planets (particularly by Neptune) once again played a significant role in this shower's behavior. Fortunately, this time was unlike the shower's failure to live up to the early predictions in 1899, when virtually nothing happened. The perturbations moved a number of thin meteoroid streams produced by the comet many orbit periods in the past into intersection range of the Earth. This circumstance produced a unique succession of strong peaks covering a span of 7 years. There was a steady increase in the levels of the Leonids starting in 1996, then peaking with an extraordinary number of fireballs in 1999. In those first years (see illustration), a steady rain of fireballs was seen for more than a day worldwide. After 1999, when early predictions had forecast a return to pre-1995 levels, the Leonids kept coming back. They appeared to

decline a bit in 2000, but in 2001 they rose to two spectacular peaks, one rivaling the 1999 peak in Asia, to put on a very spectacular display for both coasts of North America. There was a repeat of two peaks in 2002, but this was not seen in North America. In 2003 two peaks separated by 3 days were predicted, but the observed levels were more modest than those in prior years. The scientific results of this extended display were much greater than anyone had hoped.

Scientific results from the Leonids. While meteor trails had been previously recorded from the space shuttles and other spacecraft, in 1997 the first above-atmosphere, far-ultraviolet spectrum of a bright meteor was recorded during the Leonid shower. Resonance emission lines due to magnesium and ionized magnesium were most prominent in the spectrum.

Spectra obtained from the ground also yielded new information. Both the hydroxyl radical (OH) in the near-ultraviolet and the oxygen molecule in the near-infrared were positively identified for the first time. However, since these are both seen in the airglow of the Earth's upper atmosphere, both may represent the "air" radiation commonly seen in the radiation of high-speed meteors. Such radiation is thought to be produced by constituents in the air surrounding the incoming bodies, rather than from the meteoroids themselves. Searches for the cyanogen radical (CN) have been negative, but now they allow rather strict limits to be placed on its presence and its production in the meteor plasma. In the meteor wake left after the meteoroid passes through the air, infrared spectra show the presence of iron monoxide (FeO). This is likely to have been formed due to the passage of hot meteoric iron through the part of the atmosphere where oxygen is in atomic form (in the mesosphere).

For a number of years, there have been claims of the direct observation of meteoric impacts on the lunar night side, but during several of the 1997–2003 Leonid peaks the number of such claims arose dramatically. For such a claim to be considered seriously, it is not enough to simply record a flash on the Earth-lit side of the Moon. Such flashes can be easily produced by cosmic rays hitting image surfaces of charge-coupled-device (CCD) cameras or photographic film even at ground level. What is needed is confirmation by another image of the same lunar surface, taken simultaneously by another instrument at a different geographic location, showing the flash at the same location on the Moon. During the Leonids, several events satisfying this minimum criterion were reported and so must be considered seriously. What must be sorted out is whether such events were peculiar to the Leonid periods or not. It is possible that detection will not be easily duplicated with other showers with a lesser percentage of fireballs and at a much lesser impact velocity.

The large number of fireballs in the Leonid streams enabled many details of the ablation processes at meteoroid incoming velocities that were higher than average to be recorded with high-speed cameras. Among the more important results is the discovery of a previously unseen shock-wave structure that has little or no "stand-off" distance in front of the mete-

oroid itself. A second surprising result is that the meteor wake is really two separate structures subject to local air turbulence effects rather than being a single structure with a hollow central zone. The split wakes seen at high resolution appear to be the result of meteoroid fragmentation into two nearly equal parts at heights far above where wake formation takes place.

David D. Meisel

Bibliography. N. Bone, *Meteors*, Sky Publishing, 1993; N. Bone, *Observing Meteors, Comets, Supernovae and Other Transient Phenomena*, Springer, 1998; V. A. Bronshten, *Physics of Meteoric Phenomena*, D. Reidel, Dordrecht, 1983; A. B. C. Lovell, *Meteor Astronomy*, 1954; D. W. R. McKinley, *Meteor Science and Engineering*, 1961; J. M. Pasachoff and W. Tirion, *Peterson Field Guide to the Stars and Planets*, 4th ed., updated, Houghton Mifflin, 2004; C. Sumners and C. Allen, *Cosmic Pinball: The Science of Comets, Meteors, and Asteroids*, McGraw-Hill, 2000.

Meteorite

A naturally occurring solid object from interplanetary space that survives impact on a planetary surface. While in space, the object is called a meteoroid, and a meteor if it produces light or other visual effects as it passes through a planetary atmosphere. Various sounds, including hissing and thunderous detonations, have also been reported for large meteors arriving at Earth. Explosive surface impacts by large meteorites are believed to have created the plethora of craters on the solid planets and moons of the solar system. Meteor Crater, Arizona, is Earth's most famous example of an impact crater. *See* METEOR; MICROMETEORITE.

A meteorite seen to strike a surface is known as a fall, whereas a meteorite discovered by chance is known as a find. In both cases, meteorites are named after their geographic places of recovery.

The major classification of meteorites is into chondrites and nonchondrites (or nonchondritic meteorites), and major classes of the nonchondrites are achondrites, stony-irons, and irons, in recognition of their compositions that are dominated by silicate minerals and iron-nickel alloys either alone or as admixtures. Within each of the major categories, detailed classifications are based on distinctive mineralogical and chemical compositions and physical structures (**Table 1**). Group names for unusual types have been chosen based on the established names of first-recognized specimens. For example, shergottites are so named because they belong to the same variety as the meteorite that fell at Shergotty, India, in 1865.

Meteorites represent the most ancient rocks known. Their ages, as determined by radiometric dating, extend to more than 4.5×10^9 years, which is thought to be near the time of solar system formation. As samples of primordial material, stony meteorites known as chondrites are studied for clues about how the solar system formed. In contrast,

TABLE 1. Classification of meteorites¹

| | | Nonchondrites ² | | |
|-----------------------|---------------------------|----------------------------|-------------------------|--------------------|
| | | Igneous or differentiated | | |
| Chondrites | Primitive | Achondrites | Stony-irons | Irons |
| Carbonaceous | Acapulcoites ⁵ | HED ⁷ | Pallasites ⁸ | IAB ⁹ |
| CI (1) ³ | Lodranites ⁵ | Howardites | Mesosiderites | IC |
| CM (1,2) | Winonaites ⁶ | Eucrites | | IIAB |
| CO (3) | | Diogenites | | IIC |
| CR (1,2) | | Angrites | | IID |
| CB (3) | | Aubrites | | IIIE |
| CH (3) | | Brachinites | | IIF |
| CV (3,4) | | Ureilites | | IIIAB |
| CK (3-6) ⁴ | | Martian | | IIICD ⁹ |
| Enstatite | | Shergottites | | IIIE |
| EH (3-6) | | Nakhlites | | IIIF |
| EL (3-6) | | Chassignites | | IVA ⁹ |
| Ordinary | | Orthopyroxenites | | IVB |
| H (3-6) | | Lunar | | Ungrouped |
| L (3-6) | | Feldspathic highlands | | |
| LL (3-6) | | breccias | | |
| Other | | Mare basalts | | |
| R (3-6) | | Mixed breccias | | |
| K (3) | | | | |

¹Compiled by Edward R. D. Scott; modified by Michael E. Lipschutz. Not listed are the ungrouped meteorites including 14 carbonaceous chondrites, 85 irons, and a few other igneous meteorites.
²Also called nonchondritic meteorites.
³Chondrites are also assigned numbers 1–6 called petrologic types, which provide a measure of the degree of low-temperature (0–100°C), aqueous alteration (1,2) or high-temperature, thermal metamorphism (4–6). Type 3 chondrites are the least modified.
⁴There are also numerous C (3–6) samples that are not CK.
⁵These meteorites probably come from the same asteroid.
⁶Closely related to IAB irons with silicate inclusions.
⁷HED meteorites probably come from the asteroid Vesta.
⁸Pallasites are divided into the main group, which are probably related to IIIAB irons, the Eagle Station types, and pyroxene pallasites.
⁹Includes several silicate-rich irons.

achondrites, stony-irons, and irons are samples of melt products formed during processing of solid material in planetary or preplanetary bodies. Chemical analyses of meteorites by geochemists in the 1930s through 1950s supplied first knowledge about abundances of chemical elements (other than hydrogen and helium) in the solar system. Research now includes dissection of meteorites into their many complex components and subsequent analyses of their mineral, microchemical, and isotopic compositions, with the aim of learning when, where, and how the meteorites formed. *See* DATING METHODS; SOLAR SYSTEM.

Meteorite specimens have been recovered from glacial “blue ice” localities in Antarctica that seem to favor surface concentration of meteorites that have fallen on that continent over the past million years. Asteroids are believed to be the sources of most meteorites. In 1982, however, it was conclusively demonstrated that a small achondrite found in Antarctica in 1981 was from the Moon—apparently propelled to Earth at some undetermined time by a large lunar impact event. Since then, many specimens of lunar rocks have been recovered as meteorites from Antarctica and elsewhere, and identified (as of early 2006) with at least 39 meteoroids (many of which fragmented into two or more of the recovered stones on passage through the atmosphere). Even more exciting is the prospect that 34 (as of early 2006) closely related achondrites (24

shergottites, seven nakhlites, two chassignites, and one orthopyroxenite), from various recovery locations around the world, are from Mars; one of them contains trapped gases that are nearly identical to those measured for the Martian atmosphere by the Viking lander in 1976. *See* ASTEROID; MARS; MOON.

Meteoritics, the study of meteors and meteorites, is a premier example of interdisciplinary science, involving chemists, physicists, geologists, and astronomers allied through international research projects. Many advanced laboratory methods, especially for isotope-ratio measurements, have been motivated in large part by meteorite research problems.

James L. Gooding

Chondrites

Chondrites are the most abundant sort of meteorite, over 15,000 being known. They constitute about 82% of all meteorites. Chondrites include three main categories: ordinary, carbonaceous (C), and enstatite (E). (A few Kakangari- and Rumuruti-like chondrites exist; what is known of these 3 K and 19 R chondrites does not modify the overall chondrite picture.) Chondrite categories are based on mineral abundances and compositions; mineral shapes, sizes, and relationships (that is, textures); and bulk chemical compositions. Chondrites are so-named because nearly all contain small (generally 0.5–2 mm) bead-like chondrules. Only CI chondrites are essentially chondrule-free; their matrix texture and chemical

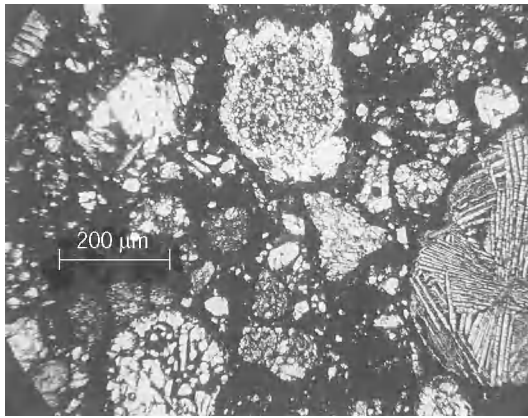


Fig. 1. Typical unequilibrated chondrite: many chondrules are present in a fine-grained mineral matrix.

composition closely relate them to the other chondrule-containing carbonaceous chondrites. The melting of chondrite-like material in parent bodies, followed by segregation and solidification of the resulting iron-nickel metal and silicates, produced the igneous meteorites.

Ordinary chondrites. These constitute 94% of all classified chondrites. Their main constituent minerals are olivine $[(\text{Mg,Fe})_2\text{SiO}_4]$, low-calcium pyroxene $[(\text{Mg,Fe})\text{SiO}_3]$, plagioclase $[(\text{Na,Ca})\text{Al}(\text{Si,Al})\text{Si}_2\text{O}_3]$, iron-nickel (Fe-Ni) metal, and troilite (FeS). They may contain silicate glass. *See* FELDSPAR; OLIVINE; PYROXENE; PYRRHOTITE.

Chondrules and chondrite matrixes consist of the same minerals: the difference is textural. Chondrule minerals crystallized within molten droplets, and show a variety of shapes consistent with a molten origin involving very rapid (minute-scale) heating and cooling. Matrix minerals are small and granular (**Fig. 1**).

Olivine and low-calcium pyroxene are iron (Fe^{2+}) and magnesium (Mg^{2+}) solid-solution minerals (that is, homogeneous crystalline phases with Fe^{2+} or Mg^{2+} ions located at random lattice points and existing in a range of concentrations). Their compositions in ordinary chondrites vary from fayalite (Fe_2SiO_4) 16 [olivine with 16% Fe^{2+} ions by number of moles, symbolized Fa16] to Fa33, and from ferrosilite (FeSiO_3) 15 (low-calcium pyroxene with 15 mole % Fe^{2+} , symbolized Fs15) to Fs25. The variation is discontinuous, and two distinct 1–2% Fa gaps in olivine composition separate ordinary chondrites into three groups: the H group (Fa16 to Fa20; high total and metallic iron), the L group (Fa22 to Fa25; low iron), and the LL group (Fa26 to Fa33; low-low iron). In addition to olivine composition, these groups are characterized by the ratio of iron in metallic minerals (iron-nickel metal and troilite) to total iron (including iron in silicate and oxide minerals). This ratio is 0.6 in the high-iron group, 0.3 in the low-iron group, and 0.1 in the low-low-iron group. High-iron and low-iron chondrites each constitute 40–44% of all chondrites; low-low-iron chondrites are about 9%. *See* SOLID SOLUTION.

Within each ordinary chondrite group, a sequence of textural and mineralogical changes suggesting metamorphism are designated numerically: H3–H6,

L3–L6, and LL3–LL6. Each subgroup number connotes similar characteristics. For example, type 3 ordinary chondrites exhibit abundant, very sharply defined chondrules containing fine-grained minerals (olivine, low-calcium pyroxene, with or without metal and troilite) aggregated in silicate glass (**Fig. 1**). Compositionally, the glass varies from chondrule to chondrule; however, it is plagioclase-like. Within chondrules, and in the surrounding granular matrix, olivine and low-calcium pyroxene exhibit large grain-to-grain compositional differences. Since adjacent grains have different fayalite and ferrosilite contents, these meteorites are called unequilibrated, a characteristic of type 3 ordinary chondrites. The nickel contents of metal grains also vary. Crystalline feldspar is absent; only chondrule glass, compositionally similar to plagioclase, exists.

The characteristics of each subgroup grade into those of the next subgroup. This is particularly apparent in type 3 ordinary chondrites, the subgroups with the most disparate properties. Each subgroup (H3, L3, and LL3) is divided into 10 subtypes (3.0–3.9) using thermoluminescence sensitivity (a measure of plagioclase crystallization from glass), variability of fayalite in olivine and of cobalt in metal, extent of matrix recrystallization, and fayalite and ferrosilite values in matrix relative to whole-rock averages. Thus, subtype 3.0 chondrites, which differ most from those of type 4, have the lowest thermoluminescence, greatest fayalite and cobalt variabilities in olivine and metal, respectively, least-recrystallized matrix, and highest fayalite and ferrosilite values in matrix relative to whole-rock. Ultimately, type 6 chondrites contain no glass but contain crystalline plagioclase. Olivine or low-calcium pyroxene grains are compositionally uniform; thus, type 5 and 6 chondrites are equilibrated. Average grain sizes are larger in such chondrites and chondrules become nearly unrecognizable (**Fig. 2**). Dominant subgroups differ: H5 and LL6 are the most common high-iron and low-low-iron chondrites, respectively, while L6 are the majority of the low-iron group.

Only collisional impacts can disrupt asteroids or planets. A variety of textural alterations, including localized or generalized melting, serve as barometers to estimate peak shock pressures. These can

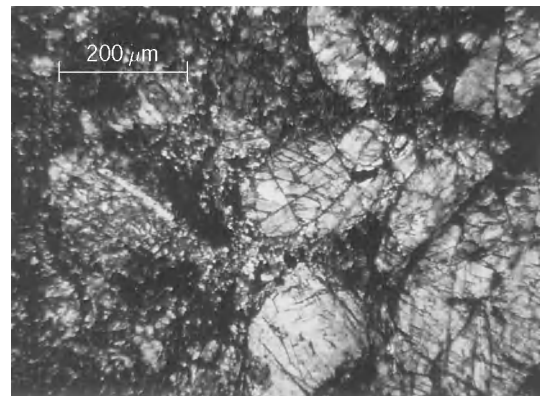


Fig. 2. Typical equilibrated chondrite. No complete chondrules are evident, and the matrix consists of relatively coarse mineral grains with possible poorly defined chondrule fragments.

exceed 500,000 atm (50 GPa), with resultant residual temperatures up to 1500°C (2730°F), persisting for months to years because collisional debris can be large, cooling slowly. About one-third of L4–L6 chondrite falls have been strongly shocked (300,000–500,000 atm); H4–H6, LL4–LL6, and other chondrites exhibit smaller percentages ($\leq 14\%$).

Carbonaceous chondrites. These constitute less than 5% of chondrite falls and are the least evolved meteorites. The same major chemical elements compose all chondrites, but carbonaceous chondrites can contain substantial carbon, hydrogen, and nitrogen, present at trace levels in ordinary chondrites. In addition to chondrules, they contain many other types of inclusions. Most important, the constituent minerals of chondrules and inclusions in carbonaceous chondrites differ in composition and kind from those composing the surrounding matrix.

In carbonaceous (C) chondrite classification, prototypical resemblance is important, occasioning a second letter. Most carbonaceous chondrites belong to classes CI (mainly types 1 and 2), CV (mainly types 3 and 4), CM, and CO (whose prototypes are meteorites named Ivuna, Vigarano, Mighei, and Ornans, respectively), so these classes are the best studied.

In normal type 1 chondrites, chondrules are extremely rare. Their matrix is 99% clay minerals and similar layer-lattice silicates with water as a structural component; CI chondrites average about 20% water by weight. *See* CLAY MINERALS; SILICATE MINERALS.

The CI matrix contains about 1% small (<100 micrometer) olivine and pyroxene crystals that are refractory minerals formed at high temperatures (750–1500°C or 1380–2730°F). Metal is entirely absent. The clay matrix necessarily formed at low temperatures (less than 120°C or 250°F). Thus, these meteorites are extremely unequilibrated. Epsomite and gypsum veins run through the CI matrix. The matrix includes a complex of organic molecules, most being poorly characterized polymers (designated kerogen), formed nonbiologically. These compounds and graphite account for the high CI carbon content, 3.6%. *See* EPSOMITE; GRAPHITE; GYPSUM; KEROGEN.

CM chondrites consist of about 50% clay and layer-lattice silicates (like those in CI chondrites) containing water, associated with organic compounds. The remaining half are refractory minerals in several forms, including single crystals, crystal fragments, chondrules as sharply defined as those in type 3 chondrites, and irregularly shaped inclusions (Fig. 1 and Fig. 3). Olivine and low-calcium pyroxene are the most common refractory minerals, and chondrules constitute only about 2% by volume. Metal is very rare. The textures of some inclusions suggest that they were loose grains aggregated before incorporation into the matrix. A small number of inclusions consist of extremely refractory minerals, composed of titanium, aluminum, and calcium oxides and silicates, formed at temperatures above 1500°C (2730°F). The combination of clay-rich matrix and refractory inclusions means these meteorites are highly unequilibrated (Fig. 3).

C3 chondrites, mainly CO and CV, contain much

less water (~1%) and carbon (~0.5%) than do C1 or C2 chondrites, because clays and layer-lattice silicates are rare in them. Instead, 35–40% (by volume) of their matrix consists of submicrometer blades of high-fayalite olivine (~Fa50). The remaining 60–65% consist of larger single crystals, fragments, chondrules, and inclusions of refractory minerals (olivine and low-calcium pyroxene). Highly refractory inclusions are relatively abundant. Thus, these meteorites are also unequilibrated. C4–C6 chondrites are rare, they constitute less than 15% of all carbonaceous chondrites, mainly metamorphosed CK chondrites.

Enstatite chondrites. These make up not more than 2% of all chondrites and consist of 60–80% enstatite (Fs0, that is, MgSiO₃) with 10–30% metal and 5–15% troilite. Enstatite chondrites range from those with many distinct chondrules to those having nearly unrecognizable chondrules. Thus, like ordinary chondrites, they fall into subgroups E3–E6. Contents of major elements differ, with EH (mainly types 3 and 4) containing more iron, and so forth, than EL (mainly types 5 and 6). E3 chondrites contain glass but no plagioclase, while E5 and E6 chondrites contain plagioclase but no glass; E4 is transitional. Common minor components are cristobalite, tridymite, and quartz. *See* ENSTATITE.

Minor and trace minerals in the matrix are quite extraordinary. Enstatite chondrites contain oldhamite (CaS), sinoite (Si₂ON₂), and osbornite (TiN) while, in other chondrites, calcium, silicon, and titanium are always combined with oxygen as silicate and oxide minerals. These unusual minerals can exist only under conditions of low oxygen activity, that is, extremely reducing environments. This is consistent with low-calcium pyroxene in them being Fe²⁺-free (Fs0). The matrix contains abundant microscopic graphite (0.4–0.8% by weight); organic compounds are absent.

Origin. Chondrites contain components from two genetic environments: nebular (those formed from dispersed materials in space) and planetary (those formed within parent bodies).

Carbonaceous chondrites contain the most obvious nebular components. Refractory mineral inclusions in them formed by direct condensation from a hot gas cloud surrounding the primitive Sun. During

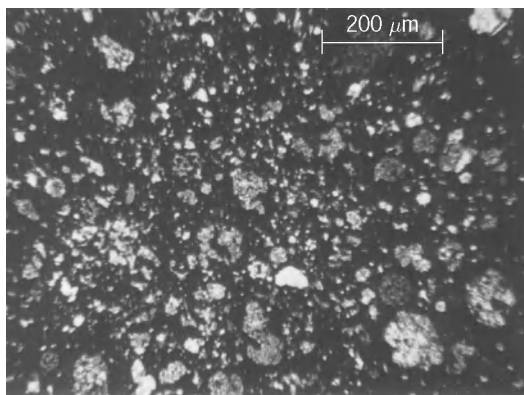


Fig. 3. Type 2 carbonaceous chondrite: numerous small grains of refractory minerals occur in a dense matrix of clay and other layer-lattice silicate minerals.

late-formation stages, as temperatures in the nebular gas fell, lower-temperature minerals formed and accreted as matrix, into asteroidal bodies together with refractory minerals. Thus, of all meteorites, carbonaceous chondrites are generally richest in volatile chemical elements. Most (about 90%) were not subsequently metamorphosed at high temperatures in their parent bodies, which would have erased their primitive characteristics. Carbonaceous chondrites, however, are not entirely pristine. Many experienced some low-temperature aqueous alteration; how much was nebular and how much was planetary is unknown.

Petrologic and chemical alterations in artificially heated chondrites demonstrate that about 10% of the C1-C3 chondrites were thermally metamorphosed at temperatures up to 900°C (1650°F) as open systems (CKs as closed systems) in parent body interiors. Impacts excavated these materials and transported them elsewhere. Spectroscopic studies show that such metamorphosed material comprise surfaces of four asteroid types (C, G, B, and F). A texturally and compositionally unique C1-like chondrite that fell on January 18, 2000, in the Yukon is spectroscopically linked to previously unsampled D-class asteroids. Its orbit could be determined because it was imaged from two locations during atmospheric passage. Like orbits determined previously for an enstatite and five ordinary chondrites, its maximum orbital distance was in the asteroid belt.

No consensus exists about the nebular condensation and accretion histories of ordinary and enstatite chondrite parent materials into asteroid-sized bodies but accretion (planetary) processes are broadly understood. The progressive equilibration from type 3 through type 6 apparently reflects planetary metamorphism that obscured chondrules and homogenized mineral grain compositions. Most ordinary chondrites and some enstatite chondrites exhibit brecciation or more severe impact-shock effects caused by parent-body collisions. The one or more collisions involving the one or more L chondrite parents occurred 500 million years ago and must have been exceptionally severe, since radiogenic ^4He and ^{40}Ar were lost then as were many volatile trace elements. See BRECCIA.

Spectroscopically, only the rare S IV type asteroids seem viable parents for the plentiful ordinary chondrites. By consensus, this paradoxical situation seems due to space weathering, mineral spectral alteration by solar particle impacts on asteroid surfaces.

Each chondritic group contains minerals with characteristic stable oxygen isotopic compositions so that each derives from its own batch of nebular material. It is not known how the very numerous LL, L, H, and E (or very rare R and K) groups are specifically related to each other. The LL-chondrite minerals formed under the highest oxygen activity, and the E-chondrite minerals, the lowest. It is tempting to interpret this as indicating formation at different distances from the primitive Sun: E-chondrites formed nearest the Sun, and the H, L, LL, and carbonaceous chondrites formed progressively farther away. While

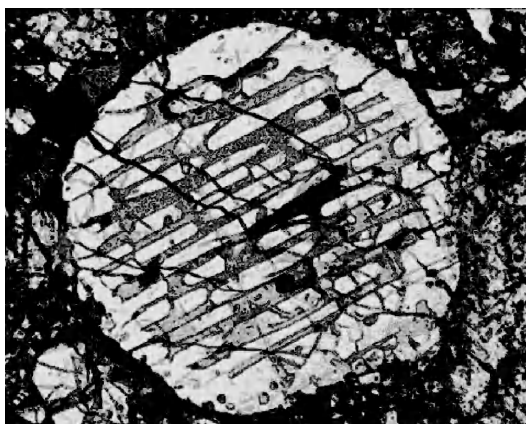
this is hypothetical, contents of volatile trace elements are lower in mildly shocked (less than 300,000 atm or 30 GPa) H4-H6 chondrites than in analogous L4-L6 samples, suggesting that H4-H6 parent material condensed at higher temperatures, that is, closer to the Sun. Two conclusions are certain: all chondrites could not have come from a single parent, and chondritic groups record different dominant genetic episodes. Michael E. Lipschutz

Chondrules. These are the most abundant particles in chondrites, with diameters up to about 1 mm (0.04 in.). They are distinguished from other particles (such as unmelted aggregates and calcium-aluminum-rich inclusions or CATs) in the chondrite matrix by textural and compositional criteria, being iron-magnesium silicate spherules with igneous textures. Chondrites themselves, when not modified much by heating in their asteroidal parent bodies, have sedimentary textures. The chondritic asteroids are aggregates of melted droplets, which cooled in the protoplanetary accretion disk (solar nebula). Some major astrophysical (or possibly geological) process clearly operated in the early solar system and caused extensive melting. Chondrules are therefore a key to preplanetary history, but understanding them is complicated by the fact that there are several different kinds, which may not all have formed in the same way.

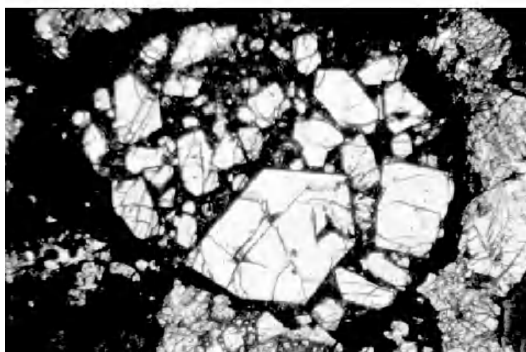
Mineralogy and textures. The most abundant minerals in chondrules are magnesian to intermediate olivine (fayallite 0 to Fa60), pyroxene (mainly inverted protopyroxene), glass rich in a feldspar component, iron metal, and iron sulfide (FeS). Chondrules are called type I if they have magnesian silicates (Fa < 10) and type II if they have more ferroan silicates (Fa > 10). The less common chondrules with compositions intermediate between the common iron-magnesium-rich kind and refractory inclusions (CAIs) have one or more aluminum-rich mineral such as spinel, plagioclase feldspar, or aluminum-rich pyroxene (fassaite). Many olivine crystals in chondrules are compositionally zoned, with enrichment in iron, calcium, and other elements toward their rims. Some pyroxene solid solutions exsolved on cooling, but the precipitates can be observed only in the electron microscope.

The main textural types (**Fig. 4**) are granular (including porphyritic on a submicroscopic scale), porphyritic (more strictly, microporphyritic), barred (parallel-plate crystals), radial/excentroradial, and cryptocrystalline/glassy, based on abundance, size, and shape of crystals. (Porphyritic texture is characterized by relatively large crystals dispersed in a matrix of finer-grained or glassy material.) A shorthand notation for chondrules combines the texture and one or more dominant minerals; thus there are BO (barred olivine), RP (radial pyroxene), and POP (porphyritic olivine-pyroxene) chondrules. The presence of glass, euhedral and skeletal crystals, and inhomogeneous crystals indicates crystallization of liquids at moderate cooling rates, on a time scale of hours, similar to thin lava flows. See IGNEOUS ROCKS.

Chondrules were not all single totally melted droplets. There are rare CAIs and forsteritic



(a)



(b)

Fig. 4. Examples of chondrules, showing different textures. (a) Barred olivine chondrule. (b) Type IIA porphyritic olivine chondrule.

ameboid olivine aggregates inside chondrules. Some chondrules enclose other chondrules. Others are aggregates of patches of somewhat different grain size. Some are layered in a way that suggests accretion. Some have igneous rims. Whether chondrule formation generally involved dustballs that were melted, or a spray of fine droplets and grains that accreted, the occurrence of these complexities on the whole suggests formation of chondrules by incomplete melting of preexisting solids. However, a number of cryptocrystalline chondrules may have formed by condensation of liquid from the gas.

Some chondrules contain grains not crystallized from the chondrule melt. Forsteritic (type D) olivine grains are common in type II chondrules and relatively ferroan (type ID) olivine with iron metal inclusions in type I chondrules. The two types of relict olivine grains suggest several generations of type I and type II chondrules, whose debris is incorporated into later chondrules.

Chemical and isotopic data. Type II chondrules, besides being ferroan, have approximately chondritic elemental abundances, with the volatile sodium and potassium present in higher concentrations relative to silicon than in the solar or CI chondrite composition. Type I chondrules are depleted in volatiles and enriched in refractory elements. They have a composition very similar to the primitive Earth mantle and may have been important in its accretion. Both type I and type II chondrules exist in silicon-

poor and silicon-rich forms, that is, in olivine-rich and pyroxene-rich varieties, given the suffixes A and B, respectively. Thus, there are type IA and IIA PO chondrules, type IAB POP chondrules, and so forth. Some cryptocrystalline chondrules are depleted in both refractory and volatile elements relative to the major element, silicon. Chondrule compositions do not vary like those of planetary igneous rocks, but are controlled by volatility-related processes, such as evaporation and condensation.

The occurrence of very fine grained type I chondrules rich in sulfur and other volatile elements whose concentrations decrease in coarser-grained similar chondrules suggests evaporative loss with more extended heating. However, the occurrence of olivine-rich cores in some chondrules, surrounded by pyroxene-rich mantles and sometimes silicon-rich rims, suggests condensation of silicon. At chondrule melting temperatures, silicate vapor rather than melt is stable, and chondrules formed in the solar nebula would therefore be expected to at least begin to evaporate. Free evaporation into a vacuum causes isotopic (Rayleigh) mass fractionation, and this process is observed in some refractory inclusions, but not in chondrules. Condensation at the pressures of the solar nebula would normally produce crystals rather than liquids.

The age of the solar system is given by the lead isotopic age of CAIs to be 4567.2 ± 0.6 million years. Similar measurements, and also the ^{26}Al - ^{26}Mg relative ages, indicate that chondrule formation continued for about 2 million years after this. See COSMOCHEMISTRY.

Chondrules in carbonaceous chondrites are enriched in ^{16}O relative to the Earth but less so than CAIs and ameboid olivine inclusions. Those in ordinary chondrites are depleted. The highest ^{16}O contents in chondrules may be related to the presence of forsteritic olivine relict grains, which have been less modified than the rest of the chondrules by exchange with the nebular gas.

Crystallization experiments. Chondrule simulation has been concerned mostly with duplicating their textures and olivine zoning by melting pellets of pressed mineral powder suspended on platinum wires. Cooling at $10\text{--}100^\circ\text{C/h}$ ($18\text{--}180^\circ\text{F/h}$) from temperatures a little lower than those which give total melting, about $1500\text{--}1600^\circ\text{C}$ ($2700\text{--}2900^\circ\text{F}$), is ideal for growing porphyritic textures. The grain size of the synthetic chondrule is very sensitive to the grain size of the starting material, because partial melting leaves a number of nuclei proportional to the initial number of crystals. Chondrule textures have also been produced after total melting of the charge by injecting mineral grains into the melt and inducing crystallization, and by spraying mineral dust into the hot zone of the furnace and collecting the resulting droplets on wire holders.

When chondrule analog are heated at low pressures, about 10^{-5} atm (1 Pa), as in the protoplanetary disk, sulfur is lost almost immediately, and sodium is severely depleted in the first few minutes, but ferrous oxide (FeO) is lost over a period of about 18 h. Iron-poor chondrules would therefore be expected

to be free of sodium, but they are not. At these low pressures, evaporative loss reproduces part of the composition trend of natural chondrules: we get first type IIA compositions, then IAB, and finally IA with up to 18 h heating at about 1550°C (2800°F), all with porphyritic textures. However, the charges show isotopic mass fractionation due to this free evaporation, while natural chondrules do not. When chondrule analog are heated in an atmosphere containing sodium, potassium, or silicon, loss of these elements has been stopped or even reversed. Pyroxene-rich rims on olivine-rich charges were formed by silicon condensation, as in some type I chondrules. Evaporation in the presence of a vapor containing the lithophile (rock-forming) elements greatly reduces their isotopic mass fractionation.

Texture-composition-temperature relationships. Chondrules with skeletal crystals can be reproduced in experiments with melting at near-liquidus temperatures (temperatures of total melting), and porphyritic chondrules with temperatures about 50°C (90°F) below the liquidus. Chondrules have a wide range of bulk compositions and therefore a wide range of liquidus temperatures (~1200–1800°C), but are dominantly porphyritic. If chondrules had formed by closed-system melting, this would be surprising: the melting temperature would appear to have been chosen to be always about 50°C (90°F) below the liquidus for each composition. Open-system melting resolves this problem. A porphyritic texture is established during heating and the composition changes during evaporation, giving a liquidus about 200°C (360°F) higher than the peak heating temperature.

Origin. Chondrules formed either by the melting of rock, dustballs, or free dust grains, or by the condensation of gas, and possibly by all of these mechanisms. Melting in the nebula should give mass-fractionated residues and condensation should give crystals. Condensation of liquid and evaporation to yield isotopically normal residues are both possible in the presence of an atmosphere highly enriched (about 100-fold) in rock-forming elements. Such an atmosphere may be achieved by concentrating solids into small domains and then evaporating them. Dust would be totally evaporated before larger aggregates, and both evaporation and condensation or recondensation would be expected. The concentration may possibly be due to accretion of planetesimals, to convection currents in the nebula that concentrate grains in stagnant zones between eddies, or to the delivery of mass to the inner edge of the disk as part of solar accretion. It seems that chondrules must form in less than 1% of the nebula in order to achieve the nonnebular gas composition. This local rather than nebula-wide formation is consistent with the relatively rapid cooling rates of chondrules.

Formation of a solar system involves numerous sources of energy that could have been responsible for chondrule melting. Mechanisms that have been considered include flares from the young Sun, lightning in the nebula, and collisions between hot planetesimals. During solar accretion, much mass is ejected in bipolar outflows, and particles melted by

solar radiation might have been thrown back into the disk. The physical conditions during these events are not sufficiently known to see the implications for the properties of chondrules formed in this way. Chondrules might have been formed by heating when shock waves propagated through the disk encountered concentrations of chondrule precursors. This process has been modeled in detail and is the mechanism whose temperature distributions best explain the relative proportions of different kinds of chondrules. How these shocks were generated needs to be established; spiral arms in an unstable massive disk may provide the mechanism. Roger H. Hewins

Achondrites

These are stony meteorites that have few, if any, chondrules and differ chemically from chondrites. They constitute about 8% of all meteorite falls and 1% of all finds. Although achondrites can be divided into several distinct groups based on chemical and isotopic composition, they are generally believed, based on aspects of their textures and composition, to have formed as the result of igneous processes on asteroidal or planetary bodies. Much of the interest in these meteorites derives from the fact that they provide clues into the nature of igneous processes and planetary differentiation early in the history of the solar system on planetary bodies outside the Earth–Moon system and on bodies presumed to be much smaller than the Earth and Moon.

Basaltic achondrites. The eucrites, howardites, and diogenites—often collectively referred to as the basaltic achondrites—are the most abundant achondritic meteorites. They appear to be samples of a series of related igneous rocks and of regolith breccias composed of fragments of these igneous rocks. They (along with the stony-iron meteorites, that is, mesosiderites and certain pallasites, and the iron meteorites) define a coherent group in terms of their oxygen isotope compositions, suggesting they are closely related. With ages near 4.5 billion years, they are products of igneous activity from the earliest history of planetary bodies in the solar system. They have been studied vigorously since the early days of the Apollo program, for they contain the most detailed record of early planetary differentiation, of igneous activity outside the Earth–Moon system, and of igneous activity on asteroid-sized bodies, from which they are presumed to have been derived.

Eucrites are dominantly composed of low-calcium pyroxene (usually pigeonite) and anorthitic plagioclase, with subordinate amounts of chromite, iron-rich metal, high-calcium pyroxene, silica polymorphs, and other minor phases. They are often, but not invariably, brecciated; most of the brecciated samples are monomict, but many of the more recently discovered samples from Antarctica are polymict. Many of the eucrites have pyroxene and plagioclase in ophitic and subophitic textures similar to those observed in terrestrial and lunar basalts. The textures of most of these eucrites and aspects of their chemical compositions, which cover a small

range compared to terrestrial and lunar basalts, suggest that they are samples of crystal-poor magmas. Some of these eucrites have vesicles, suggesting an extrusive or hypabyssal origin. A small number have gabbroic textures and appear to be cumulates from liquids related to the more common eucrites. Many eucrites show features attributed to shock metamorphism and to thermal metamorphism. *See* METAMORPHISM; PIGEONITE.

The diogenites are composed nearly entirely of magnesian orthopyroxene, with subordinate amounts of olivine, silica polymorphs, plagioclase, metal, chromite, and other minor phases. They are typically monomict breccias, but some are unbreciated, and others contain minor amounts of eucritic material. They range texturally from coarse- to fine-grained and show evidence of thermal metamorphism. They are generally believed to represent cumulates from magmas related to, but more magnesian than, known eucrites.

Based on comparison with lunar regolith breccias, with which they have some similarities, the howardites are generally considered to be samples of the regoliths of the parent planets of the eucrites and diogenites. They are polymict breccias that consist dominantly of angular to subrounded basaltic and pyroxenitic clasts and mineral fragments set in a finer matrix. They also contain chondritic fragments and rare glass beads (presumed to result from impact melting). Most of the fragments found in the howardites resemble eucritic and diogenitic material, and howardite bulk chemical compositions can be approximated as mixtures of these two types of meteorites. However, howardites also contain igneous fragments with textures and chemical compositions unknown among eucrites and diogenites. Many lithic and mineral clasts in howardites experienced a complex range of shock and metamorphic processes prior to incorporation into howardites. Ages of clasts in howardites extend from close to those of the oldest eucrites to 2×10^9 years younger; the young ages are usually interpreted as due to secondary, nonigneous processes.

The eucrites, diogenites, and individual fragments in howardites contain the record of extensive early igneous activity on what are generally assumed to be asteroidal parent bodies. The magmas from which this series of igneous rocks formed are thought to have been produced by partial melting of volatile-depleted, metal-depleted, olivine-rich planetary interiors. Although the relative roles of differing degrees of partial melting and fractional crystallization in the formation of the suite of igneous rocks sampled by these meteorite groups are controversial, it is clear that igneous processes as they are known from study of terrestrial and lunar rocks were active on small bodies very soon after their formation. The heat source for such igneous activity is still under investigation, but it could be the decay of the aluminium isotope ^{26}Al or perhaps heating by electric currents induced in small planets by the passage of an intense solar wind associated with a very active early Sun (T-Tauri phase). The reflectance spectrum

of the surface of the asteroid 4 Vesta closely resembles those of eucritic meteorites, and it has been suggested that this could be the source of the basaltic achondrites, although there are dynamical difficulties with such a source. *See* BASALT; IGNEOUS ROCKS; SOLAR WIND.

Shergottites, nakhlites, and chassignites. These are rare meteorites, but there is considerable interest in them because it has been suggested that they may have come from Mars. The 24 shergottites (as of early 2006) are composed primarily of pigeonite and augite pyroxenes plus glass with the composition of an intermediate plagioclase feldspar. The plagioclase glass is known as maskelynite, formed from crystalline plagioclase by shock metamorphism. The shergottites also contain titanomagnetite and rare hydrous amphibole; some contain olivine. The textures are similar to terrestrial diabases, and they have been interpreted as partial cumulates. Determination of their ages is controversial and may be complicated by the effects of shock metamorphism, but the ages are generally between 1.8×10^8 years and 1.3×10^9 years. Aspects of the chemical and isotopic composition of the shergottites suggest that they are derived from a parent body that experienced a complex, multistage history extending over much of the history of the solar system. Although they are basaltic like the eucrites, the shergottites are distinguished by their much higher volatile contents, the presence in them of iron(III) ion (Fe^{3+} , as opposed to metallic iron in the eucrites), and their much younger ages. In these respects the shergottites are more similar to terrestrial basaltic rocks. *See* AMPHIBOLE; DOLERITE; ROCK AGE DETERMINATION.

The seven nakhlites (as of early 2006) are augite-rich rocks whose textures strongly suggest they are cumulates. Augite appears to be the only cumulus phase; other phases include iron-rich olivine, pigeonite, sodic plagioclase feldspar, alkali feldspar, and titanomagnetite. The nakhlites contain hydrous alteration phases that may be preterrestrial in origin. Like the shergottites, the nakhlites have relatively young crystallization ages ($\sim 1.3 \times 10^9$ years) and appear to have formed in oxidizing, volatile-rich, Earth-like environments compared to the basaltic achondrites. Unlike the shergottites, the nakhlites show only minor effects of shock metamorphism.

The Chassigny meteorite (the prototype of the two chassignites, as of early 2006) is a lightly to moderately shocked dunite. Like the shergottites and nakhlites, it is volatile-rich and oxidized and also contains hydrous amphibole. Its age is not well constrained, but a single potassium-argon age is consistent with the young ages of the shergottites and nakhlites. *See* DUNITE.

The young ages of the shergottite, nakhlite, and chassignite meteorites (often referred to as the SNC group), plus the similarity of the bulk compositions of the shergottites to that of the Martian soil as determined by the Viking landers, first led to the suggestion that these meteorites could be derived from Mars. It is difficult to conceive of a heat source for endogenous igneous activity (these meteorites have no

features resembling known impact melts) on an asteroidal parent body at about 1.3×10^9 years ago; and given the limited choice of available larger planets, Mars seemed the most likely choice. The similarity of relative noble gas and nitrogen abundances and isotopic ratios in the Martian atmosphere and shock-produced glass in one shergottite provide strong support for this hypothesis. The very low paleomagnetic intensities of the shergottites are also consistent with a Martian origin. It is still a subject of controversy whether fragments of sufficient size to explain measured cosmic-ray exposure ages could be ejected more or less intact from Mars by impact and subsequently delivered to Earth. It is generally accepted that the question of whether or not these meteorites are from Mars will be resolved only after a sample-return mission to Mars. *See* PALEOMAGNETISM.

Angrites. Until the discovery of two additional examples of this class of meteorite in Antarctica, Angra dos Reis was the only meteorite of this type. Nevertheless, it has been of great interest because of its unusual mineralogy and very primitive isotopic characteristics. Angra dos Reis is composed nearly entirely of an iron-bearing, aluminous clinopyroxene (fassaite), with minor amounts of plagioclase, spinel, calcic olivine, and a variety of other phases. The more recently discovered members of this group are less enriched in fassaite, but their mineralogies are similar. Angra dos Reis is generally interpreted as having formed by accumulation of fassaite from a relatively silica-poor magma, while the Antarctic examples have been interpreted as more closely resembling magmatic compositions. With an age of 4.55×10^9 years and evidence for extinct short-lived radionuclides, Angra dos Reis provides another case of magmatic activity from the very earliest history of the solar system. However, based on the inferred characteristics of the parent liquids of the angrites, this igneous activity was distinct from that recorded by the basaltic achondrites. Due to their rarity, it has been difficult to constrain the details of their petrogenesis, but efforts to do so have generally suggested complex igneous histories. *See* MAGMA.

Ureilites. The dominant silicates of this strange meteorite type are olivine and low-calcium pyroxene (typically pigeonite), usually present as millimeter-sized mosaic aggregates. The silicates are enclosed in a dark matrix that consists mainly of metallic iron, graphite, diamond, lonsdaleite (another high-pressure form of carbon), plus other minor phases. Where the matrix and silicates are in contact, they have reacted at low pressures to form iron-rich metal and more magnesian silicates. Some of the ureilites also exhibit long, oriented cracks or voids. In terms of their oxygen isotopic compositions, the ureilites are distinct from other achondrites, but are similar to certain carbonaceous chondrites.

The silicates often are preferentially oriented. This, along with their relatively coarse grain size, suggests that they are plutonic rocks. However, it is generally difficult to distinguish between accumulations of crystals settled from a magma chamber and residual crystals left behind after the extraction of par-

tial melts from planetary interiors, and both origins have been suggested for the ureilites. In either case, aspects of their trace-element and isotopic compositions suggest a complex igneous history for the ureilites if these features are all igneous in origin. The carbon in the ureilites could be a primary igneous feature, or it could have been introduced subsequent to their igneous history, perhaps during the shock events that led to the mosaicism of the silicates and the presence of the high-pressure phases diamond and lonsdaleite in the matrix. Aspects of the textures of these meteorites tend to favor the former possibility. If so, the coexistence of graphite plus iron-bearing silicates during igneous activity suggests somewhat elevated pressures (that is, more than a few hundred bars). Although no direct connection has been made, many researchers in this field believe that the parent bodies of the ureilites were related to those of carbonaceous chondrites.

Aubrites. These meteorites, also known as enstatite achondrites, are composed almost entirely of polymorphs of enstatite (MgSiO_3), with grain sizes sometimes exceeding several centimeters. Most are monomict breccias, but unbrecciated and polymict examples are known. At least some of these meteorites appear to be regolith breccias, based on their textures and the presence of foreign meteorite clasts and solar rare gases. Like the enstatite chondrites, they contain silica polymorphs, metallic iron with dissolved silicon, and an assortment of rare, highly reduced minerals, some of which are known only from these meteorites. The assemblages found in these meteorites indicate that they formed under conditions more reducing than expected in a gas of solar composition.

It is generally accepted that the aubrites are related to the enstatite chondrites, but the relationship is not clear. They are chemically similar to enstatite chondrites, but they are distinctive; in most cases they extend to the more extreme fractionations the same trends observed among the enstatite chondrites. The very coarse grain size of the aubrites suggests that they formed as cumulates, perhaps during differentiation of an enstatite chondrite parent body. Alternatively, it has been suggested that they originated by fractionation processes occurring in the solar nebula and that they represent a continuation of the same processes by which the various nonigneous enstatite chondrites are related. According to this view, the coarse grain size of the aubrites would have resulted from annealing in high-temperature nebular environments. *See* NEBULA.

Edward Stolper

Iron and Stony-Iron Meteorites

Iron and stony-iron meteorites are made largely or partly of metallic iron-nickel and come from deep inside asteroids that were melted around 4560 million years ago. In most of these bodies, the metal formed dense cores that were surrounded by silicate mantles. Iron and stony-iron meteorites are much stronger than other meteorites, so they can survive longer in space. At least 12 of the 20 craters on Earth under 1.2 km (0.75 mi) in diameter were made by iron or stony-iron meteoroids.

Iron meteorites. Iron meteorites are pieces of once molten metallic cores and pools in asteroids that were fragmented by impacts after they cooled slowly. About 700 different iron meteorites have been identified; 40 were seen to fall, and the rest fell during the last million years. The smallest iron meteorites, which weigh only 5–30 g (0.18–1.1 oz), were found in Antarctica and are aerodynamically shaped like certain tektites. The largest single iron meteorite weighs about 60 metric tons (66 tons) and still lies in Namibia. The second largest is from Greenland and weighs 31 metric tons (34 tons). It is on display in the American Museum of Natural History in New York. *See* TEKTITE.

Nine much larger iron masses hit the Earth during the last million years, forming numerous craters 50 to 1200 m (150 to 3900 ft) in diameter. However, the surviving fragments from these meteoroids weigh less than a ton. The largest crater, which is in Arizona, was formed about 50,000 years ago by the impact of a meteoroid measuring about 50 m (150 ft) across. The impact released energy equivalent to about 20–60 megatons of TNT.

Mineralogy. When iron meteorites are sawed, ground, polished, and etched, they typically show a striking geometrical array of oriented crystals known as a Widmanstätten pattern. This pattern results from the crystallographically controlled formation of kamacite plates parallel to the four sets of octahedral planes in the precursor taenite crystals (**Fig. 5**). Kamacite and taenite are the mineral names for the body-centered cubic and face-centered cubic crystal structures of iron alloys. Iron meteorites that show such an oriented array of kamacite plates in taenite are called octahedrites and contain 6–15% nickel. A few iron meteorites known as hexahedrites contain 5–6% nickel and are almost entirely composed of kamacite. Even rarer are the ataxites, which have more than 15% nickel and contain only microscopic kamacite plates. *See* CRYSTAL STRUCTURE.

The large sizes of the precursor taenite crystals and the oriented kamacite plates indicate that iron meteorites were hot when they formed and that they cooled at an extremely slow rate. Thus at one time iron meteorites must have been buried deep inside a large volume of poorly conducting rocky material. The cooling rate in the temperature range 700–400°C (1290–750°F) can be estimated from the thickness of the kamacite plates, the bulk nickel concentration, knowledge of the stability fields of kamacite and taenite, and the rate at which iron and nickel atoms diffuse in these minerals. From these data, computer models have been developed for the diffusion-controlled growth of kamacite plates; these models yield typical cooling rates of 10–100°C (18–180°F) per million years. Therefore, most iron meteorites must have been buried under tens of kilometers of rocky material when they cooled from 700 to 400°C (1290 to 750°F).

Most iron meteorites contain small amounts of a wide variety of other minerals; sulfides, phosphides, carbides, oxides, phosphates, and silicates together make up no more than a few percent by volume of most iron meteorites (**Table 2**).

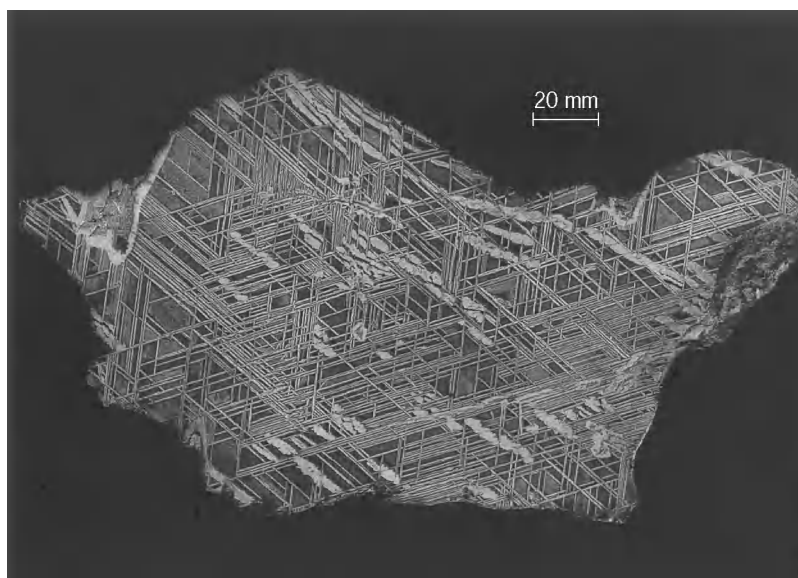


Fig. 5. A piece of the Edmonton (Kentucky) iron meteorite that has been sliced, polished, and etched to show the Widmanstätten pattern. The plates of kamacite 0.3 mm (0.01 in.) wide are oriented parallel to the faces of an octahedron. (Smithsonian Institution)

Although the mineralogy of iron meteorites largely reflects slow cooling at depth, some features result from shock waves of high intensity caused by impacts between asteroids at speeds of several kilometers (1 km = 0.6 mi) per second. Kamacite in many iron meteorites has a distorted structure caused by transient formation of a hexagonal close-packed structure at shock pressures above 13 gigapascals (130 kilobars). Two iron meteorites contain diamonds formed from graphite by shock. Those in the Canyon Diablo crater formed during the impact responsible for the 1200-m (3900-ft) crater in Arizona, whereas diamonds in a closely related octahedrite from Antarctica probably formed during an impact on the parent asteroid. Another iron meteorite contains stishovite, which formed from silica during a collision between asteroids. *See* DIAMOND; SHOCK WAVE; STISHOVITE.

Chemical composition. The chemical compositions of iron meteorites provide important clues to their classification and the manner in which they solidified in molten asteroids. The trace elements gallium and germanium, which are present at concentrations between 0.01 and 1000 parts per million by weight, are

TABLE 2. Common minerals in iron meteorites

| Mineral | Composition |
|---------------|--|
| Kamacite | Fe-Ni; Ni <7.5% |
| Taenite | Fe-Ni; Ni >30% |
| Tetrataenite | FeNi (ordered) |
| Troilite | FeS |
| Daubreelite | FeCr ₂ S ₄ |
| Schreibersite | (Fe,Ni) ₃ P |
| Cohenite | Fe ₃ C |
| Haxonite | Fe ₂₃ C ₆ |
| Graphite | C |
| Carlsbergite | CrN |
| Chromite | FeCr ₂ O ₄ |
| Whitlockite | Ca ₃ (PO ₄) ₂ |
| Chlorapatite | Ca ₅ (PO ₄) ₃ Cl |

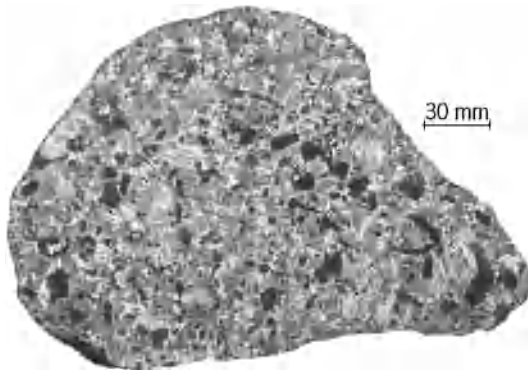


Fig. 6. A slice of the Four Corners (southwestern United States) iron meteorite showing angular black silicate inclusions separating regions with independently oriented Widmanstätten patterns. Like other group IAB members, this meteorite did not form in the core of an asteroid. A large impact probably mixed silicate with molten metallic iron-nickel long before the Widmanstätten pattern formed. (Smithsonian Institution)

especially useful in classifying iron meteorites. Most iron meteorites (84%) belong to one of 13 chemical groups, all but two of which have concentrations of gallium and germanium that vary by less than a factor of 2. Each of the 13 groups, which are identified as IAB, IC, IIAB, and so forth, has between 5 and 210 members. This chemical classification of iron meteorites has not made the older, structural classification obsolete, because most iron meteorites (perhaps 70%) can be classified into chemical groups on the basis of structure alone. For example, nearly all hexahedrites belong to group IIA, the medium octahedrites (those having kamacite bandwidths of 0.5 to 1.3 mm or 0.02 to 0.068 in.) largely belong to group IIIAB, and ataxites to group IVB.

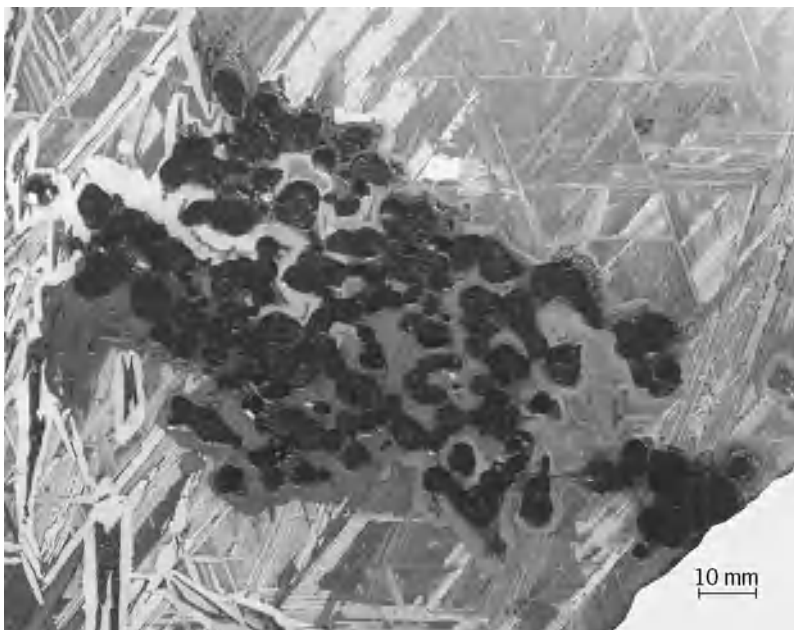


Fig. 7. A part of a slice of the Brenham (Texas) pallasite, showing a cluster of rounded, dark crystals of olivine enclosed in iron-nickel metal that shows a Widmanstätten pattern. Pallasites formed at the core-mantle boundary of asteroids that melted. (Smithsonian Institution)

Analyses for other trace elements, such as iridium, gold, and tungsten, and the minor elements phosphorus and nickel can also be used to classify iron meteorites, but they show larger variations within groups. For example, iridium concentrations vary by a factor of 6000 in group IIAB alone. The iridium variations within groups, which are very much larger than the variation in chondritic metal, were produced when molten metal solidified. Thus iridium, tungsten, and other elements that preferentially concentrate in solid metal were enriched in the first metal to solidify, whereas elements that are concentrated in molten metal, such as nickel, phosphorus, and gold, were enriched in the last metal that solidified.

Laboratory measurements of the concentrations of these elements in coexisting solid and liquid metallic iron-nickel allow the chemical trends found within the groups to be accurately modeled. Thus, nearly all of the 13 groups of iron meteorites formed from separate asteroids; the approximately 100 ungrouped iron meteorites probably come from another 40-odd bodies. Cooling rates for samples from a single metallic pool should be uniform, and this is observed for most groups.

Origins. The chemical and mineralogical evidence discussed above shows that iron meteorites formed from molten pools of metal that solidified and then cooled over many millions of years. This evidence is consistent with an origin for iron meteorites in the cores of asteroids that melted and differentiated. When an asteroid is partly melted, iron-nickel and iron sulfide, being denser than the associated silicates, will begin to sink to the center. With sufficient heating, a core of molten sulfur-rich metal will form. Since most iron meteorites contain no silicates and most achondrites have only trivial amounts of metal, it is likely that metallic cores are the source of many iron meteorites.

Two groups of iron meteorites (IAB and IIE) contain several volume percent of silicate inclusions and probably did not form in cores of asteroids. Instead these meteorites are probably derived from many metallic pools distributed within asteroids. Group IAB iron meteorites have angular silicate inclusions (Fig. 6), and it is likely that impacts mixed their metal and silicate components while they were hot. In addition, if the planetesimals from which the asteroids formed by accretion had already melted and solidified, it is possible that several metallic cores could have been combined into one asteroid.

At least one group of irons, IVA, formed in an asteroid that probably suffered a catastrophic impact before it had cooled below $\sim 900^{\circ}\text{C}$ (1600°F). Evidence for the timing and nature of more recent impacts comes from the cosmic-ray exposure ages of the meteorites. Since galactic cosmic rays can penetrate only to depths of about a meter, meteoroids or asteroidal fragments are not significantly exposed to cosmic rays until they are a meter or less in size. Exposure ages of most iron meteorites are $0.1\text{--}1 \times 10^9$ years, but two groups show a significant clustering of ages: group IIIAB and IVA iron meteorites

have ages of 6.5×10^8 and 4.5×10^8 years, respectively. Thus about 40% of all irons are fragments of two giant impacts in the asteroid belt that occurred in the last billion years. See COSMIC RAYS.

Astronomers believe they have identified many metallic asteroids 1 to 200 km (0.6 to 120 mi) in diameter, but none has been linked yet to a specific group of iron meteorites.

Stony-iron meteorites. There are two major types of stony-iron meteorites: pallasites and mesosiderites. They are among the rarest groups of meteorites, together constituting 1% of meteorite falls. About 50 specimens of each type have been found.

Pallasites consist of 5–80% by volume of olivine $[(\text{Mg,Fe})_2\text{SiO}_4]$ embedded in metallic iron-nickel, with minor amounts of chromite $(\text{FeCr}_2\text{O}_4)$, troilite (FeS) , and pyroxene $[(\text{Mg,Fe})\text{SiO}_3]$ (Fig. 7). Three groups are distinguished: the main group, the Eagle Station types, and the pyroxene-rich pallasites. All three groups formed at the core-mantle boundaries of differentiated asteroids when molten metal was intruded into the surrounding olivine-rich mantle. Most pallasites, which belong to the main group, have metal compositions that match those of the nickel-rich members of the IIIAB group of iron meteorites, implying that they formed in the same asteroid. Molten metal and silicates were largely mixed at the core-mantle boundaries of asteroids possibly as a result of impacts that gouged holes in the crust about 4500 million years ago. Impacts between large asteroids were common in the asteroid belt then as there were vastly more bodies than exist today.

Mesosiderites are enigmatic breccias consisting of fragments of basalt and gabbro mixed with roughly equal amounts of iron-nickel and troilite (FeS) . About 50 specimens are known including the Eltanin samples, which came from a 1–5 km (0.6–3 mi) asteroid that was one of the largest to hit the Earth in the last 2 million years. The Eltanin samples were identified in cores of sediment drilled from the southeast Pacific Ocean.

Mesosiderites, like pallasites, formed by mixing of molten metallic iron and nickel with solid silicate. But the impact that created the mesosiderites caused vast quantities of molten metal to spew over fragments of rock from the surface and interior. The silicate-metal mixtures were deeply buried after the impact and cooled at a rate of about 0.4 K per million years, slower than any other group of meteorites.

Olivine-rich meteorites from the mantles of asteroids that melted are much rarer than stony-iron meteorites. Similarly, olivine-rich asteroids are rarer than metal-rich ones. The impacts that excavated the deeply buried iron and stony-iron meteorites and brought fragments to Earth probably crushed most of the olivine-rich mantles to fine dust.

Edward R. D. Scott

Isotopic Anomalies in Meteorites

Primitive meteorites show a large variety of isotopic anomalies, that is, deviations from the average solar system composition (called “normal” composition)

that cannot be explained by chemical fractionation and radioactive decay taking place today. They are rather of presolar origin or the result of irradiation from an early active Sun or in interplanetary space. These anomalies provide information about the nucleosynthetic sources of the material that formed the solar system. See ISOTOPE.

It is well established that carbon and all the heavier elements are produced in stars by nuclear processes (stellar nucleosynthesis). These elements are ejected into the interstellar medium either from explosions of massive stars (supernovae) or as stellar winds and planetary nebulae from low-to-intermediate mass stars at the end of their evolution. Stars of different masses and ages produce elements with very different isotopic ratios. The solar system was formed from material originating from many stars. However, until the early 1970s it was believed that all this material was thoroughly homogenized into a mix of uniform isotopic composition in a hot solar nebula. Isotopic measurements of samples from the Earth, the Moon, and various meteorites seemed to confirm this belief. Exceptions were well-understood excesses of certain isotopes that were the daughters of long-lived (they still exist today) radioactive isotopes or isotopic fractionation effects as the result of physical and chemical processes. See NUCLEOSYNTHESIS.

This belief in isotopic homogeneity was shattered by the discovery of isotopic anomalies in several elements in 1973; and today, thanks to increases in the precision of isotopic analysis and the capability to measure small samples, a plethora of isotopic anomalies are known in meteorites. The largest anomalies are found in small samples, where the effects are not diluted by isotopically normal material.

Isotopic anomalies in meteorites and interplanetary dust particles (IDPs) can be divided into four classes.

(1) *Mass-dependent fractionation caused by physicochemical processes (diffusion, evaporation, condensation, and chemical reactions).* Certain physical and chemical processes can also lead to non-mass-dependent fractionation that might mimic effects of nuclear origin.

(2) *Effects from the decay of radioisotopes.* In addition to effects from the decay of long-lived isotopes, meteorites show also the effects of short-lived, now extinct isotopes.

(3) *Nuclear effects reflecting nucleosynthetic processes in stellar sources.* These effects are found in samples that formed in the solar system but inherited some nucleosynthetic components, as well as in presolar grains, bona fide stardust, that formed in stellar atmospheres and thus represent the isotopic compositions of their parent stars.

(4) *Effects due to irradiation of material by an early active Sun or irradiation of meteorites by galactic and solar cosmic rays.* These effects provide information on the exposure history of meteorites, both on their parent bodies and in interplanetary space during their travel between their asteroidal sources and the Earth.

TABLE 3. Short-lived radioisotopes[†]

| Radioisotope | Half-life, million years | Daughter isotope | Reference isotope | Initial abundance |
|-------------------|-----------------------------|---------------------|----------------------|---------------------------|
| ⁴¹ Ca | 0.10 | ⁴¹ K | ⁴⁰ Ca | 1.5×10^{-8} |
| ²⁶ Al | 0.74 | ²⁶ Mg | ²⁷ Al | 5×10^{-5} |
| ¹⁰ Be | 1.5 | ¹⁰ B | ⁹ Be | $\sim 5 \times 10^{-4}$ |
| ⁶⁰ Fe | 1.5 | ⁶⁰ Ni | ⁵⁶ Fe | $\sim 1.5 \times 10^{-6}$ |
| ⁵³ Mn | 3.7 | ⁵³ Cr | ⁵⁵ Mn | $\sim 2 \times 10^{-5}$ |
| ¹⁰⁷ Pd | 6.5 | ¹⁰⁷ Ag | ¹⁸⁷ Pd | 4.5×10^{-5} |
| ¹⁸² Hf | 9 | ¹⁸² W | ¹⁸⁰ Hf | 2×10^{-4} |
| ¹²⁹ I | 16 | ¹²⁹ Xe | ¹²⁷ I | 1×10^{-4} |
| ²⁴⁴ Pu | 81 | Fission Xe | ²³⁸ U | $(4-7) \times 10^{-3}$ |
| ¹⁴⁶ Sm | 103 | ¹⁴² Nd | ¹⁴⁴ Sm | 8×10^{-3} |

[†] Short-lived radioisotopes for which evidence has been found in meteorites.

Isotopic fractionation effects. Refractory inclusions (CAIs) from primitive meteorites (mostly carbonaceous and unequilibrated ordinary chondrites) exhibit mass-dependent fractionation effects in the elements oxygen, magnesium, silicon, calcium, and titanium. These effects must be due to evaporation and condensation during the formation of the CAIs and in some cases indicate multistage processing. Fractionation effects are largest in a class of CAIs named FUN inclusions because, in addition to fractionation (F), they show nuclear effects (called UN effects, because they were unknown at the time of discovery) in many elements. It is still not clear, however, why F and UN effects are associated with one another.

Interstellar cloud material. Carbonaceous and unequilibrated ordinary chondrites and interstellar dust particles have large excesses in deuterium (ranging up to 50-fold in certain interstellar dust particles) and ¹⁵N. These anomalies appear to be associated with organic matter. Because deuterium is destroyed in stars (it was produced only in the big bang), deuterium excesses cannot be of nucleosynthetic origin. They and the ¹⁵N excesses are most likely due to the incorporation of interstellar cloud material into meteorites and interstellar dust particles. This material acquired its isotopic anomalies through fractionation during ion-molecule exchange reactions that took place at very low temperatures (10 to 30 K or -442 to -405°F) in dense molecular clouds. Extremely large deuterium excesses (up to 1 million-fold) relative to the interplanetary medium are observed astronomically in simple molecules from interstellar clouds. See DEUTERIUM; INTERSTELLAR MATTER; MOLECULAR CLOUD.

Oxygen. The discovery of large ¹⁶O excesses in refractory minerals from the Allende carbonaceous chondrite in 1973 refuted the hot solar nebula model. Large ¹⁶O variations dominate the oxygen isotopic effects in meteorites with only minor variations in the ¹⁷O/¹⁸O ratio. CAIs were formed from a reservoir enriched in ¹⁶O by about 5% relative to the average isotopic composition of other solar system materials. This enrichment might have a nucleosynthetic origin (supernovae produce huge amounts of ¹⁶O); however, among presolar grains (discussed below) almost no oxygen-rich grains with a supernova origin have been detected. Alternatively, the ¹⁶O excesses could be the result of non-mass-dependent physicochemical processes or of self-shielding, a photo-

chemical effect in the early solar system that separated ¹⁶O from the other two isotopes because of the large difference in the abundance of ¹⁶O from those of ¹⁷O and ¹⁸O. In the case of self-shielding, the Sun is predicted to be enriched in ¹⁶O relative to the inner planets and asteroids. Analysis of solar wind samples collected by the *Genesis* mission may settle this issue, despite contamination caused by its crash landing in September 2004. See SUPERNOVA.

Short-lived isotopes in the early solar system. Meteorites contain evidence of the presence of short-lived isotopes. These isotopes have half-lives so short that today they are extinct. The initial presence of a short-lived isotope is indicated by excesses in the daughter isotope that are proportional to the parent-to-daughter elemental ratios. There is now solid evidence for a series of short-lived isotopes (Table 3). Aluminum-26 has been studied most thoroughly. Its initial presence has been established for many CAIs, most of which have an initial inferred ²⁶Al/²⁷Al ratio of 5×10^{-5} . Like other short-lived isotopes (⁵³Mn and ¹²⁹I are prominent examples), ²⁶Al can in principle serve as a fine-scale chronometer of relative ages of early solar system events. It (and ⁶⁰Fe) also has been considered as a possible heat source for melting small planetesimals. Both of these functions require its widespread and uniform distribution in the solar nebula. The fact that relative ages between CAIs and other meteoritic samples determined with the ²⁶Al clock agree with age differences obtained with the absolute uranium-lead chronometer indicates that this is indeed the case.

The abundances of the radioisotopes ⁵³Mn, ¹⁸²Hf, and those with longer half-lives can be explained by continuous equilibrium production by supernovae in the Galaxy. However, the presence of ²⁶Al and ⁶⁰Fe, and especially that of ⁴¹Ca with its much shorter lifetime, in early solar system solids requires production in a stellar event immediately (less than a million years) before solar system formation. The possible candidates are a supernova, whose explosion could have triggered the collapse of the molecular cloud from which the solar system formed, or an asymptotic giant branch (AGB) star, a low-to-intermediate mass (with a mass of less than 8 solar masses) star in the late stages of its evolution. Beryllium-10 is the only short-lived isotope that is not produced by stellar nucleosynthesis but by irradiation of heavier

elements (carbon and oxygen) by energetic nuclear particles, most likely from an early active Sun. See RADIOISOTOPE (GEOCHEMISTRY); STELLAR EVOLUTION.

Nuclear anomalies. Many CAIs were found to carry nuclear anomalies. These anomalies are especially large in FUN inclusions. These inclusions show isotopic patterns in the heavy elements (elements heavier than iron) that carry the signatures of distinct nuclear processes. The heavy elements are made either by neutron capture starting from iron, either in the *s* process (slow neutron addition in stellar environments with low neutron density) or in the *r* process (rapid neutron addition in a high-neutron-density environment), or by the *p* process that is responsible for the production of proton-rich isotopes. Whereas the solar (normal) isotopic composition represents a mixture of these different components, FUN inclusions show isotopic signatures of individual components. **Figure 8** shows the isotopic ratios of barium, neodymium, and samarium measured in one FUN inclusion, EK 1-4-1. The patterns show excesses in the isotopes that are produced by the *r* process, and for samarium also an excess in a *p*-process isotope. These patterns are compared with the isotopic patterns measured in presolar silicon carbide (SiC) grains (discussed below). The latter are the inverse of the *r*-process patterns of the FUN inclusions; they indicate production by the *s*-process. These complementary patterns are clear evidence that different stellar sources contributed material to the solar system.

Refractory inclusions have also anomalies in the neutron-rich isotopes of elements in the vicinity of iron in the chart of nuclides (iron-peak elements). These isotopes include ^{48}Ca , ^{50}Ti , ^{54}Cr , ^{58}Fe , ^{64}Ni , and ^{66}Zn . Correlated excesses and depletions are observed. Effects are pronounced for FUN inclusions but, for ^{48}Ca and ^{50}Ti , are largest in inclusions containing the refractory mineral hibonite ($\text{CaO}[\text{Al}_2\text{O}_3]_6$). The elements of the iron peak are all produced in supernovae, but the correlated excesses and depletions of the neutron-rich isotopes indicate contributions from different types of these stellar explosions. Different supernova types are distinguished by astronomical observations.

Improvements in analytical capabilities have led to the discovery of numerous isotopic anomalies in different meteorites and in their chemical separates (components obtained through chemical treatment with acids or other solvents). Although nucleosynthetic components for these anomalies have not been identified in all cases, it is clear that different stellar sources produced the elements and that these components were not completely mixed in the solar system.

Presolar dust grains. Whereas all the solids with isotopic anomalies mentioned so far formed in the solar system and contain only surviving traces of presolar isotopic signatures, primitive meteorites contain also true stellar grains. These grains formed in stellar atmospheres, were ejected into the interstellar medium, survived the formation of the solar system, and were incorporated into meteorites. They

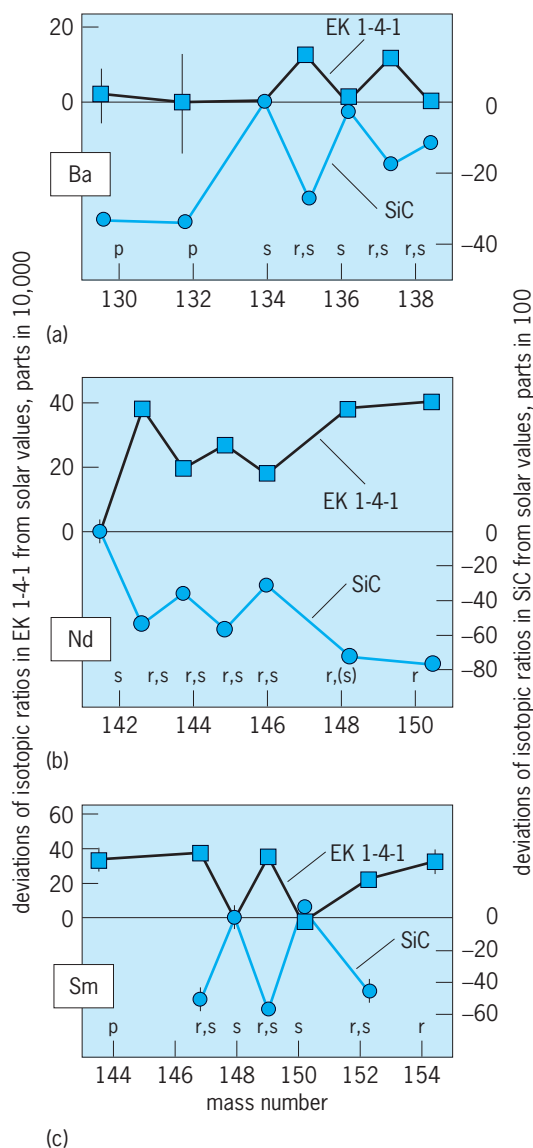


Fig. 8. Isotopic anomalies in (a) barium (Ba), (b) neodymium (Nd), and (c) samarium (Sm) measured in the FUN inclusion EK 1-4-1 and in presolar silicon carbide (SiC). Production of isotopes in the nuclear *s*, *r*, or *p* processes is indicated by letters at the bottoms of the graphs. Whereas EK 1-4-1 exhibits *r*-process patterns, presolar SiC exhibits complementary *s*-process patterns. (Adapted from F. Begemann, *Isotope abundance anomalies and the early solar system: MuSiC vs. FUN*, in N. Prantzos, E. Vangioni-Flam, and M. Cassé, eds., *Origin and Evolution of the Elements*, pp. 517–526, Cambridge University Press, 1993)

can be located in and extracted from their parent meteorites and studied in detail in the laboratory. Their stellar origin is indicated by their isotopic compositions, which are completely different from those of the solar system and, for some elements, cover extremely wide ranges. For example, in Fig. 8 the anomalies carried by presolar silicon carbide are more than a factor of 100 larger than those in the FUN inclusion EK 1-4-1.

The first presolar grains were discovered because they carry anomalous noble gas components in neon (Ne-E, almost pure ^{22}Ne) and xenon (Xe-S, exhibiting an *s*-process pattern in the form of excesses in ^{128}Xe ,

TABLE 4. Presolar grain types

| Grain type | Abundance [†] , parts per million | Size, μm | Stellar sources [‡] |
|---|--|---------------------|------------------------------|
| Silicates in IDPs* | ~400 | ≤ 1 | RG, AGB |
| Silicates in meteorites | 180 | ≤ 0.5 | RG, AGB |
| Spinel (MgAl_2O_4) | 1.5 | 0.15–2 | RG, AGB, SNe |
| Corundum (Al_2O_3) | 0.15 | 0.15–3 | RG, AGB, SNe |
| Nanodiamonds | 1500 | 0.002 | SNe |
| Mainstream SiC | 14 | 0.3–20 | AGB |
| SiC type A + B | 0.5 | 0.5–5 | J stars |
| SiC type X | 0.2 | 0.3–5 | SNe |
| Graphite | 1 | 1–20 | SNe, AGB |
| Nova grains | 0.001 | ~ 1 | Novae |
| Silicon nitride (Si_3N_4) | 0.002 | ≤ 1 | SNe |
| Titanium carbide (TiC) | ~0.001 | 0.01–0.5 | SNe, AGB |

*Silicates are pyroxene [$(\text{Mg,Fe})_2\text{Si}_2\text{O}_6$] and olivine [$(\text{Mg,Fe})_2\text{SiO}_4$]. IDPs = interplanetary dust particles.

[†]Abundances vary with meteorite type. Shown here are the maximum values.

[‡]RG = red giants; AGB = asymptotic giant branch stars; SNe = supernovae; J stars are carbon stars with very low $^{12}\text{C}/^{13}\text{C}$ ratios.

^{130}Xe , and ^{132}Xe ; and Xe-HL, exhibiting excesses in the heavy xenon isotopes ^{132}Xe , ^{134}Xe , and ^{136}Xe , and the light Xe isotopes ^{124}Xe , ^{126}Xe , and ^{128}Xe). These grains turned out to be of carbonaceous nature: diamond, graphite, and silicon carbide. With the exception of diamonds, which are too small, single grains can be analyzed for their isotopic compositions by secondary ion mass spectrometry (SIMS) in the ion microprobe or by resonance ionization mass spectrometry (RIMS), and were found to be anomalous in all their isotopic ratios. This, in addition to the

size of the anomalies, identifies them as samples of stellar material. Subsequently, other presolar grain types, not tagged by exotic noble gases, were identified by isotopic analysis of individual grains in the ion microprobe. See INERT GASES; RESONANCE IONIZATION SPECTROSCOPY; SECONDARY ION MASS SPECTROMETRY (SIMS).

Table 4 lists the types of presolar grains for which isotopic ratio measurements have been obtained. The laboratory study of these grains has developed into a new branch of astrophysics because it provides information on stellar evolution and nucleosynthesis, galactic chemical evolution, stellar atmospheres, conditions in interstellar space, the early solar system, and meteorite parent bodies.

Silicon carbide has been studied in greatest detail because it can relatively easily be isolated from meteorites, many grains are several micrometers in size, and it contains many trace elements. The isotopic compositions of the noble gases from helium to xenon, and of the elements carbon, nitrogen, magnesium, silicon, potassium, calcium, titanium, strontium, zirconium, molybdenum, ruthenium, rubidium, barium, neodymium, samarium, and dysprosium, have been measured in presolar silicon carbide. Figure 9 shows the nitrogen and carbon and Fig. 10 the silicon isotopic ratios of individual silicon carbide grains. These ratios have been used to classify the grains into different types whose relative abundances are indicated in Fig. 9. Different stellar sources have been identified for these grains. Mainstream grains are from carbon stars, asymptotic giant branch stars that became carbon-rich during their evolution. Y and Z grains are likely also from carbon stars but from stars with low metallicity, that is, depleted in the elements heavier than helium. Grains of type X are from supernovae. Proof of such an origin comes from evidence of initial ^{44}Ti in many X grains. Titanium-44, which has a half life of only 60 years and decays into ^{44}Ca , is produced only in supernovae in an interior layer that consists of almost pure ^{28}Si , in agreement with the ^{28}Si excesses (depletions in ^{29}Si and ^{30}Si) found in X grains (Fig. 10). The low $^{12}\text{C}/^{13}\text{C}$ ratios defining A + B grains are observed in J-type carbon stars, but it is presently not

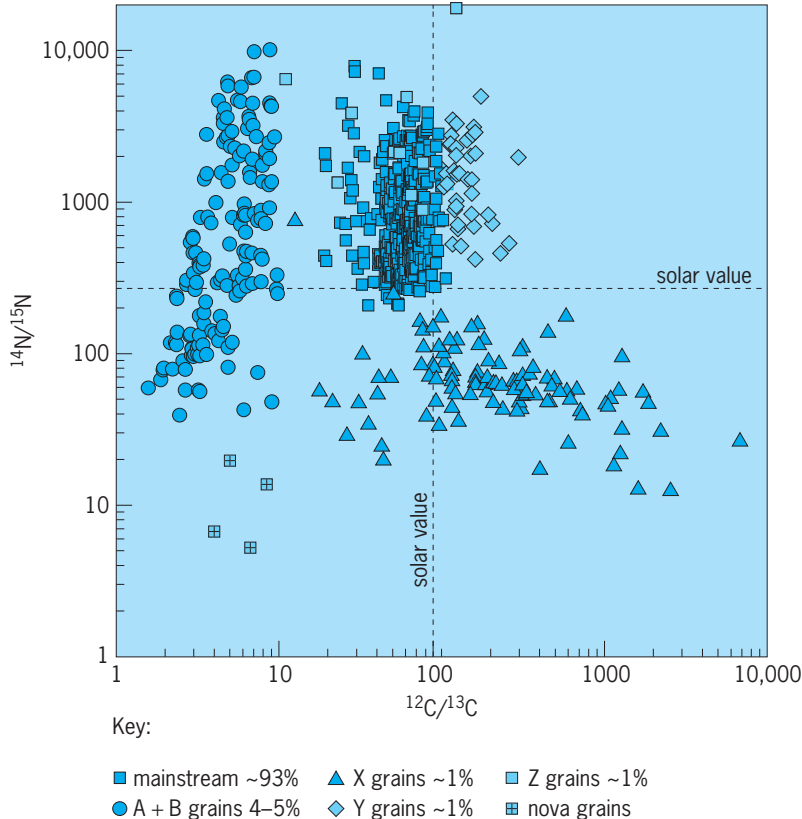


Fig. 9. Nitrogen and carbon isotopic ratios measured in individual presolar SiC grains. The enormous ranges spanned by these ratios are evidence for a stellar, presolar origin of these grains. The nitrogen and carbon isotopic ratios (shown here) and silicon isotopic ratios (shown in Fig. 10) are used to distinguish between different types of grains. The relative abundances of these types are indicated in the key.

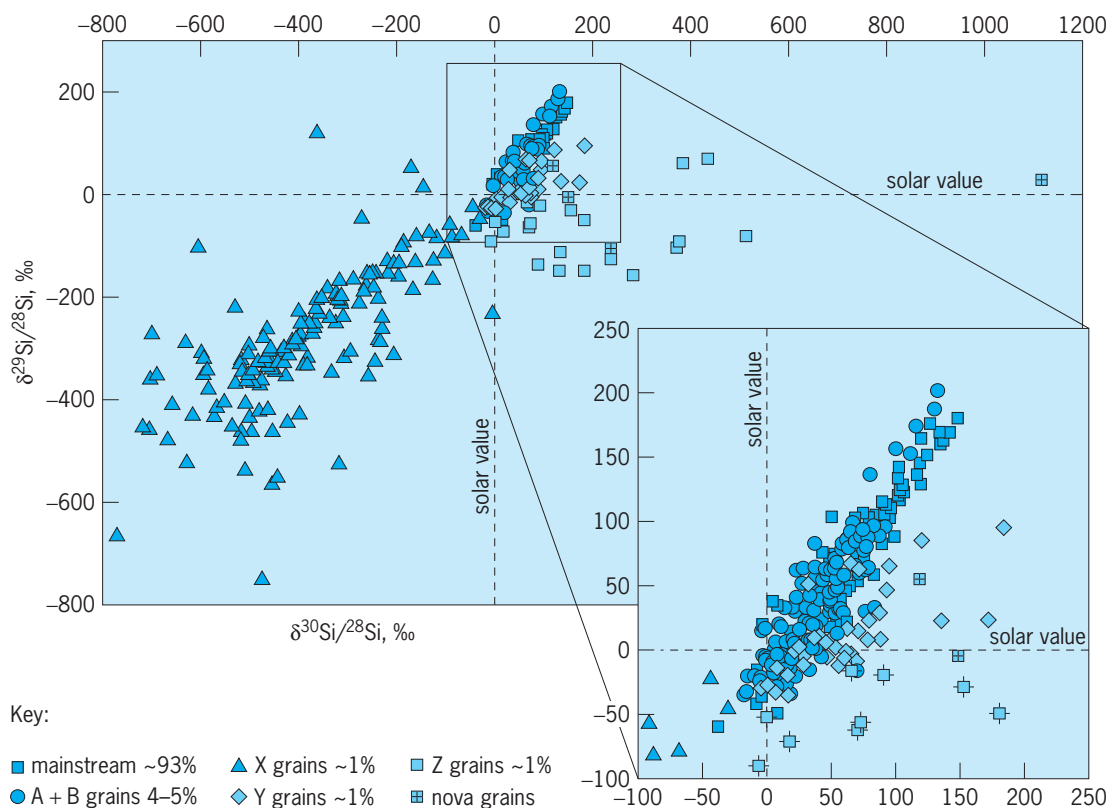


Fig. 10. Silicon isotopic ratios measured in individual presolar SiC grains. The ratios are expressed as delta values, that is, deviations (δ) from the solar values in parts per thousand (‰).

understood how these stars acquired their isotopic compositions. Finally, a handful of silicon carbide grains come from novae, white dwarf stars onto which a companion star dumps material that periodically ignites in a nuclear explosion. *See* CARBON STAR; CATAclysmic VARIABLE; NOVA.

Low-density graphite grains also have large ^{44}Ca excesses from ^{44}Ti decay, indicating a supernova origin. Many of these grains have large ^{18}O excesses, which is also a supernova signature originating from a layer where ^{14}N , the product of hydrogen burning (capture of protons) in the carbon-nitrogen-oxygen (CNO) cycle, is turned into ^{18}O by helium burning (capture of helium nuclei). *See* CARBON-NITROGEN-OXYGEN CYCLES.

Whereas all silicon carbide grains found in primitive meteorites are of presolar origin, this is the case for only a small fraction of oxygen-rich grains. The reason is that in the solar system the abundance of oxygen is higher than that of carbon, leading to the formation of oxygen-rich minerals such as silicates, but not to the formation of carbonaceous phases. As a consequence, oxygen-rich presolar grains have to be identified from isotopic measurements in the ion microprobe. The presolar minerals discovered in this way include corundum, spinel, hibonite, and the silicates olivine and pyroxene (Table 4). Most oxygen-rich presolar grains have excesses in ^{17}O and depletions in ^{18}O . This signature has been explained by hydrogen burning in deep, hot layers during the main-sequence phase of the parent star. After exhaustion of hydrogen in the interior of the star, a

special mixing process, called the first dredge-up, brought material from these layers to the stellar surface, where grains formed during the subsequent red giant phase when the expanding star lost material in the form of a stellar wind. Some oxygen-rich presolar grains have ^{18}O depletions that are much larger than those that can be explained by the first dredge-up. They indicate additional circulation of envelope material to hot regions close to the thin layer where hydrogen burns during the red giant or asymptotic giant branch phase of evolution. This is one example of isotopic measurements of presolar grains in the laboratory leading to new insights into stellar processes.

Cosmogenic nuclides. During travel from their parent bodies to Earth, meteorites are bombarded by galactic and solar cosmic rays. In some cases, when the meteorites were not buried deeply before their ejection, bombardment also took place on the parent asteroids. This bombardment produces various isotopes, called cosmogenic nuclides, through nuclear reactions. Cosmogenic nuclides include both radionuclides and stable nuclides. Their measurement provides information on exposure ages of meteorites in interplanetary space, terrestrial ages (time between Earth entry and collection), burial history on parent asteroids, and preterrestrial meteorite sizes (most stony meteorites break up upon entry into the terrestrial atmosphere). Whereas the concentrations of radionuclides used to be determined by detecting their radioactive decay, nowadays isotopes such as ^{10}Be , ^{26}Al , ^{36}Cl , and ^{53}Mn

are measured by accelerator mass spectrometry. Stable radionuclides are detected as isotopic anomalies. For example, the three neon isotopes ^{20}Ne , ^{21}Ne , and ^{22}Ne are produced by cosmic rays in approximately equal amounts. However, because in the solar system the abundance ratio of $^{21}\text{Ne}/^{20}\text{Ne}$ is only 2.4×10^{-3} , any cosmogenic neon shows up as an excess in ^{21}Ne relative to the other two neon isotopes. Similarly, low-abundance isotopes of other noble gases (for example, ^3He and ^{126}Xe) are identified as cosmogenic nuclides from isotopic anomalies. See COSMOGENIC NUCLIDE.

Ernst Zinner

Meteorite Impact

The process of impact cratering was of fundamental importance for the accumulation of planets in the early solar system, the formation of planetary landscapes, and the Archean geology of the Earth. In addition, meteorite impacts are implicated in the Moon's origin and the extinction of the dinosaurs. However, recognition of the importance of impacts has been achieved only since the 1970s. The dominant role of impact cratering in sculpting the Moon's surface was not widely understood until after the Apollo landings in 1969. The spectacular plumes raised above Jupiter's cloud decks by the impact of fragments of Comet Shoemaker-Levy 9 in July 1994 have led to new investigations of the interaction between meteorites and atmospheres. In July 2005, the National Aeronautics and Space Administration's *Deep Impact* mission created an artificial crater about 200 m (650 ft) in diameter on Comet Tempel 1. See COMET.

Cratering mechanics. The precise outcome of a planetary collision depends on the size of the meteorite and conditions on the target planet. Small meteorites striking planets such as the Earth or Venus dissipate most of their energy in the atmosphere and do not strike the surface at high speed. In general, if the mass of the meteorite is small compared to the mass of atmospheric gases displaced during its entry,

it will not create an impact crater. On airless bodies such as the Moon, there seems to be no lower limit on impact crater size: Craters as small as a few micrometers in diameter have been discovered on the lunar rocks. On Earth, the atmosphere prevents stony meteorites or comets from making craters smaller than a few kilometers in diameter, and even iron meteorites cannot make high-speed impact craters smaller than a few hundred meters in diameter.

Penetration of large meteorites into the Earth's atmosphere is limited by two effects. The less important is slowing by air drag. More important is crushing and dispersion of incoming objects by the large difference in pressure between the front of the fast-moving meteorite and its rear. Objects ranging in size from about 10 cm (4 in) to 100 m (300 ft) in diameter are broken into many fragments in the atmosphere. When they do finally fall to the surface, these fragments form a characteristic "strewn field" of small craters created by the impact of decelerated fragments that fall at their terminal velocity. Craters in strewn fields on the Earth spread over an elliptical region a few kilometers wide and up to 10 km (6 mi) long, with the largest crater at the downrange end. (On Venus, with its much denser atmosphere, the pattern is the same, but these dimensions are multiplied by a factor of 10.) Most meteorites recovered on Earth originally fell in clusters of this type.

When a large meteorite does penetrate a planet's atmosphere, it initiates a series of swift but orderly processes that eventually create a characteristic landform, an impact crater. Three principal stages are recognized in this process.

Stage 1: contact and compression. The meteorite first plunges into the surface rocks at high speed, compressing the underlying rocks and converting its initial kinetic energy into both heat and kinetic energy of the surface rocks. The high pressures produce a series of characteristic mineralogical changes in the surrounding rocks (**Table 5**) that often permit verification of the impact origin of a suspected crater. The

TABLE 5. Petrographic shock indicators

| Material | Indicator | Pressure, GPa |
|-----------------------------------|-------------------------------|---------------|
| Tonalite (igneous rock) Quartz | Shatter cones | 2–6 |
| | Planar elements and fractures | 5–35 |
| | Stishovite | 15–40 |
| | Coesite | 30–50 |
| | Melting | 50–65(?) |
| Plagioclase | Planar elements | 13–30 |
| | Maskelynite | 30–45 |
| | Melting | 45–65(?) |
| Olivine | Planar elements and fractures | 5–45 |
| | Ringwoodite | 45 |
| | Recrystallization | 45(?)–65(?) |
| | Melting | <70 |
| Clinopyroxene | Mechanical twinning | 5–40(?) |
| | Majorite | 13.5 |
| | Planar elements | 30(?)–45 |
| | Melting | 45(?)–65(?) |
| Graphite | Cubic diamond | 13 |
| | Hexagonal diamond | 70–140 |

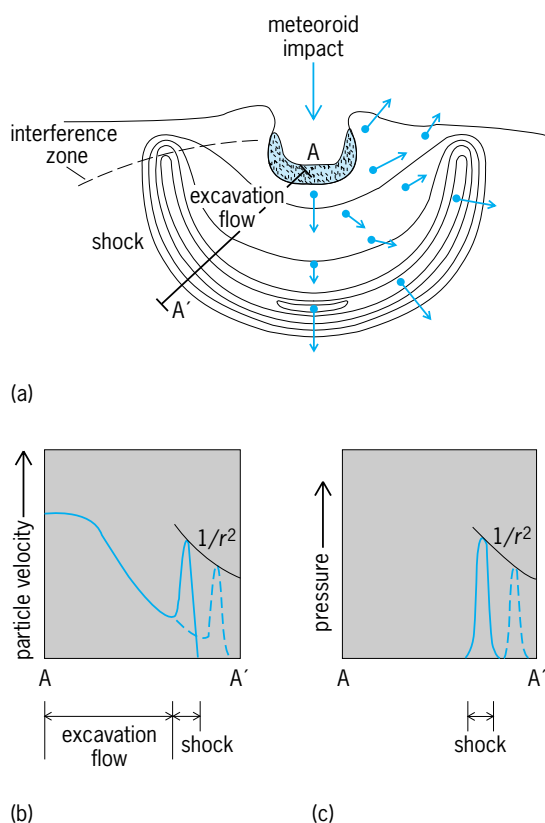


Fig. 11. Cratering by a meteorite. (a) Expanding shock wave and excavation flow following a meteorite impact; contours represent pressure at some particular time after impact, with the region of high shock pressure isolated or “detached” on an expanding hemispherical shell. (b) Profile of particle velocity and (c) profile of pressure along the section AA'. The broken lines show particle velocity and pressure some time later than those shown by the solid lines, and the solid curves connecting the peaks are portions of the “envelopes” of peak particle velocity and peak pressure.

duration of this compression stage is short, however, lasting only as long as it takes the meteorite to travel a distance equal to its own diameter.

Stage 2: excavation. Subsequently, the pent-up pressures in the compressed rocks create an explosion, blasting aside the surrounding rocks as a strong shock wave radiates away from the impact site (Fig. 11). The explosive nature of impacts was first recognized in 1924 by A. C. Gifford, who realized that the kinetic energy per unit mass of a meteorite striking at only 3 km/s (1.8 mi/s) is equivalent to the chemical energy of TNT. Because kinetic energy increases as the square of velocity, faster impacts resemble powerful explosions. This explains why most impact craters are circular, even though the most probable angle of impact is 45° . Only very oblique impacts, at angles of less than about 6° to the surface, produce elliptical craters.

The nearly hemispherical shock wave from the impact expands and weakens as time passes, leaving behind outward-moving rock debris that eventually excavate the crater. The time required for a crater of diameter D to open is given roughly by $(D/g)^{1/2}$, where g is the surface acceleration of gravity. Thus, the 1-km diameter (0.6-mi) Arizona Meteor Crater

(also known as Barringer Crater) was excavated in about 10 s, whereas the 1000-km diameter (600-mi) Imbrium basin on the Moon took about 13 min to open.

Impact craters are not very efficient at excavating deeply buried material. Although most fresh craters have depth/diameter ratios of about 1/5, the maximum depth of excavation is only about 10% of the crater’s diameter, so even the 180-km diameter (110-mi) Chicxulub crater in Yucatan did not excavate rocks from the Earth’s mantle. Most debris ejected from an impact crater falls within about one crater diameter of the crater rim, although small quantities of material may be flung thousands of kilometers, or even entirely off the target planet, as demonstrated by the SNC meteorites that were ejected from the surface of Mars by large impacts.

The immediate result of crater excavation is called the transient crater. This is a relatively deep, steep-walled crater that begins to collapse as soon as it forms. The diameter of a transient crater is a function of the velocity of the impact, the angle between the approaching meteorite and the ground, the diameter of the meteorite, and the density of the projectile and of the target surface. Several websites use well-established mathematical relations among these variables to quickly compute transient crater sizes.

Stage 3: modification. Small transient craters are quickly filled by a lens of broken rock that forms from debris that slides down from the rim and pools at the bottom of the crater. Such bowl-shaped craters floored by broken rock are called simple craters.

In larger craters the floor rises as the rim sinks, producing craters with central mounds that are thinly veneered with broken and melted rock. The rims of such craters are scalloped and terraced with great blocks of slumped rock. Termed complex craters, they form at diameters greater than a certain transition size that depends on the gravity and composition of the target planet. The transition from simple to complex craters occurs at about 3 km (1.8 km) on the Earth and about 15 km (9 mi) on the Moon. Still larger craters exhibit circular mountainous rings instead of central peaks.

The very largest impact structures, particularly on the Moon, are surrounded by inward-facing, roughly circular (but often incomplete) mountain rings that probably formed well outside the crater cavity by a process of inward flow and slumping in the fluid asthenosphere beneath the crater. They are termed multi-ring basins, and the ratio between the diameters of the rings is often said to be close to a multiple of $\sqrt{2}$. Multi-ring basins on the Moon range in size from the 410-km-diameter (250-mi) basin Grimaldi to the 2600-km-diameter (1600-mi) South Polar basin. These enormous structures dominate the Moon’s surface and form the principal stratigraphic markers on that body. A somewhat different variant of multi-ring basin, in which the center is surrounded by dozens of rings, is observed on the Jovian satellites Callisto and Europa, where the impact occurred in a relatively thin water ice layer overlying a liquid water ocean.

Impact cratering and planetary evolution. It is now believed that the planets formed out of the same cloud of gas and dust from which the Sun formed 4.5 billion years ago. Dust settled toward the mid-plane of the early solar nebula and accreted into kilometer-scale clumps by some still poorly understood process. The kilometer-size clumps were large enough to perturb each other gravitationally and collide. These collisions were at first gentle, but became more violent as the objects grew and gravitational perturbations became stronger. The last stages of planetary growth are thought to have been characterized by truly titanic collisions between the early protoplanets and objects up to half their diameter. Such collisions must have caused widespread melting and vaporization of the protoplanets and could have blasted material from the surface of the protoplanet, or the colliding object into orbit. One such collision is thought to have created the Moon, when a Mars-size protoplanet (most of which is now merged with the Earth) struck the protoearth and ejected a Moon-sized quantity of material into close orbit about it. Study of extinct radionuclide chronometers dates this event to about 30 million years after the first meteorites formed. The Moon subsequently receded from the Earth as a result of tidal friction, to take up its present orbit about 60 Earth radii away.

The rain of cosmic debris onto the surfaces of the planets diminished after accretion was largely complete, but did not end completely. Many of the most heavily cratered surfaces in the solar system—on the Moon, the southern highlands of Mars, and Mercury—were created during the era of late heavy bombardment, a time between 4.6 and about 3.8 billion years ago, when the cratering rate was hundreds of times larger than at present. Many lunar geologists believe that most of the craters dating from this era formed during a brief but intense cratering event called the terminal cataclysm about 3.9 billion years ago.

In the present era, comets from the outer solar system and asteroids straying from the main asteroid belt occasionally cross the path of the major planets and collide. There are about 170 impact craters currently recognized on the Earth, and new ones are discovered each year. The current cratering rate for the whole Earth is about 2.5 craters greater than 20 km (12 mi) in diameter per million years.

The most dramatic impact event in recent geologic history occurred 65 million years ago, when a 15-km-diameter (9-mi) asteroid or comet struck what is now the northern Yucatan Peninsula and created the 180-km-diameter (110-mi) Chicxulub crater. The impact sprayed vaporized projectile material to a thickness of a few millimeters over the entire Earth, rained melted rock debris over most of North and South America, initiated massive tsunamis and underwater landslides, ignited global wildfires, and may have produced long-term climatic changes by introducing massive amounts of water, carbon dioxide, and sulfates into the upper atmosphere. These sudden and extreme environmental changes caused a tremendous number of animal species to become extinct: Dinosaurs disappeared forever, along with the

marine ammonites and many other groups. Marine planktonic communities changed drastically, with the disappearance of many species of foraminifera. See EXTINCTION (BIOLOGY); TSUNAMI.

Because at least one major biological extinction was caused by an impact, a vigorous debate has ensued about how large a hazard they pose to humanity. Efforts have begun to detect and chart the orbits and sizes of all asteroids larger than about 1 km (0.6 mi) in diameter that cross the Earth's orbit. To date, about 600 asteroids larger than 1 km in diameter have been charted, and it is estimated that there are another 600 yet to be discovered. Whether or not asteroid or comet impacts are an immediate threat to humanity, it is clear that meteorite impact is far more important than previously supposed for the ancient development of the solar system, the origin of planetary landscapes, and the evolution of life on Earth.

H. J. Melosh

Bibliography. T. J. Bernatowicz and E. Zinner (eds.), *Astrophysical Implications of the Laboratory Study of Presolar Materials*, American Institute of Physics, New York, 1997; A. Bevan and J. R. de Laeter, *Meteorites: A Journey Through Space and Time*, Smithsonian Institution Press, 2002; W. F. Bottke, Jr., et al. (eds.), *Asteroids III*, Arizona University Press, 2002; V. F. Buchwald, *Handbook of Iron Meteorites: Their History, Distribution, Composition and Structure*, 3 vols., University of California Press, 1975; R. N. Clayton, R. W. Hinton, and A. M. Davis, Isotopic variations in the rock-forming elements in meteorites, *Phil. Trans. Roy. Soc. London*, A325:483–501, 1988; A. M. Davis (ed.), *Treatise on Geochemistry*, vol. 1: *Meteorites, Comets, and Planets*, Elsevier, 2004; R. T. Dodd, *Thunderstones and Shooting Stars*, 1986; R. A. F. Grieve, Terrestrial impact: The record in the rocks, *Meteoritics*, 26:175–194, 1991; F. Heide and F. Wlotzka, *Meteorites: Messengers from Space*, Springer-Verlag, New York, 1995; R. H. Hewins, R. H. Jones, and E. R. D. Scott (eds.), *Chondrules and the Protoplanetary Disk*, Cambridge University Press, 1996; R. Hutchison, *Meteorites: A Petrologic, Chemical and Isotopic Synthesis*, Cambridge University Press, 2004; R. Hutchison, *The Search for Our Beginning*, 1983; J. F. Kerridge and M. S. Matthews (eds.), *Meteorites and the Early Solar System*, 1988; D. S. Lauretta and H. Y. McSween (eds.), *Meteorites and The Early Solar System II*, University of Arizona Press, 2006; M. E. Lipschutz and L. Schultz, Meteorites, in P. Weissman, L.-A. McFadden, and T. V. Johnson (eds.), *Encyclopedia of the Solar System*, pp. 629–671, Academic Press, 1998; V. Mannings, A. P. Boss, and S. S. Russell (eds.), *Protostars and Planets IV*, University of Arizona Press, Tucson, 2000; H. Y. McSween, Jr., *Meteorites and Their Parent Planets*, 2d ed., 1999; K. Mark, *Meteorite Craters*, 1987; H. J. Melosh, *Impact Cratering: A Geologic Process*, 1989; O. R. Norton, *Rocks from Space*, 2d ed., Mountain Press Publishing, 1998; O. R. Norton, *The Cambridge Encyclopedia of Meteorites*, Cambridge University Press, 2002; A. E. Rubin, *Disturbing the Solar System*, Princeton University Press, 2002; S. R. Taylor, *Solar System Evolution*, 2d ed., Cambridge University Press, 2001;

J. T. Wasson, *Meteorites: Their Record of Early Solar System History*, 1985; D. E. Wilhelms, *The Geologic History of the Moon*, USGS Prof. Pap. 1348, 1987; D. E. Wilhelms, *To a Rocky Moon*, 1994; B. Zanda and M. Rotaru (eds.), *Meteorites: Their Impact on Science and History*, Cambridge University Press, 2001.

Meteorological instrumentation

Devices that measure or estimate properties of the Earth's atmosphere. Meteorological instruments take many forms, from simple mercury thermometers and barometers to complex observing systems that remotely sense winds, thermodynamic properties, and chemical constituents over large volumes of the atmosphere.

The method for monitoring the atmosphere depends strongly on the application. The atmosphere is observed in order to unravel the mysteries of storms, to describe current weather for aviation operations, to provide warnings of tornadoes, hail, and floods, to predict the occurrence of wind-driven oceanic waves, to estimate monsoon precipitation, to anticipate abrupt climate change, as well as to deal with other issues. Types of instruments include thermometers, barometers, hygrometers, anemometers, rain gages, rawinsondes, radiometers, spectrometers, radars, and lidars.

The atmosphere is observed on time scales from millennia (climate change) to milliseconds (aircraft turbulence) and on spatial scales from planetary waves down to the microstructure of ice crystals and the specks of dust (nuclei) on which ice crystals form. A very important middle scale, or mesoscale, of atmospheric circulations contains organized patches of fair or foul weather. At this scale, circulations are particularly adept at suppressing or producing precipitation, large temperature changes, and rapidly varying winds. Mesoscale weather events include temperature fronts, rain and snow storms, thunderstorms, hurricanes, tornadoes, downbursts, mountain-valley circulations, and sea breezes. *See HURRICANE; MESOMETEOROLOGY; THUNDERSTORM.*

Most weather is confined to the troposphere, which is nominally the lowest 8–16 km (5–10 mi) of atmosphere. Just above the Earth's surface and within the lower troposphere is the planetary boundary layer, with unique properties that require special meteorological instrumentation. Most routine atmospheric observations are made near the surface, and most of these are located over land. To fill the data voids, meteorologists use mobile and remote platforms on which to place instrumentation. Examples include satellites, balloons, crewed aircraft, robotic aircraft, ships, buoys, and even oil rigs. *See TROPOSPHERE.*

Variables, processes, and phenomena. The standard meteorological variables are air temperature, air pressure, humidity, winds, and precipitation. Temperature is a measure of molecular kinetic energy. Air pressure is the columnar weight of atmosphere per unit area extending from the altitude of observa-

tion to the outer limits of the Earth's atmosphere; it is usually expressed in kilopascals (kPa) or columnar inches of mercury (in. Hg). Humidity is the amount of water vapor in the air, expressed in several ways. Absolute humidity is the mass of water vapor per unit volume of air. Specific humidity is a dimensionless mass known as mixing ratio; for example, X grams (lb) of water vapor per Y grams (lb) of air mixture. The most familiar expression of atmospheric water vapor content is relative humidity, the percentage of water vapor in the air compared to the maximum possible. Instruments measure humidity several ways, including absolute, specific, and relative humidity. *See AIR PRESSURE; AIR TEMPERATURE; HUMIDITY.*

Precipitation is measured as a flux of water mass from the atmosphere to the surface. The water-mass flux is referred to as rainfall rate (mm/h or in./h). The density of ice-phase precipitation is highly variable, depending on the conditions of ice-particle growth in clouds and in the air below. Therefore, when precipitation is in the form of snow or hail, the mass flux is weighed and converted to a rainfall depth equivalent. *See PRECIPITATION (METEOROLOGY); SNOW SURVEYING.*

Wind measurements are critical to nearly all meteorological research and societal applications of meteorology. Measurement of the vertical wind speed, combined with information on humidity and temperature, allows computation of cloud condensate production. Estimating vertical air speed is crucial to quantitative forecasts of precipitation. Horizontal wind is much easier to measure than vertical wind, typically being of order 100 times larger. Precise knowledge of the horizontal wind field allows accurate computation of the vertical wind field. *See WIND.*

Beyond these basic variables, there are many other atmospheric quantities of interest which are measured. Among these are gases that vary in concentration, such as carbon dioxide, methane, ozone, oxides of nitrogen and sulfur, and hydrocarbons.

Particle measurements, including dry aerosols and hydrometeors, are an important category of meteorological instrumentation. Such particles affect the fair-weather electric field and determine the properties of clouds. Nearly all cloud droplets and ice crystals begin their growth on aerosols. Only a small fraction of aerosols are active as cloud condensation nuclei or ice nuclei. *See AEROSOL.*

Hydrometeor measurements are at the heart of meteorology. Types of hydrometeors include cloud water droplets, raindrops, individual ice crystals, snowflakes (which are aggregates of crystals), frozen cloud droplets, sleet (which is frozen raindrops), graupel (which are aggregates of frozen cloud droplets), and hail. Measuring the mass of hydrometeors is an extremely important piece of information for reasons related to buoyancy and determining the amount of heat released into the atmosphere by the processes of condensation from vapor to water, freezing from water to ice, and deposition from vapor to ice. *See CLOUD PHYSICS; HAIL; SNOW.*

In weather forecasting and in meteorological

research, the object of individual variable measurements is often the detection, quantification, and understanding of complex phenomena, such as storms, temperature fronts, sea breezes, jet streams, and a host of terrain-induced circulations. To detect and understand the evolution of such phenomena, it is necessary to measure the air motions, kinematics, and to estimate the thermodynamic structure, including temperature deviations, pressure perturbations, and humidity inhomogeneities. When storms produce precipitation, hydrometeor or microphysical measurements complete the characterization of weather systems. The combination of kinematic, thermodynamic, and microphysical information provides the meteorologist with a basis to detect, quantify, understand, and forecast weather events. See WEATHER FORECASTING AND PREDICTION.

Sampling considerations. There is no instrument designed to cover the full range of atmospheric temporal or spatial scales. If the purpose is to detect evidence of abrupt global climate warming, it might be sufficient to place mercury thermometers at 50 representative locations throughout the world and measure temperature at a few standard times per day for 20 years. Choosing representative locations is a challenge because of the effects of urbanization and heat island effects. If the purpose is to directly measure the fluxes of sensible heat inside a thunderstorm, independent measurements of temperature might be made from several research aircraft every 0.01 s for a period of 1–2 h. Both objectives require about one million measurements of temperature with similar accuracy and precision. Since the frequency, spacing, duration, and sampling domains differ radically, so do the measurement technologies. See URBAN CLIMATOLOGY.

Some research measurements are best conducted under highly controlled conditions in the laboratory, such as detailed microphysical measurements of water and ice nucleation, vapor deposition growth of ice crystals, and rime growth of graupel from the collection of supercooled cloud droplets. Other characteristics such as hydrometeor terminal fallspeed, orientation, and binary interactions among particles are also measured in laboratories. Slowly reacting trace gases and aerosol distributions are spectroscopically analyzed in the laboratory after air samples are collected from research aircraft.

Instruments, observing systems, and platforms. A few broad categories of apparatus are traditional and modern immersion devices and passive and active remote sensors. Traditional immersion devices make contact with the air and are nearly always mechanical. They rely on the expansion properties of materials, on the wind pushing parts, and on mechanical or volumetric measurement of mass. Modern immersion instruments are less mechanical and typically rely on the electrically resistive, capacitive, and radiative properties of sensor materials and of the atmosphere itself (Fig. 1). Passive remote-sensing devices usually receive naturally emitted electromagnetic radiation from the atmosphere. Most often, these instruments are referred to as radiometers. Active remote-sensing devices, such as radars, transmit



Fig. 1. Portable, automated weather station which operates on solar power. Propeller anemometers are at the top of the mast (10 m or 33 ft). The thermometer/psychrometer is located in a beehive radiation shield. Sensor electronics, including aneroid barometer, are in the white box. The geostationary satellite data uplink antenna is above the solar panels. (National Center for Atmospheric Research)

electromagnetic or acoustic radiation at specific frequencies. After a brief delay to allow for electromagnetic wave propagation, these instruments receive a small fraction of the transmitted energy, reflected, refracted, or scattered by the atmosphere. The intensity, frequency, phase, and polarization of the received energy reveal various microphysical, kinematic, and thermodynamic properties of the atmosphere. See ACOUSTIC RADIOMETER; METEOROLOGICAL RADAR; METEOROLOGICAL SATELLITES.

Traditional weather station measurements provide a description of conditions near the ground. In addition to the average regional conditions, these measurements also provide local information on mesoscale phenomena such as cold fronts, sea breezes, and disturbed conditions resulting from nearby thunderstorms. Traditional thermodynamic instruments are mechanical or heat-conductive devices relying on the expansion and contraction of metallic and nonmetallic liquids or solid materials as a function of temperature, pressure, and humidity. Among these are the mercury, alcohol, and bimetallic thermometers for measurement of temperature, mercury and metallic bellow (aneroid) barometers for measurement of pressure, human hair hygrometers, and wet/dry-bulb thermometers (called psychrometers) for measurement of relative humidity. Mercury barometers are simply weighing devices that balance the mass of the atmospheric column against the mass of a mercury column. On average, a column of atmosphere weighs the same as 76 cm (29.92 in.) of mercury. Psychrometers measure humidity by means of the wet-bulb depression technique. A moist thermometer is cooled by evaporation when relative humidity is less than 100%. The

temperature difference between wet and dry thermometers is referred to as the wet-bulb depression, a well-known function of relative humidity at standard airflow speeds. A related method of humidity measurement is the chilled mirror technique (dewpointer). A polished surface is cooled to the temperature of water vapor saturation, at which point the cooled surface becomes fogged. Dewpoint saturation uniquely defines humidity at a known temperature and pressure. *See* BAROMETER; DEW POINT; HYGROMETER; PSYCHROMETER; TEMPERATURE MEASUREMENT; THERMOMETER.

Modern in-place thermodynamic measurements usually employ electrical responses to temperature and humidity. A so-called thermocouple naturally generates electric current in a loop between the junctions of two dissimilar metals at two locations of different temperature. The temperature dependence of electrical resistance in metals, such as platinum, or semiconductors is often employed for accurate and fast measurement of temperature. Because water vapor is absorbed by porous materials, electrical resistance or capacitance of substances, such as carbon or thin film polymers, is a strong function of relative humidity. Increasingly, pressure is measured by thin silicon piezo-resistive membranes, which stretch in response to pressure change, thereby changing their electrical resistance properties. Quasi-spectroscopic methods are employed for very fast response applications in measurement of water vapor fluxes, density, and temperature. Energy from calibrated sources of infrared or ultraviolet radiation are absorbed by water vapor along short transmission paths. The energy losses are in direct proportion to the number of water vapor molecules in the path, and these are measured at the receiver. *See* THERMOCOUPLE.

Traditional precipitation measurement devices may be described as precision buckets, which measure the depth or weight of that which falls into them. These gages work best for rainfall, but they are also used in an electrically heated mode for weighing snow. Rulers are routinely used for measurement of snow depth. Time-resolved measurements of rainfall are traditionally made by counting quantum amounts (0.01 in. or 0.25 mm) of rain with a small, mechanically controlled tipping bucket located beneath a large collecting orifice. Modern rain measuring is sometimes performed along short paths via drop-induced scintillations of infrared radiation, which is emitted by a laser. When the raindrop size distribution is needed, optical-shadowing spectrometers are employed, as are momentum-measuring impact distrometers, devices that measure the number density versus the size distribution of raindrops or other hydrometeors (**Fig. 2**). *See* SNOW GAGE.

Traditional wind measurements are performed by anemometers, some of which use wind-driven spinning cups for wind speed determination. Vanes are used in conjunction with cups for indication of wind direction. Alternatively, three-axis propeller anemometers may be employed to provide orthogonal components of the three-dimensional wind vector. Many hybrids of these basic approaches continue to be successfully employed. Fast-response



Fig. 2. Hydrometeor optical spectrometers and temperature sensors on a research aircraft. Optical spectrometers use lasers to create particle shadow images in two dimensions. Particles are imaged as they flow between the blunt arms ahead of the electronics pod beneath the wing. (*University of Wyoming*)

sonic anemometers employ ultrasound transmission, where the apparent propagation speed of sound is measured (**Fig. 3**). The difference between this measured speed and the actual speed for a fluid at rest is the wind speed. Such measurements are made on a time scale 0.01 s and are used to determine the fluxes of momentum, water vapor, sensible heat, and other scalars in the planetary boundary layer. *See* ANEMOMETER; WIND MEASUREMENT.

Balloon-borne vertical profiles or soundings of temperature, humidity, and winds are central to computerized (numerical) weather prediction. Such observations are made simultaneously or synoptically worldwide on a daily basis. The temperature and humidity sensors are lightweight expendable versions of traditional surface station instruments. Balloon drift during ascent provides the wind measurement. The preferred method of tracking these rawinsondes is to use global navigation aid systems such as Omega, Loran-C, and the Global Positioning System.

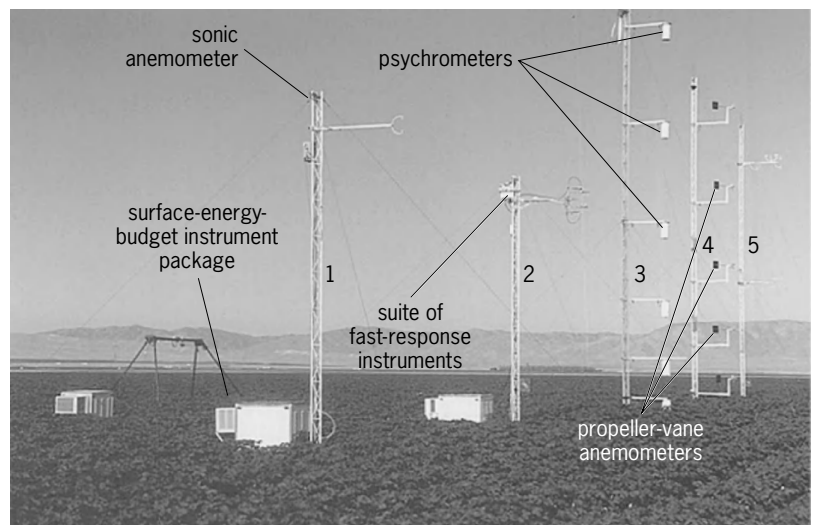


Fig. 3. Atmospheric turbulent energy exchange research facility. Tower 3 makes psychrometric profiles of mean temperature and humidity. Tower 4 measures mean wind profiles with propeller vanes. Towers 1, 2, and 5 support sonic anemometers for high-speed turbulent wind components together with various fast scaler measurements, including temperature, humidity, O_3 , and CO_2 . The black stand in the background supports visible/infrared radiation measurements. (*After W. F. Dabberdt et al., Atmosphere-surface exchange measurements, Nature, 260:1472-1481, 1993*)

Parachute-borne dropsondes are often released from aircraft in data-sparse regions. *See* LORAN; SATELLITE (ASTRONOMY).

Remote sensing. Remote sensing, principally via electromagnetic radiation, is a mainstay of modern meteorology. Such devices typically operate in the optical, infrared, millimeter-wave, microwave, and high-frequency radio regions of the electromagnetic spectrum.

Passive radiometers typically operate at infrared and microwave frequencies; they are used for estimates of temperature, water vapor, cloud heights, cloud liquid water mass, and trace-gas concentrations. These observations are made from the ground, aircraft, and satellites, usually measuring naturally emitted radiation. Passive instruments are especially effective in Earth orbit because electrical power consumption is low and global coverage is unique. Some radiometers simply estimate columnar amounts of water vapor averaged over the whole troposphere. Multiple-frequency infrared and millimeter-wave radiometers can resolve smoothed profiles of water vapor. More complex radiometers are able to scan and image a horizontal field of view. Advanced radiometers are able to resolve radiation over narrow frequency bands that correspond to molecular absorption lines for various gases, including water vapor, ozone, and methane. Other radiometers can receive radiowaves from human sources that are transmitted by the Global Positioning System of satellites to estimate temperature and water vapor from wave propagation characteristics such as signal phase and angles of refraction over known distance. Microwave radiometers can be tuned to oxygen

absorption lines to measure temperature profiles, because the mixing ratio of oxygen in the free troposphere is essentially uniform and known with high precision. Imaging optical radiometers are more commonly known as cameras, and these are routinely mounted on satellites to track the movement of clouds and storm systems. *See* SATELLITE NAVIGATION SYSTEMS.

Radarlike, active remote-sensing devices are among the most powerful tools available to meteorology. Collectively, these instruments are capable of measuring kinematic, microphysical, chemical, and thermodynamic properties of the troposphere at high spatial and temporal resolution. Active meteorological remote sensors are principally deployed on land, ships, and aircraft platforms, as well as aboard satellites. Unlike passive instruments, active remote sensors can precisely resolve the distance at which a measurement is located.

At optical frequencies, lidars measure conditions in relatively clear air. Capabilities include determining the properties of tenuous clouds; determining concentrations of aerosol, ozone, and water vapor; and measuring winds through the Doppler frequency-shift effect. Millimeter-wave radars are used to probe opaque, nonprecipitating clouds. Polarimetric and Doppler techniques reveal hydrometeor type, water mass, and air motions. *See* LIDAR.

The best-known meteorological remote sensor is the microwave weather radar. In addition to measuring rainfall and tracking movement of storms, powerful and sensitive meteorological radars can measure detailed flow fields in and around storms by using

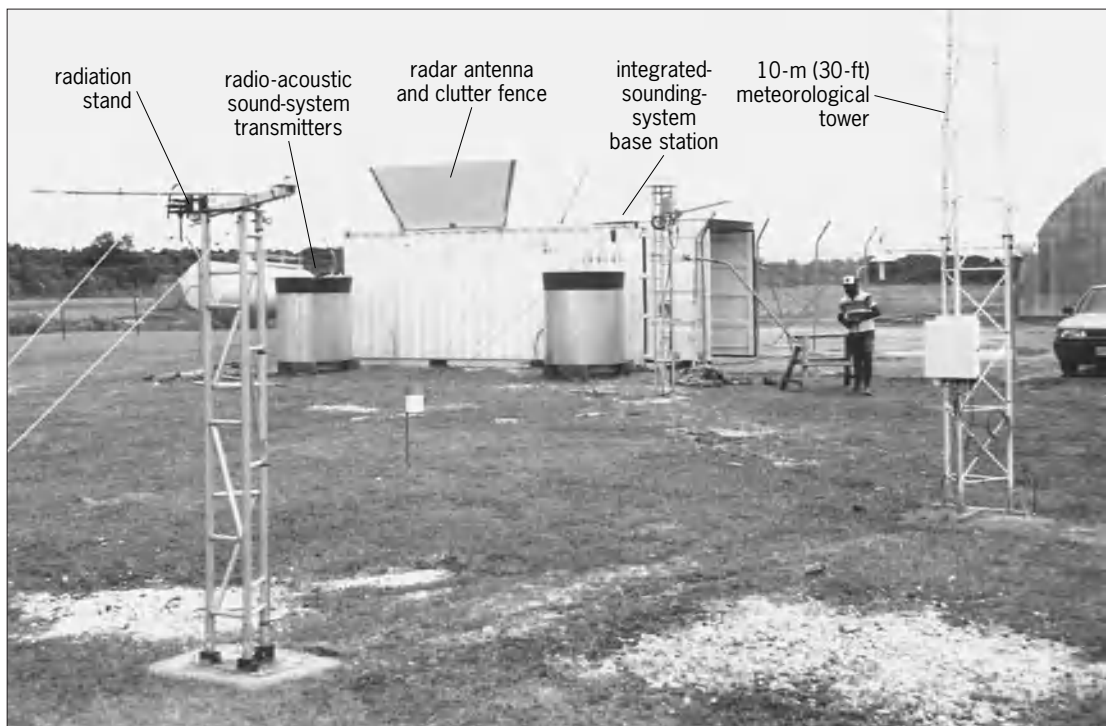


Fig. 4. Integrated sounding/profiling system which contains an ultrahigh-frequency radar wind profiler, acoustic transmission for radio-acoustic air density profiling, a conventional Omega balloon sonde capability (in seatainer), and tower-mounted meteorological instruments. (After D. W. Parsons et al., *The sounding system: Description and preliminary from TOGA COARE*, *Bull. Amer. Meteorol. Soc.*, 75:553–567, 1994)

hydrometeors, insects, and blobs of water vapor as reflective targets. Employing polarimetric methods, these radars can also distinguish between rain, hail, and snow. When Doppler measurements are combined with the atmospheric equations of motion, thermodynamic perturbation fields, such as buoyancy, are revealed inside violent convective storms. At ultrahigh and very high radio frequencies, radars known as wind profilers measure the mean wind as a function of height in the clear and cloudy air (Fig. 4). Superior to infrequent weather balloons, radio wind profiling methods permit continuous measurement of winds with regularity and high accuracy. When radio wind profilers are colocated with acoustic transponders, the speed of sound is easily measured through radar tracking of the acoustic wave. This permits the computation of atmospheric density and temperature profiles, on which the speed of sound is strongly dependent. See DOPPLER RADAR; METEOROLOGY; RADAR METEOROLOGY; REMOTE SENSING.

Richard E. Carbone

Bibliography. D. Atlas (ed.), *Radar in Meteorology*, 1990; W. F. Dabberdt et al., Atmosphere-surface exchange measurements, *Nature*, 260:1472-1481, 1993; T. P. DeFelice, *An Introduction to Meteorological Instrumentation and Measurement*, 1997; D. H. Lenschow (ed.), *Probing the Atmospheric Boundary Layer*, 1986; D. W. Parsons et al., The integrated sounding system: Description and preliminary observations from TOGA COARE, *Bull. Amer. Meteorol. Soc.*, 75:553-567, 1994; S. Raghavan, *Radar Meteorology*, 2003; R. Ware et al., GPS sounding of the atmosphere from low Earth orbit: Preliminary results, *Bull. Amer. Meteorol. Soc.*, 77:19-40, 1996.

Meteorological optics

The study of optical phenomena occurring in the atmosphere. Many light effects can be seen by looking skyward, and all of them, resulting from the interaction of light with the atmosphere, lie in the province of atmospheric optics or meteorological optics. The subject also includes the effect of light waves too long or too short to be detected by the human eye—light-type radiation in the infrared or ultraviolet regions of the spectrum. Light interacts with the different components of the atmosphere by a variety of physical processes, the most important being scattering, reflection, refraction, diffraction, absorption, and emission. Some other processes involving photochemistry and ionization are not considered in this article. See ATMOSPHERE; OPTICS.

Scattering. An observer of light in the sky, while looking in a direction away from the Sun, will see evidence that some process has changed the path of the sunlight to direct it to the observer's eye. For most of the light in the sky this process is that of scattering, by which some of the incident light is sent off in all directions. Scattering by dust in the air makes visible the beam of light coming into a room through a window. Scattering by dust particles or liquid droplets in the atmosphere can add

to the brightness of the sky, but the sky would appear light even if the atmosphere were free of all such particles. On a submicroscopic scale, air is not a continuous, uniform fluid but is composed of molecules, which are smaller than the wavelength of visible light by a factor of about a thousand. Such small particles scatter light but do not scatter all wavelengths with equal efficiency. They scatter the short light waves (blue) more strongly than the long waves (red), with the result that the clear sky appears blue.

The Sun's disk, seen high in the sky on a clear day, appears white. As it moves closer to the horizon, its color changes to yellow, and to orange, and perhaps to red. When the low Sun is observed, the light rays that reach the eyes travel a much longer path through the Earth's atmosphere than when the Sun is overhead. Throughout this path, the shorter (blue) wavelengths are selectively scattered in all directions, with the result that the unscattered light that reaches the observer from the Sun is depleted in the blue end of the spectrum and appears orange or red. Thus the blue sky and the red setting Sun are both consequences of the same scattering process.

Larger particles in the atmosphere scatter light more strongly than the gas molecules. If they are of sizes less than the wavelength of visible light, they will selectively scatter the shorter, visible wavelength and contribute to red sunsets. Smoke from large forest fires or fine ejected material from volcanic eruptions can enhance the colors of sunsets thousands of miles away. In fact, following major volcanic explosions, sunsets may be enhanced over the entire Earth as fine ejected material mixes with the atmosphere and circles the Earth with a fallout time of up to a few years.

White clouds are composed of either small transparent water droplets or small transparent ice crystals, but cloud particles are small on a scale different from the scatterers of the blue sky. The droplets are typically 50 times larger than the wavelength of visible light, and so would be considered large particles in light scattering. Such large particles scatter light of all visible wavelengths equally well so that the scattered light would have the same spectral distribution as the incident sunlight, which is known as white light. Even if there is some wavelength dependence of the scattered light from one particle, multiple scattering from many particles in a cloud of the appropriate thickness will result in the scattered light appearing white. If the atmosphere of the Earth were made thicker (that is, if air were added), the sky would appear whiter. This effect can be seen even on a clear day: the sky near the horizon is whiter than the sky overhead. There is more air along a sighting path near the horizon than along a vertical path through the atmospheric layer.

The blue color of distant mountains, hazes, or smogs demonstrates the wavelength dependence of light scattering from small particles. Because red or infrared radiation is scattered less by some hazes, red filters or infrared-sensitive film or detectors are sometimes used in aerial photography or satellite image recording for better penetration of haze layers.

See AERIAL PHOTOGRAPH; REMOTE SENSING; SCATTERING OF ELECTROMAGNETIC RADIATION.

Reflection. When light encounters the smooth surface of an ice crystal, some of the light is reflected from the surface and some is transmitted. Sometimes many small ice crystals grow in the atmosphere and slowly fall through the air. Their presence is seen as clouds or as a haze or (ice) fog. The very complicated, many branched ice crystals known as snowflakes usually occur at ground level in temperate regions; simple ice-crystal forms can be found at higher elevation in the sky.

As hexagonal flat-plate crystals fall through the air, they tend to orient with their wide, flat surfaces nearly horizontal. Reflection of sunlight from the nearly horizontal surfaces of such crystals can give rise to a vertical column of light that appears above or below the Sun (or both), when the Sun is near the horizon. These sun pillars are similar in origin to the streak of sunlight (called a glitter path) reflected from the slightly rough surface of a lake or ocean.

Sun pillars can also be produced by hexagonal-column crystals that have the shape of a wooden pencil (before sharpening). Such crystals tend to fall with their long axes horizontal, and light reflection from the long side faces produce sun pillars. Reflection from the vertically oriented faces of falling crystals (the side faces of falling plates or the end faces of falling pencils) can result in the parhelic circle, a band of light parallel to the horizon, passing through the Sun and extending all the way around the sky. Many other effects can result from a combination of reflection and refraction in falling ice crystals or raindrops.

A particular form of a sun pillar, called the subsun, can be seen when looking down from an airplane flying over a layer of ice-crystal clouds (see **illustration**). When the Sun is high in the sky and the crystals are well oriented, a somewhat elongated, bright spot can be seen. It is just a reflected image of the Sun in the nearly horizontal ice-crystal surfaces. See REFLECTION OF ELECTROMAGNETIC RADIATION.

Refraction. When a ray of light passes from one transparent medium to another, it is refracted at the surface, that is, its direction changes abruptly. Light from the Sun is thus refracted as it enters an ice crystal and again as it leaves. As a result of these two refractions, light deviates from its original direction. For a hexagonal ice crystal the deviation can bring sunlight to an observer's eye from different directions, producing a variety of spots, arcs, or streaks of light intensity in the sky. Sun dogs and the 22° halo are two of the more commonly observed effects that result from light refracted by falling ice crystals. Light that enters an ice crystal and is internally reflected from one or more crystal faces before emerging can produce a wide variety of halo effects that are observable with the naked eye. See HALO; SUN DOG.

In the case of water, where surface tension acts as an elastic skin, squeezing a small falling drop into a spherical shape, light rays that enter the transparent



The subsun, resulting from sunlight reflected off the nearly horizontal faces of falling flat-plate ice crystals. (From R. Greenler, *Rainbows, Halos, and Glories*, Cambridge University Press, 1980)

sphere are reflected internally before they emerge to produce a rainbow. See RAINBOW.

When a light ray passes through air whose density is not uniform but changes gradually with position, the ray gradually changes its direction, traveling in a smooth curve. The normal variation in atmospheric pressure (and hence of air density) with height causes stars near the horizon to appear elevated above their true position. The curved paths of the light rays enable observation of the Sun, apparently sitting just above the horizon, when it actually is located geometrically just below the horizon. The density of air also changes as a result of the air temperature. Temperature variation of air near the Earth's surface can produce a number of optical distortions that are referred to as mirages. Small-scale variations and temporal fluctuations of air density resulting from air turbulence or temperature variations result in the twinkling of stars or the shimmering of distant scenes. See MIRAGE; REFRACTION OF WAVES; TWINKLING STARS.

Diffraction. Diffraction depends on both the wavelength of the light and the size of the particles. If the Moon is viewed through a thin cloud containing water droplets or small ice crystals, it seems to be surrounded by colored rings. The rings are a diffraction effect, with the long (red) wavelengths giving rise to a larger set of rings than the shorter (blue) wavelengths. The resulting display is called the corona. A similar display, called the glory, can be seen by looking in exactly the opposite direction from the Sun

or Moon; an observer flying in the sunlight over a cloud layer can see the colored rings of glory around the shadow of the airplane on the clouds below. See DIFFRACTION; SOLAR CORONA.

Emission and absorption. Small particles in the air, and the gas molecules that constitute the air, can absorb and emit light. In a heavily polluted atmosphere the color of the sky may be affected by the selective color absorption of smoke particles, but in general there is not much absorption or emission of visible light in the atmosphere. The absorption and emission of infrared radiation, however, are very important processes in establishing the temperature of the Earth's atmosphere and the Earth's surface. See ABSORPTION OF ELECTROMAGNETIC RADIATION; AIR POLLUTION; GREENHOUSE EFFECT. Robert Greenler

Bibliography. C. F. Bohren and D. R. Huffman, *Absorption and Scattering of Light by Small Particles*, 1983; K. L. Coulson, *Polarization and Intensity of Light in the Atmosphere*, 1988; R. Greenler, *Rainbows, Halos, and Glories*, 1989; K. N. Liou, *Radiation and Cloud Physics Processes in the Atmosphere*, 1992; M. Minnaert, *Light and Colour in the Open Air*, 1954; W. Tape, *Atmospheric Halos*, 1994; W. Tape and J. Moilanen, *Atmospheric Halos and the Search for Angle X*, 2005; R. A. R. Tricker, *Introduction to Meteorological Optics*, 1970.

Meteorological radar

A remote-sensing device that transmits and receives microwave radiation for the purpose of detecting and measuring weather phenomena. Radar is an acronym for radio detection and ranging. Today, many types of sophisticated radars are used in meteorology, ranging from Doppler radars, which are used to determine air motions (for example, to detect tornadoes), to multiparameter radars, which provide information on the phase (ice or liquid), shape, and size of hydrometeors. Airborne Doppler radars play a vital role in meteorological research. Additionally, weather radar is now orbiting the Earth on the NASA *Tropical Rainfall Measuring Mission* satellite, launched in November 1997. Radars are also used to detect hail, estimate rainfall rates, probe the clear-air atmosphere to monitor wind patterns, and study the electrification processes in thunderstorms that generate lightning discharges.

Doppler radar. Commonly used, pulsed Doppler radar operates in the microwave region, with standard wavelengths of 10, 5, and 3 cm, referred to as S-, C-, and X-band radars, respectively. The electromagnetic radiation is focused into a narrow beam by illuminating a parabolic dish reflector with microwave energy provided by the radar transmitter. S-band radars require the use of large antennas (**Fig. 1**) to generate a narrow beam of microwave energy; transmit high power (peak power of 1 megawatt); and suffer relatively little attenuation as the radar beam passes through regions of heavy rain and hail. X-band radars use much smaller antennas to achieve similar narrow beams, and are highly

portable. However, X-band radars suffer from attenuation when used to probe precipitation, which significantly limits their range. Attenuation results when the radar energy is either absorbed by the raindrops or reemitted from the raindrops in directions other than toward the radar. See ANTENNA (ELECTROMAGNETISM); DOPPLER RADAR; MICROWAVE.

A pulsed Doppler radar typically emits 1000 electromagnetic pulses per second. These individual pulses are typically 1 microsecond (10^{-6} s) in duration. The pulsing rate defines the pulse repetition frequency (PRF) of the radar, which in turn determines its maximum unambiguous range. For a PRF of 1000 Hz, the maximum range is 150 km (90 mi). Strong echoes beyond this range can still be observed, but are said to be range-folded. The Doppler radar also provides information on the target's velocity, either toward or away from the radar when viewed along the radar beam. The Doppler shift, which is measured as a small difference between the frequency of the transmitted pulse and the frequency of the energy backscattered to the radar, provides a measure of the scatterer's radial motion. The Doppler shift frequency Δf is given by Eq. (1),

$$\Delta f = 2v/\lambda \quad (1)$$

where λ is the wavelength of the radar expressed in meters and v is the radial velocity. Scatterers in the case of meteorological radar include raindrops, ice particles (snowflakes), hailstones, and even insects, providing clear air returns. A Doppler radar

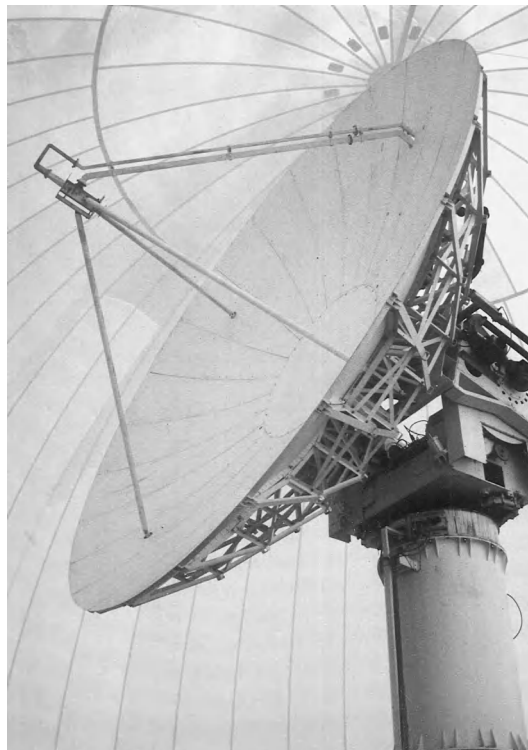


Fig. 1. Parabolic dish antenna of the CSU-CHILL 11-cm multiparameter Doppler radar operating at Colorado State University. The antenna is housed within a large, inflatable radome. (P. Kennedy, Colorado State University)

also detects the amplitude of the backscattered signal, which can be used as a measure of storm intensity and as a means of estimating rainfall rates. When the radar wavelength is large compared to the diameter of the meteorological targets, Rayleigh scattering conditions occur, and thus the energy backscattered from a raindrop or ice particle is proportional to the sixth power of the particle diameter. Under Rayleigh scattering conditions, the reflectivity factor Z is defined by Eq. (2), where $N(D)\Delta D$ is the

$$Z = \sum_0^{\infty} N(D)D^6 \Delta D \quad (2)$$

concentration of scatterers (m^{-3}) and D is the particle diameter expressed in millimeters (mm^6). See PRECIPITATION (METEOROLOGY); SCATTERING OF ELECTROMAGNETIC RADIATION; STORM; STORM DETECTION.

Airborne Doppler radar. In the early 1980s, researchers began using aircraft equipped with Doppler radar to study storms. Several advantages are provided by airborne Doppler systems compared to surface-based systems, including portability and the immediate integration of other meteorological sensors with the Doppler data, thereby providing comprehensive sampling of storms. However, because of limitations of weight and antenna size, airborne Doppler radars operate at wavelengths of 3 cm or smaller. Since attenuation is significant at these ranges, airborne Doppler radars provide usable data out to about 50 km (30 mi). Currently, airborne Doppler radars are used for meteorological research on aircraft operated by the National Oceanic and Atmospheric Administration (NOAA), NASA, the National Center for Atmospheric Research (NCAR), and the University of Wyoming (the last two facilities are under the sponsorship of the National Science Foundation). NCAR maintains a sophisticated airborne Doppler radar, known as ELDORA (Electra Doppler Radar). ELDORA actually consists of two complete radars (each transmitting five different frequencies) driving a dual-antenna system to direct radar beams both forward and rearward of a line perpendicular to the aircraft's flight path. The antennas are housed within a large, rotatable radome affixed to the tail of the aircraft (Fig. 2). These forward- and rearward-looking beams intersect at



Fig. 2. Electra aircraft; the antenna radome. (R. Carbone, National Center for Atmospheric Research)

various distances from the aircraft, and at these points of intersection, both the east-west (u) and north-south (v) components of the horizontal wind can be determined. When the horizontal gradients of u and v are combined with the equation of mass continuity, the vertical wind speed may also be estimated. In this way, a single airborne radar can provide information on the three-dimensional airflow and precipitation structure in storms. The resolution of the ELDORA radar is about several hundred meters, providing unprecedented views of tornadoes and other severe weather.

Multiparameter radar. During the late 1970s and early 1980s, researchers began to experiment with more advanced Doppler radar systems involving dual-wavelength and polarimetric techniques. Dual-wavelength radar transmits electromagnetic energy at two wavelengths, and it also receives energy at both wavelengths. Typically, S- and X-band wavelengths are used. Dual-wavelength techniques were originally proposed to detect large hail. At S-band, hail is usually a Rayleigh target, whereas at X-band, hail is considered a Mie scatterer. Since radar energy is scattered in various directions by a Mie target, the power returned at the X-band wavelength is reduced relative to that at S-band. The presence of large hail is interpreted on the basis of the ratio of backscattered power at X-band to that at S-band. Dual-wavelength techniques are also used to estimate rainfall rates by comparing the backscattered power at a nonattenuating wavelength (S-band) to that at attenuating wavelengths (X-band). See HAIL.

Considerable research is being done with radars that use polarization diversity. Dual-polarization radar is able to transmit and receive both horizontally and vertically polarized radiation (the polarization of a radar beam is defined by the orientation of the electric field vector that comprises an electromagnetic wave). These radars are now used in meteorological research and have largely superseded dual-wavelength radars. A suite of multiparameter variables is being used to infer information on particle phase (ice or water), size, orientation, and shape. One key variable is the differential reflectivity Z_{dr} (decibels, dB), defined by Eq. (3),

$$Z_{dr} = 10 \log_{10}(Z_{hh}/Z_{vv}) \quad (3)$$

where Z_{hh} is the reflectivity for horizontal transmit to horizontal receive and Z_{vv} is the reflectivity for vertical transmit to vertical receive. For raindrops larger than 1 mm in diameter (D), aerodynamic drag forces act to distort the drops into flattened spheres. Since the reflectivity is proportional to D^6 , a flattened raindrop is larger, and therefore backscatters more power, when viewed with horizontally polarized radiation compared to vertical radiation. For raindrops 4 mm (0.16 in.) in diameter, the ratio of the horizontal axis to the vertical axis is about 1.2, which results in a differential reflectivity of approximately 5 dB. Ice particles such as graupel and hail do not deform, and therefore $Z_{dr} = 0$. By examining reflectivity and

differential reflectivity simultaneously, it is possible to distinguish hail from heavy rain, and locate water and ice regions within storms. See POLARIZATION OF WAVES.

Another important multiparameter variable is the differential phase Φ_{dp} , and the gradient in range of differential phase, K_{dp} , which is called the specific differential phase. Differential phase is large when flattened targets are present. Differential phase is a measure of the phase shift (measured in degrees) between horizontal and vertical polarization as the waves pass through a region containing flattened particles (such as raindrops). The specific differential phase (degrees per kilometer) is used to provide accurate estimates of intense rainfall rates in storms containing both rain and hail. Since the hail can dominate the reflectivity signature by virtue of its large size, reflectivity alone can overestimate rainfall rates. However, hail does not contribute to differential phase due to its quasi-spherical nature (only oriented particles or flattened raindrops contribute to differential phase shifts). Irregularly shaped hail tumbles as it falls, and appears to the radar as if it were spherical. Multiparameter radars provide a powerful means to study cloud, precipitation, and electrification processes in storms.

Space-borne radar. Radar operating at a wavelength of 2 cm is in low Earth orbit (350 km or 210 m above the surface) and is used for mapping tropical precipitation. Understanding the amount and distribution of tropical rainfall is crucial for better understanding the Earth's climate. This space-borne radar and associated satellite was jointly developed by the United States (NASA) and Japan (National Space Development Agency); it is known as the *Tropical Rainfall Measuring Mission (TRMM)* satellite (Fig. 3). Space-borne radar presents many challenging problems, including cost, size constraints, reliability issues, and temporal sampling. It is obviously impossible to continuously sample every precipitating cloud in the tropics from radar orbiting the Earth. But the *TRMM* satellite will help scientists develop a statistical distribution of rain rates within a certain area, and calculate the probability of a specific

rain rate occurring. Based on this information, it will be possible to generate monthly mean rain amounts within areas of 10^5 km^2 . Such information will be vital for the verification of climate models. See SATELLITE METEOROLOGY.

Clear-air and optical radars. Radars used to probe the clear air, or regions devoid of clouds, are known as profilers. A profiler is essentially a Doppler radar that operates at much longer wavelengths compared to weather radar. Wavelengths of 6 m, 70 cm, and 33 cm are commonly used. In the case of a profiler, the reflected power is not only from hydrometeors but also from gradients in the index of refraction of air, which are caused by turbulent motions in the atmosphere. These turbulent motions in turn cause small fluctuations in air temperature and moisture content, which also change the index of refraction. A profiler can determine the airflow in the cloud-free atmosphere, roughly up to 10 km above the Earth's surface. Optical radars, called lidars, use lasers as the radiation source. At these short wavelengths (0.1–10 μm), the laser beam is scattered by small aerosol particles and air molecules, allowing air motions to be determined, especially in thin, high tropospheric clouds and in the Earth's boundary layer (approximately the lowest 1 km or 0.6 mi of the Earth's atmosphere). See AEROSOL; HYDROMETEOROLOGY; LASER; LIDAR.

Operational radars. In the late 1990s, the weather radars used by the National Weather Service to provide warnings of impending severe weather were updated from antiquated WSR-57 and WSR-74 noncoherent radars to NEXRADs (Next Generation Weather Radars). NEXRAD (WSR-88D) radars are state-of-the-art Doppler radars operating at a wavelength of 10 cm. Using NEXRAD's Doppler capability, weather forecasters are able to warn the public sooner of approaching tornadoes and other severe weather. In severe storms, a mesocyclone first develops within the storm. The mesocyclone may be 10 km (6 mi) or more wide, and represents a deep rotating column of air within the storm. Severe and long-lasting tornadoes are often associated with mesocyclones. The mesocyclone is readily detected by a Doppler radar such as NEXRAD. The entire continental United States is covered by the NEXRAD network, consisting of more than 100 radars. NEXRADs along the Gulf Coast, in Florida, and along the eastern seaboard provide warning information on land-falling hurricanes. About 60 of the nation's busiest airports are also equipped with Doppler radars. These radars (operating at a wavelength of 5 cm, known as Terminal Doppler Weather Radars) provide weather-related warnings to air-traffic controllers and pilots. One particularly dangerous weather condition is wind shear, which often occurs as a microburst or intense downdraft. Microbursts can severely affect the flight of landing and departing aircraft, and have been identified as a factor in many aircraft accidents. See RADAR; RADAR METEOROLOGY; TORNADO; WEATHER FORECASTING AND PREDICTION.

Steven A. Rutledge

Bibliography. D. Atlas (ed.), *Radar in Meteorology*, 1990; V. N. Bringi and V. Chandrasekar,

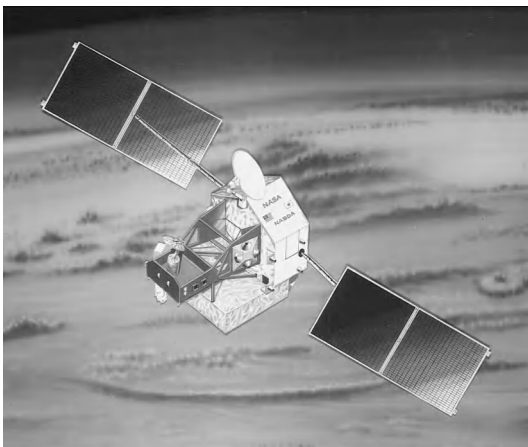


Fig. 3. *Tropical Rainfall Measuring Mission (TRMM)* satellite. (O. Thiele, National Aeronautics and Space Administration)

Polarimetric Doppler Weather Radar: Principles and Applications, 2005; R. J. Doviak and D. S. Zrnic, *Doppler Radar and Weather Observations*, 2d ed., 1993; P. Meischner (ed.), *Weather Radar: Principles and Advanced Applications*, 2005; R. Meneghini and T. Kozu, *Spaceborne Weather Radar*, 1990; R. E. Rinehart, *Radar For Meteorologists*, 4th ed., 2004; M. I. Skolnik, *Radar Handbook*, 2d ed., 1990.

Meteorological rocket

A small rocket system used for extending observations of the atmosphere above feasible limits for balloon-borne and telemetering instruments. Synoptic exploration of the middle-atmospheric circulation (20–95 km or 12–60 mi altitude) through use of these systems (also known as rocketsondes) matured in the 1960s into a highly productive source of information on atmospheric structure and dynamics. Many thousands of small meteorological rockets have been launched in a coordinated investigation of the wind field and the temperature and ozone structures in the middle atmosphere region at 25–55 km (16–34 mi) altitude.

These data produced dramatic changes in the scientific view of this region of the atmosphere, with a resulting alteration of the structural concepts—which had previously developed without adequate measurements—into a space-age atmospheric model that is primarily characterized by intense dynamics.

The availability of inexpensive, small meteorological rocket systems prompted the idea for simultaneous observations from the many different North American rocket ranges. These simultaneous, or synoptic, observations were scheduled for only a few days each season, but rapidly evolved into three to five launchings a week from many launch ranges. The coordination of these rocket launchings began in 1959 with the formation of the Meteorological Rocket Network (MRN). As other countries (for example, France, Japan, the United Kingdom, and the Soviet Union) developed their particular rocket system and capability, the launch activity became more specific and followed the schedule recommended by recognized scientific organizations such as the Committee on Space Research of the International Council of Scientific Unions.

Meteorological Rocket Network. The Meteorological Rocket Network and other international launch ranges have been very important and extremely effective in obtaining the much-needed temperature and wind data in the middle atmosphere. These data contributed significantly to the knowledge of stratospheric (15–55 km or 10–34 mi) and mesospheric (55–80 km or 34–50 mi) climatology, and they have been used widely in the development of reference atmosphere models. Reference atmospheres have important engineering applications for the design and testing of large launch vehicles such as the space shuttle.

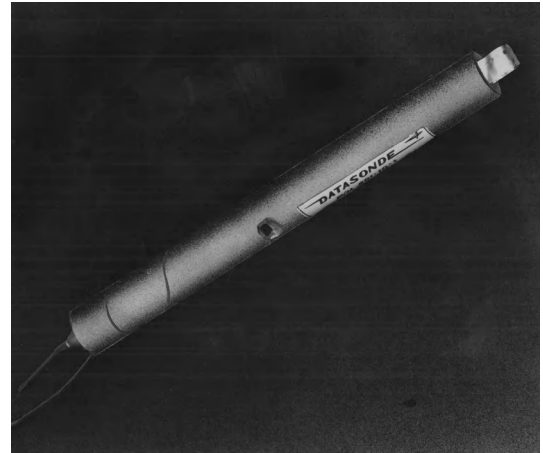


Fig. 1. Transmitter-bead thermistor payload for the Starute.

The development of the small meteorological rocket began in 1959 with a rocketsonde system known as ARCAS. ARCAS permitted a relatively large payload volume to be employed, but its low initial acceleration and slow speed made its trajectory sensitive to low-level winds during launch. The maximum altitude reached by this system was about 60 km (37 mi). A less wind-sensitive rocketsonde system, the Loki-Datasonde (PWN-8B), replaced the ARCAS system during the early 1970s. It was soon replaced with the Super Loki-Datasonde (PWN-11D). The PWN-11D rocketsonde motor burns for 2 s before separation from its inert dart and payload, which are thereby propelled to about 80 km (50 mi) altitude, where the payload is ejected. The payload consists of a small bead thermistor temperature sensor attached to a radio transmitter (Fig. 1) that sends the temperature data to a ground receiver, and a Starute parachute. The meteorological measurements are made during payload descent. At launch, the Super Loki-Datasonde (Fig. 2) has an overall weight of 31 kg (68 lb), and its length is approximately 4 m (13 ft).

The validity of rocketsonde measurements is highly dependent on the instrument precision. Tests show that the Datasonde measurements have a precision of 1°C (1.8°F) up to 55 km (34 mi) and about 1.5°C (2.7°F) at 60 km (37 mi). Wind speed accuracy is dependent on the quality of the radar tracking system, but it usually is better than 3 m/s (6.6 mi/h).

The total observational system includes a ground station composed of a radar for tracking the metallized Starute and a receiving station for the temperature telemetry data. Again, the equipment of these stations over the Meteorological Rocket Network is highly variable, the best being the precision tracking and telemetry systems found at the missile range.

Because of the increasing cost of rocket motors and payloads, launch equipment, and launch range operations, the Meteorological Rocket Network slowly was dissolved as funds needed to

maintain a viable launch schedule no longer were available. By the latter half of the 1980s, only nine launch ranges remained active. By 1998 only five facilities were still launching small meteorological rockets, but they were no longer using a coordinated schedule. As the number of launch sites was reduced, the coordinated launch schedule was replaced with launch schedules designed to meet unique launch range requirements or to further special research. Meteorological rocket data continue to be needed for verification and calibration check of remotely measuring instruments, that is, satellites and lidar. See LIDAR; METEOROLOGICAL INSTRUMENTATION; TELEMETERING.

Results. Meteorological Rocket Network synoptic exploration of the stratospheric circulation revealed an upper atmosphere markedly different from the quiescent, static, drab, and uninteresting place characterized in early models.

A prime example of the synoptic-scale dynamical systems that originate in the stratospheric region is illustrated by the "explosive warmings" first detected in high-altitude balloon flights over Berlin. Explosive warmings are wintertime phenomena, and they appear strongest in the upper stratosphere (30–40 km or 18–24 mi) where large temperature increases of 50°C (90°F) or more occur within a few days. Another phenomenon observed was the 26-month, or quasi-biennial oscillation. This oscillation is noticeable in the east-west wind flow over the Equator. The wind flows from the west for approximately 13 months and then reverses direction and flows from the east for about 13 months. This phenomenon was first observed with balloon soundings; but the rocketsonde observations were, and still are, the only method possible to define its vertical dimension and wind speeds.

Possibly of equal significance at the other end of the turbulence spectrum is the fact that the sensitive sensors of the rocketsonde have revealed a tremendous amount of small-scale variability in the upper atmosphere. These data make clear that assumptions of an inviscid fluid in the atmosphere are highly suspect; the values of the viscous terms in the equations of motion are probably larger than assumed in prior analysis.

Observational data have shown that the small-scale variability is largest in high latitudes during the winter. Temperature variations larger than 10°C (18°F) have been observed to occur within 2–4 h. The source of these extreme perturbations still is vague, but evidence points to gravity-wave activity.

Synoptic-scale circulation systems in the upper atmosphere are demonstrated by the rocketsonde data to be very obviously keyed to the geographic and orographic structures of the Earth's surface. In winter, oceanic regions characteristically have poleward extensions of ridges of high pressure, and continental regions have shifty troughs of low pressure extending equatorward over them. This intimate relationship between the surface and 50 km (31 mi) is most likely the direct result of tur-

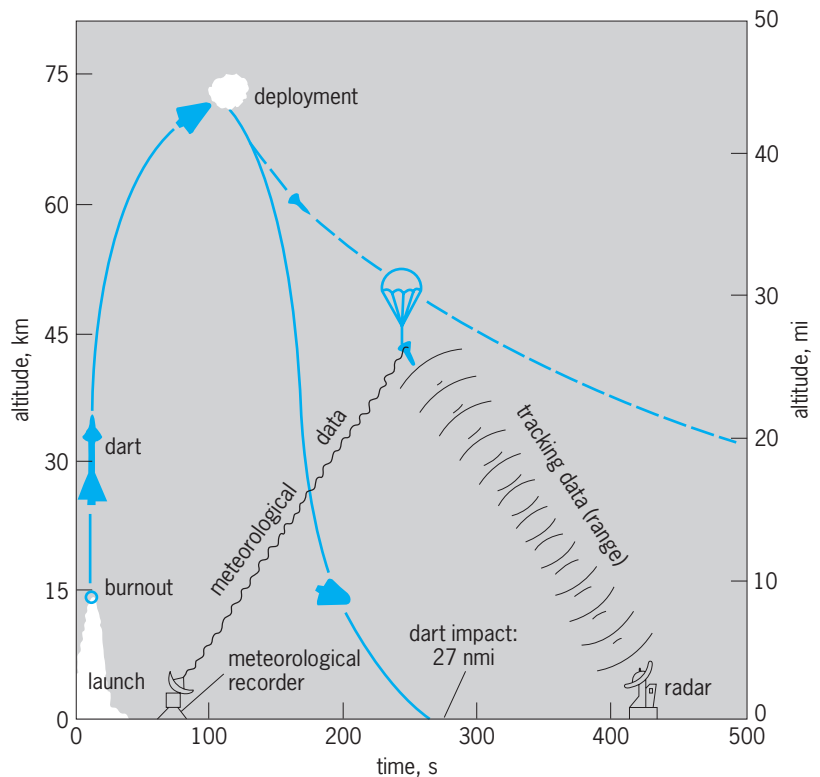


Fig. 2. Diagram of typical rocketsonde launch. Booster impact is within 400 m (1300 ft) under normal wind conditions. (NASA)

bulent energy transport in the vertical direction, and a total understanding of the entire atmospheric system cannot be realized until these factors are incorporated. See ATMOSPHERIC GENERAL CIRCULATION; STRATOSPHERE; UPPER-ATMOSPHERE DYNAMICS.

Willis L. Webb; Francis J. Schmidlin

Bibliography. J. M. Moran and M. D. Moran, *Meteorology: The Atmosphere and the Science of Weather*, 4th ed., 1994; F. J. Schmidlin, Intercomparisons of temperature, density, and wind measurements from *in situ* and satellite techniques. *Adv. Space Res.*, 4(6):101–110, 1984; F. J. Schmidlin, Rocket techniques used to measure the middle atmosphere, in Middle Atmosphere Program, *Handbook for MAP*, vol. 19, 1986; World Data Center A (Asheville, NC), *Meteorological Rocket Network Firings*, vols. 1–97.

Meteorological satellites

Satellites dedicated to the observation of meteorological phenomena and atmospheric or surface properties used for weather forecasting. Operational meteorological satellites provide routine observations of weather conditions as well as an ever-expanding range of environmental properties, such as aerosol, dust and ash clouds from volcanic eruptions, ozone, and land vegetation cover. For this reason, they are known in the United States as operational environmental satellites. See METEOROLOGY; SATELLITE (SPACECRAFT); SATELLITE METEOROLOGY.

The United States launched the first meteorological satellite in low Earth orbit, *TIROS 1 (Television and Infra-Red Observation Satellite)*, in April 1960. For the first time, complete pictures of clouds associated with large weather systems were seen by forecasters. A total of 10 *TIROS* were launched, followed by a series of progressively more sophisticated *ESSA* (Environmental Science Services Administration) and *NOAA* (National Oceanic and Atmospheric Administration) satellites, while the National Aeronautics and Space Administration (NASA) was experimenting with new sensors on a succession of *Nimbus* research satellites. In parallel, the Department of Defense initiated the Defense Meteorological Satellite Program (DMSP) used by the Air Force and the Navy for global weather surveillance and forecasting applications. At the same time the Soviet Union implemented its national operational *Meteor* satellite series, which continues to this day. Current plans in the United States are to merge both civilian and military low-altitude meteorological satellite programs into a National Polar-orbiting Operational Environmental Satellite System (NPOESS), beginning around 2008. Altogether, three *NPOESS* Satellites will likely occupy three polar orbits staggered in time to provide global coverage at approximately 4-hour intervals. See STORM DETECTION; WEATHER FORECASTING AND PREDICTION.

NASA pioneered Earth observation from geostationary orbit on its first experimental *Application Technology Satellite* launched in 1966, soon to be followed by the operational *SMS (Synchronous Meteorological Satellite)* and *GOES (Geostationary Operational Environmental Satellite)* series occupying two stations at 75° and 135° west longitude. Similarly, the European Space Agency and Japan National Space Development Agency developed their own equally capable geostationary satellite systems, *Meteosat* operated by EUMETSAT at 0° longitude and *Himawari (Chrysanthemum)* operated by the Japan Meteorological Agency at 140° east longitude, respectively. India and later China have attempted to fill the remaining gap over Asia and the Indian Ocean with their *Insat* and *FY* geostationary satellites.

Optical imaging sensors. Historically, the first recognized application of orbital observation was the visual exploitation of cloud images associated with weather systems. The preferred design of their imaging instruments emphasized spatial resolution and contrast over radiometric accuracy, as opposed to sensors that provide quantitative measurements. The difference in instrument design is beginning to blur in recent instruments, such as the Advanced Very High Resolution Radiometer (AVHRR) on *NOAA* satellites, which is used for a variety of quantitative applications, such as remote sensing of sea surface temperature, monitoring changes in land vegetation, and discriminating between different kinds of clouds. There is a pervasive trend to increase the number of spectral bands in imaging sensors, from 5 channels in the current AVHRR to 36 channels in the experimental Moderate-resolution Imaging Spec-

troradiometer (MODIS) developed by NASA. These channels sample the full spectrum of backscattered solar radiation in the visible, near-infrared, and long-wave infrared, and a good part of the emitted terrestrial radiation spectrum (thermal infrared). This multiplicity of spectral bands allows the detection of a wide variety of features, from aerosols and smoke in the atmosphere to chlorophyll in the ocean. See CLOUD; REMOTE SENSING; TERRESTRIAL RADIATION; WEATHER.

Except for observing polar regions, or providing meteorological support to operations in remote locations worldwide, the ideal platforms for cloud imaging are those in geosynchronous equatorial orbit, also known as geostationary orbit, at the precise altitude (35,900 km) where the orbital period matches the period of rotation of the Earth, so that the satellite appears to hover over a fixed location at the Equator. The international system of four to six geostationary meteorological satellites provides uninterrupted visibility of the global tropics and midlatitudes (up to 60° north and south at the satellite longitude) with the ability to monitor fast-developing weather systems that often are the most dangerous. The sharpness of cloud images (1-km picture elements in the visible), as well as the ability to scan the same scene repeatedly at time intervals as short as 5 minutes, allow for tracking the apparent motion of clouds, deducing wind velocity, and instantaneously assessing the strength of developing storms, a valuable capability in warm climate regions (Fig. 1). See EARTH ROTATION AND ORBITAL MOTION; TROPICAL METEOROLOGY.

Imaging from geostationary satellites is the principal source of observation that has allowed the National Weather Service to monitor the evolution of strong convective storms and extend the warning time for tornado strikes to about 8 minutes. In general, geostationary satellites provide meteorologists with a view from an ideal vantage point, a perfect synoptic coverage of regional weather, and the ability to continuously track the progression of weather disturbances and extrapolate their arrival time at a particular location.

Imaging microwave radiometers. Also interesting is the detection of diverse atmospheric properties and surface features using multifrequency microwave radiometers with small antenna beams. Water molecule absorption of microwave radiation emitted by the ocean provides an accurate estimation of total precipitable water in the atmospheric column. Microwave radiation emitted by the relatively homogeneous moist atmosphere below is scattered in a recognizable way by waterdrops and ice particles in rain clouds, thus providing an indirect means to estimate precipitation rates. Microwave radiation contrast discriminates ice floes from open ocean water, and wet from dry soil. Microwave radiometry enables diagnostics of sea state and wind strength over the surface of the ocean, or the sea surface temperature. The principal design constraint of imaging microwave radiometers is the diffraction limit of the sensor—large apertures are desirable, but

bulky antennas are a problem because mechanical scanning is needed to preserve radiometric accuracy. In order to achieve reasonably small footprints, microwave sensors are currently deployed in low Earth orbit, such as the Special Instrument/Microwave Imager on *DMSP* satellites and Conical Microwave Imager/Sounder on future *NPOESS* satellites. See MICROWAVE; PRECIPITATION (METEOROLOGY); RADIOMETRY.

Sounding sensors. The principal challenge of meteorological observation from space is obtaining quantitative information about atmospheric temperature and pressure, and the concentration of water vapor and other minor constituents, especially the vertical profile of these quantities in the lower 10–20 km.

Radiation emitted by the atmosphere at a given wavelength originates principally from the layer where the optical depth of the overlying atmosphere is equal to unity, meaning that a photon emitted from that region has about one chance out of three to escape. The explanation is that photons emitted much lower in the atmosphere are mostly reabsorbed and few escape to space, while photons emitted much higher in the atmosphere are few because the density of the emitting material decreases exponentially with altitude (Fig. 2). The pressure level where the optical depth equals one is determined by the mass of the overlying atmosphere and its specific absorption for the particular wavelength—the stronger the absorption, the higher the level. Thus, the spectrum of terrestrial radiation received by a satellite contains information on thermal emission of the atmosphere at different levels, determined by atmospheric temperature and absorption at different wavelengths. See ATMOSPHERE; PHOTON.

The retrieval of temperature profile and water vapor information from spectral data is a difficult and not a fully determined mathematical problem. The solutions are highly sensitive to spectral resolution and small errors in radiometric measurements. Nevertheless, considerable progress has been made since the early days of the *Nimbus* program. The latest Atmospheric Infra-Red Sounder (AIRS) instrument developed by NASA is expected to yield temperature profiles as accurate as balloon measurements, 1°C within each successive 1-km-thick layer of the lower atmosphere. See HYDROMETEOROLOGY; INFRARED RADIATION.

Atmospheric sounders operate in the thermal infrared, using the absorption bands of carbon dioxide molecules (3.7–4.9 μm and 13–15 μm), and in the microwave spectrum, using the 54-GHz absorption band of oxygen. Emitted radiation is much weaker and atmospheric sounders correspondingly less sensitive in the microwave region. However, nonprecipitating clouds are largely transparent to such relatively long wavelengths, thus allowing all-weather albeit less accurate observations.

Measurements of temperature and moisture are used mainly to update numerical weather prediction computations that forecast the circulation of the global atmosphere several days in advance. For this quantitative application, a delay of a few hours is



Fig. 1. On May 17, 2000, the first visible image was broadcast by the NOAA Geostationary Operational Environmental Satellite (*GOES 11*) launched 2 weeks earlier. A huge storm is visible over the central United States. It dropped 30 cm of snow in Wyoming and 5 cm of hail in Colorado and spawned many tornadoes across Nebraska.

immaterial but homogeneous global coverage is essential. Thus, atmospheric sounders are principally deployed on Sun-synchronous polar orbits. The parameters of these circular low Earth orbits are selected from a discrete set of altitudes (800–1000 km) and inclinations (retrograde quasi-polar) that allow

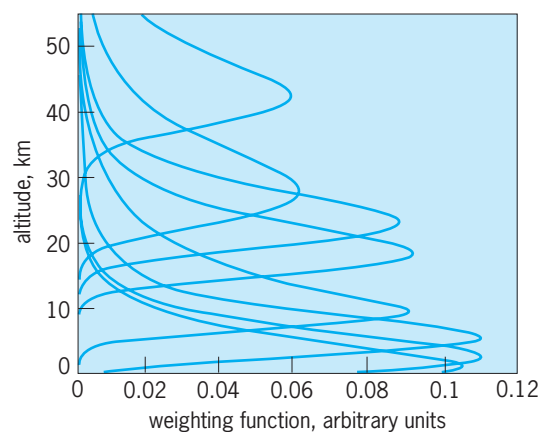


Fig. 2. Outgoing terrestrial radiation observed in space originates from different altitudes in the atmosphere. Depending upon the wavelength, the atmosphere is more or less absorbing. Emitted infrared radiation that is most effectively absorbed emerges only from the uppermost range of altitudes. Shown is a sample of altitude probability distributions or weighting functions for different wavelengths that apply to the advanced Atmospheric Infra-Red Sounder (AIRS) developed by NASA.

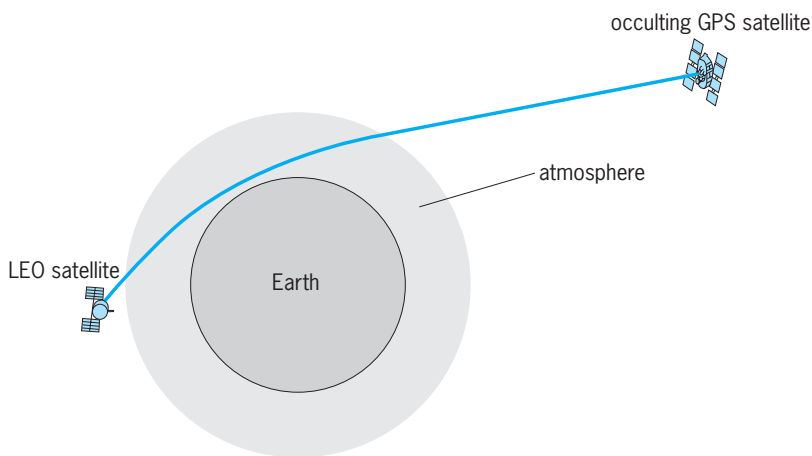


Fig. 3. Air density, pressure, and temperature data are derived from measurements of the increasing microwave refraction index of the atmosphere as altitude above the Earth surface decreases. The index of refraction is deduced from the propagation delay of radio-electric signals between any member of the Global Positioning System (GPS) satellites and a low-altitude orbiting satellite. The effect is significant near the moment of occultation, when the path is close to the limb of the Earth.

the orbital plane to drift by about 1° of longitude per day and match the change in Sun-Earth direction. Thus, a Sun-synchronous satellite crosses the Equator at (nearly) the same local time on every successive orbit.

Active sensors. Satellite sensors are not limited to receiving reflected solar radiation or emission from the atmosphere and clouds. Orbital systems are now powerful enough to probe the atmospheric medium or the surface with beams of electromagnetic radiation generated in space. The first operational sensor of this kind was a coarse radar or scatterometer that measured microwave radiation backscattered by the ocean surface. Backscatter is sensitive to surface roughness and thus provides a measurement of vector wind speed over the ocean (as well as a coarse all-weather mapping of sea ice). NASA developed the NSCAT and Seawinds microwave scatterometers for deployment on experimental Japanese environmental satellites *ADEOS 1* and *2*. A similar operational instrument will be carried by European *METOP* satellites.

Various radar altimeters have been used to map the changing topography of the ocean surface (principally to reconstruct the oceanic circulation from measured altitude gradients). Higher-frequency experimental radar and lidar systems are being tested to profile the distribution and optical properties of aerosol and cloud ice particles and waterdrops. The first demonstration of a space-borne precipitation radar is being conducted with the United States-Japan *Tropical Rain Measuring Mission (TRMM)* launched in 1997. Rain rate can be deduced from the three-dimensional distribution of ice particles and water in rain clouds, as observed by the *TRMM* satellite. See LIDAR; METEOROLOGICAL RADAR; RADAR METEOROLOGY.

Yet another promising technology will determine wind velocity in clear air from direct measurements of the frequency shift (Doppler effect) of multiple

laser pulses backscattered by aerosol and other diffusive particles. NASA planned to conduct the first flight demonstration of this technique on the space shuttle in year 2001. Global wind measurements will provide an invaluable enhancement of the worldwide meteorological observing network, especially at low latitudes where the wind field cannot be deduced from atmospheric pressure. See DOPPLER RADAR.

Bistatic systems. Planetary fly-by missions have provided opportunities to deduce the atmospheric density structure by measuring the propagation delay of a radioelectric signal traversing increasingly longer paths through a planet's atmosphere. Exceedingly precise radioelectric occultation measurements of this type are also possible on Earth by monitoring the propagation delay of navigation signals transmitted by the constellation of Global Positioning System (GPS) satellites and received by low-altitude orbiting satellites whenever the propagation path comes near the Earth limb (apparent outer edge) [Fig. 3]. The first GPS-MET flight demonstration, sponsored by the National Science Foundation, accurately retrieved atmospheric density, pressure, and temperature from about 25 km down to about 5 km above the Earth's surface. NASA will continue to test this new observing method with an experimental constellation of GPS-equipped satellites. Future operational satellites such as *NPOESS* and *METOP* will carry advanced versions of these GPS receivers. See OCCULTATION; SATELLITE NAVIGATION SYSTEMS.

In the lower atmosphere, the refractive effect of unknown amounts of water vapor invalidates the radioelectric method, unless humidity can be otherwise determined. Alternatively, the method can be used to determine total precipitable water along the propagation path when dry air pressure is known. This meteorological application is being developed for a number of GPS-equipped geodetic stations already in existence to monitor Earth crust deformations.

Pierre Morel

Bibliography. Committee on Earth Observation Satellites, *Towards an Integrated Global Observing Strategy*, European Space Agency, 1997; S. Q. Kidder and T. H. Von der Haar, *Satellite Meteorology: An Introduction*, 1995; W. P. Menzel and J. F. W. Purdom, Introducing GOES-1, The first of a new generation of Geostationary Operational Environmental Satellites, *Bull. Amer. Meteorol. Soc.*, 75:757-781, 1994; National Research Council, *Continuity of NOAA Satellites*, National Academy Press, 1997; P. K. Rao (ed.), *Weather Satellites: Systems, Data and Environmental Applications*, American Meteorological Society, Boston, 1990.

Meteorology

A discipline involving the study of the atmosphere and its phenomena. Meteorology and climatology are rooted in different parent disciplines, the former in physics and the latter in physical geography. They

have, in effect, become interwoven to form a single discipline known as the atmospheric sciences, which is devoted to the understanding and prediction of the evolution of planetary atmospheres and the broad range of phenomena that occur within them. The atmospheric sciences comprise a number of interrelated subdisciplines. See CLIMATOLOGY.

Subdisciplines. Atmospheric dynamics (or dynamic meteorology) is concerned with the analysis and interpretation of the three-dimensional, time-varying, macroscale motion field. It is a branch of fluid dynamics, specialized to deal with atmospheric motion systems on scales ranging from the dimensions of clouds up to the scale of the planet itself. The activity within dynamic meteorology that is focused on the description and interpretation of large-scale (greater than 1000 km or 600 mi) tropospheric motion systems such as extratropical cyclones has traditionally been referred to as synoptic meteorology, and that devoted to mesoscale (10–1000 km or 6–600 mi) weather systems such as severe thunderstorm complexes is referred to as mesometeorology. Both synoptic meteorology and mesometeorology are concerned with phenomena of interest in weather forecasting, the former on the day-to-day time scale and the latter on the time scale of minutes to hours. See DYNAMIC METEOROLOGY; MESOMETEOROLOGY.

The complementary field of atmospheric physics (or physical meteorology) is concerned with a wide range of processes that are capable of altering the physical properties and the chemical composition of air parcels as they move through the atmosphere. It may be viewed as a branch of physics or chemistry, specializing in processes that are of particular importance within planetary atmospheres. Overlapping subfields within atmospheric physics include cloud physics, which is concerned with the origins, morphology, growth, electrification, and the optical and chemical properties of the droplets within clouds; radiative transfer, which is concerned with the absorption, emission, and scattering of solar and terrestrial radiation by aerosols and radiatively active trace gases within planetary atmospheres; atmospheric chemistry, which deals with a wide range of gas-phase and heterogeneous (that is, involving aerosols or cloud droplets) chemical and photochemical reactions on space scales ranging from individual smokestacks to the global ozone layer; and boundary-layer meteorology or micrometeorology, which is concerned with the vertical transfer of water vapor and other trace constituents, as well as heat and momentum across the interface between the atmosphere and the underlying surfaces and their redistribution within the lowest kilometer of the atmosphere by motions on scales too small to resolve explicitly in global models. Aeronomy is concerned with physical processes in the upper atmosphere (above the 50-km or 30-mi level). See AERONOMY; ATMOSPHERIC CHEMISTRY; ATMOSPHERIC ELECTRICITY; ATMOSPHERIC GENERAL CIRCULATION; ATMOSPHERIC WAVES, UPPER SYNOPTIC; CLOUD PHYSICS;

METEOROLOGICAL OPTICS; MICROMETEOROLOGY; RADIATIVE TRANSFER; TERRESTRIAL RADIATION.

Although atmospheric dynamics and atmospheric physics in some circumstances can be successfully pursued as separate disciplines, important problems such as the development of numerical weather prediction models and the understanding of the global climate system require a synthesis. Physical processes such as radiative transfer and the condensation of water vapor onto cloud droplets are ultimately responsible for the temperature gradients that drive atmospheric motions, and the motion field, in turn, determines the evolving, three-dimensional setting in which the physical processes take place.

The atmospheric sciences cannot be completely isolated from related disciplines. On time scales longer than a month, the evolution of the state of the atmosphere is influenced by dynamic and thermodynamic interactions with the other elements of the climate system, that is, the oceans, the cryosphere, and the terrestrial biosphere. A notable example is the El Niño–Southern Oscillation phenomenon in the equatorial Pacific Ocean, in which changes in the distribution of surface winds force anomalous ocean currents; the currents can alter the distribution of sea-surface temperature, which in turn can alter the distribution of tropical rainfall, thereby inducing further changes in the surface wind field. On a time scale of decades or longer, the cycling of chemical species such as carbon, nitrogen, and sulfur between these same global reservoirs also influences the evolution of the climate system. Human activities represent an increasingly significant atmospheric source of some of the radiatively active trace gases that play a role in regulating the temperature of the Earth. See BIOSPHERE; MARITIME METEOROLOGY; TROPICAL METEOROLOGY.

Vertical profiles of pressure and density. The decrease with height in the pressure (p) and density (ρ) in planetary atmospheres is approximately exponential (**Fig. 1**). This behavior is characteristic of atmospheres composed of ideal gases, in which absolute temperature does not vary strongly with height. If it is assumed that the vertical component of the acceleration of the air is small in comparison to gravity (g) and that the weight of a differential layer of unit horizontal cross section and thickness (dz) in the vertical is equal to the difference between the pressure on its upper and lower surfaces, the hydrostatic equation (1) results. This relationship is valid

$$\frac{dp}{dz} = -\rho g \quad (1)$$

throughout the atmosphere except in the most vigorous convective clouds. Substituting for ρ from the equation of state (2) yields Eq. (3), where R is the

$$p = \rho RT \quad (2)$$

$$d(\ln p) = -\frac{g}{RT} dz \quad (3)$$

gas constant appropriate to the chemical composition of the atmosphere and T is the temperature in

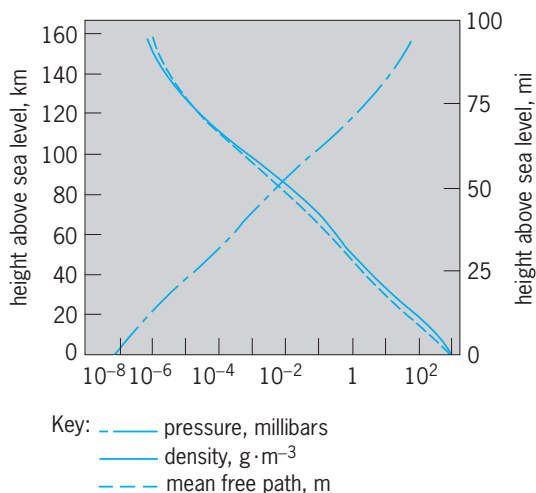


Fig. 1. Vertical profile of pressure, density, and mean free path for typical conditions in the Earth's atmosphere. 1 millibar = 10^2 Pa. 1 m = 3.28 ft. (After R. C. Weast, ed., *CRC Handbook of Chemistry and Physics*, 70th ed., CRC Press, 1989)

kelvins. If T is assumed to be independent of height, this expression can be integrated from a reference level z_0 , at which $p = p_0$, to level z and is expressed in the form shown in Eq. (4), where $H \equiv (RT)/g$

$$p = p_0 \exp -\frac{z}{H} \quad (4)$$

is known as the scale height, that is, the height over which pressure decreases by a factor of e (the base of natural logarithms). By combining Eqs. (2) and (4), it is readily verified that density exhibits the same functional dependence. The gas constant R is obtained by dividing the universal gas constant $R^* = 8313 \text{ J} \cdot \text{kmol}^{-1} \cdot \text{K}^{-1}$ by the molecular effective weight μ of the mixture of gases of which the atmosphere is composed. For Earth, up to a level of 120 km (72 mi), $\mu = 28.97$ as explained below, $R = 287 \text{ J} \cdot \text{kg}^{-1} \cdot \text{K}^{-1}$, $g = 9.8 \text{ m} \cdot \text{s}^{-2}$, and $T = 240 \text{ K}$ (to within $\pm 15\%$), and so, consistent with the slopes of the pressure and density curves in Fig. 1, $H \cong 7 \text{ km}$ (4 mi). Mean sea-level pressure averaged over the Earth is 1013 millibars (101 kilopascals), and the mean density is on the order of $1.25 \text{ kg} \cdot \text{m}^{-3}$. In relating pressure to height in the Earth's atmosphere in the remainder of this section, it is convenient to note that pressure and density drop off by a factor of 10 over a vertical distance of $H \ln(10) \cong 16 \text{ km}$; and so the 100-mbar (10-kPa) level corresponds to roughly 16 km (10 mi) above sea level, the 10-mbar (1-kPa) level to 32 km (20 mi), the 1.0-mbar (100-Pa) level to 48 km (30 mi), and so forth. Since the pressure at any level is equal to the mass per unit area above that level, divided by g , it follows that 10% of the mass of the Earth's atmosphere lies above the 100-mbar (10-kPa) level, 1% above the 10-mbar (1-kPa) level, and so forth. See GAS.

Composition of atmosphere. The atmospheres of the planets are believed to have originated from the outgassing of volatile substances [mainly water (H_2O), carbon dioxide (CO_2), and nitrogen com-

pounds] from their interiors. Most of the water vapor condensed out immediately. On Earth (in contrast to Venus and Mars) nearly all the carbon dioxide has dissolved in the oceans and subsequently been incorporated into carbonate deposits in the crust by shell-forming species of plankton. Most of the nitrogen remains in the atmosphere in the form of molecular nitrogen (dinitrogen; N_2). Photosynthesis by plant life that was buried and fossilized before it had time to decay has generated large quantities of molecular oxygen (dioxygen; O_2).

From the surface up to about 100 km (60 mi), macroscale fluid motions keep the atmosphere well mixed, so that the relative proportions of its major gaseous constituents are nearly constant. Nitrogen, oxygen, and argon are the dominant constituents of dry air: they account for roughly 78%, 21%, and 1% of the molecules, respectively (hence the effective molecular weight of 28.97). Important atmospheric trace constituents include water vapor (up to 4% by volume and highly variable in space and time), ozone (O_3 ; up to 15 parts per million at the 25-km level), and carbon dioxide (approximately 350 ppm in 1990 and increasing at a rate of about 12 ppm per decade). The latter are triatomic species, which exhibit strong absorption bands associated with rotational-vibrational transitions in the infrared part of the electromagnetic spectrum, the implications of which will be discussed below.

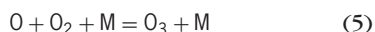
At 100 km (60 mi) above the Earth's surface, the mean free path between collisions reaches 1 m (3.3 ft; Fig. 1) and the characteristic time scale for molecular diffusion becomes comparable to the time scale for mixing by macroscale fluid motions. Above this so-called turbopause, the various atmospheric constituents exhibit gravitational settling under the influence of molecular diffusion, the concentration of each species decreasing exponentially, as shown in Eq. (4), with scale height H inversely proportional to its molecular weight, as if it were the only constituent present. At 500 km (300 mi) the atmosphere is composed primarily of atomic oxygen (O), and above 1000 km (600 mi) helium and hydrogen are the dominant species.

Effects of x-ray and ultraviolet radiation. In Fig. 2 the marked enhancements of the spectrum of solar radiation relative to the blackbody curve at the very short and very long wavelengths are due to the emission from the solar chromosphere and corona. Ultraviolet radiation with wavelengths (λ) shorter than 0.31 micrometer accounts for less than 2% of the energy emitted by the Sun; radiation with $\lambda < 0.24 \mu\text{m}$, less than 0.1%; and x-rays with $\lambda < 0.1 \mu\text{m}$, only about 3 parts in 10^6 —yet the radiation in these wavelength bands has profound influences on the structure and composition of the upper atmosphere.

Several types of photochemical reactions occur within the Earth's atmosphere (Fig. 3). X-rays with $\lambda < 0.1 \mu\text{m}$ are sufficiently energetic to ionize whatever atomic species happen to be present. Peak ion concentrations (about 10^6 cm^{-3}) are observed about 300 km (180 mi) above the Earth's surface. Virtually all the solar radiation in this wavelength band is

absorbed before it reaches the 60-km (36-mi) level (the base of the ionosphere), by which time ion concentrations have declined by four orders of magnitude relative to those at 300 km (180 mi). See IONOSPHERE; PHOTOCHEMISTRY.

Ultraviolet radiation with $\lambda < 0.24 \mu\text{m}$ is sufficiently energetic to photodissociate molecular oxygen. As a consequence of this reaction, most of the oxygen is in the atomic form at levels above 120 km (72 mi). Radiation at these wavelengths is sufficiently depleted by the time it penetrates to the 20–50-km (12–30-mi) level that it produces only trace amounts of atomic oxygen. At these levels the air is sufficiently dense that atomic oxygen quickly combines with molecular oxygen in three-body reaction (5), where



M represents another molecule, to form ozone. The ozone molecules created in this reaction are dissociated by ultraviolet solar radiation with wavelengths $< 0.31 \mu\text{m}$, as in reaction (6). The resulting oxygen



atom quickly recombines with O_2 [reaction (5)].

The net result of reactions (5) and (6) is the absorption of a photon of solar radiation with $0.24 < \lambda < 0.31 \mu\text{m}$, and the heating of the other molecules (M) involved in the three-body collision [reaction (5)]. Through this mechanism, the creation of a single odd oxygen molecule can ultimately result in the absorption of millions of photons of ultraviolet radiation which would otherwise be lethal to life on the surface of the planet. The photochemical reactions associated with ozone chemistry are most active in the 30–60-km (18–36-mi) layer, the so-called ozone layer. Under certain conditions, chlorine radicals created by the photodissociation of anthropogenically produced chlorofluoromethanes at these levels appear to be capable of destroying odd oxygen through catalytic reactions. See STRATOSPHERIC OZONE.

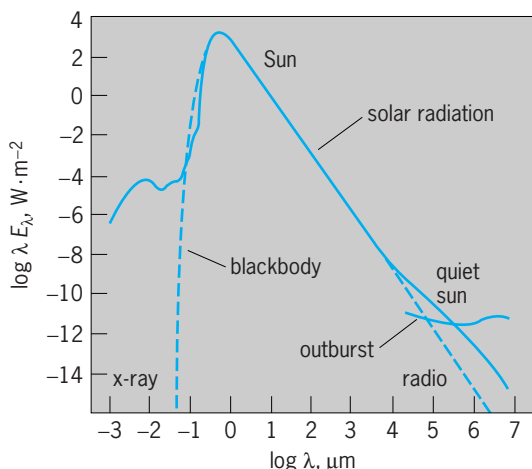


Fig. 2. Spectrum of solar radiation as compared with the spectrum of a blackbody at 5780 K (9934°F). (After C. W. Allen, *Solar radiation*, *Quart. J. Roy. Meteorol. Soc.*, 84:311, 1958)

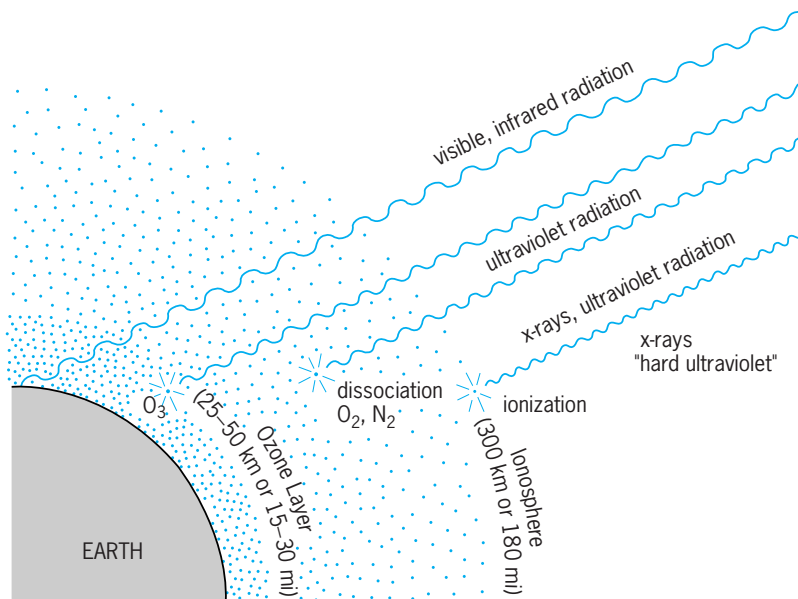


Fig. 3. Schematic representation of the processes responsible for the absorption of solar radiation in the Earth's atmosphere.

Photoionization and photodissociation are strong heat sources above the 100-km (60-mi) level. Even though only a few parts per million of the solar radiation incident on the top of the atmosphere is absorbed at these levels, the energy input is large because the molecules present at these levels account for less than one-millionth of the mass of the atmosphere. Energy absorption per unit mass increases with height in this layer and, in addition, the molecules become less efficient at disposing of energy by emitting infrared radiation as the frequency of molecular collisions decreases. Hence, above about 80 km (50 mi), temperature increases with height, and the downward flux of heat by down-gradient molecular diffusion plays an important role in the energy balance. This outermost layer of the Earth's atmosphere is known as the thermosphere. Analogous layers are observed in the atmospheres of the other planets. Solar output in the x-ray part of the spectrum varies strongly in response to sunspots and solar flares, whose frequency exhibits a remarkable 11-year cycle. During the active part of the cycle, temperatures in the outer thermosphere reach values of 2000 K (3140°F; compared to about 500 K or 440°F in the quiet sun years), and a significant fraction of the hydrogen atoms attain velocities high enough to allow them to escape from the Earth's gravitational field. Over the lifetime of the solar system, appreciable quantities of hydrogen are believed to have escaped from the Earth's atmosphere, and it has been proposed that the hydrogen atoms in the water outgassed from Venus might have been lost in this manner. See SOLAR SYSTEM; SUN.

The absorption of ultraviolet radiation by ozone molecules in the 30–60-km (15–36-mi) layer gives rise to a distinct peak in the vertical profile of temperature (Fig. 4). It is notable that the atmospheres of Mars and Venus, which lack the oxygen necessary

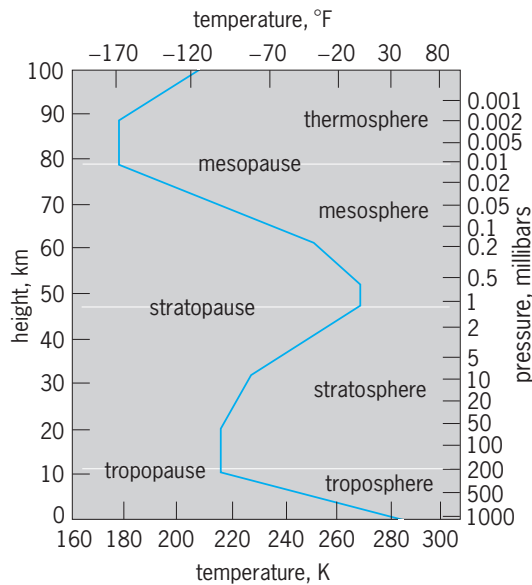


Fig. 4. Typical temperature profile in middle latitudes. 1 millibar = 10^2 Pa. 1 km = 0.6 mi. (After J. M. Wallace and P. V. Hobbs, *The Atmosphere: An Introductory Survey*, Academic Press, 1977)

to form an ozone layer, do not exhibit such an intermediate temperature maximum between the surface and the thermosphere. The peak in the temperature profile divides the Earth's middle atmosphere into an upper layer (the mesosphere) and a lower layer (the stratosphere). The temperature maximum corresponds to the stratopause, that is, the top of the stratosphere. See MESOSPHERE; STRATOSPHERE.

The base of the stratosphere corresponds to a distinct temperature minimum that marks the transition to a lower layer of the atmosphere in which the temperature distribution is maintained by the upward flux of energy from the Earth's surface by macroscale motions. Analogous layers, known as tropospheres, are observed on Venus and Mars. The temperature minimum in the Earth's atmosphere corresponds to the tropopause (that is, the top of the troposphere). In the atmospheres of Venus and Mars a deep isothermal (constant-temperature) layer extends from the tropopause to the base of the thermosphere. See TROPOPAUSE; TROPOSPHERE.

Electrical properties. Among the processes that contribute to the production of charged particles in the atmosphere are photoionization of neutral atoms by solar x-rays and ultraviolet rays, and charge separation that takes place when ice particles (or ice and water particles) collide within clouds. The former process gives rise to the ionosphere, which extends upward from the 60-km (36-mi) level to the outermost reaches of the atmosphere, and the latter is responsible for maintaining the fair weather electric field, which exists within the lowest few kilometers above the Earth's surface, and the much stronger fields that sometimes exist locally within and near clouds. Although only a minute fraction of the atoms that make up the Earth's atmosphere are electrically charged, their presence accounts for a wide range of geophysical phenomena.

Ion concentrations increase monotonically with height from the base of the ionosphere to a maximum near 300 km (180 mi). The increase with height tends to be concentrated in a series of layers, labeled (in order of increasing height) D, E, and F. The lower layers are present only during the daytime: in the absence of the Sun's ionizing radiation, the charged particles quickly recombine to form neutral particles. Higher in the ionosphere, recombination is slower because the mean free path between collisions is much longer. Collisions between electrons and neutral particles within the D layer are effective at absorbing AM radio waves propagating upward. When the D layer disappears during the night, radio waves are free to propagate into the upper layers of the ionosphere, where they are reflected to the ground, causing interference and sometimes permitting the reception of distant stations.

The fair weather electric field is strongest in the lowest 100 m (330 ft) of the atmosphere, where it averages 120 V/m (36 V/ft) in the vertical: the atmosphere carries a positive charge relative to the Earth's surface. The Earth and its atmosphere may be viewed as a capacitor whose inner conductor is the Earth and whose outer conductor, which encompasses most of the atmosphere, is referred to as the electrosphere. The conductivity within the electrosphere below 60 km (36 mi) is primarily due to the presence of charged particles generated by cosmic rays colliding with air molecules. Relatively few charged particles are generated within the lowest few kilometers of the atmosphere, and they tend to be immobilized by the presence of large, slow-moving particles. This poorly conducting layer serves as the dielectric of the capacitor. The upward flux of electrons through this "leaky dielectric" would discharge the electrosphere within a matter of minutes were it not continually being recharged by lightning and point discharge currents from the ground within thunderstorms, whose cloud bases carry a strong negative charge relative to the ground because of the charge separation going on within them. See SFERICS; STORM ELECTRICITY.

Lapse rates and vertical mixing. The various temperature layers (Fig. 4) have distinctly different dynamical properties, which in turn affect the physical and chemical processes that take place within them. In order to understand why the thermal structure discussed in the previous section is so important, it is necessary to understand the concept of static stability. An idealized air parcel is free to expand and contract in response to the hydrostatic pressure changes [as defined in Eq. (1)] that it encounters as it goes up and down in the atmosphere, and it is thus capable of doing work on its environment as it rises and expands, or having work done on it as it sinks and is compressed; but it is assumed that this air parcel does not exchange heat with its environment. If the parcel does not have any heat source or sink of its own, its temperature will change adiabatically (that is, without the addition or subtraction of heat) as it rises and sinks. The rate of temperature change with height (dT/dz) can be inferred from the first law of thermodynamics, which for an ideal gas

can be written in the form shown in Eq. (7), where

$$dq = c_v dT + p d\alpha \tag{7}$$

dq is the differential amount of heat added to the parcel, $c_v dT$ is the increase in the internal energy of the parcel, $p d\alpha$ is the work done by the parcel on the surrounding air as it expands, c_v is the specific heat of air at constant volume ($717 \text{ J} \cdot \text{kg}^{-1} \cdot \text{K}^{-1}$), and $\alpha = 1/\rho$ is the specific volume of the parcel expressed in $\text{m}^3 \cdot \text{kg}^{-1}$. The second term on the right-hand side can be rewritten as $d(p\alpha) - \alpha dp$. Substituting from Eq. (2) for $d(p\alpha)$, Eq. (7) can be rewritten as Eq. (8).

$$dq = (c_v + R) dT - \alpha dp \tag{8}$$

For an isobaric (constant-pressure) process, $dp = 0$, and therefore the specific heat at constant pressure can be defined by expression (9), which is equal to

$$c_p \equiv c_v + R \tag{9}$$

$717 + 287 = 1004 \text{ J} \cdot \text{kg}^{-1} \cdot \text{K}^{-1}$. Substituting into Eq. (8) gives Eq. (10), which is the form of the first

$$dq = c_p dT - \alpha dp \tag{10}$$

law more frequently used in atmospheric thermodynamics. The first term on the right-hand side of Eq. (10) is referred to as an incremental change in enthalpy or sensible heat. Equation (11) results from

$$-\left(\frac{dT}{dz}\right)_{\text{adiab.}} = \frac{g}{c_p} \tag{11}$$

setting $dq = 0$ and substituting for α from Eq. (1), and this is defined as the dry adiabatic lapse rate, that is, the rate at which the temperature of an air parcel that is not saturated with water vapor drops with increasing height under adiabatic conditions. (Note that the minus sign is implicit in the definition of the term lapse rate.) In the Earth's atmosphere, the dry adiabatic lapse rate has a numerical value of $9.8 \text{ K} \cdot \text{km}^{-1}$. See ENTHALPY; THERMODYNAMIC PRINCIPLES.

If the environmental lapse rate is the same as the adiabatic lapse rate, an air parcel forcibly displaced upward or downward from its original level would always remain at the same temperature as its environment, and so it would be neutrally buoyant. However, if the environmental lapse rate is less than the adiabatic lapse rate, a rising air parcel will find itself colder (and therefore more dense) than its environment (and vice versa), and so it will encounter a restoring force that will push it back toward its equilibrium level. The larger the difference in lapse rates, the larger the restoring force for a given vertical displacement. The restoring force per unit vertical displacement can be regarded as a measure of the static stability of the atmosphere at that level. Unstable lapse rates (that is, environmental lapse rates larger than the dry adiabatic value) are only very rarely observed in planetary atmospheres, because free convection produces a strong upward transfer of heat whenever the lapse rate reaches the adiabatic value. An isothermal lapse rate ($dT/dz = 0$)

represents quite a stable stratification, and a so-called inversion ($dT/dz > 0$) represents an even stronger one.

Thus the stratosphere is much more stably stratified than the troposphere (hence the name), and therefore it is a layer in which vertical mixing of trace substances is strongly suppressed. The characteristic residence time of air parcels in the stratosphere (that is, the time elapsed since they were in the troposphere) ranges from a few months just above the base to many years at the stratopause level. Hence, the stratosphere functions as a long-term reservoir for certain types of pollutants, such as debris from nuclear tests, which are quickly cleansed from tropospheric air by processes discussed below. The high static stability of the stratosphere also limits the height to which plumes of rising air in severe thunderstorms or volcanic eruptions can rise before they become negatively buoyant.

Atmospheric motions. The levels and temperatures of the tropopause, stratopause, and mesopause vary with latitude and season (Fig. 5). The stratopause is warmest at the summer pole, where the solar heating is strongest, and coolest at the winter pole, which is in darkness. However, the temperature distribution at the mesopause is the reverse of what would be expected on the basis of arguments based on radiative transfer; the summer pole is much colder than the winter pole. An equally strange distribution (from the point of view of radiative transfer) is observed at the tropopause, which is coldest (-80°C or -112°F or less) on the Equator and warmest at the summer pole and at middle latitudes of the winter hemisphere. At these levels the radiative heating is relatively weak, and dynamical processes are capable of driving temperatures far from their radiative equilibrium values. For example, the extreme coldness of the summer mesopause and the equatorial tropopause is maintained by adiabatic expansion associated with large-scale upward motion.

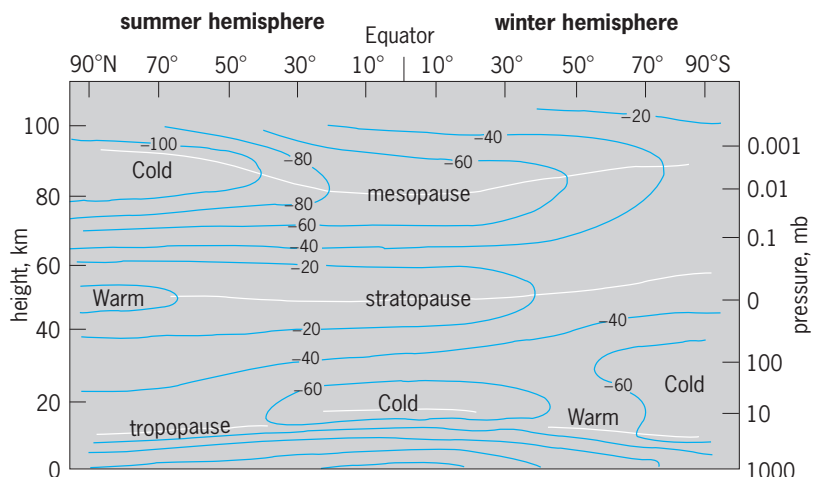


Fig. 5. Meridional cross section of the longitudinally averaged temperature ($^\circ\text{C}$) at the time of the solstices. $^\circ\text{F} = (^\circ\text{C} \times 1.8) + 32$. 1 millibar = 10^2 Pa . 1 km = 0.6 mi. (After J. M. Wallace and P. V. Hobbs, *The Atmosphere: An Introductory Survey*, Academic Press, 1977)

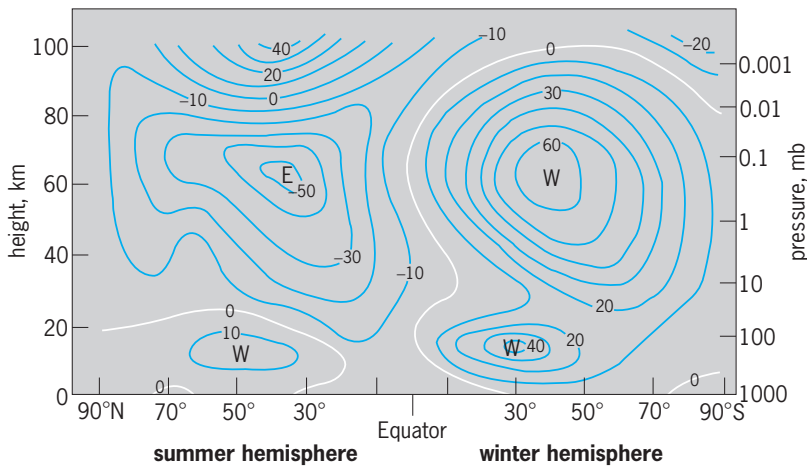


Fig. 6. Meridional cross section of the longitudinally averaged zonal wind ($\text{m} \cdot \text{s}^{-1}$) at the time of the solstices. Positive zonal winds denote flow from west (W) to east (E). $1 \text{ m} \cdot \text{s}^{-1} = 2.2 \text{ mi/h}$. (After J. M. Wallace and P. V. Hobbs, *The Atmosphere: An Introductory Survey*, Academic Press, 1977)

There is a corresponding distribution of the zonal (west-to-east) component of the wind (**Fig. 6**). Wind maxima are observed at the tropopause level (the so-called tropospheric jet streams) and in the mesosphere. The tropospheric jet stream, which is present throughout the year, blows from west to east. It is stronger and somewhat farther equatorward during winter. In the mesosphere the winds blow from the west during winter and from the east during summer. These zonal winds are in thermal wind balance, with the temperature distribution shown in **Fig. 6**: wherever temperature decreases (increases) with latitude, the zonal wind is becoming more (less) westerly with increasing height. See GEOSTROPHIC WIND; JET STREAM.

Winds in the equatorial lower stratosphere (not shown in **Fig. 6**) exhibit a remarkable 27-month quasiperiodicity, with alternating periods of remarkably persistent easterly and westerly winds, which appear first near the 10-mb or 1-kPa (30-km or 18-mi) level and gradually descend, over the course of a year or so, to the 70-mb or 7-kPa (18-km or 11-mi) level. The peak-to-peak amplitude of the so-called quasi-biennial oscillation reaches $45 \text{ m} \cdot \text{s}^{-1}$ (100 mi/h). The westerly polar vortex in the winter hemisphere is distorted by planetary waves that propagate energy upward from below. In the Northern Hemisphere, these disturbances sometimes become so intense during midwinter that they produce a so-called sudden warming, that is, the disappearance of the cold temperatures normally found over the polar regions at this time of year and the westerly vortex that encircles them.

The tropospheric circulation exhibits a complex array of disturbances on a wide range of space and time scales. Prominent among them are baroclinic waves, whose signature on synoptic charts at the Earth's surface is characterized by migrating extratropical cyclones and anticyclones. Because of the relatively lower static stability at these levels, tropospheric disturbances are characterized by much larger vertical motions than stratospheric distur-

bances. Lifting is often sufficient to produce widespread condensation of water vapor, which gives rise to clouds and precipitation. See DYNAMIC INSTABILITY; UPPER-ATMOSPHERE DYNAMICS; WIND.

Temperature at Earth's surface. With the notable exception of Jupiter, which is emitting substantial amounts of energy released by gravitational compression, the individual planets can be regarded as being in radiative equilibrium with the Sun. They intercept $\pi R^2 S$ units of solar radiation, where R is the radius of the planet and S is the solar irradiance, which decreases with distance from the Sun in accordance with the inverse-square law. Of the intercepted solar radiation, the fraction A , defined as the planetary albedo, is reflected to space; the remaining fraction $(1 - A)$ is absorbed by the planet and its atmosphere. An equal amount of radiation is emitted to space by the planet and its atmosphere. The planetary radiation may be expressed as the amount of radiation emitted by a blackbody at the effective temperature T_E of the planet. T_E can be determined from the balance between incoming and outgoing radiation, Eq. (12), where σ is the Stefan-Boltzmann constant

$$(1 - A)\pi R^2 S = \sigma T_E^4 (4\pi R^2) \quad (12)$$

$5.67 \times 10^{-8} \text{ W} \cdot \text{m}^{-2} \cdot \text{K}^{-4}$. For the Earth, the solar irradiance S is $1380 \text{ W} \cdot \text{m}^{-2}$, $A \approx 0.30$, for which the solution of Eq. (12) yields $T_E = 255 \text{ K}$. It is readily verified that if nothing else changed, $dT_E/T_E = 1/4 dS/S$ so that, for example, a 1% increase in solar irradiance would raise the effective temperature of the Earth by about 0.64 K (1.2°F). Because of its high planetary albedo (0.78), Venus exhibits an effective temperature of only 227 K (−51°F) despite the fact that it intercepts more than three times as much solar radiation per unit surface area than the Earth does. The effective temperature of Jupiter, as measured from space, is about 125 K (−235°F), compared with a value of 105 K (−271°F) computed from Eq. (12). See ALBEDO; HEAT RADIATION; PLANETARY PHYSICS; SOLAR RADIATION.

The peak wavelength λ_m of planetary radiation can be estimated by applying the Wien displacement law to a blackbody at the effective temperature of the planet [Eq. (13)]. For the Earth, it is readily verified

$$\lambda_m = \frac{2897}{T} \quad (13)$$

that $\lambda_m \approx 15 \mu\text{m}$. Normalized blackbody curves for solar and terrestrial radiation are shown in **Fig. 7** on a logarithmic scale. Since the overlap between the two curves is minimal, solar and terrestrial radiation may be treated separately. See BLACKBODY.

The fact that the mean surface temperature of the Earth is about 33 K (59°F) warmer than the effective temperature is due to the absorption and subsequent downward reemission of substantial amounts of terrestrial radiation by water vapor, CO_2 , and O_3 , and cloud layers, all of which absorb strongly in the infrared (the so-called greenhouse effect). **Figure 7b** shows position of the absorption bands of the atmosphere's major gaseous constituents in relation to

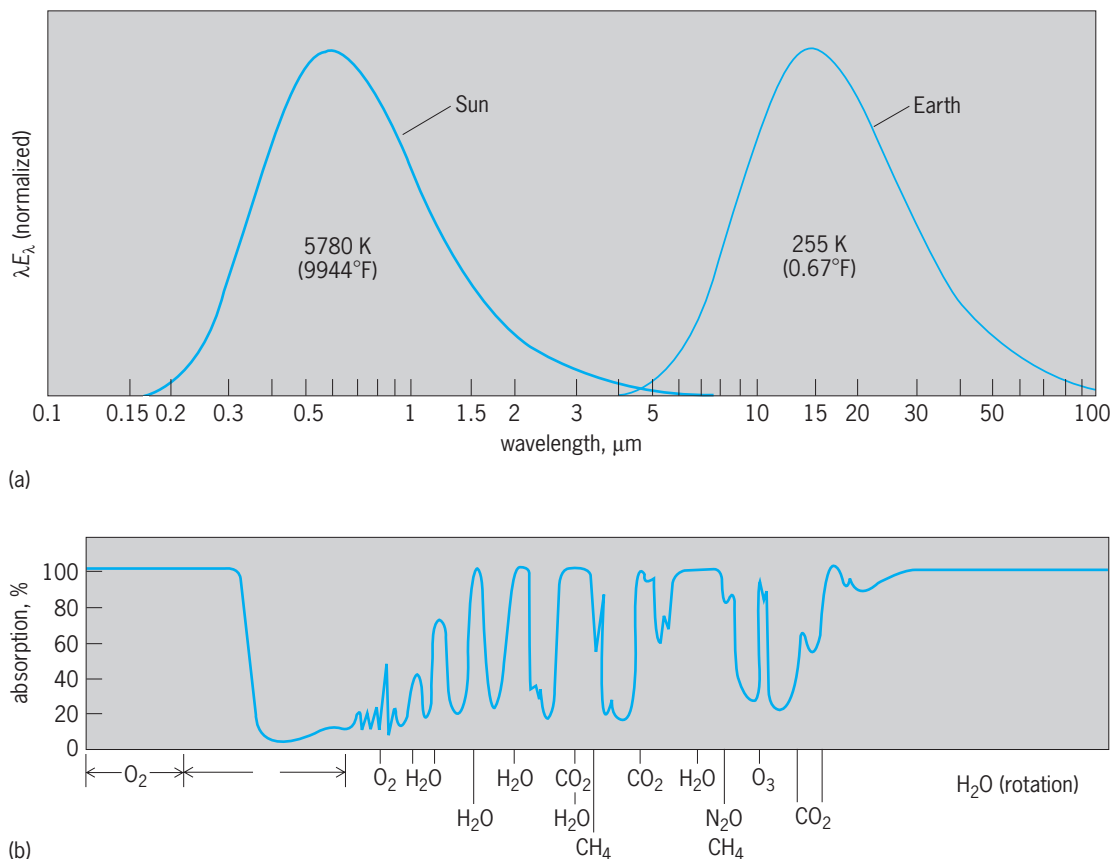


Fig. 7. Radiation curves. (a) Normalized blackbody spectra representative of the Sun and the Earth plotted on a logarithmic scale. (b) Absorption spectrum for the Earth's atmosphere as a whole. The ordinate denotes the fraction of the radiation incident upon the Earth's atmosphere (either from above or below) that is absorbed during its passage through the atmosphere; the gaseous constituents primarily responsible for the absorption of the radiation at various wavelengths are indicated at the horizontal axis. (After R. M. Goody, *Atmospheric Radiation*, Oxford University Press, 1964)

the spectra of solar and terrestrial radiation. The broad spectral window that enables most of the incident solar radiation (apart from that reflected by clouds) to reach the Earth's surface is readily apparent. The main spectral window through which terrestrial radiation escapes to space is in the vicinity of 10 μm. The increasing concentration of CO₂ in the atmosphere due to the burning of fossil fuels is causing its absorption bands to encroach on the long-wavelength end of this window, leading to predictions of a substantial global greenhouse warming. Methane (CH₄), nitrogen oxides (NO_x), and chlorofluorocarbons, whose concentrations are also increasing in the Earth's atmosphere at least partly as a result of human activity, exhibit absorption bands closer to the middle of this window. Hence, even though their concentrations are orders-of-magnitude smaller than that of CO₂, they are also a source of concern. Based on current projections, increases in these constituents over the next 50 years could cause a global warming comparable to that resulting from a doubling of CO₂. See GREENHOUSE EFFECT; HALOGENATED HYDROCARBON.

The mean surface temperature of the Earth is determined by a rather delicate balance between the globally averaged fluxes of solar and terrestrial radi-

ation, sensible heat, and moisture. The 100 units of incoming radiation shown in Fig. 8 represents the solar irradiance passing through the Earth's orbit (S) times the cross-sectional area of the Earth (πR^2), divided by the total surface area of the Earth ($4\pi R^2$): hence, it is given by $S/4 = 345 \text{ W} \cdot \text{m}^{-2}$. The combined effects of the reflection of solar radiation by clouds, air molecules, and the Earth's surface amount to 30 of the 100 units of incoming solar radiation, which accounts for the planetary albedo of 30% alluded to above. Absorption of solar radiation by water-vapor molecules, ozone molecules in the stratosphere, clouds, and aerosols together amounts to 19% of the incoming solar energy. The remaining 51 units are absorbed at the Earth's surface and eventually are returned to the atmosphere by the processes indicated on the right-hand side of the diagram.

The Earth's surface can be regarded as a blackbody at 288 K (15°C or 59°F), for which the Stefan-Boltzmann law predicts in upward irradiance of $390 \text{ W} \cdot \text{m}^{-2}$, or 113 of the units in Fig. 8. The irradiance of downward infrared radiation emitted by the atmosphere is 92 units, leaving a net upward irradiance of only 21 units, as shown. If the Earth's surface were in radiative equilibrium (in which case it would have to dispose of the full 51 units of solar

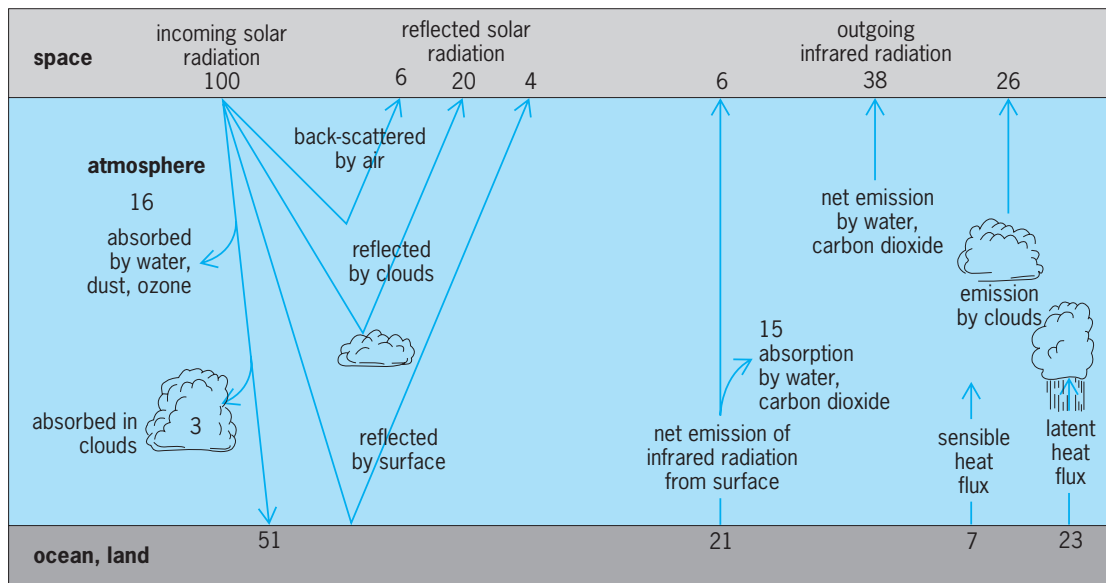


Fig. 8. Annual mean global energy balance of the Earth–atmosphere system. Numbers are given as percentages of the globally averaged solar radiation incident upon the top of the atmosphere. (After *Understanding Climatic Change, U.S. National Academy of Sciences, 1975*)

radiation that it absorbs by a net upward irradiance of infrared radiation), its mean temperature would have to be on the order of 340 K (152°F). The remaining 30 units are transferred to the atmosphere through the fluxes of sensible heat ($c_p T$) and the latent heat of vaporization of water evaporated at the Earth's surface, which is eventually transferred to the surrounding air molecules when the vapor condenses in clouds. From Fig. 8 it is evident that the effective temperature of the Earth, as viewed from space, is a weighted average of the temperatures of the surface ($^6/_{70}$), the greenhouse gas molecules ($^{38}/_{78}$), and cloud tops ($^{26}/_{70}$). The role of water vapor and clouds will be discussed in more detail in the next section.

The estimates presented in Fig. 8 refer to annual and globally averaged quantities. The latitudinal seasonal variation net incoming solar radiation at the top of the atmosphere is shown in Fig. 9. The polar night region is shaded. The small deviations from equatorial symmetry are a consequence of the ellipticity of the Earth's orbit: the Earth is closest to the Sun in January. The equator-to-pole heating gradient is very strong in the winter hemisphere, and it disappears or even reverses for about a month centered on the summer solstice. In the annual average, the polar regions absorb only about 30% as much solar radiation per unit area as the equatorial belt when the high reflectivity of the ice and snow and the persistent cloud cover in those regions are taken into account. The outgoing infrared radiation through the top of the atmosphere (not shown) is proportional to the fourth power of the local effective temperature T_E at which the radiation is emitted. Annual average surface temperature varies from nearly 300 K (80°F) in the tropics to about 250 K (−10°F) in the polar regions, which would be consistent with a 2:1 ratio of outgoing radiation between Equator and poles. Because cloud tops and water vapor, one of the major

gaseous sources of terrestrial radiation, extend to higher levels in the tropical atmosphere than in the polar regions and are therefore relatively colder, the actual ratio is closer to 1.3:1.

Averaged over the entire globe and over the whole year, the net downward solar radiation through the top of the atmosphere must very nearly equal the upward infrared radiation. Since the meridional gradient of the former is stronger, it follows that at low latitudes the incoming solar radiation must exceed the local outgoing infrared radiation, while at higher latitudes the opposite situation must prevail. Hence, the atmosphere and oceans must transport energy poleward from a low-latitude source to a high-latitude sink. This north-south heating gradient, which is much larger in the winter hemisphere than in the summer hemisphere, is the main driving force for large-scale atmospheric motions, the so-called general circulation. Land-sea contrasts (differences in heat capacity, thermal conductivity, reflectivity, and availability of moisture between land and sea and between different types of land surfaces) also contribute to the large observed spatial and temporal variability of the local energy balance. The resulting east-west (zonal) heating gradients drive the monsoon circulations in the tropics and subtropics and the stationary planetary waves at higher latitudes. See MONSOON METEOROLOGY.

The energy balance arguments used in this section should not be interpreted as indicating that the Earth's climate is static. The climate system has, in fact, emerged from a major ice age less than 20,000 years ago, and it has warmed substantially since the 1880s. However, it can be shown that the energy fluxes required to account for the observed changes in the volume of the polar icecaps and the heat storage in the atmosphere and oceans on these time scales are much smaller than the uncertainties in the

estimates of the current energy balance in Fig. 8. See GLACIAL EPOCH.

Impact of hydrologic cycle. Water in its various phases exerts a profound influence not only upon the biosphere but also upon many aspects the behavior of the atmosphere. It is evident in Fig. 8 that evaporation is an important heat sink for the oceans and vegetated land surfaces, and that condensation of water vapor in clouds is the atmosphere's largest single heat source. Hence, the hydrological cycle transfers massive amounts of heat from the Earth's surface into the atmosphere. On a more local scale, condensation plays a major role in generating the buoyancy and the horizontal temperature gradients that drive hurricanes, explosively deepening extratropical cyclones, and severe thunderstorms. It occurs selectively within rising air parcels, which tend to be warmer than the environmental air at the same level, and it causes them to cool more slowly as they expand than they would under adiabatic conditions. Observed lapse rates in the troposphere rarely exceed this so-called moist adiabatic lapse rate.

Condensation of water vapor within the atmosphere involves a complex array of physical processes. Submicrometer aerosol particles known as cloud condensation nuclei play an essential role in the initial formation of droplets of liquid water, which subsequently grow by many orders of magnitude on a time scale of minutes to hours, until the speeds at which they fall become large enough to enable them to reach the ground as rain or snow. The principal mechanisms through which this growth occurs involve the coalescence of smaller droplets into larger ones through collisions and the so-called Bergeron mechanism—the freezing of a small fraction of the droplets and their subsequent growth at the expense of the remaining supercooled (below-freezing) liquid droplets through the diffusion of water-vapor molecules from one to the other.

Liquid water droplets and ice particles tend to be concentrated in macroscale clouds, for which a detailed classification scheme has been devised. Among the most common cloud types are extensive, long-lived cloud layers (cirrostratus, altostratus, stratus, stratocumulus), which cool the Earth through their contribution to the planetary albedo and warm it through their contribution to the greenhouse effect (Fig. 8). On balance, upper-tropospheric (cirrostratus) cloud layers, which emit radiation to space at low effective temperatures, tend to warm the Earth; and low cloud layers (stratus, stratocumulus) tend to cool it. Because of this compensation, the phenomenon of cloud-climate feedbacks represents one of the largest sources of uncertainty in projections of future global climate change due to the buildup of greenhouse gases in the atmosphere.

Cloud droplets provide a hospitable environment for certain types of chemical reactions that would not otherwise take place in the atmosphere. Such reactions figure prominently in the destruction of the stratospheric ozone over the polar regions and in the formation of acid rain downstream from indus-

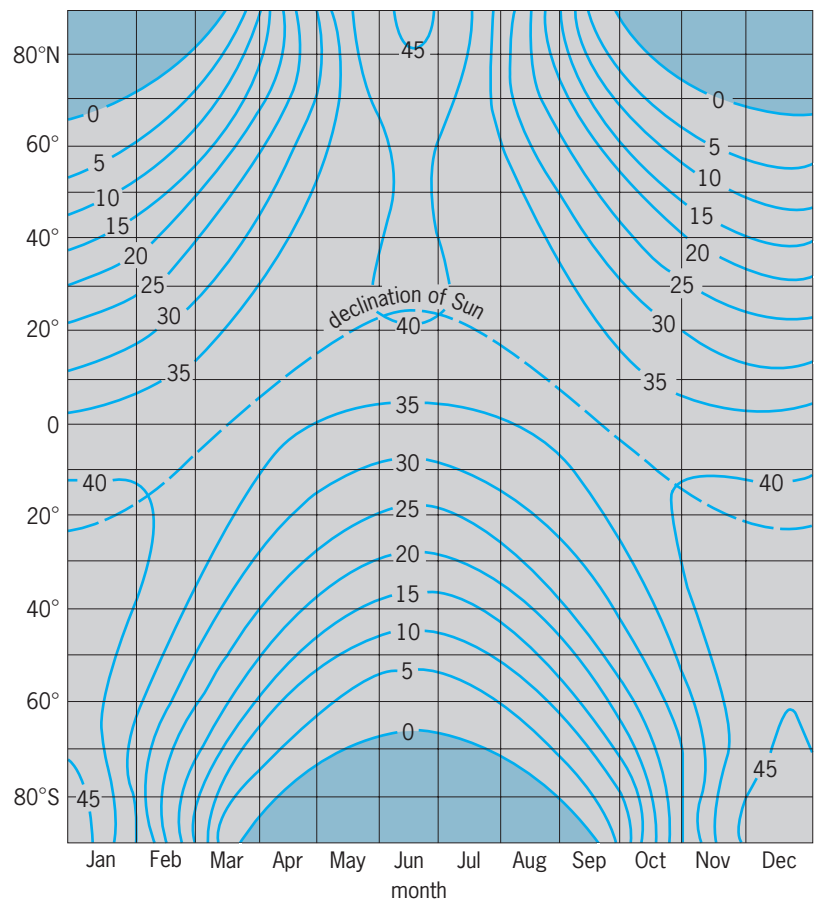


Fig. 9. Solar radiation incident upon a unit horizontal surface at the top of the atmosphere as a function of latitude and calendar date; contours are given in $10^9 \text{ J} \cdot \text{m}^{-2} \cdot \text{day}^{-1}$. (After R. J. List, ed., *Meteorological Tables*, 6th ed., Smithsonian Institution, 1951)

trial sources of sulfur dioxide (SO_2). Scavenging by cloud droplets that subsequently fall out as precipitation is the primary mechanism for cleansing dust, smoke, and other particles from the troposphere. Charge separation by cloud droplets and ice particles within clouds also has implications for the global atmospheric electric field, and it leads to the buildup of the locally strong potential gradients responsible for lightning discharges. See LIGHTNING.

The 23 units of latent heat flux from the Earth's surface in Fig. 8 are based on an observed evaporation rate of about 0.25 cm (0.1 in.) of liquid water per day averaged over the Earth's surface, which must be equal to the average rate of precipitation. The average precipitable water in the atmosphere is on the order of about 2 cm (0.8 in.) of liquid water, and thus the mean residence time for water-vapor molecules in the atmosphere must be on the order of a week. The mass of liquid water and ice present in the atmosphere in the form of cloud droplets is several-orders-of-magnitude smaller than the amount present in the vapor state. In contrast to the tropospheric air, which is often saturated with water vapor, stratospheric air tends to be remarkably dry because it enters the stratosphere by way of the extremely cold equatorial tropopause (Fig. 5), where most of its water vapor is condensed out.

Stratospheric cloud layers are observed only in the polar night region where temperatures can drop to -80°C (-112°F). See CLOUD; HYDROMETEOROLOGY; PRECIPITATION (METEOROLOGY).

Atmospheric prediction. Throughout the atmospheric sciences, prediction is a unifying theme that sets the direction for research and technological development. Prediction on the time scale of minutes to hours is concerned with severe weather events such as tornadoes, hail, and flash floods, which are manifestations of intense mesoscale weather systems, and with urban air-pollution episodes; day-to-day prediction is usually concerned with the more ordinary weather events and changes that attend the passage of synoptic-scale weather systems such as extratropical cyclones; and seasonal prediction is concerned with regional climate anomalies such as drought or recurrent and persistent cold air outbreaks. Prediction on still longer time scales involves issues such as the impact of human activity on the temperature of the Earth, regional climate, the ozone layer, and the chemical makeup of precipitation. See CLIMATE MODELING; DROUGHT; HAIL; TORNADO.

The evolution of the atmospheric sciences from a largely descriptive field prior to World War II to a mature, quantitative physical science discipline is apparent in the development of vastly improved predictive capabilities based upon the numerical integration of specialized versions of the Navier-Stokes equations, which include sophisticated parametrizations of physical processes such as radiative transfer, latent heat release, and microscale motions. The so-called numerical weather prediction models have largely replaced the subjective and statistical prediction methods that were widely used as a basis for day-to-day weather forecasting as recently as the 1950s. The state-of-the-art numerical models exhibit significant skill for forecast intervals as long as about a week. See NAVIER-STOKES EQUATION.

A distinction is often made between weather prediction, which is largely restricted to the consideration of dynamic and physical processes internal to the atmosphere, and climate prediction, in which interactions between the atmosphere and other elements of the climate system are taken into account. The importance and complexity of these interactions tend to increase with the time scale of the phenomena of interest in the forecast. Weather prediction involves shorter time frames (days to weeks), in which the information contained in the initial conditions is the dominant factor in determining the evolution of the state of the atmosphere; and climate prediction involves longer time frames (seasons and longer), for boundary forcing is the dominant factor in determining the state of the atmosphere.

As in many other systems governed by nonlinear equations, the uncertainties inherent in the definition of the initial state of the atmosphere grow exponentially with time during the forecast until they become as large as typical differences between two arbitrarily chosen states of the atmosphere. Deterministic prediction based solely upon the information contained in the initial conditions is impossible

beyond this time frame. The characteristic predictability time of large-scale atmospheric circulation patterns is believed to be on the order of 2 weeks. Prediction on longer time scales such as months or seasons exploits the memory (autocorrelation) and extended predictability of the more slowly varying components of the climate system, which force the atmosphere at its lower boundary. On these longer time scales, prediction is focused on climate anomalies (that is, departures of statistics such as seasonal mean temperature and precipitation from their climatological mean values), which persist or develop in response to the slowly varying boundary forcing. Deterministic prediction of sequences of day-to-day weather changes is not feasible on these extended time scales.

Atmospheric prediction has benefited greatly from major advances in remote sensing. Geostationary and polar orbiting satellites provide continuous surveillance of the global distribution of cloudiness, as viewed with both visible and infrared imagery. These images are used in positioning of features such as cyclones and fronts on synoptic charts. Cloud motion vectors derived from consecutive images provide estimates of winds in regions that have no other data. Passive infrared and microwave sensors aboard satellites also provide information on the distribution of sea-surface temperature, sea state, land-surface vegetation, snow and ice cover, as well as vertical profiles of temperature and moisture in cloud-free regions. Improved ground-based radar imagery and vertical profiling devices provide detailed coverage of convective cells and other significant mesoscale features over land areas. Increasingly sophisticated data assimilation schemes are being developed to incorporate this variety of information into numerical weather prediction models on an operational basis. See ATMOSPHERE; CLIMATIC PREDICTION; CYCLONE; FRONT; RADAR METEOROLOGY; SATELLITE METEOROLOGY; WEATHER FORECASTING AND PREDICTION.

John M. Wallace

Bibliography. S. Ackerman and J. A. Knox, *Meteorology: Understanding the Atmosphere*, 2002; C. D. Ahrens, *Meteorology Today*, 2d ed., 2006; W. S. Broecker, *How to Build a Habitable Planet*, 1985; F. K. Lutgens et al., *The Atmosphere: An Introduction to Meteorology*, 9th ed., 2003; V. J. Schaefer and J. Day, *A Field Guide to the Atmosphere*, 1981; R. B. Stull, *Meteorology for Scientists and Engineers*, 2d ed., 1999; J. M. Wallace and P. V. Hobbs, *The Atmosphere: An Introductory Survey*, 1977.

Methane

The simplest compound of carbon and hydrogen. Each molecule of methane contains four atoms of hydrogen bound in a tetrahedral arrangement to one atom of carbon (see **illustration**). At room temperature, methane is a gas less dense than air. The gas liquefies at -164°C (-263°F) and solidifies at -183°C (-297°F). It is not very soluble in water. Methane is combustible, and mixtures of about 5–15% in



Molecular model of methane.

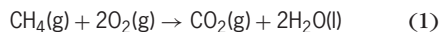
air are explosive. Complete combustion of methane produces carbon dioxide and water. Methane is not toxic when inhaled, but it can produce suffocation by reducing the concentration of oxygen. *See* ALKANE.

Methane is widely distributed in nature. It is the principal component of natural gas, a mixture that by volume contains about 75% methane (CH₄), 15% ethane (C₂H₆), and 5% other hydrocarbons such as propane (C₃H₈) and butane (C₄H₁₀). The combustible gas found in coal mines is chiefly methane. Marsh gas, which is produced under water by anaerobic bacterial decomposition of plant and animal matter, is also methane. Some plants emit methane, and many animals emit methane as a by-product of food digestion. It is estimated that domestic animals account for about 15% of annual methane emissions. Although methane constitutes only about 0.00017% by volume of the atmosphere, its concentration has more than doubled in the last 250 years. Methane is one of the most powerful greenhouse gases in the atmosphere, being more than 60 times as effective as carbon dioxide. *See* COALBED METHANE; GREENHOUSE EFFECT; NATURAL GAS.

Natural gas occurs in reservoirs beneath the surface of the Earth, and it is often found in conjunction with petroleum deposits. Before it is commercially distributed, natural gas usually undergoes some sort of processing in which the heavier hydrocarbons (propane and butane) and nonhydrocarbon gases (such as hydrogen sulfide) are removed. The cleaned gas is then distributed through pipelines. Natural gas is also distributed worldwide in liquid form (liquefied natural gas, or LNG), which is produced by chilling the gas to below its boiling point. Because an undetected natural gas leak could result in an explosion or asphyxiation, local utilities add an odorant to make the gas readily detectable by smell. The odorant is a small amount of a mixture of organic sulfur compounds such as tertiarybutyl mercaptan [(CH₃)₃CSH] and dimethyl sulfide (CH₃—S—CH₃). *See* LIQUEFIED NATURAL GAS (LNG).

Methane is synthesized commercially by distilling bituminous coal and by heating a mixture of carbon and hydrogen. It can be produced in the laboratory by heating sodium acetate with sodium hydroxide and by the reaction of aluminum carbide (Al₄C₃) with water.

Uses. The principal use of methane is as a fuel. The combustion of methane is highly exothermic [reaction (1)]. The energy released by the combus-



$$\Delta H = -891 \text{ kJ}$$

tion of methane, in the form of natural gas, is used directly to heat homes and commercial buildings. It is also used in the generation of electric power. In 1996–2005, natural gas accounted for one-fifth of the total energy consumption worldwide and about one-third in the United States.

In the chemical industry, methane is a raw material for the manufacture of methanol (CH₃OH), formaldehyde (CH₂O), nitromethane (CH₃NO₂), chloroform (CH₃Cl), carbon tetrachloride (CCl₄), and some halogenated hydrocarbons (compounds containing carbon and fluorine, and perhaps chlorine and hydrogen). The reactions of methane with chlorine and fluorine are triggered by light, and when exposed to bright visible light, mixtures of methane with chlorine or fluorine can react explosively. Industrially, methane is converted (reformed) catalytically by the Fischer-Tropsch process to a mixture of carbon monoxide and hydrogen gases called syngas [reaction (2)]. Via other catalysts, syngas is converted



to liquid diesel fuel. The hydrogen gas produced in the reformation of methane is used in another catalytic process to make ammonia which, in turn, is converted to important fertilizers. *See* FISCHER-TROPSCH PROCESS; SYNTHETIC FUEL.

Hydrates. Methane hydrates are solids consisting of methane and water, and these white solids are less dense than water. Methane hydrates are examples of clathrate hydrates, in which “cages” of water molecules, held together by hydrogen bonds, trap gases (such as N₂, O₂, CO₂, CH₄, H₂S) at high pressure and low temperature. On average, methane hydrates contain about one mole of methane to almost six moles of water. Methane hydrates are found in the oceans at several hundred feet below sea level, where they are trapped in mud and rocks. When brought to the surface, they decompose to water and methane, and the methane burns—hence, the name ice-on-fire for the phenomenon. Methane hydrates are being investigated as a commercial source of methane. *See* CLATHRATE COMPOUNDS; HYDRATE.

Planets. The detection of methane on other planets is of major scientific interest. Whether on Mars or the cold surface of Saturn’s largest moon, Titan, the presence of methane and the conjecture of riverlike channels due to liquid methane have intrigued scientists and spurred closer investigations.

Bassam Z. Shakhshiri; Rodney Schreiner

Bibliography. S. Lee, *Methane and Its Derivatives*, 1996; R. T. Morrison and R. N. Boyd, *Organic Chemistry*, 6th ed., 1992; E. D. Sloan, *Clathrate Hydrates of Natural Gases*, 2d ed., 1998.

Methanogenesis (bacteria)

The microbial formation of methane. It is confined to anaerobic habitats where the production of hydrogen, carbon dioxide, formic acid, methanol, methylamines, or acetate—the major substrates used by methanogenic microbes (methanogens)—occurs.

Habitats. In freshwater or marine sediments, in the intestinal tracts of animals, or in habitats engineered by humans such as sewage sludge or biomass digesters, these substrates are the products of anaerobic bacterial metabolism. Methanogens are terminal organisms in the anaerobic microbial food chain—the final product, methane, being poorly soluble, anaerobically inert, and not in equilibrium with the reaction which produces it. In hydrothermal vents, hydrogen formation is a product of geochemistry. In methanogenic habitats, the reducing potential may approximate the hydrogen electrode. The reducing potential in hydrothermal vents is maintained by geochemically produced hydrogen sulfide; and in sediments, especially marine sediments, hydrogen sulfide is a product of bacterial reduction of sulfate. In freshwater sediments and in the intestinal tract of animals, reducing potential is mainly maintained by the hydrogenases of fermentative anaerobes that produce molecular hydrogen. *See* HYDROTHERMAL VENT.

Two highly specialized digestive organs, the rumen and the cecum, have been evolved by herbivores to delay the passage of cellulose fibers so that microbial fermentation may be complete. In these organs, large quantities of methane are produced from hydrogen and carbon dioxide or formic acid by methanogens. From the rumen, which may hold 26 gallons (100 liters) of fermenting biomass, an average cow may belch 26 gallons of methane per day. Methanogens may form endosymbiotic or exosymbiotic relationships with cellulose-ingesting protozoa in the rumen, cecum, hind gut of termites, or sediments.

Isolation and cultivation of methanogens. Methanogens are sensitive to oxygen and cannot be cultivated by ordinary bacteriological techniques; exclusion of oxygen is essential but not sufficient. To plate methanogens on solid agar media in petri dishes, an anaerobic chamber is required for routine bacteriological techniques. Colonies on plates must be examined in an anaerobic chamber. Cell walls of methanogens lack the typical peptidoglycan of bacteria, so methanogens are not sensitive to the common cell-wall antibiotics such as cycloserine and penicillin. Addition of such antibiotics to growth media inhibits the growth of typical bacteria and aids in the selective isolation of methanogens from crude inocula.

Organisms. Methanogens are the only living organisms that produce methane as a way of life. The biochemistry of their metabolism is unique in the biological world and definitively delineates the group. *See* ARCHAEA.

Metabolic reactions are energy-yielding reactions

that are carried out in nature. Two reductive biochemical strategies are employed: an eight-electron reduction of carbon dioxide to methane or a two-electron reduction of a methyl group to methane. All methogens form methane by reducing a methyl group.

The major energy-yielding reactions used by methanogens utilize substrates such as hydrogen, formic acid, methanol, acetic acid and methylamine. Dimethyl sulfide, carbon monoxide, and alcohols such as ethanol and propanol are substrates that are used less frequently.

Methanococcus vannielii is a motile coccus, isolated from marine sediment, that grows at 20–40°C. In contrast, *Methanocaldococcus jannaschii* (formerly known as *Methanococcus jannaschii*), isolated from a deep-sea hydrothermal vent, grows at 48–94°C with optimal growth at 85°C.

Of the rod-shaped organisms, *Methanobrevibacter ruminantium*, isolated from a bovine rumen, is a short rod with tapering ends that grows at 37–39°C. The organism has complex nutritional requirements, including B vitamins, coenzyme M, amino acids, acetate, and 2-methylbutyrate. *Methanobacterium formicicum*, isolated from sediment, is a slightly bent rod that grows at 25–37°C. *Methanothermobacter thermoautotrophicus* (formerly known as *Methanobacterium thermoautotrophicum*), isolated from a sewage sludge digester, grows at 50–75°C. This species is easily cultivated in kilogram quantities at 65°C on hydrogen and carbon dioxide in a mineral salts medium. Most of the biochemistry of carbon dioxide reduction to methane was revealed through study of this organism.

Methanothermus fervidus, isolated from a geothermal area in Iceland, is a short rod that grows at 60–97°C. *Methanosphaera stadtmaniae*, although closely related to the rod-shaped methanogens, has a spherical morphology and is unique because it cannot reduce carbon dioxide to methane. Instead, hydrogen is oxidized and a methyl group is reduced to methane. The blunt rods of *Methanosaeta concilii* are encased in a sheath. The organism is widely distributed in nature and grows only by an acetoclastic reaction. *Methanosarcina barkeri*, the most metabolically diverse methanogen, also has the ability to use the acetoclastic reaction. Coccoid cells form large clumps. Most isolates are mesophilic; a few will grow at 50°C.

Methanogenium cariaci, a marine coccoid organism that is not closely related to *Methanococcus*, grows optimally at 40°C. *Methanospirillum hungatei*, isolated from a sewage sludge digester, grows at 35–40°C and is the only spiral-shaped methanogen. The most distantly related of all methanogens is the extreme thermophile *Methanopyrus kandleri*, which was isolated from a hydrothermal vent and has an upper limit of growth at 110°C. *See* BACTERIAL PHYSIOLOGY AND METABOLISM; METHANE.

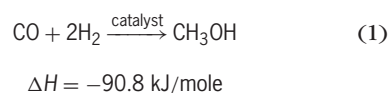
Ralph S. Wolfe

Bibliography. J. G. Ferry (ed.), *Methanogenesis*, 1993.

Methanol

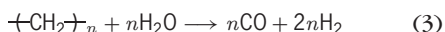
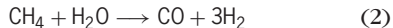
The first member of the homologous series of alcohols, with the formula CH_3OH . This compound was originally obtained by the destructive distillation of wood as a by-product in the preparation of charcoal, hence the older name wood alcohol. Methanol is a highly flammable liquid, boiling point 64.7°C (149°F), and is miscible with water and most organic liquids. It is a highly poisonous substance; sublethal amounts can cause permanent blindness. See ALCOHOL.

Methanol is one of the major industrial organic chemicals. It is produced commercially from a mixture of carbon monoxide (CO) and hydrogen (H_2), generally known as synthesis gas, according to reaction (1), where ΔH = change in enthalpy. The pro-



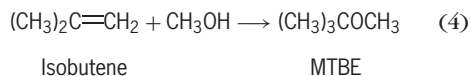
cess is carried out at $200\text{--}300^\circ\text{C}$ ($390\text{--}570^\circ\text{F}$) at moderate pressure (50 atm or 5 MPa) with a copper-based catalyst. See CATALYSIS; ENTHALPY.

The synthesis gas used to produce methanol can be produced from coal, natural gas, or petroleum fractions. The major source of synthesis gas is the steam reforming of methane (CH_4) or other hydrocarbons. Reactions (2) and (3) represent the basic



steps; the overall process also includes a partial oxidation step to supply heat and adjust the CO/ H_2 ratio. See REFORMING PROCESSES.

Methanol has the usual chemical properties of a primary alcohol, undergoing traditional reactions to give methyl acetals, amines, ethers, esters, and halides. Addition to isobutene gives methyl *t*-butyl ether (MTBE) [reaction (4)], which is produced on



large scale as an octane enhancer to replace lead additives in gasoline.

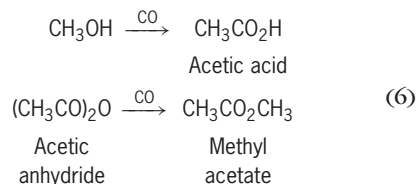
The major industrial use of methanol is conversion to formaldehyde (CH_2O). This process is carried out by the reaction of methanol with a limited amount of oxygen over a silver catalyst [reaction (5)].



See FORMALDEHYDE.

Since the 1970s, several manufacturing processes have been developed based on transition metal-catalyzed carbonyl insertion reactions. One of the most important of these is the carbonylation of methanol to acetic acid. A rhodium-based catalyst system was the first commercial application, but a less costly process using a nickel catalyst has been

reported. Similar carbonylation converts methyl acetate to acetic anhydride [reactions 6].



Carbonyl insertion can be coupled with oxidation or reduction, thus leading to an extensive suite of compounds, including such important products as vinyl acetate and ethylene glycol. Methanol or its immediate derivatives formaldehyde and methyl formate are central to these processes, all of which are based on synthesis gas as feedstock. Since the technology exists for obtaining synthesis gas from coal also, this so-called C_1 chemistry provides an important alternative to petroleum- and ethylene-based raw materials for the chemical industry. See COAL GASIFICATION; ORGANIC SYNTHESIS.

The potential of methanol as an alternative fuel has been widely discussed. The volatility, ease of transport and distribution, and low combustion emissions combine to make its use as an automotive fuel a realistic possibility. See ALCOHOL FUEL. James A. Moore

Bibliography. D. R. Fahey (ed.), *Industrial Chemicals via C_1 Processes*, Amer. Chem. Soc. Symp. Ser. 328, 1987; J. Falbe, *Chemical Feedstocks from Coal*, 1982; C. L. Gray and J. Alson, *Moving America to Methanol*, 1985; R. A. Sheldon, *Chemicals from Synthesis Gas*, 1983; Society of Automotive Engineers, *Fuel Methanol: A Decade of Progress*, 1990.

Methods engineering

A technique used by industrial engineers to improve productivity and quality and to reduce costs in both direct and indirect operations of manufacturing and service organizations. Methods engineering is applicable in any enterprise requiring human effort. It can be defined as the systematic procedure for subjecting all direct and indirect operations to close scrutiny in order to introduce improvements that will make work easier to perform while maintaining or improving quality, and will allow work to be done more smoothly, in less time, with less energy, effort, and fatigue, and with less investment per unit. The ultimate objective of methods engineering is increasing profits, but it is also important in improving worker health and safety.

The terms methods engineering, operation analysis, and work design are frequently used synonymously. In most cases, methods engineering refers to a technique for increasing the production per unit of time or decreasing the cost per unit output—in other words, productivity improvement. However, methods engineering entails analysis work at two different times during the history of a product. Initially, the methods engineer is responsible for designing and developing the various work centers

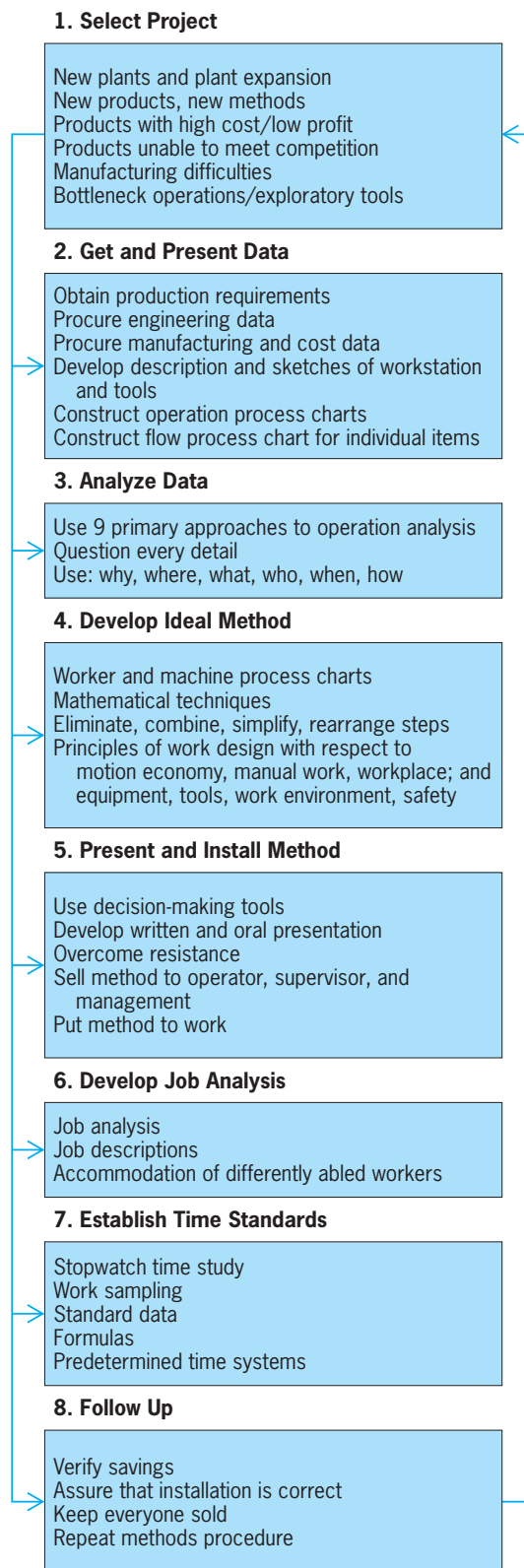


Fig. 1. Principal steps in a methods engineering program.

where the product will be produced. Next, he or she continually restudies the work centers to find a better way to produce the product and/or improve its quality. More recently, this second analysis has been called reengineering or lean manufacturing.

See LEAN MANUFACTURING; OPERATIONS RESEARCH; PRODUCTIVITY.

Steps. Methods engineers use a systematic procedure (Fig. 1) to develop a work center, to produce a product, or to provide a service. This procedure is outlined below. Note that Steps 6 and 7 are not strictly part of a methods study but are necessary in a fully functioning work center.

1. *Select the project.* Typically, projects selected represent either new or existing products that have a high cost of manufacture and a low profit. Also, products that have quality problems and difficulty in meeting competition are logical methods engineering projects.

2. *Get and present the data.* Assemble all the important facts relating to the product or service. These include drawings and specifications, quantity requirements, delivery requirements, and projections about the anticipated life of the product or service. Once all the important information has been acquired, record it in an orderly form for study and analysis. The development of process charts at this point is very helpful.

3. *Analyze the data.* Utilize the primary approaches to operations analysis to decide which alternative will produce the best product or service. These primary approaches include purpose of operation, design of part, tolerances and specifications, materials, manufacturing process, setup and tools, working conditions, materials handling, plant layout, and principles of motion economy.

4. *Develop the ideal method.* Select the best procedure for each operation, inspection, and transportation by considering the various constraints associated with each alternative, including productivity, ergonomic, and health and safety implications.

5. *Present and install the method.* Explain the proposed method in detail to those responsible for its operation and maintenance. Consider all details of the work center to assure that the proposed method will provide the results anticipated.

6. *Develop a job analysis.* Make a job analysis of the installed method to assure that the operator or operators are adequately selected, trained, and rewarded.

7. *Establish time standards.* Establish a fair and equitable standard for the installed method.

8. *Follow up the method.* At regular intervals, audit the installed method to determine if the anticipated productivity and quality are being realized, if costs were correctly projected, and if further improvements can be made.

As a part of developing or maintaining the new method, it is also important to use the principles of work design to ergonomically fit the task and workstation to the human operator. Unfortunately, work design is typically forgotten in the quest for increased productivity. Far too often, overly simplified procedures result in machinelike repetitive jobs for the operators, leading to increased rates of work-related musculoskeletal disorders. Any increases in productivity and reduced costs are more than offset by the

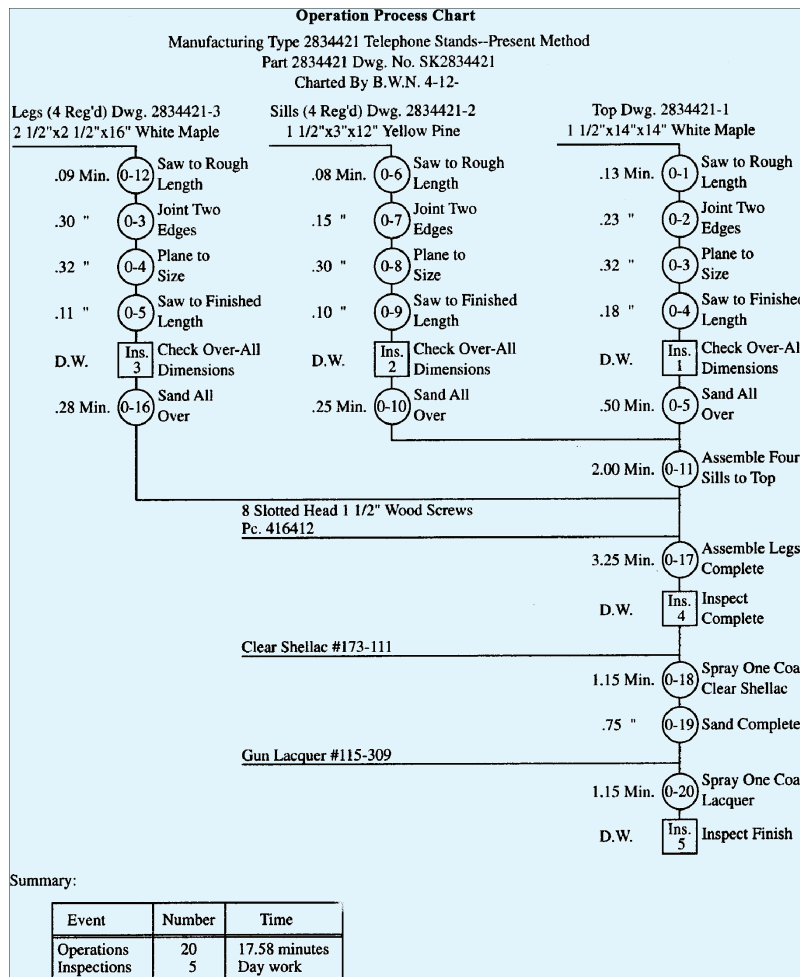


Fig. 2. Operation process chart.

increased medical and workers' compensation costs, especially considering today's ever-escalating health-care trends. Thus, it is necessary for the methods engineer to incorporate the principles of work design into a new method so that the method will be not only more productive but also safe and injury-free for the operator.

Process chart analysis. A process chart is a graphic representation of events occurring during a series of actions or operations. It is most useful in Step 2 (Get and present data) for obtaining an initial understanding of a process, and in Step 3 (Analyze data) and in Step 4 (Develop the ideal method) for analyzing the process and developing a better one. The four process charts most often used are the operation process chart, flow process chart, worker-machine process chart, and operator process chart. The operation process chart and flow process chart are used for analyzing processes involving a number of events or operations. The multiple activity process chart and operator process chart are used to analyze single operations in detail.

An operation process chart shows the chronological sequence of all the operations (and sometimes inspections) in a manufacturing process or service

operation (Fig. 2). Vertical lines indicate the general flow of the process, while horizontal lines indicate materials or components feeding into the main process. Sometimes, time values for the processes are included. The completed chart helps analysts visualize the present method in hopes of eliminating unnecessary steps. It is particularly useful in connection with plant layout studies.

A flow process chart (Fig. 3) represents the sequence of all operations, transportations, inspections, delays, and storages occurring during a process or procedure, and includes the information needed for analysis, such as time required and distance moved. It is used to follow either materials or workers through a process and is an essential tool for materials-handling studies.

A worker-machine process (Fig. 4) is used to study, analyze, and improve one workstation at a time. The chart shows the exact time relationship between the working cycle of the person and the operating cycle of the machine. These facts can lead to fuller utilization of both work and machine time, with the goal of having one worker operate more than one machine, known as machine coupling. If both worker and machines are operating on regular cycles, the relationship is known as synchronous servicing, and the

| Flow Process Chart | | | | | Page 1 of 1 | | | |
|---|-----------|-------------------|--------------------|---|---------------|---------|----------|---------|
| Location: Dorben Co. | | | | | Summary | | | |
| Activity: Field Inspection of LUX | | | | | Event | Present | Proposed | Savings |
| Date: 4-17-97 | | | | | Operation | 7 | | |
| Operator: T. Smith | | Analyst: R. Ruhf | | | Transport | 6 | | |
| Circle appropriate Method and Type: | | | | | Delay | 2 | | |
| Method: (Present) Proposed | | | | | Inspection | 6 | | |
| Type: (Worker) Material Machine | | | | | Storage | 0 | | |
| Remarks: | | | | | Time (min) | 32.60 | | |
| | | | | | Distance (ft) | 375 | | |
| | | | | | Cost | | | |
| Event Description | Symbol | Time (in Minutes) | Distance (in Feet) | Method Recommendation | | | | |
| Leave vehicle, walk to front door, ring bell. | ○ ◊ D □ ▽ | 1.00 | 75 | Call home in advance to reduce waiting delays. | | | | |
| Wait, enter home. | ○ ◊ D □ ▽ | | | | | | | |
| Walk to field reservoir. | ○ ◊ D □ ▽ | .25 | 25 | | | | | |
| Disconnect field reservoir from unit. | ○ ◊ D □ ▽ | .35 | | | | | | |
| Inspect for dents, cracks in shroud, cracked glass or missing hardware. | ○ ◊ D □ ▽ | 1.25 | | This can be done while walking back to vehicle. | | | | |
| Clean unit with approved cleaner and disinfectant. | ○ ◊ D □ ▽ | 2.25 | | This can be done more effectively at vehicle. | | | | |
| Return to vehicle with empty tank. | ○ ◊ D □ ▽ | 1.00 | 75 | | | | | |
| Unlock vehicle, place empty tank in fixture and connect hardware. | ○ ◊ D □ ▽ | 1.75 | | | | | | |
| Open valve; begin fill. | ○ ◊ D □ ▽ | .25 | | | | | | |
| Wait for tank to fill. | ○ ◊ D □ ▽ | 12.00 | | Clean unit while being filled. | | | | |
| Check humidifier for proper function. | ○ ◊ D □ ▽ | .5 | | Eliminate. No need to do this twice. | | | | |
| Check pressure (indicator). | ○ ◊ D □ ▽ | .2 | | | | | | |
| Check reservoir contents (indicator). | ○ ◊ D □ ▽ | .2 | | | | | | |
| Return to patient with filled tank. | ○ ◊ D □ ▽ | 1.10 | 100 | | | | | |
| Hook up filled tank. | ○ ◊ D □ ▽ | 1.00 | | | | | | |
| Check humidifier for proper function. | ○ ◊ D □ ▽ | .75 | | | | | | |
| Wait for patient to remove nasal cannula or face mask. | ○ ◊ D □ ▽ | 2.00 | | | | | | |
| Install new nasal cannula or face mask. | ○ ◊ D □ ▽ | 2.50 | | | | | | |
| Check flows with patient. | ○ ◊ D □ ▽ | 2.25 | | | | | | |
| Affix a dated, initialed inspection sticker. | ○ ◊ D □ ▽ | 1.00 | | Perform this while unit being filled. | | | | |
| Return to vehicle. | ○ ◊ D □ ▽ | 1.00 | 100 | | | | | |

Fig. 3. Flow process chart, worker type.

optimum number of machines assigned to a worker can be found with the formula

$$N \leq \frac{l + m}{l + w}$$

where N is the number of machines assigned to worker, l is the operator loading/unloading time (both worker and machine interacting), m is the machine running time (independent of worker), and w is the worker time (not directly with the machines).

An operator process chart, sometimes referred to as a two-hand process chart, is a motion-study tool (Fig. 5). This chart shows all movements and delays made by the right and left hands, and the relationship between the hands. Each task element performed by the operator can be subdivided into basic motion elements, or therbligs (Gilbreth spelled backward). The therbligs are generally partitioned into effective ones (reach, move, grasp, release, preposition, use, assemble, and disassemble) and ineffective ones (search, select, plan, position, hold, avoidable delay, unavoidable delay, and rest to overcome fatigue). In general, the first eight accomplish neces-

sary work, while the next seven retard the progress of work and should be eliminated when possible. A summary of the effective and ineffective times (corresponding to effective and ineffective therbligs) for left and right hands yields a starting point for methods improvements, the goal being the reduction of the ineffective therbligs. Frequently, therblig analysis can evolve into the setting of standard times using predetermined time systems, most of which have evolved from these same therbligs.

Operations analysis. There are 10 major points of operation analysis that are applied in methods study. Listed in the order in which they usually are considered, they are purpose of operation, design of part, tolerances and specifications, material, process of manufacture, setup and tools, materials handling, plant layout, working conditions, and ergonomics and principles of motion economy.

Where work has not been previously studied in detail, industrial engineers repeatedly find that much of what is being done is unnecessary. Sometimes whole operations are eliminated merely by asking why something is being done and recognizing that

there is no sound answer. The primary technique used to discover such situations is through the analysis of the process charts, considering the methods study approach of purpose of operation.

When the methods analyst considers the design of the part, the objective is to discover manufacturing economics that the design engineer may have overlooked. Considering tolerances and specifications will often suggest better ways of obtaining necessary quality and product reliability. Questioning the materials of which the product is made and the way that the supply materials are used in performing the operations required to make the product will frequently lead to economies. There are usually many different ways to perform a given operation. By questioning the process of manufacture, the methods analyst will consider all alternative processes to determine if the best ones are being used. Software can be developed to evaluate alternative processes based on information such as material being processed, quantity to be produced, geometric configuration of the desired product, and tolerance and finish constraints. Just as there usually are several alternative processes to be considered, there are alternative tooling and setup procedures. The larger the quantity of parts to be produced, the more advanced the tooling should be. As quantity requirements change, existing setup and tooling may become inefficient. Under numerical control, tool setting can be done off-line while the facility is producing. Usually, preassembled fixtures are much better than disassembling and re-assembling fixtures during setup. Consequently, the methods analyst should always consider this side of operations analysis. See PRODUCTION ENGINEERING.

Materials handling should always be given serious attention. Within a factory or other enterprise, handling of materials adds to the cost of a product without adding to its salable value. Good materials handling provides for the delivery of an adequate inventory of material at the proper time and in the proper condition to the point of use at the least cost. The operation and flow process charts are especially used to give a clear picture of present materials-handling methods and to point the way to new and improved ones. See MATERIALS HANDLING.

The physical layout is an important element of an entire production system. The methods analyst will strive to have an effective layout that permits the manufacture of the desired number of products of the desired quality at the least cost. Workstations and facilities, including service facilities, should be arranged to permit the least travel and the most efficient processing of a product with a minimum of handling. Computer programs are used to assist in the development of a sound plant layout. See INDUSTRIAL FACILITIES.

Analysis of the working conditions of the immediate environment of each workstation can be another source of savings, productivity, and quality improvement. The methods analyst will study the working conditions of each work center shown on the operation process chart to determine the common possibilities for job improvement. See HUMAN-FACTORS

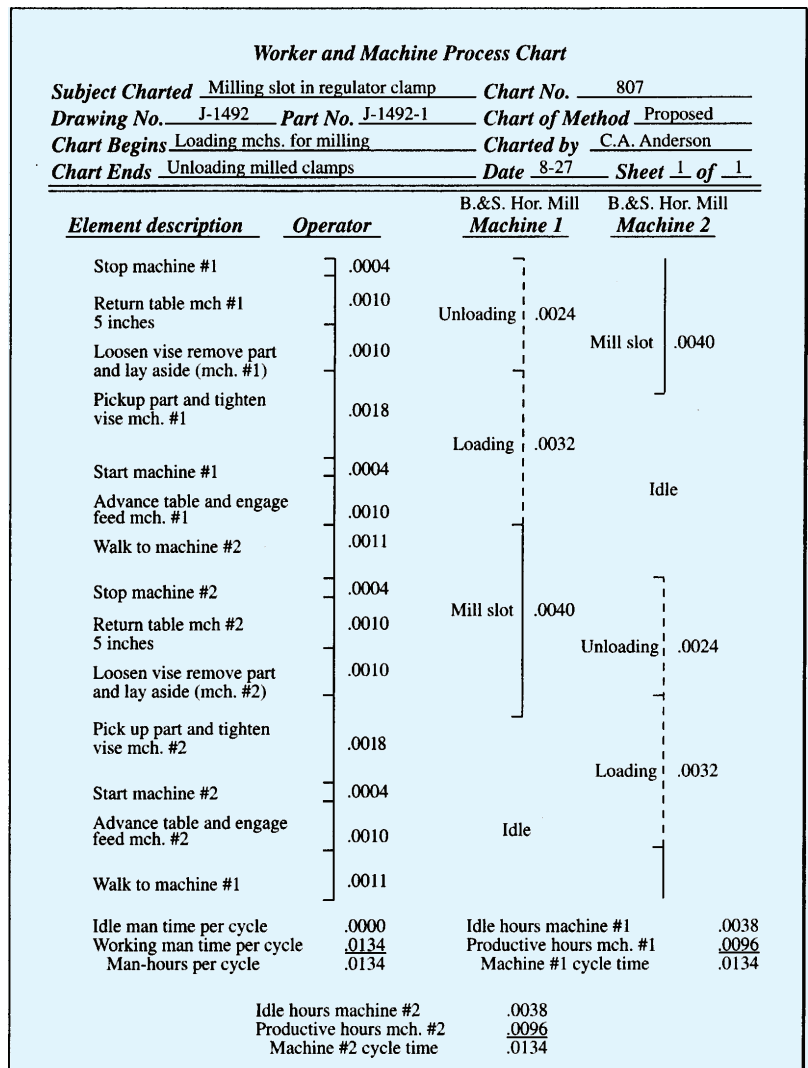


Fig. 4. Worker-machine process chart.

ENGINEERING; OCCUPATIONAL HEALTH AND SAFETY.

Ergonomics and principles of motion economy involve the study of both manual and eye movements that occur in an operation or work cycle in order to eliminate wasted movements and establish a better sequence of less fatiguing, coordinated movements.

In making a motion study, the methods engineer studies each individual motion in detail and tries to shorten it, combine it with others, or eliminate it altogether. Ineffective motions are identified and replaced, if possible, with effective motions. As a result of this intensive study, easier, more effective, less fatiguing, and quicker ways of doing the work are developed. Motions may be studied individually by direct observation, or they may be observed and timed in groups during the making of a stopwatch time study. When these data are properly arranged on a multiple-activity process chart, possibilities for improvement are often revealed.

In summary, methods engineering can be defined as the systematic close scrutiny of all direct and indirect operations to find improvements, making work easier to perform and allowing work to be done in

| Two-hand process chart | | | | Page 1 of 1 | | |
|--|-------------|---------------|-------------------------|-------------------|------------------------|-------------------------|
| Operation: Assemble cable clamps | | Part: SK-112 | | Summary | Left hand | Right hand |
| Operator name and no.: J.B. #1157 | | | Effective time: | 2.7 | 11.6 | |
| Analyst: G. Thuring | | Date: 6-11-98 | | Ineffective time: | 11.6 | 2.7 |
| Method (circle choice) Present Proposed | | | Cycle time = 14.30 sec. | | | |
| Sketch: | | | | | | |
| | | | | | | |
| Left hand description | Sym- bol | Time | Time | Sym- bol | Right hand description | |
| Get U-bolt (10") | RE G | 1.00 | | 1.00 | RE G | Get cable clamp (10") |
| Place U-bolt (10") | M P | 1.20 | | 1.20 | M P RL | Place cable clamp (10") |
| Hold U-bolt | H | 11.00 | | 1.00 | RE G | Get first nut (9") |
| | | | | 1.20 | M P | Place first nut (9") |
| | | | | 3.40 | U RL | Run down first nut |
| | | | | 1.00 | RE G | Get second nut (9") |
| | | | | 1.20 | M P | Place second nut (9") |
| | | | | 3.40 | U RL | Run down second nut |
| Dispose of assembly | M RL | 1.10 | | 0.90 | UD | Wait |

Fig. 5. Operator process chart.

less time with less investment per unit. Thus, the main objective of methods engineering is profit improvement.

Andris Freivalds
Bibliography. L. S. Aft, *Work Measurement & Methods Improvement*, Wiley, 2000; S. Konz and S. Johnson, *Work Design: Occupational Ergonomics*, 6th ed., Holcomb Hathaway, 2004; B. W. Niebel and A. Freivalds, *Methods, Standards and Work Design*, 11th ed., McGraw-Hill, 2003; J. A. Tomkins et al., *Facilities Planning*, Wiley, 2003.

Metric system

A system of measurement units used in scientific work and for everyday applications by most countries, with the notable exception of the United States. A prime advantage of the metric system is its international acceptance as a standard of measurement, providing a common measurement language for most of the world's population.

The metric system was developed in France in the late eighteenth century when Louis XVI authorized scientific investigations aimed at the reform of

French weights and measures. These investigations led to the development of the first "metric" system. After slow initial acceptance by France, the metric system spread through most of Europe in the nineteenth century and to most of the world in the twentieth century.

The meter-kilogram-second (MKS) and centimeter-gram-second (cgs) systems were major variations of the metric system. Both are now superseded by the SI-metric system adopted in 1960 by the 11th General Conference on Weights and Measures (CGPM). The new SI-metric system, given the official abbreviation SI for *Système International d'Unités*, is also called the modernized metric system. Most of the original definitions of the SI base units have been superseded as technological advances have permitted more precise determinations. Throughout all these changes, however, the units have remained the same size, only being specified more precisely with each new definition.

SI-metric units have unique symbols, not abbreviations, allowing the units to be recognized in spite of language differences that may affect the spelling of the unit names (such as meter and metre). The symbols represent both the singular and plural of the unit and do not need an added "s" when more than one is meant; nor are the symbols to be followed by periods except at the end of a sentence.

Qualities of the metric system include being decimal and coherent. The decimal nature makes it easy to manipulate these units with the decimal number system used throughout the world. Coherency is a quality of the derivation of SI units from the base units of the system. There is a one-to-one relationship between the equations of physics and the SI units for the quantities expressed in an equation. Metric units and symbols can be manipulated algebraically in the same manner as the physical quantities they represent. There is no need for either multipliers or conversion factors, as is often the case with nonmetric systems of units.

International System of Units. The SI-metric system has seven base units from which all other units are derived (Table 1). Two of the base units are named after prominent scientists, A. M. Ampère and Lord Kelvin. These seven units are independent quantities, six of which are now defined in terms of reproducible experimental procedures. Only the kilogram is still defined using an artifact, a platinum-iridium cylinder kept under glass at the International Bureau of Weights and Measures (BIPM) in France, copies

TABLE 1. SI base units

| Quantity | Unit | Symbol |
|---------------------------|----------|--------|
| Length | meter | m |
| Mass | kilogram | kg |
| Time | second | s |
| Electric current | ampere | A |
| Thermodynamic temperature | kelvin | K |
| Amount of substance | mole | mol |
| Luminous intensity | candela | cd |

TABLE 2. SI prefixes

| Prefix | Symbol | Multiplication factor |
|--------|--------|----------------------------|
| yotta | Y | 10^{24} |
| zetta | Z | 10^{21} |
| exa | E | 10^{18} |
| peta | P | 10^{15} |
| tera | T | 10^{12} |
| giga | G | 1,000,000,000 10^9 |
| mega | M | 1,000,000 10^6 |
| kilo | k | 1000 10^3 |
| hecto | h | 100 10^2 |
| deka | da | 10 10^1 |
| deci | d | 0.1 10^{-1} |
| centi | c | 0.01 10^{-2} |
| milli | m | 0.001 10^{-3} |
| micro | μ | 0.000,001 10^{-6} |
| nano | n | 0.000,000,001 10^{-9} |
| pico | p | 10^{-12} |
| femto | f | 10^{-15} |
| atto | a | 10^{-18} |
| zepto | z | 10^{-21} |
| yocto | y | 10^{-24} |

of which are available to other countries. For definitions of the SI-metric base units, used as the foundation for all other physical measurements, *see* PHYSICAL MEASUREMENT.

Twenty SI prefixes, from plus to minus the 24th power of 10, cover a range of measurement from the extremely large to the extremely small (Table 2). The prefixes are spaced in increments of 10 to the power 3, except for additional powers of ± 1 and ± 2 . Prefixes for other powers are no longer to be used. The use of the proper SI prefix with a quantity allows values to stay between 1 and 1000. The symbols for prefixes greater than kilo are represented by capital letters, others by lowercase letters, and all symbols are Latin letters except for the Greek mu for micro, 10 to the -6 power. In the case of the kilogram, which for historical reasons already contains a prefix, other prefixes are added to the root word gram or the symbol g.

Derived units in the metric system are formed by combining base and other derived units in a manner similar to the algebraic relationships that link the corresponding quantities. Likewise, the symbols for those derived units are obtained by use of the same relationships. Examples of the many derived units without special names are speed, acceleration, and density. There are, however, 22 SI derived units with special names (Table 3). The first two units in the table are dimensionless derived quantities that until 1995 were classified as supplementary units. Seventeen of the derived units are named after prominent scientists. These, together with the two base units ampere and kelvin already mentioned, total 19 SI-metric units that borrow names from scientists. When the metric unit is meant, these names are not capitalized (except for Celsius), but the symbols for these units are always capitalized.

Some other units are in use with SI. These include units that are already used worldwide for time measurement, and for plane angles in degrees, minutes, and seconds (Table 4). In addition, there are specially named units for area, volume, and mass

that arose from a combination of need and common usage. The liter in particular is different from other metric units in that it has two acceptable symbols, the lowercase l and the uppercase L. The uppercase L, needed to avoid confusion with the number 1, is the preferred symbol in the United States.

Finally, there are yet other units that are to be used only temporarily with SI (Table 5). These units could be replaced by metric units, but they are retained because of wide usage. It may be a long time before these units disappear. Complete elimination of these units at this time could be the cause of confusion and possibly dangerous situations, especially for radiation exposure units. *See* PHYSICAL MEASUREMENT.

More recent adoptions. The metric system has become the preferred measurement system for most of the world, starting with Europe and spreading into countries where nonmetric units were previously established. In particular, the Imperial system of weights and measures was used predominantly by British Commonwealth nations until the 1970s. Increases in international trade led to the realization that two competing measurement systems cannot be economically tolerated. One system had to be abandoned in favor of a common and universal measurement language. At first there was some resistance, but eventually the leaders of many of the Commonwealth nations saw the need to change to the metric system.

Notable examples of rather swift and complete transitions to the metric system occurred in Australia, New Zealand, and South Africa. Each country decided on a definite plan and timetable for conversion. In other countries as well, there was a need not only for a decimal measurement system but also for currency systems to match the decimal nature of the base-10 number system used by the entire world. Thus, the change to the metric system in many countries was preceded by adoption of decimal currencies, leaving behind the British system of coinage based on the pound divided into 12 shillings and the shilling divided into 20 pence.

Canada, the United Kingdom, and the United States also embarked on metric transition plans, but with somewhat less success. Many more visible aspects of the metric transition have occurred in Canada and the United Kingdom than in the United States. In the United States, the Metric Conversion Act of 1975 was voluntary in nature, with no conversion deadlines. This is in contrast to the adoption by the United States of the world's first decimal currency when it developed its own monetary system in 1795. Also, the metric system has been legal in the United States since 1866, paving the way for its use alongside conventional units that evolved from, but are significantly different from, those of the Imperial system. In particular, United States liquid measures are not the same as the Imperial namesakes. One of the reasons for adopting metric units was to eliminate major differences that existed between the British and American measurement systems.

The Metric Conversion Act created the U.S. Metric Board (USMB) to coordinate the transition to the

TABLE 3. SI derived units with special names and symbols

| Quantity | Name | Symbol | Expression in terms of other SI units |
|--|----------------|--------|---------------------------------------|
| Angle, plane | radian* | rad | m/m = 1 |
| Angle, solid | steradian* | sr | m ² /m ² = 1 |
| Celsius temperature | degree Celsius | °C | K |
| Electric capacitance | farad | F | C/V |
| Electric charge, quantity of electricity | coulomb | C | A·s |
| Electric conductance | siemens | S | A/V |
| Electric inductance | henry | H | Wb/A |
| Electric potential difference, electromotive force | volt | V | W/A |
| Electric resistance | ohm | Ω | V/A |
| Energy, work, quantity of heat | joule | J | N·m |
| Force | newton | N | kg·m/s ² |
| Frequency (of a periodic phenomenon) | hertz | Hz | 1/s |
| Illuminance | lux | lx | lm/m ² |
| Luminous flux | lumen | lm | cd·sr |
| Magnetic flux | weber | Wb | V·s |
| Magnetic flux density | tesla | T | Wb/m ² |
| Power, radiant flux | watt | W | J/s |
| Pressure, stress | pascal | Pa | N/m ² |
| Activity (referred to a radionuclide) | becquerel | Bq | 1/s |
| Absorbed dose, specific energy imparted, kerma | gray | Gy | J/kg |
| Dose equivalent, ambient dose equivalent, directional dose equivalent, personal dose equivalent, organ dose equivalent | sievert | Sv | J/kg |
| Catalytic activity | katal | kat | mol/s |

*The radian and steradian, previously classified as supplementary units, are dimensionless derived units that may be used or omitted in expressing the values of physical quantities.

metric system for Americans. However, after many years of little real progress, the USMB was disbanded in 1982. Metric transition activities of the United States government were handed over to the National Institute of Standards and Technology (NIST) Metric Program office. Private conversion activities are being handled by the U.S. Metric Association, formed in 1916, and dedicated to helping the United States transition to the SI-metric system used by virtually the entire world.

The Omnibus Trade and Competitiveness Act of 1988 strengthened the Metric Conversion Act by making the metric system the preferred system of measurement for the United States government. A Presidential Order issued in 1992 was an additional incentive for federal agencies to convert. However, reasons not to use metric units were still found, such as those contained in a clause stating that metric units need not be used "to the extent that such use

is impractical or is likely to cause significant inefficiencies or loss of markets for U.S. firms." As a result, the United States is the least converted of the world's major economies. But the metric system is not dead in the United States, as behind the scenes many consumer products are being designed, built, or packaged in metric.

The units used by the United States have been defined in terms of metric units since 1893. Both the yard and the pound were given equivalent values in terms of metric units by legislation of the U.S. Congress called the Mendenhall Order. In addition, the inch was redefined as exactly 25.4 millimeters in 1959. This change in the definition standardized the nonmetric units in the major inch-using nations of that time: the United States, Canada, and the United Kingdom. Before that change, each country already defined the inch in terms of the metric system as the standard, but differences existed in the sixth

TABLE 4. Units in use with SI

| Quantity | Unit | Symbol | Value in SI units |
|-------------|--------------------------------|--------|---|
| Time | minute | min | 1 min = 60 s |
| | hour | h | 1 h = 60 min = 3600 s |
| | day | d | 1 d = 24 h = 86,400 s |
| Plane angle | week, month, etc. | | |
| | degree* | ° | 1° = (π/180) rad |
| | minute* | ' | 1' = (1/60)° = π/10,800 rad |
| | second* | " | 1" = (1/60)' = (π/648,000) rad |
| Area | revolution, turn | r | 1 r = 2π rad |
| | hectare | ha | 1 ha = 1 hm ² = 10 ⁴ m ² |
| Volume | liter [†] | L, l | 1 L = 1 dm ³ = 10 ⁻³ m ³ |
| Mass | metric ton, tonne [‡] | t | 1 t = 10 ³ kg |

*Decimal degrees should be used for division of degrees, except for fields such as astronomy and cartography.

[†]The symbol L is preferred for use in the United States.

[‡]Metric ton or tonne is restricted to commercial usage.

TABLE 5. Units in use temporarily with SI

| Name | Symbol | Value in SI Units |
|---------------|----------------------|--|
| Nautical mile | | 1 nautical mile = 1852 m |
| Knot | | 1 nautical mile per hour = (1852/3600) m/s |
| Bar* | bar | 1 bar = 100 kPa |
| Barn | b | 1 b = 100 fm ² = 10 ⁻²⁸ m ² |
| Curie | Ci | 1 Ci = 3.7 × 10 ¹⁰ Bq |
| Roentgen | R | 1 R = 2.58 × 10 ⁻⁴ C/kg |
| Rad | rad, rd [†] | 1 rad = 1cGy = 10 ⁻² Gy |
| Rem | rem | 1 rem = 1cSv = 10 ⁻² Sv |

*Use is limited to meteorology.
[†]When there is risk of confusion with the symbol for radian, rd may be used as the symbol for rad.

decimal place between the conversion factors used by the three countries, enough so that manufacturers using precision measurements were experiencing difficulty in the exchange of inch-based machinery. This problem arose in particular during World War II, but was finally solved by the international agreement to define the inch similarly.

Ever-increasing international cooperation and trade will probably drive future changes to metric units. No countries that have converted to metric have reverted to their former units. Although a cost is involved in the transition to metric, there are also costs in not going metric, such as the costs of dual inventories of products, and the loss of market share for products that are increasingly not accepted in an otherwise metric world.

Donald W. Hillger

Bibliography. Bureau International des Poids et Mesures (BIPM), *Le Système International d'Unités (SI): The International System of Units (SI)*, 7th ed., 1998; Institute of Electrical and Electronic Engineers and American Society for Testing and Materials, *Standard for Use of the International System of Units (SI): The Modern Metric System*, IEEE/ASTM SI 10, 2002; National Institute of Standards and Technology, *Guide for the Use of the International System of Units (SI)*, NIST Spec. Publ. 811, 1995; National Institute of Standards and Technology, *The International System of Units (SI)*, NIST Spec. Publ. 330, 2001; U.S. Metric Association, *Guide to the Use of the Metric System (SI Version)*, 15th ed., 2002.

Metzgeriales

An order of liverworts in the subclass Jungermanniidae. Twelve families make up the Metzgeriales, which are also known as the Anacrogynae because of archegonia produced behind the growing apex. The gametophyte plant body is flat, elongated, and usually thallose, with no tissue differentiation or surface pores; less commonly there is a stem with two rows of leaves. The archegonia are generally protected by an involucre formed external to the calyptra and sometimes by additional structures. Capsules dehisce by valves. See BRYOPHYTA; HEPATICOPSIDA; JUNGERMANNIADA. Howard Crum

Mica

Any one of a group of hydrous aluminum silicate minerals with platy morphology and perfect basal (micaceous) cleavage.

Structure. Sheets can be produced from silicate (SiO₄) tetrahedra by having each tetrahedron share each of its three basal oxygen atoms with a different adjacent SiO₄ tetrahedron. The overall stoichiometry of these sheets will be (Si₂O₅)²⁻. Such sheets can also be considered to be produced by the linking together of double chains of SiO₄ tetrahedra such as occur in the amphibole structure. The structures of trioctahedral phyllosilicates can be thought of as being derived from that of brucite [Mg₃(OH)₆], a three-layer structure in which a layer of octahedrally coordinated magnesium ions (Mg²⁺) is sandwiched between layers of hydroxyl (OH⁻) groups. If the sheets described above replace those two OH⁻ layers in brucite, the mineral talc [Mg₃(Si₄O₁₀)(OH)₂] is produced; in this structure the Mg²⁺ ions remain octahedrally coordinated, this time by the unshared apical oxygen ions of the sheet-forming tetrahedra and by the remaining hydroxyl ions between the sheets. In the micas, the net negative charge on the sheets is increased by the substitution of one aluminum ion (Al³⁺) for one of every four silicon ions (Si⁴⁺), and this is compensated for by the addition of one alkali metal ion (for each Al³⁺ ion) between the two layers of basal oxygen ions that form the top and bottom of two three-layer sandwiches stacked on top of each other. If potassium ion (K⁺) is the alkali, the trioctahedral mica phlogopite [KMg₃(AlSi₃O₁₀)(OH)₂] results. The structure of dioctahedral micas is related in an analogous way to that of the three-layer structure of gibbsite [Al₂(OH)₆], giving such phases as muscovite [KAl₂(AlSi₃O₁₀)(OH)₂]. In the brittle micas, there is a divalent ion (for example calcium, Ca²⁺) between the sandwiches. Each mica can exist in several structural types or polytypes. This occurs because the sheets can be rotated relative to one another, and different stacking sequences can result from repeating layers that have been rotated by specific amounts at regular intervals, that is, at every so many layers. Micas crystallize in the monoclinic system, but lepidolite can be trigonal. See AMPHIBOLE; CRYSTAL STRUCTURE; LEPIDOLITE; MUSCOVITE; PHLOGOPITE; TALC.

Chemistry. The most common micas are muscovite [KAl₂(AlSi₃O₁₀)(OH)₂], paragonite [NaAl₂(AlSi₃O₁₀)(OH)₂], phlogopite [K(Mg,Fe)₃(AlSi₃O₁₀)(OH)₂], biotite [K(Fe,Mg)₃(AlSi₃O₁₀)(OH)₂], and lepidolite [K(Li,Al)_{2.5-3.0}(Al_{1.0-0.5}Si_{3.0-3.5}O₁₀)(OH)₂]. Calcium (Ca), barium (Ba), rubidium (Rb) and cesium (Cs) can substitute for sodium (Na) and potassium (K); manganese (Mn), chromium (Cr), and titanium (Ti) for magnesium (Mg), iron (Fe), and lithium (Li); and fluorine (F) for hydroxyl (OH).

Physical properties. Mica is commonly found as small flakes or lamellar plates without a crystal outline. Muscovite and biotite sometimes occur in thick books, tabular prisms with a hexagonal outline that can be up to several feet across. The prominent

basal cleavage is a consequence of the layered crystal structure. Thin cleavage sheets of micas, particularly muscovite and phlogopite, are flexible, elastic, tough, and translucent to transparent (isinglass). They have low electrical and thermal conductivity and high dielectric strength. Percussion figures may be developed on cleavage plates by striking the surface sharply with a dull-pointed tool. This yields a six-rayed star, the rays of which are parallel to certain crystallographic directions.

Micas have Mohs hardnesses of 2–3 and specific gravities of 2.8–3.2. Upon heating in a closed tube, they evolve water. They have a vitreous-to-pearly luster. Muscovite is colorless to pale shades of brown, green, or gray. Paragonite is colorless to pale yellow. Phlogopite is pale yellow to brown. Biotite is dark green, brown, or black. Lepidolite is most often pale lilac, but it can also be colorless, pale yellow, or pale gray. *See* BIOTITE.

Occurrence. The three major species, muscovite, biotite, and phlogopite, are widely distributed rock-forming minerals, occurring as essential constituents in a variety of igneous, metamorphic, and sedimentary rocks and in many mineral deposits.

Muscovite is found in regionally metamorphosed aluminous rocks that formed under a wide range of physical conditions. In igneous rocks, it occurs in some types of granites, in aplites, and as books in pegmatites. It is a characteristic phase of greisens which are produced when fluorine and other volatiles are introduced from granitic melts into adjacent rocks. Sericite is the name given to fine-grained white mica, usually muscovite. This mineral is a widespread gangue mineral in many hydrothermal ore deposits, either in the deposits themselves or in the adjacent altered wall rocks. Paragonite occurs in schists, gneisses, quartz veins, and fine-grained sediments. Phlogopite is a product of regional metamorphism of impure magnesian limestone. It is also characteristic of mantle-derived kimberlites and inclusions in kimberlites. Of all the micas, biotite occurs in the widest range of geological settings. It is common in the thermally metamorphosed rocks adjacent to granitic intrusions. In regionally metamorphosed rocks, it occurs in schists with chlorite, garnet, staurolite, kyanite, and sillimanite, although not all of these simultaneously. In plutonic igneous rocks, biotite is most common in intermediate and acid rocks, but it even occurs in some norites. Biotite books are found in pegmatites. Lepidolite is the most common lithium-bearing mineral and occurs almost exclusively in pegmatites with beryl, topaz, tourmaline, and other lithium minerals such as spodumene and amblygonite. It is also found occasionally in granites and aplites. *See* APLITE; GRANITE; ORE AND MINERAL DEPOSITS; PEGMATITE.

Uses. Commercial mica is of two main types: sheet, and scrap or flake. Sheet muscovite, mostly from pegmatites, is used as a dielectric in capacitors and vacuum tubes in electronic equipment. Lower-quality muscovite is used as an insulator in home electrical products such as hot plates, toasters, and

irons. Scrap and flake mica is ground for use in coatings on roofing materials and waterproof fabrics, and in paint, wallpaper, joint cement, plastics, cosmetics, well drilling products, and a variety of agricultural products. *See* ELECTRIC INSULATOR; SILICATE MINERALS.

Lawrence Grossman; Steven Simon
Bibliography. W. A. Deer, R. A. Howie, and J. Zussman, *An Introduction to the Rock-Forming Minerals*, 2d, ed., 1992; C. Klein, *Manual of Mineralogy*, revised 22nd ed., 2001; J. J. Papike, Chemistry of the rock-forming silicates: Multiple-chain, sheet, and framework structures, *Rev. Geophys.*, 26:407–444, 1988.

Mice and rats

The names associated with a great number of species of mammals in a number of different families of the order Rodentia. The rodents constitute the largest order of mammals, and they presently include 2277 species, or about 42% of the known mammalian species of the world. Most of the smaller rodents are called rats or mice. There is no fundamental taxonomic difference between mice and rats; they differ only in size. (Smaller species are generally called mice, larger ones rats.) Both mice and rats are often in the same family, and there are mice and rats in a number of different rodent families. *See* MAMMALIA; RODENTIA.

Some of the rats are of great economic significance to humans, particularly the Norway rat, *Rattus norvegicus*, and the black or roof rat, *R. rattus*. In addition to harboring many diseases transmissible to humans, such as bubonic plague, endemic typhus, rat-bite fever, and infectious jaundice, infested rats can transmit trichinosis to swine. *See* PLAGUE.

Some species of mice and rats are used for research in biology and medicine. In addition to their use in studying the mechanisms of genetics, they are important in the study of carcinogenesis, effects of drugs, and virology. They are also important experimental animals in studying cell physiology, such as for cell and tissue culture research, and in animal behavior.

There is much current discussion and disagreement concerning the higher classification of rodents. The present classification is as follows. The Rodentia are divided into five suborders, and there are mammals called mice and rats in four of the five. All the animals in one large suborder, the Myomorpha, are rats and mice. The four suborders with mice and rats (Sciuromorpha, Castorimorpha, Myomorpha, Hystricomorpha) are indicated below, with the families including animals classed as mice and rats listed below each of the suborders.

Sciuromorpha

The Sciuromorpha are squirrel-like rodents, including the dormice.

Gliridae. The dormice make up this family. They are one of the oldest living families of rodents,

appearing in Eocene deposits, and they are widespread in the Old World, their 28 species occurring in much of Eurasia and Africa below the Sahara. Dormice are squirrel-like, but unlike most squirrels are generally nocturnal. They live in hollow limbs, rocky crevices, in burrows of other species, and in buildings. They could have been considered with squirrels, but are included here since they are called "mice." They have long, mostly rather bushy tails. Most have large eyes and rounded ears. Most are good climbers and have short, curved claws adapted for climbing. The dental formula is I 1/1 C 0/0 Pm 1/1 M 3/3 = 20. They are generally smaller than squirrels, but larger than mice. The fat or edible dormouse of Europe, *Glis glis*, is about 140–240 mm (5–9 in.) in total length and weighs about 70–180 g (2.5–6.3 oz), which is much smaller than the smallest North American squirrels. This species has been introduced into England. It lives in forests or orchards, usually in cavities in trees, but also it will live in many other protected areas. Some dormice hibernate, but wake occasionally and eat stored food. Food consists of fruit, nuts, insects, eggs, and even small vertebrates, which is similar to that of North American flying squirrels (*Glaucomys*). Most species produce two litters per year, but *Glis* probably produces only one because of the long hibernation, that is, short active period. The meat of *Glis* is edible as the name suggests, and has been a gourmet food in parts of Europe, although the species has become rare in many areas because of destruction of forests.

The African Dormice, subfamily Graphiurinae, with 14 species, are all in the genus *Graphiurus* and all occur in Africa. *Graphiurus murinus* looks much like a deermouse (*Peromyscus*) but with a bushy tail. Some African dormice in southern Africa hibernate. They make twittering sounds and a fairly loud "shriek." The subfamily Leithiinae includes 12 Eurasian and mideastern species: *Chaetocauda*, the Sichuan dormouse of China; *Dryomys*, forest dormice, 3 species in Turkey, Pakistan, and Russia; *Eliomys*, the garden dormice, with 3 species; *Muscardinus*, the hazel dormouse, *Myomimus*, the mouse-tailed dormice, 3 species; and *Selevinia*, the desert dormouse (the latter 4 genera are all of Eurasia). The subfamily Glirinae includes only two species: *Glirulus japonicus*, the Japanese dormouse, and *Glis glis*, the fat or edible dormouse of Eurasia.

Castorimorpha

This suborder comprises the single family Heteromyidae.

Heteromyidae. The Heteromyidae comprise kangaroo rats and mice and pocket mice, with 6 genera and 60 species. The dental formula is I 1/0 C 0/0 Pm 1/1 M 3/3 = 18. The 19 species of kangaroo rats (*Dipodomys*) occur in desert regions of North America. They have elongate hindlegs and hop rather than run, thus resembling a kangaroo. The short front legs are used for holding food, and the long hairy tail acts as a balancing organ or prop when the an-

imal hops or rests. These rodents collect seeds into shallow sand-covered pits for drying by the heat of the sun; the seeds are then carried into the burrows for winter use. These animals do not usually drink water, but obtain fluid from food, an adaptation for water balance in desert environments. Over a period of months with only dry food, there is no change in the water content of their bodies. These species lack sweat glands, the feces have a low water content, and the kidneys are efficient, with urine highly concentrated and excreted in small amounts. Water is conserved and body heat is regulated by these adaptations. When fed high-protein diets with the resulting formation of large amounts of urea, these animals will drink water, even seawater which would poison most animals. However, the kidneys can handle the high salt concentration.

Twenty-six species of pocket mice (*Perognathus*, 9 species, and *Chaetodipus*, 17 species) occur in the United States, with the center of distribution being in the southwest. The ears are small, the tail is moderately long, and a fur-lined cheek pouch opens externally on each side of the mouth. The hindlimbs and hindfeet are moderately enlarged; there are four digits on the forefeet, and five on the hindfeet. Some species are the plains pocket mouse (*P. flavescens*), the Great Basin pocket mouse (*P. parvus*), and the desert pocket mouse (*C. penicillatus*).

There are two species of kangaroo mice (*Microdipodops*), small rodents restricted to the desert areas of the western United States. These are *M. megacephalus*, the dark kangaroo mouse, and *M. pallidus*, the pale kangaroo mouse. Their hindlegs are long and forelegs relatively short, and the mice move like a kangaroo, hopping on the hindlegs and using the tail as a balancing organ. They do not drink regularly, but obtain sufficient moisture from their food.

There are five species of spiny pocket mice (*Liomys*). *Liomys irroratus* enters the United States from its center of distribution in the thorny bushlands of the Rio Grande Valley in Mexico. The ears are small, and the tail is as long as, or longer than, the body. The fur is coarse and consists principally of stiff, flattened spines.

Myomorpha

This suborder includes the families Dipodidae, Platacanthomyidae, Spalacidae, Callomyscidae, Nesomyidae, Cricetidae, and Muridae.

Dipodidae. The Dipodidae now includes three subfamilies: the jumping mice, birchmice, and jerboas.

The jumping mice for years were in their own family, Zapodidae, but now are considered as a subfamily, Zapodinae, of the Dipodidae. The Zapodinae contains three genera and five species. All of the species are entirely North American, except *Eozapus setchwanus*, which is confined to eastern Asia. The meadow jumping mice (*Zapus*, 3 species) occur throughout much of North America. Jumping mice have long tails for balance, grooved upper incisors, and colorful fur. These rodents are adapted for

jumping, with long legs and large feet. *Napaeozapus* can jump up to 10 ft (3 m), *Zapus* usually not more than 2–3 feet (0.6–0.9 m). However, the meadow jumping mice (genus *Zapus*) more often run for cover or move by short hops rather than long leaps. By day, jumping mice hide in clumps of vegetation. Nests may be underground, in a hollow log, or otherwise protected. They hibernate for 6–9 months (the duration varies by species, locality, and elevation) in a nest at the end of a burrow in a bank or other raised area. Gestation is about 17 or 18 days, longer if the female is lactating. Meadow jumping mice eat many seeds, especially of grasses, but species of seeds vary greatly during the season. Subterranean fungi related to *Endogone* form about 35% of the volume of food in the woodland jumping mouse, and about 12% in meadow jumping mice. Among animal foods, the most important are moth larvae (caterpillars, often cutworms) and snout and ground beetles.

Zapus has 18 teeth and the dental formula is $1/1 \text{ C } 0/0 \text{ Pm } 1/0 \text{ M } 3/3$ in. *Napaeozapus insignis*, the woodland jumping mouse, occurs in Canada and eastern United States. It has only 16 teeth; it lacks the upper incisors that appear as a small tooth at the front end of the molariform toothrow in *Zapus*.

The feet of birchmice (subfamily Sicistinae) are only slightly enlarged, and their upper incisors lack the grooves of jumping mice. Their tail and legs are shorter than those of jumping mice, but jumping is their major mode of locomotion. When climbing into bushes, they use their outer toes to cling to vegetation. Birchmice make nests of herbaceous vegetation underground, where they hibernate for about half of the year. Birchmice are active primarily at night, and their main foods are seeds, berries, and insects. Gestation is probably about 18–24 days, with parental care lasting another 4 weeks.

Jerboas (subfamily Dipodinae) are built for jumping. Their hindlegs are at least four times as long as their front, which gives them the ability to escape from predators, although they will move by slow hops. Only the hindlegs are used in moving, allowing the front feet to gather food. Some jerboas can jump 5–10 ft (1.5–3 m), and desert jerboas, *Jaculus*, can jump nearly 3 ft (0.9 m) vertically. Jerboas use their long tails as props when standing upright and as balancing organs when jumping. Jerboa species in sandy areas have tufts of hairs on the undersides of the feet which serve as “snowshoes” to maintain traction and to kick sand backwards when burrowing. Other tufts of hair keep sand out of their ears, and some jerboas, for example, *Jaculus*, have a fold of skin which can be pulled forward over the nose when burrowing. Jerboas feed primarily on seeds, sometimes on succulent vegetation, and also on insects. Jerboas, like pocket mice, manufacture “metabolic water” from food. Some jerboas hibernate during winter, and some enter torpor during hot or dry periods. Northern jerboas mate shortly after emergence from hibernation, but most female jerboas usually

produce two litters per year of two to six young. See JERBOA.

Platacanthomyidae. The members of this family are the spiny dormice. There are two species, and they are similar to dormice. *Platacantbomys lasiurus* of southern India weighs about 75 g (2.6 oz). Its dorsum is covered with sharp flat spines intermixed with thin underfur. The basal part of the tail is scaly with few hairs and the distal portion forms a brush. It has a pointed snout, medium-sized eyes, and prominent ears. The hindfeet are broad and strong. There are no premolars; the tooth formula is the same as that of the murids, $1/1 \text{ O}/0 \text{ 3}/3 \text{ O}/0$. Spiny dormice live in rock clefts on rocky hills and in hollows in forested valleys. The long tail is useful as a balancing organ as the animal leaps about in trees. Natives in some areas call it the “pepper rat” because it commonly feeds on peppers, but it also eats fruits, seeds, grains, and roots. The Chinese pygmy dormouse, *Typhlomys cinereus*, occurs in heavy forests in mountainous regions in China and Vietnam. It is mouselike with a long thin tail, prominent ears, and relatively small eyes. The tooth formula is as in *Platacantbomys*. Little is known of its habits, but natives claim that cats will not eat it.

Spalacidae. The members of this family are the mole rats. The family Spalacidae includes six genera: *Eospalax* and *Myospalax*, the mole rats or zocors, with five species in Russia, Mongolia, and China; two genera of bamboo rats, *Cannomys* and *Rhizomys*, with four species in Asia; the blind mole rats (*Spalax*) with 13 species in the Russian region and the mideast; and *Tachoryctes* with 11 species of African mole rats.

Mole rats are short and chunky with short powerful legs. The forefeet have long, heavy claws for digging which are doubled under when the animals walk. They have no external ears and the eyes are tiny and hidden in the fur. Mole rats are fossorial (adapted for digging) and can dig into the ground like moles. They are found in a variety of habitats. Their burrows are very long and include a deep nest chamber, a food storage chamber, and an area for defecation. There are also one to four food tunnels that approach the surface below the food source. Unlike true moles, which feed on animal matter, mole rats feed on plant material such as roots, tubers, and plant stems. They may cause damage in potatoes and sugar beets. Some species of *Spalax* form “breeding mounds” in wet weather in fall and winter. These are elaborate structures that are usually about 160 cm long, 135 cm wide, and 40 cm in height (roughly 5×4 ft and 1.3 ft high). The African mole rats have stiff tactile hairs on the face. These animals resemble American pocket gophers. They live in open areas with much rain. They also have burrows and mounds, but burrowing is apparently done with the incisors, with the hindfeet used as braces. The forefeet are used to push the dirt under the animal. Then when ample soil has accumulated, the animal kicks it on back with its forefeet, and finally turns around and pushes the dirt, bulldozer style, using one side of the face and one

forefoot, behaviors all similar to those of true moles. Reproduction may occur throughout the year, but is concentrated during the wet season. *See* BURROWING ANIMALS; MOLE (ZOOLOGY).

Bamboo rats (*Rhizomys* and *Cannomys*, including four species) also resemble American pocket gophers (but lack cheek pouches). They have long naked tails, small eyes and ears, and large orange upper incisors outside the lips. They are generally found in bamboo thickets. They have several burrows and come above ground to feed on their main food—bamboo. They may emerge at night and even climb the bamboo. They will eat other plant foods, and do not drink water if there is adequate moisture in their food. Some species will feed on cultivated tapioca and sugarcane, and in turn are captured by natives for food.

Callomyscidae. This family includes one genus, *Callomyscus*, with eight species, occurring in the Middle East, Syria, and Azerbaijan. They are best called “mouse-like hamsters” or just *Callomyscus*. These animals appear most like jerboas or heteromyids, or even *Peromyscus*, with long tufted tails, big ears, and big eyes. There are no cheek pouches. Active mainly at night in summer, they eat seeds, flowers, and leaves, but will readily eat animal food.

Nesomyidae. The nesomyids include 38 species in 12 genera from Africa called the pouched rats (three genera, eight species), swamp mouse (1 species), African climbing mice (3 genera, 14 species), African desert mouse (1 species), African tree mouse (1 species), African fat mice (1 genus, 8 species), African white-tailed rat (1 species), and the pygmy rock mice (4 species). In addition there are 23 species in nine genera in this family all occurring in Madagascar. These include the antsangy (1 species), short-tailed rats (2 species), tufted-tailed rats (10 species), the voalavoanala, votsovotaa, and voalavo (1 species each), the big-footed mice (2 species), and the nasomys (3 species). The African giant rat, *Cricetomys gambianus*, is large, measuring as much as 32 in. (81 cm) from nose to tail tip. It has large cheek pouches and is at times referred to as the Gambian pouched rat. It is a vegetarian and occurs in forest regions, where it nests near the base of trees. It harbors a large, wingless, ectoparasitic cockroach, *Hemimerus talpoides*, which is similar to a louse and is apparently host-specific.

Cricetidae. The major families of mice and rats (Muridae with 730 species and Cricetidae with 130 genera and 681 species), like the higher taxonomic categories of rodents, have also been the subject of much debate and rearrangement. In the early part of the last century, most of the murid mice and rats were placed in two families, the Cricetidae or “New World rats and mice,” and the Muridae, or “Old World mice and rats,” although the voles were included in Cricetidae and occurred in both New and Old Worlds. However, in the latter part of the last century almost all of these animals were placed in the

Muridae. Today, once again, most of the New World mice are placed in the family Cricetidae, but they are now divided into different subfamilies. The present arrangement of the cricetids follows Carleton and Musser with the major North American genera indicated to help understand the new system. The Cricetidae currently is a very large family of mice, rats, and voles, and includes six subfamilies, as follows.

The Arvicolinae are the voles, lemmings, bog lemmings, and muskrats and were long known as “microtines.” They include the typical “meadow mice.” Most are burrowers, many with short tails, and have relatively small ears and eyes. Most of these animals feed on green vegetation, a food item that wears teeth down very rapidly. They have evolved two major adaptations for dealing with the tough cellulose in plant food: rootless teeth and flat grinding surfaces on the molariform teeth. Rootless teeth tend to grow throughout the life of the animal and therefore continue to grow from below as they are worn off at the crowns. The flat surfaces help in grinding the food. *See* MUSKRAT.

There are 28 genera and 151 species currently in the Arvicolinae. Arvicolinae are holarctic (that is, they inhabit northern parts of the Old and New Worlds). There are 19 Eurasian genera, including mountain voles (*Aticola*, *Neodon*, *Phaimys*), water voles (*Arvicola*), snow voles (*Chionomys* and *Dineromys*), mole voles (*Ellobius*), Sichuan voles (*Volemys*), and wood lemmings (*Myopus*). The genera *Arborimus*, *Lemmyscus*, *Neofiber*, *Ondatra*, *Phenacomys*, and *Synaptomys* are North American.

Four genera occur in both eastern Eurasia and in North America, *Microtus* (62 species), *Myodes* (was *Clethrionomys*, 12 species), *Lemmus* (5), and *Dicrostonyx* (8). The last two genera are the brown and collared lemmings, which occur in the far north.

Microtus is the largest genus of voles (often called meadow mice), with 62 species. These animals are usually brown and most often live in grassy fields and feed almost entirely on green vegetation. Many species leave characteristic piles of cuttings of grass or other plant stems often including the rachis and seeds of grasses. Some of these animals are very prolific, producing a number of litters per year.

The Cricetinae include the hamsters, with 7 genera and 18 species all of Eurasia. *Cricetus cricetus* is the common hamster often kept as a pet. It has a very short tail, nearly hairless. It lives in burrows and hibernates in winter. It feeds on seeds, roots, green parts of plants, and other wild and cultivated plant items and stores food for use in winter. In addition hamsters will feed on insect larvae and frogs. Hamsters are solitary, with only one animal per burrow. Although they may produce litters throughout the year in captivity, they normally produce two litters of 4–12 young per year in the wild.

The subfamily Lophiomyinae includes one species, the maned rat, *Lophiomyys imbausi*, found in Africa often in forests, but not restricted to them. The maned rat is a large species with long hair, and

a long furry tail. The animal is more like a guinea pig than rat. It is variable in color, predominantly gray, but often with black and white spots. The tip of the tail is white. There is a mane of erectile hair from the top of the head to the first quarter of the tail. It is nocturnal and feeds on leaves and tender shoots.

The Neotominae (along with the Arvicolinae) are the big group of North American rats and mice, including 16 genera and 124 species. Mice and rats in this suborder spend more time aboveground than do the voles. Consequently they are more aware of their environment and generally have longer tails, bigger ears, and bigger eyes than the voles and allies.

Seven of the genera occur in North America, the woodrats (*Neotoma*, 22 species), golden mice (*Ochrotomys*, one), grasshopper mice (*Onychomys*, 3), pygmy mice (*Baiomys*, 2), deer mice (*Peromyscus*, 56), Florida deer mice (*Podomys*, 1), and harvest mice (*Reithrodontomys*, 20).

Woodrats, *Neotoma*, are the largest of this group and are often called packrats, as they have a habit of collecting things: food items, pieces of bone, leaves, sticks, etc. They will often carry a bright object to their nests, frequently replacing it with pebbles or other material; because of this practice they are also called trade rats. Most species have long thin tails, except the bushy-tailed woodrat, *Neotoma cinerea*, a western species, common in the Rocky Mountain region, while *N. magister* occurs in the northeast south to the Tennessee River. *Neotoma floridana* ranges along the South Atlantic and Gulf coasts, north to the Tennessee River, but interestingly to extreme southern Illinois. Wood rats are found in timbered areas and desert regions. Most are nocturnal.

The golden mouse, *Ochrotomys nuttalli*, is found in the southeastern United States. It is a beautiful golden-cinnamon above, white below. It is highly arboreal, and often climbs trees to 30 ft or more. Its long prehensile tail is used for balance and support. This species often occurs in dense brushy undergrowth such as honeysuckle, greenbriar, grapes, or, when present, in Spanish moss. It makes large bulky nests in the undergrowth and is gregarious.

Three species of grasshopper mice, *Onychomys*, occur in the grassland and desert areas of the western United States and Mexico.

Baiomys, the pygmy mice, include two species, *B. musculus* of Mexico and *B. taylori* of the southwestern United States. Pygmy mice are small and short-tailed. They occur in grassy or brushy areas.

There are numerous species, races, varieties, and subspecies of deer mice (*Peromyscus*). They feed on seeds and insects. One of the common species is *P. maniculatus*, the deer mouse, with about 65 subspecies found throughout the United States (see illustration). These rodents are small with moderate to long tails and large ears. Another is *P. leucopus*, the white-footed mouse, which occurs in eastern woodlands. The Florida deer mouse, *Podomys floridana*, is very similar to *Peromyscus*, but is much larger. It occurs only in Florida. It seldom makes its own burrow, but usually uses that of the gopher tortoise or pocket gopher.



Peromyscus maniculatus; deer mouse. (Photo by Glenn and Martha Vargas; © 2002 by California Academy of Sciences)

Harvest mice (*Reithrodontomys*) are small to very small rodents with a long tail, prominent ears, and long, narrow hindfeet. Five of the 17 living species are found in the United States. The eastern harvest mouse (*R. humulis*) occurs in the deciduous forests of the southeastern states, while *R. montanus* is common in the central grasslands. The salt marsh harvest mouse (*R. raviventris*) is found only in the salt marshes near San Francisco Bay.

Many of the genera are primarily Mexican: *Habromys* (6 species), *Hadomys* (1), *Megadontomys* (3), *Nelsonia* (2), *Neotomodon* (1), and *Osgoodomys* (1). Farther to the south are *Isthmomys* (2 species) and *Scotinomys* (2).

Included in the subfamily Sigmodontinae are 74 genera and 374 species, mostly South American, although a few genera reach into Central America, and two genera, *Sigmodon*, the cotton rats, and *Oryzomys*, the rice rats, occur in the United States. Both of these genera are more common in Central and South America. In much of the northern hemisphere, voles are the dominant small herbivores, whereas in the deep south, cotton rats, *Sigmodon*, fill a similar role. Cotton rats are compact-bodied, coarse-haired rodents, with short rounded ears. *Sigmodon hispidus* is the common cotton rat throughout the southern United States, especially in grassy fields. *Sigmodon* will also feed on eggs and chicks of ground nesting birds such as quail, and will eat crayfish and fiddler crabs. Cotton rats are very prolific. Breeding activity occurs from late winter through late fall, and sometimes throughout the year in the more southern areas. Gestation is about 27 days and up to 15 young may occur in a litter.

The marsh rice rat, *Oryzomys palustris*, lives in the southeastern United States, and it and its relatives occur through east Texas and south through much of Central America. Rice rats usually live in wet marshy areas. They are good swimmers, and feed heavily on rice and other aquatic grasses, but they also will eat insects, and small crabs. They may produce litters of up to six offspring from February to November.

The Galápagos Islands have a relatively impoverished fauna as they are so far from the mainland (about 600 miles west of Ecuador). However, rice rats (*Oryzomys*) probably reached the Galápagos Islands from South America some 3 to 3.5 million

years ago, differentiated into *Nesoryzomys*, and, like "Darwin's finches," spread to the different islands and further differentiated, *N. darwini* on Santa Cruz Island (probably extinct), *N. fernandina* on Fernandina Island, *N. indefessus* (probably extirpated in part of its range), and *N. swarthi* on San Salvador Island. The Galápagos giant rat (*Megaoryzomys*) also occurred on the Galápagos, but it too is extinct.

There are many other Sigmodontinae in the South American region, the "Akodonts" *Abrothrix* (8 species), *Akodon* (40), *Bibimys* (3), *Chelemys* (3), *Deltamys* (1), *Geoxus* (1), *Juscelinomys* (3), *Lenoxus* (1), *Nectomys* (9), *Notiomys* (1), *Pearsonomys* (1), *Phaenomys* (1), *Podoxymys* (1), *Tbalpomys* (1), and *Thaptomys* (1). Likewise, the "Pericotes" include the genera, *Andalgalomys* (3), *Auliscomys* (3), *Galenomys* (1), *Graomys* (4), *Loxadontomys* (2), *Paralomys* (1), and *Phyllotis* (13).

The Tylominae, vesper and climbing rats, occur in Mexico and Central America. There are four genera and 10 species.

Muridae. The members of this family are the Old World rats and mice, a huge family (150 genera and 730 species). Except for some introduced species, all occur in the Old World. There is a great biodiversity in this group. They are placed in five subfamilies, including Deomyinae (42 species in four genera), Gerbillinae (105 species in 14 genera), Leimacomomyinae (1 species, 1 genus), and Otomyinae (23 species in 3 genera). The rest are in the subfamily Murinae, which includes *Mus* and *Rattus*, which have been introduced into much of the rest of the world. The common house mouse is one of the oldest known species of domestic rodent pests. They begin to breed at 3 months of age, have a gestation of about 3 weeks, and have 4–6 litters of four to eight young per year. Adults have a pointed snout, compact body about 3 in. (8 cm) long, and an equally long tail. The ears are fairly large, as are the legs. These animals have 16 teeth and the dental formula I 1/1 C 0/0 Pm 0/0 M 3/3. House mice are often thought to exist primarily in buildings. However, they are often very common in croplands of the midwestern United States, but require much ground cover. They disappear when the crops are harvested. The other main small mammal species in these croplands is the prairie deer mouse, *Peromyscus maniculatus bairdii*. However, this latter species uses the soil as its cover and is the only species that remains when the crops are harvested and even when the ground is plowed. House mice are omnivorous, and will eat grains and other vegetable foods. However, they eat huge quantities of weed seed, foxtail grass, moth larvae such as army worms, and many others. Of the 44 species of *Mus* known, only one species occurs in the United States, and it has become wild in some parts of the country. The familiar white laboratory mouse is an albino form of the house mouse and is used extensively in laboratory research.

Rats (*Rattus*) are usually active at night, feeding on nearly every type of food. These animals have the same dental formula as *Mus*. The teeth are modified for gnawing, with sharp and chisel-like incisors that

grow continually. The genus *Rattus* contains about 137 species with many varieties and races throughout the world. Fecundity is high with as many as six litters of six to eight young per year; the gestation period is about 21 days; breeding begins at about 3 months. Average lifespan is 2–3 years. Rats have a maximum weight of slightly over 1 lb (0.5 kg). The thin tail is covered with overlapping scales and nearly devoid of hair. The ears and eyes are relatively large, and cheek pouches are lacking. Many species live in burrows, others infest dwellings.

Hystricomorpha

The Hystricomorpha are referred to as the South American rodents, although some of the species occur in Africa.

Octodontidae. The octodonts are South American. Most are not called rats but they are ratlike, and two genera each with two species are called rats, *Octomys* (Viscacha rats), which look like Dipodidae, and *Aconaemys* (rock rats), which look like voles. They make complex burrows and feed on green vegetation.

Abrocomidae. The members of this family are the chinchilla rats. There is one genus and two living species in this family, *Abrocoma bennetti* and *A. cinerea*, both South American. They look similar to chinchillas, but the head and ears are longer, giving them a ratlike appearance. Chinchilla rats have very long intestinal tracts and a huge cecum. These characteristics indicate that they are herbivorous.

Echimyidae. The members of this family are the spiny rats. This group contains 17 genera and 69 species and thus helps fill the mouse and rat niche of South America. They include various sized animals varying from 21 g (0.7 oz), which would easily be called a "mouse," up to 450 g (1 lb). Tail length varies greatly, from less than one-quarter of the body length to more than the body length, and they otherwise exhibit the great morphological variation one would expect from rodents living in a variety of habitats with a variety of niches. An adult female of one species of echimyid weighed 640 g (1.4 lb), which would be a very large rat, and other species in the family can weigh even more. Although the family, collectively, is referred to as the spiny rats, not all are called "rats." For example, members of the genus *Isotbrix* are called toros, *Dactylomys* are coro-coros, and *Chaetomys* are thin-spined porcupines. Members of this family are said to be vegetarian, but most are little known and it is likely that some species would be omnivorous or even primarily insectivorous.

Thryonomyidae. The members of this family are the cane rats. The one living genus, *Thryonomys*, contains two species, both in Africa. These are very large for rats, weighing from 4 to 9 kg (9–20 lb). One of the species is semiaquatic (*T. swinderianus*) and lives in marshes, whereas the other lives on dry ground. Both are herbivorous.

Petromuridae. There is but one species, *Petromurinus typicus*, the Dessie rat, and it is African. The

fur is soft, but underfur is absent. Hairs grow in clusters of three to five, and these animals have long black whiskers. They live in tiny crevices in rocky areas on hillsides, and their bodies can be greatly compressed.

Bathyergidae. These are the African mole-rats. This family contains 5 genera and 16 species, and all are restricted to Africa. They are molelike in having small ears and eyes and spend all or most of their time in underground burrows. Also, they throw up mounds to get rid of excess dirt when extending their burrows. Their front legs may be enlarged somewhat, but they do not have developed forelimbs as in true moles. Also, true moles feed primarily on earthworms and other invertebrates, whereas mole-rats feed mainly on bulbs and roots and store some in underground chambers for later use. Zech's mole rat (*Cryptomys zechi*) has large protruding incisors, short legs, powerful feet, and a reduced tail, features which make it well adapted for its burrowing existence. Like many desert species, it obtains sufficient water from food.

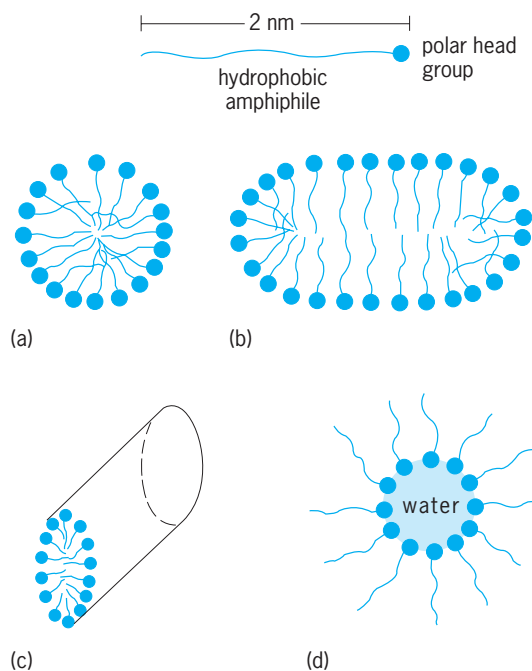
The strangest and smallest member of this family is *Heterocephalus glaber*, the naked mole-rat of Africa, measuring about 4 in. (10 cm). It lives in hot sandy regions of Africa. It is not totally naked as there are a few light-colored but long hairs on its body and tail and it has well-developed vibrissae. External ears are lacking, and it is blind. It weighs about 30–80 g (1–2.8 oz). This species is colonial and is completely fossorial, digging a complex burrow system with a mound above. This species has the poorest temperature regulation of any known mammal. Another particularly interesting trait is that they will form a chain to dig dirt from deeper areas, with one animal digging, but another animal pushes the soil backward all the way to the surface. The other animals, in turn, will push the soil to the surface, with the “pusher” taking its place at the end of the line. The animals in the chain straddle the pusher as the earth is moved to the surface.

John O. Whitaker, Jr.

Bibliography. M. D. Carleton and G. G. Musser, Order Rodentia, pp. 745–752 in D. E. Wilson and D. M. Reeder (eds.), *Mammal Species of the World: A Taxonomic and Geographic Reference*, 3d ed., Johns Hopkins University Press, Baltimore, 2005; R. M. Nowak, *Walker's Mammals of the World*, Johns Hopkins University Press, Baltimore, 1999; J. O. Whitaker, Jr., and W. J. Hamilton, Jr., *Mammals of the Eastern United States*, Cornell University Press, Ithaca, 1998; D. E. Wilson and S. Ruff, *The Smithsonian Book of North American Mammals*, 1999.

Micelle

A colloidal aggregate of a unique number (50 to 100) of amphipathic molecules, which occurs at a well-defined concentration called the critical micelle concentration. In polar media such as water, the hydrophobic part of the amphiphiles forming the micelle tends to locate away from the polar phase while the polar parts of the molecule (head



Form of an amphiphile and several forms of micelle: (a) spherical, (b) disk, (c) rod, and (d) reversed.

groups) tend to locate at the polar micelle solvent interface.

A micelle may take several forms, depending on the conditions and composition of the system, such as distorted spheres, disks, or rods (illus. a, b, and c). The dimensions of the particle are derived from those of the amphiphile, for example, a sphere radius of about 2 nanometers or a rod cross-sectional radius of 2 nanometers. Frequently the polar head group is a salt which ionizes in polar media. The micelle may have about 30% of the amphiphiles in the ionized state, giving rise to a highly charged particle surrounded by a cloud of counterions (ions with a charge opposite to that of the micelle ions), and also counterions bound into the micelle surface. Micelles are formed in nonpolar media such as benzene, where the amphiphiles cluster around small water droplets in the system, forming an assembly known as a reversed micelle (illus. d).

When the surfactant contains two long alkyl chains, a vesicle structure rather than a micelle is formed because of geometric restraints. This entity consists of two closed concentric spherical layers of surfactant which enclose an internal volume of water. Such a structure is similar to that of a biological membrane; micelles and vesicles mimic biological systems both in structure and in many kinetic processes. See CELL MEMBRANES; MEMBRANE MIMETIC CHEMISTRY.

Micellar systems have the unique property of being able to solubilize both hydrophobic and hydrophilic compounds. They are used extensively in industry for detergency and as solubilizing agents. A strong catalytic action is often associated with these systems and is attributed to the clustering of reactants in the micelle, thereby creating high local

reactant concentration, and also to the strongly charged surface which influences the transition state of a reaction. See DETERGENT; SURFACTANT.

The locally high concentration of guest molecules or reactants in a micelle surface leads to enhanced rates of reaction over those observed when the reactants are dispersed in the bulk phase. This catalytic effect can be as large as 1000-fold and can, in some cases, be used selectively. Photolysis of dialkyl ketones in micelles leads to radicals whose local concentrations are high. Only a portion of these radicals can react in micelles, those that have experienced spin relaxation due to their particular isotopic constituency. The unreacted portion with a different isotopic constituency diffuse into the bulk phase and give alternative products. The technique has been used successfully to produce isotopic enrichment. See CATALYSIS; FREE RADICAL.

Micelles play an important role in photo-induced reactions, in particular photo-induced electron transfer reactions. Anionic micelles such as sodium dodecyl sulfate strongly promote photoionization of several molecules (for example, phenothiazine, aminopyrene, or tetramethyl benzidine) located at the micelle interface. The ionization threshold is reduced by more than 3.0 eV [>70 kilocalories (300 kilojoules) per mole], an effect not observed in homogeneous solution. The electron of the guest molecule is transferred to the aqueous phase and observed as a hydrated electron, while the cation is stabilized by the micelle. Electron transfer between two guest molecules, giving two radical ions, often occurs most efficiently on micellar systems. The high local concentration and the interface promote electron transfer and lead to efficient reaction. The charged micelle interface stabilizes the ion of opposite charge, while repelling into the bulk water phase the ion of like charge. This latter effect of the micelle surface leads both to efficient charge separation and to long-lived radical ions. The ions can be used to generate useful chemistry via subsequent reaction with other components of the system. In particular, the presence of catalysts such as colloidal platinum can lead to water breakdown, the radical anion donating an electron to the platinum which leads to the formation of hydrogen. These types of micelle-mediated processes are of use in the storage of solar energy. See PHOTOCHEMISTRY

J. K. Thomas

Bibliography. J. H. Fendler, *Membrane Mimetic Chemistry*, 1982; K. Kalyanasundaram, *Photochemistry in Microheterogeneous Systems*, 1987; D. O. Shah (ed.), *Micelles, Microemulsions, and Monolayers*, 1998; J. K. Thomas, *The Chemistry of Excitation at Interfaces*, Amer. Chem. Soc. Monog. Ser. 181, 1984.

Microbial ecology

The study of interrelationships between microorganisms and their living and nonliving environments. Microbial populations are able to tolerate and to

grow under varying environmental conditions, including habitats with extreme environmental conditions such as hot springs, salt lakes, and deep-sea thermal vents. Some microorganisms, referred to as extremophiles, grow only under conditions usually considered hostile to life. The archaeobacteria (Archaea) have many extremophiles that grow at high temperatures, high salinities, and low pH. Understanding the environmental factors controlling microbial growth and survival offers insight into the distribution of microorganisms in nature, and many studies in microbial ecology are concerned with examining the adaptive features that permit particular microbial species to function in particular habitats.

Within habitats some microorganisms are autochthonous (indigenous), filling the functional niches of the ecosystem, and others are allochthonous (foreign), surviving in the habitat for a period of time but not filling the ecological niches. Because of their diversity and wide distribution, microorganisms are extremely important in ecological processes. The dynamic interactions between microbial populations and their surroundings and the metabolic activities of microorganisms are essential for supporting productivity and maintaining environmental quality of ecosystems. Microorganisms are crucial for the environmental degradation of liquid and solid wastes and various pollutants and for maintaining the ecological balance of ecosystems—essential for preventing environmental problems such as acid mine drainage and eutrophication. See ARCHAEA; ECOSYSTEM; EUTROPHICATION.

Population interactions. The various interactions among microbial populations and between microbes, plants, and animals provide stability within the biological community of a given habitat and ensure conservation of the available resources and ecological balance. Interactions between microbial populations can have positive or negative effects, either enhancing the ability of populations to survive or limiting population densities. Sometimes they result in the elimination of a population from a habitat.

Commensal relationships, in which one population benefits and the other is unaffected, occur when one population modifies the habitat to the benefit of a second population. For example, one microbial population may produce a growth factor or may oxidize a substrate and form a metabolic product that a second population can use; one microorganism also may grow on the surface of another organism. Synergism or protocoperation between two populations occurs when both populations benefit from a nonobligatory relationship. Synergism enables microbial populations to reach higher densities in the rhizosphere (soil influenced by plant roots) than in root-free soil, and plants exhibit enhanced growth characteristics as a result of interactions with rhizosphere microbes. Mutualism, an obligatory interrelationship between two populations, allows populations to unite, thereby enabling them to occupy habitats that are unfavorable for the existence of either alone. Mutualistic relationships may lead to the

evolution of new organisms, such as occurs when a fungus unites with either an alga or cyanobacterium to form a lichen. Some animals rely on microorganisms to degrade cellulosic plant residues; ruminants, such as cattle, establish mutualistic relationships with cellulose-degrading microbial populations from which they derived their nutrition. *See* RHIZOSPHERE.

Competition for the same resources results in both populations achieving lower densities than would have occurred in the absence of competition, and prevents populations from occupying the same ecological niche. Amensalism, or antagonism, occurs when one population produces a substance inhibitory to another population, such as the production of an antibiotic by a microbe that inhibits the growth of another microbe. In parasitism, the parasite derives its nutritional requirements from the host, and as a result damages the host. Predation involves the consumption of a prey species by a predatory population for nutrition. Many protozoa nondiscriminantly prey or graze upon bacteria, and protozoa and invertebrates similarly graze on algae. *See* POPULATION ECOLOGY.

Biogeochemical cycling. The transfer of carbon and energy stored in organic compounds between the organisms in the community forms an integrated feeding structure called a food web. Primary producers that form organic matter are at the base of the food web. In food webs based on phytoplankton, algae and cyanobacteria are the primary food sources; in detrital food webs, microbial biomass produced from growth on dead organic matter serves as the primary food source for grazers, the organisms that feed upon primary producers. Grazers are eaten by predators, which in turn may be preyed upon by larger predators. Microbial decomposition of dead plants and animals and partially digested organic matter in the decay portion of a food web is largely responsible for the conversion of organic matter to carbon dioxide. *See* BIOMASS; FOOD WEB.

Biological nitrogen fixation is restricted to specific bacterial and archaeobacterial genera. In terrestrial habitats, the microbial fixation of atmospheric nitrogen is carried out by free-living bacteria, such as *Azotobacter*, and by bacteria living in symbiotic association with plants, such as *Rhizobium* or *Bradyrhizobium* living in mutualistic association within nodules on the roots of leguminous plants. In aquatic habitats, cyanobacteria, such as *Anabaena* and *Nostoc*, fix atmospheric nitrogen. The incorporation of the bacterial genes controlling nitrogen fixation into agricultural crops through genetic engineering may help improve yields. *See* NITROGEN FIXATION.

Microorganisms also carry out other processes essential for the biogeochemical cycling of nitrogen. In nitrification, chemolithotrophic microorganisms oxidize ammonium to nitrate. In soil, *Nitrosomonas* oxidizes ammonia to nitrite and *Nitrobacter* oxidizes nitrite to nitrate. The conversion of ammonia to nitrate causes leaching of nitrogen from the soil and

results in the loss of soil fertility and nitrate contamination of ground water. Denitrification, the microbial conversion of fixed forms of nitrogen to molecular nitrogen, completes the nitrogen cycle. *See* BIOGEOCHEMISTRY; NITROGEN CYCLE.

Acid mine drainage is a consequence of the metabolism of sulfur- and iron-oxidizing bacteria. When coal mining exposes pyrite ores to atmospheric oxygen, the combination of autoxidation and microbial sulfur and iron oxidation produces large amounts of sulfuric acid. The acid draining from mines kills aquatic life and renders water unsuitable for drinking or for recreation. *See* AUTOXIDATION.

Biodegradation of wastes. The biodegradation (microbial decomposition) of waste is a practical application of microbial metabolism for solving ecological problems. Solid wastes are decomposed by microorganisms in landfills and by composting. In landfills, organic matter is decomposed by anaerobic microorganisms. The products of anaerobic microbial metabolism, including methane and fatty acids, move into the surrounding soil, water, and air, causing the landfill to settle. Eventually, decomposition slows, subsidence ceases, and the land is stabilized. Alternatively, the organic portion of solid waste can be biodegraded aerobically by composting, converting noxious organic waste materials into carbon dioxide, water, and a humuslike product.

Liquid waste (sewage) treatment uses microbes to degrade organic matter, thereby reducing the biochemical oxygen demand (BOD). Otherwise, the BOD would cause oxygen depletion and fish kills when the sewage enters receiving waters. In the trickling filter system, sewage is sprayed over a bed of porous material covered with a slimy film of aerobic bacteria, such as *Zoogloea ramigera*, that degrade the organic matter and thus reduce BOD. Similarly, rotating biological contactors use a film of aerobic bacteria to reduce sewage BOD. In the activated sludge process, which uses a suspension of aerobic bacteria, sewage is introduced into an aeration tank, and sludge from a previous run that contains a massive number of microbes is added as an inoculum. Aerobic microbial metabolism results in the complete degradation of most organic compounds in the sewage. Septic tanks and anaerobic sludge digestors rely on anaerobic microbial metabolism for treating sewage.

Because fecal contamination of potable water from untreated or inadequately treated sewage promotes the rapid dissemination of pathogens, chlorination is used for disinfection following sewage treatment and chloramination is used to disinfect municipal water supplies. The degree of fecal contamination of water is monitored by testing for indicator coliform bacteria. Because the coliform bacterium *Escherichia coli* is present in far greater numbers in human fecal material than are enteropathogens, the coliform test can reliably detect potentially dangerous fecal contamination. *See* ESCHERICHIA; SEWAGE TREATMENT.

Biodegradation of environmental pollutants. Human exploitation of fossil fuels and the production of novel synthetic compounds (xenobiotics) such as plastics and pesticides have introduced many compounds into the environment that are difficult for microorganisms to biodegrade. When petroleum spills occur, for example, microorganisms are faced with the task of degrading tons of hydrocarbons with varying chemical structures. Although it cannot work quickly enough to remove coastal oil spills before beaches are coated and plants and animals are killed, microbial biodegradation will eventually degrade petroleum pollutants, preventing the oceans from being completely covered with oil. Bioremediation, that is, the use of microorganisms to remove pollutants, is used as a biotechnological treatment of oil spills. It also has great potential for the environmental cleanup of other pollutants, including polychlorinated biphenyls (PCBs). *See* BIODEGRADATION.

The ability of a microorganism to degrade an environmental pollutant is highly dependent on the chemical structure of the pollutant. Often a simple change in the substituents of a pesticide may make the difference between biodegradability and recalcitrance (complete resistance to biodegradation). When xenobiotics are recalcitrant and microorganisms fail in their role of “biological incinerators,” environmental pollutants accumulate. Synthetic organic compounds can be designed to eliminate obstacles to biodegradation. For example, some alkyl benzyl sulfonates in laundry detergents and plastics have been synthesized so that they are biodegradable. The environmentally safe use of chemical pesticides depends on microbial biodegradation to prevent persistence and biomagnification of these toxic compounds. When microorganisms fail to degrade the pesticide, serious environmental consequences can result far from the site of application, as seen in the case of dichlorodiphenyltrichloroethane (DDT). Biological control, through the regulated establishment of microbial diseases in pest populations, offers a useful alternative to chemical pesticides. *See* DETERGENT; ECOLOGY; ENVIRONMENTAL TOXICOLOGY; BIODEGRADATION; PESTICIDE.

Ronald M. Atlas

Bibliography. R. M. Atlas and R. Bartha, *Microbial Ecology: Fundamentals and Applications*, 4th ed., 1997; R. G. Burns and J. H. Slater (eds.), *Experimental Microbial Ecology*, 1982; R. E. Campbell, *Microbial Ecology*, 1983; C. Edwards, *Microbiology of Extreme Environments*, 1990; J. G. Jones (ed.), *Advances in Microbial Ecology*, vol. 13, 1994; J. M. Lynch and J. Hobbie (eds.), *Microorganisms in Action: Concepts and Applications in Microbial Ecology*, 1988.

Microbiology

The multidisciplinary science of microorganisms which began with the study of bacteria. The prefix micro- generally refers to an object sufficiently small that a microscope is required for visualization. In the

seventeenth century, Anton van Leeuwenhoek first documented observations of bacteria by using finely ground lenses. Bacteriology, as a precursor science to microbiology, was based on Louis Pasteur's pioneering studies in the nineteenth century, when it was demonstrated that microbes as minute simple living organisms were an integral part of the biosphere involved in fermentation and disease. Microbiology matured into a scientific discipline when students of Pasteur, Robert Koch, and others sustained microbes on various organic substrates and determined that microbes caused chemical changes in the basal nutrients to derive energy for growth. These observations contributed to further understanding of the microbial basis of fermentation. Modern microbiology continued to evolve from bacteriology by encompassing the identification, classification, and study of the structure and function of a wide range of microorganisms. The comprehensive range of organisms is reflected in the major subdivisions of microbiology, which include medical, industrial, agricultural, food, and dairy. *See* BACTERIOLOGY; MICROSCOPE.

Modern microbiology includes protozoa, algae, fungi, viruses, rickettsia, and parasites as well as bacteria. Based on Koch's germ theory of disease, certain microbes are classified as either animal or plant pathogens if, after invasion of the host, a specific disease is manifested. The study of host-parasite relationships gave rise to the concept that the host and parasite coexist in a delicate balance modulated in part by the host's immune system. The latter responds primarily to invasions by agents foreign to the host. Thus, the overlapping disciplines of cell biology, immunology, and modern medicine can trace their origins through microbiology. *See* ALGAE; CELL BIOLOGY; FUNGI; IMMUNOLOGY; MEDICAL BACTERIOLOGY; MEDICAL MYCOLOGY; MEDICAL PARASITOLOGY; PROTOZOA; RICKETTSIOSES; VIRUS.

Microbiology as an interdisciplinary science has directly and indirectly spawned the development of many related fields while retaining a distinct identity. This development can be traced historically from the initial theory of fermentation to microbial metabolism, to enzymology, to biochemistry, to protein chemistry, and finally to molecular biology and biotechnology. Biochemists, geneticists, cytologists, molecular biologists, and immunologists all depend on microbes as basic experimental tools in the study of fundamental processes such as metabolism, photosynthesis, enzymatic catalysis, gene action, population dynamics, and immune responsiveness. *See* BIOCHEMISTRY; BIOTECHNOLOGY; ENZYME; GENETICS; IMMUNOLOGY; METABOLISM; MOLECULAR BIOLOGY; PHOTOSYNTHESIS; POPULATION ECOLOGY.

Microbiology continues as a dynamic science as evidenced by the finding that microbes can no longer be divided into just the two groupings prokaryotes, which have a primitive and dispersed nuclear material, and eukaryotes, which display a distinct nucleus bounded by membrane. A third life form, archaeobacteria, is neither prokaryote nor eukaryote in genetic and biochemical properties. Archaeobacteria are

usually found in harsh environments such as under extreme temperature, pressure, or salt conditions. Thus, consistent with the dynamic state of microbiology another discipline has been created that will yield important information about the mechanisms whereby cells and their constituents (proteins) function under extreme conditions. See ARCHAEA; EUKARYOTAE; PROKARYOTAE.

Microbiological research, both basic and applied, is conducted in the industrial, academic, and government sectors. Industrial microbiological research was stimulated by the discovery of antibiotics of microbial origin. The expertise acquired in microbiology prepared the industrial sector for the advent of biotechnology and recombinant gene cloning, which utilizes bacterial cells for the production of enormous quantities of medically important recombinant proteins. The massive production of recombinant human insulin and bovine growth hormone by genetically engineered bacterial cells provides classical examples. See ANTIBIOTIC; GENETIC ENGINEERING; INDUSTRIAL MICROBIOLOGY.

In the academic sector, microbial agents serve as experimental systems to examine the structure and function of various macromolecules such as proteins, polysaccharides, nucleic acids, and lipids. Bacteria and bacterial phage were critical elements leading to the discovery of deoxyribonucleic acid (DNA) as the universal genetic material responsible for the synthesis of proteins. Elucidation of the intricate mechanisms involved in gene expression, control, and regulation are the foundation of modern biotechnology. See DEOXYRIBONUCLEIC ACID (DNA).

Government research has focused on the epidemiology, confinement, and eradication of a wide range of endemic diseases of microbial etiology. Federal agencies are primarily responsible for the continuous monitoring of potential epidemics and the eventual production of drugs or vaccines to combat threats posed by such microbial agents as the influenza virus. Federal laboratories also rely on microbiology to better understand food production, geochemistry, and nutrition. International agencies, such as the World Health Organization, are involved in constant efforts to monitor and eradicate microbe-related diseases such as malaria, especially in developing countries. Federal and international agencies must also identify and focus on new diseases which threaten the world's population, as in the case of the viral-induced acquired immune deficiency syndrome (AIDS). See ACQUIRED IMMUNE DEFICIENCY SYNDROME (AIDS); EPIDEMIOLOGY; MALARIA.

Many important scientific endeavors are directly traceable to basic microbiological findings and principles. With the discovery of restriction enzymes, which cleave specific sites within chromosomes, specific genes can be identified and isolated, and their structure determined for eventual use in gene cloning experiments. These capabilities led to the founding of the Human Genome Project, which is an attempt to map and determine a base se-

quence for the DNA in all 23 human chromosomes of a human prototype. Having a complete DNA sequence may allow for the diagnosis and eventual cure of large numbers of gene-linked diseases. See CLINICAL MICROBIOLOGY; INDUSTRIAL MICROBIOLOGY; MARINE MICROBIOLOGY; PETROLEUM MICROBIOLOGY; SOIL MICROBIOLOGY; TEXTILE MICROBIOLOGY. Edward W. Voss, Jr.

Bibliography. T. D. Brock et al., *Biology of Microorganisms*, 9th ed., 1999; P. R. Murray et al., *Medical Microbiology*, 3d ed., 1998.

Microbiota (human)

Microbial flora harbored by normal, healthy persons. The normal fetus is sterile, but during and after birth the infant is exposed to an increasing number of microorganisms. Subsequently, those microorganisms best adapted to survive and colonize particular sites establish themselves and become predominant.

In a healthy human, internal tissues (such as brain, blood, cerebrospinal fluid, and muscle) are normally free of microorganisms. Conversely, surface tissues (such as skin and mucous membranes) are constantly in contact with environmental microorganisms and are readily colonized by certain microbial species. The mixture of microorganisms regularly found at any anatomical site is referred to as the normal microbiota, the indigenous microbial population, the microflora, or the normal flora. For consistency, the term normal microbiota is used here. Because bacteria make up most of the normal microbiota, they are emphasized over the fungi (mainly yeasts) and protozoa. See BACTERIA; FUNGI; MICROBIAL ECOLOGY; PROTOZOA; YEAST.

Skin. The adult human is covered with approximately 2 m² (21.5 ft²) of skin. It has been estimated that this surface area supports about 10¹² bacteria. Commensal microorganisms living on the skin can be either resident (normal) or transient microbiota. Resident organisms normally grow on or in the skin. Their presence becomes fixed in well-defined distribution patterns. Those microorganisms that are temporarily present are transients. Transients usually do not become firmly entrenched and are unable to multiply.

Most skin bacteria are found on the superficial cells, on colonizing dead cells, or are closely associated with oil and sweat glands. Secretions from these glands provide the water, amino acids, urea, electrolytes, and specific fatty acids that serve as nutrients primarily for *Staphylococcus epidermidis* and aerobic corynebacteria. Gram-negative bacteria generally are found in the moister regions. The yeasts *Pityrosporum ovale* and *P. orbiculare* normally occur in the scalp. Some dermatophytic fungi may colonize the skin and produce athlete's foot and ringworm. Other normal microbiota of the skin include coagulase-negative staphylococci, diphtheroids (including *Propionibacterium acnes*), *Staphylococcus aureus*, streptococci (various species), *Bacillus*

spp., *Malassezia furfur*, *Candida* spp., and *Mycobacterium* spp. (occasionally).

Nose and nasopharynx. The normal microbiota of the nose are found just inside the nostrils. *Staphylococcus aureus* and *S. epidermidis* are the predominant bacteria present and are found in approximately the same numbers as on the skin and face.

The nasopharynx, that part of the pharynx lying above the level of the soft palate, may contain small numbers of potentially pathogenic bacteria such as *Streptococcus pneumoniae*, *Neisseria meningitidis*, and *Haemophilus influenzae*. Diphtheroids, a large group of nonpathogenic gram-positive bacteria that resemble *Corynebacterium* spp., are commonly found in both the nose and nasopharynx.

Oropharynx. The oropharynx is that division of the pharynx lying between the soft palate and the upper edge of the epiglottis. The most important bacteria found in the oropharynx are the various alpha-hemolytic streptococci (*S. oralis*, *S. milleri*, *S. gordonii*, *S. salivarius*), large numbers of diphtheroids, *Branhamella catarrhalis*, and small gram-negative cocci related to *Neisseria meningitidis*. The palatine and pharyngeal tonsils harbor a similar microbiota, except within the tonsillar crypts, where there is an increase in *Micrococcus* and the anaerobes *Porphyromonas*, *Prevotella*, and *Fusobacterium*.

Respiratory tract. The upper and lower respiratory tracts (trachea, bronchi, bronchioles, alveoli) do not have a normal microbiota, because microorganisms are removed in at least three ways. First, a continuous stream of mucus is generated by the goblet cells. This mucus entraps microorganisms, and the ciliated epithelial cells continuously move the entrapped microorganisms out of the respiratory tract. Second, alveolar macrophages phagocytize and destroy microorganisms. Finally, a bactericidal effect is exerted by the enzyme lysozyme, which is present in the nasal mucus.

Eye. At birth and throughout life, a small number of bacterial commensals are found on the conjunctiva of the eye. The predominant bacterium is *Staphylococcus epidermidis*, followed by *S. aureus*, *Haemophilus* spp., and *Streptococcus pneumoniae*.

External ear. The normal microbiota of the external ear resemble those of the skin, with coagulase-negative staphylococci and *Corynebacterium* predominating. Mycological studies show the following fungi to be normal microbiota: *Aspergillus*, *Alternaria*, *Penicillium*, *Candida*, and *Saccharomyces*.

Mouth. The normal microbiota of the mouth or oral cavity contain organisms that resist mechanical removal by adhering to surfaces like the gums and teeth. Those that cannot attach are removed by the mechanical flushing of the oral cavity contents to the stomach, where they are destroyed by hydrochloric acid. The continuous desquamation (shedding) of epithelial cells also removes microorganisms. Those microorganisms able to colonize the mouth find a very comfortable environment due to the availability of water and nutrients, the suitability of pH and tem-

perature, and the presence of many other growth factors.

The oral cavity is colonized by microorganisms from the surrounding environment within hours after a human is born. Initially, the microbiota consist mostly of the genera *Streptococcus*, *Neisseria*, *Actinomyces*, *Veillonella*, and *Lactobacillus*. Some yeasts are also present. Most microorganisms that invade the oral cavity initially are aerobes and obligate anaerobes. When the first teeth erupt, the anaerobes (*Porphyromonas*, *Prevotella*, and *Fusobacterium*) become dominant due to the anaerobic nature of the space between the teeth and gums. As the teeth grow, *Streptococcus parasanguis* and *S. mutans* attach to their enamel surfaces; *S. salivarius* attach to the buccal and gingival epithelial surfaces and colonize the saliva. These streptococci produce a glycocalyx and various other adherence factors that enable them to attach to oral surfaces. The presence of these bacteria contributes to the eventual formation of dental plaque, caries, gingivitis, and periodontal disease. See ANAEROBIC INFECTION.

Stomach. Many microorganisms are washed from the mouth into the stomach. Owing to the very acidic pH values (2 to 3) of the gastric contents, most microorganisms are killed. As a result, the stomach usually contains less than 10 viable bacteria per milliliter of gastric fluid. These are mainly *Streptococcus*, *Staphylococcus*, *Lactobacillus*, *Preptostreptococcus*, *Helicobacter*, and yeasts such as *Candida* spp. Microorganisms may survive if they pass rapidly through the stomach or if the organisms ingested with food are particularly resistant to gastric pH (mycobacteria). See GASTROINTESTINAL TRACT DISORDERS.

Small intestine. The small intestine is divided into three anatomical sections: the duodenum, jejunum, and ileum. The duodenum (the first 25 cm or 10 in. of the small intestine) contains few microorganisms because of the combined influence of the stomach's acidic juices and the inhibitory action of bile and pancreatic secretions. Of the bacteria present, gram-positive cocci and rods make up most of the microbiota. *Enterococcus faecalis*, lactobacilli, diphtheroids, and the yeast *Candida albicans* are occasionally found in the jejunum. In the distal portion of the small intestine (ileum), the microbiota begin to take on the characteristics of the colon microbiota. It is within the ileum that the pH becomes more alkaline. As a result, anaerobic gram-negative bacteria and members of the family Enterobacteriaceae become established.

Large intestine (colon). The large intestine, or colon, has the largest microbial community in the body. Microscopic counts of feces approach 10^{12} organisms per gram of wet weight. Over 400 different species have been isolated from human feces. The colon can be viewed as a large fermentation vessel, and the microbiota consist primarily of anaerobic, gram-negative, nonsporing bacteria and gram-positive, spore-forming or nonsporing rods. Not only are the vast majority of microorganisms anaerobic,

but many different species are present in large numbers. Several studies have shown the ratio of anaerobic to facultative anaerobic bacteria is approximately 300 to 1. Even the most abundant of the latter, *Escherichia coli*, is only about 0.1% of the total population.

Besides the many bacteria in the colon, the yeast *Candidia albicans* and certain protozoa may occur as harmless commensals. *Trichomonas hominis*, *Entamoeba hartmanni*, *Endolimax nana*, and *Iodamoeba butschlii* are common inhabitants.

Genitourinary tract. The upper genitourinary tract (kidneys, ureters, and urinary bladder) is usually free of microorganisms. In both the male and female, a few bacteria (*Staphylococcus epidermidis*, *Enterococcus faecalis*, and *Corynebacterium* spp.) usually are present in the distal portion of the urethra.

In contrast, the adult female genital tract, because of its large surface area and mucous secretions, has a complex microbiota that constantly changes with the female's menstrual cycle. The major microorganisms are the acid-tolerant lactobacilli, primarily *Lactobacillus acidophilus*, often called Doderlein's bacillus. They ferment glycogen produced by the vaginal epithelium, forming lactic acid. As a result, the pH of the vagina and cervix is maintained between 4.4 and 4.6.

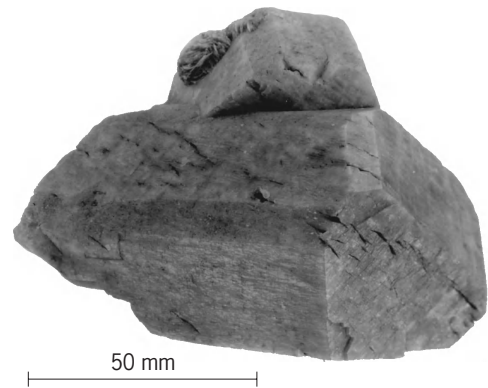
John P. Harley

Bibliography. F. Backhed et al., Host-bacterial mutualism in the human intestine, *Science*, 307:1915–1920, 2005; B. S. Drasar et al., *Intestinal Microbiology*, American Society for Microbiology, Washington, DC, 1985; P. B. Ekburg et al., Diversity of the human intestinal microbial flora, *Science*, 308:1635–1638, 2005; P. A. Mackowiak, The normal microbial flora, *N. Engl. J. Med.*, 307:83, 1982; J. Travis, Gut check, *Sci. News*, 163:344–345, 2003.

Microcline

Triclinic potassium feldspar, $KAlSi_3O_8$, that usually contains a few percent sodium feldspar (Ab = $NaAlSi_3O_8$) in solid solution. Its hardness is 6; specific gravity, 2.56; mean refractive index, 1.52; color, white (green varieties are called amazon stone or amazonite; see **illus.**). Good (001) and (010) planar cleavages are inclined at 90.4° . Microcline is found in some relatively high-grade regional metamorphic rocks, but is much more common in pegmatites, granites, and related plutonic igneous rocks. In the last, it often occurs as a microcline perthite, containing exsolved low albite intergrowths. See PERTHITE.

Microcline is usually twinned on both the Albite and Pericline laws. This fine-scale, cross-hatched (or tweed) appearance is observed only in thin rock sections examined on a polarizing light microscope. It is the result of ordering of Al and Si from a higher-temperature, monoclinic, sanidinelike arrangement, which may include orthoclase as an intermediate state. Microcline that crystallizes authigenically at relatively low temperatures ($<570^\circ\text{F}$ or 300°C), and thus in its own stability field, will lack this twinning entirely. Low microcline is completely ordered,



Microcline, variety amazon stone, Pikes Peak, Colorado. (American Museum of Natural History specimen)

with Al in one of the four nonequivalent tetrahedral sites and Si in the other three. Intermediate microclines are relatively rare in nature and have complex crystal structures with somewhat disordered Al, Si arrangements. See AUTHIGENIC MINERALS; CRYSTAL STRUCTURE; FELDSPAR; IGNEOUS ROCKS; ORTHOCLASE; SOLID SOLUTION.

Paul H. Ribbe

Microcomputer

A digital computer whose central processing unit consists of a microprocessor, a single semiconductor integrated circuit chip. Once less powerful than larger computers, microcomputers are now as powerful as the mini- and super-minicomputers of just several years ago. This is due in part to the growing processing power of each successive generation of microprocessor, plus the addition of mainframe computer features to the chip, such as floating-point mathematics, computation hardware, memory management, and multiprocessing support. See INTEGRATED CIRCUITS; MICROPROCESSOR; MULTIPROCESSING.

Microcomputers are the driving technology behind the growth of personal computers and workstations. While the hardware distinctions between personal computers and workstations have blurred, the applications they use and the operating system environments are still distinct. Traditionally, personal computers offer lower performance at a lower cost, with applications focusing on business, games, and education. In contrast, workstations use more specialized components that deliver high performance on numerically intensive scientific and engineering computations, such as simulations and three-dimensional (3-D) design.

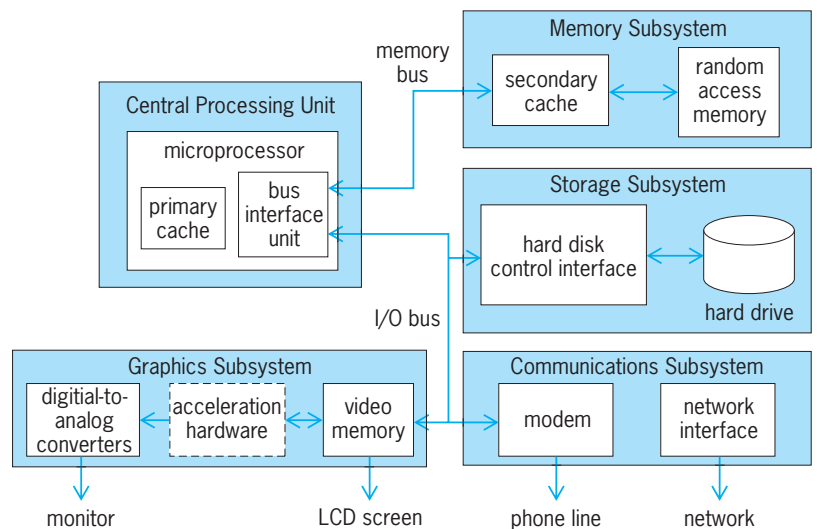
Early microcomputers relied on ASCII text (characters) for input via a keyboard or paper tape as well as output on a terminal or display. They used floppy disks for storage and had limited computational and memory capacity. Today's microcomputers employ bit-mapped graphical user interfaces on color displays and handle a wide variety of input and output devices such as mice, digitizing tablets,

digital color video cameras, and compact-disk-quality stereo sound. The memory capacities of microcomputers range from tens to hundreds of megabytes, and certain specialized systems may have a gigabyte of memory. Mass storage is handled by hard disk drives with capacities of several gigabytes or more, augmented by optical storage devices such as CD-ROM (compact-disk, read-only memory), magneto-optical storage, and DVD (digital video disk or digital versatile disk). The computational and memory capacity of microcomputers increased from 64 kilobytes in 1974 to over 4 gigabytes in 1998, an increase of four orders of magnitude. Within just several years, microprocessor clock frequencies climbed from 133 MHz to 1 GHz in 2000. The combined memory capacity and the processing speed of today's microcomputers make new applications possible. Such applications include streaming digital video for videoconferencing and entertainment, and the real-time display of three-dimensional graphics for data analysis, games, and virtual reality.

Microcomputers are employed in virtually every sector of society, including commerce, manufacturing, education, government, travel, and the home. General-purpose microcomputers, in the form of personal and professional computers, support a variety of applications, from video games to business reports to laboratory analysis. The microcomputer's network capabilities and storage capacity have fueled the growth of the Internet, a worldwide network of interconnected computers once used only by research laboratories and universities. This network formerly consisted of mainframes and supercomputers, but now microcomputers serve both as the Internet's backbone for storing and transporting information on demand and as the terminals by which information is accessed. The Internet's ease of use and ubiquity have created new ways to gather information and conduct commerce that are still evolving. See INTERNET.

The capabilities of today's microprocessors in combination with reduced power consumption have created a new category of microcomputers: handheld devices. Some of these devices are actually general-purpose microcomputers: They have a miniature keyboard and liquid-crystal-display (LCD) screen and use an operating system that runs several general-purpose applications. Many others, however, serve a fixed purpose. Examples of the latter are cellular telephones that provide a display for receiving text-based pager messages and automobile navigation systems that use satellite-positioning signals to plot the vehicle's position. See LIQUID CRYSTALS; MOBILE COMMUNICATIONS; RADIO PAGING SYSTEMS; SATELLITE NAVIGATION SYSTEMS.

Hardware. The microcomputer is an electronic device. The microprocessor acts as the microcomputer's central processing unit (CPU), performing all the operations necessary to execute a program. In early microcomputer designs, the microprocessor also managed simple input/output (I/O) devices such as the graphics display and keyboard. Today's microcomputers are complex machines having sev-



Elements of a microcomputer. The various subsystems are controlled by the central processing unit. Some designs combine the memory bus and bus input/output into a single system bus. The graphics subsystem may contain optional graphics acceleration hardware.

eral distinct subsystems (see *illus.*). The microprocessor still makes up the system's CPU.

A memory subsystem uses semiconductor random-access memory (RAM) for the temporary storage of data or programs. The memory subsystem may also have a small secondary memory cache that improves the system's performance by storing frequently used data objects or sections of program code in special high-speed RAM.

The graphics subsystem consists of hardware that displays information on a color monitor or LCD screen: a graphics memory buffer stores the images shown on the screen, digital-to-analog convertors (DACs) generate the signals to create an image on an analog monitor, and possibly special hardware accelerates the drawing of two- or three-dimensional graphics. (Since LCD screens are digital devices, the graphics subsystem sends data to the screen directly rather than through the DACs.) See DIGITAL-TO-ANALOG CONVERTER.

The storage subsystem uses an internal hard drive or removable media such as floppies, Zip™ or Jazz™ cartridges, or magneto-optical platters for the persistent storage of data.

The communications subsystem consists of a high-speed modem or the electronics necessary to connect the computer to a network. This last subsystem was considered optional in early microcomputer designs; however, the ability to share data among coworkers and to connect to the Internet has become so important that it has become a standard part of microcomputer design.

Architecture. The microprocessor is built of the same components as any digital computer. It has a control unit, an arithmetic and logic unit (ALU), special temporary storage units called registers, and a bus interface unit (BIU). The control unit orchestrates the operation of all the other microprocessor units, having them perform their dedicated tasks simultaneously and at very high speed. The control

unit works by ordering the BIU to fetch instructions from semiconductor memory. It decodes them, typically several at a time, and has the appropriate units perform a sequence of operations that completes the task as directed by the fetched instructions. For example, a set of instructions may have the BIU fetch more data from memory into the microprocessor's registers. Next, the control unit commands the ALU to perform a mathematics calculation (perhaps an add operation) using the values found in two of the registers. Finally, the BIU writes the result of this action back into memory, where it can be used by subsequent instructions. In many microprocessors, a small amount of on-chip memory called the primary cache stores the most recently used instructions or data. If these items are still resident in this cache when the control unit requires them, the BIU can draw the information from the high-speed cache, rather than from the slower off-chip RAM. The performance of microcomputers is often measured in millions of instructions per second (MIPS). Many personal computers operate in the range of 200–400 MIPS.

The BIU thus handles all of the signals necessary for the microprocessor to transfer data between its internal registers and semiconductor memory in the memory subsystem. This electronic interface is known as the memory bus. The BIU also implements the data routes that connect the microprocessor to the other input/output subsystems. Data sent or received from any input/output device in the system typically traverse the memory bus. However, certain special-purpose processor designs use separate buses for memory and input/output, for the sake of design simplicity (to achieve a low cost in a consumer product) or performance (the microprocessor is free to simultaneously decode a compressed video image stored in RAM, while it reads more images from a DVD drive). *See* COMPUTER ARCHITECTURE; DIGITAL COMPUTER.

Word size. An important factor contributing to a microcomputer's processing power is the width of its data paths. These paths usually determine the size of the data elements with which the microcomputer works. The width also determines how fast data can move through the system. The data path width is commonly called the machine's word size and is measured in bits, where each bit represents a single data path. Early microcomputers had 8- or 16-bit data paths, but microcomputers today commonly use 32-bit data paths. Advanced microprocessors using 64-bit data paths are commonly found in workstations or supercomputers (where they make up the elements of a multiprocessor array). Eight- and sixteen-bit processors are still in widespread use as controllers in consumer products such as household appliances. However, 32-bit embedded processors are appearing that meet the processing needs of sophisticated products such as set boxes and smart office copiers. *See* BIT; EMBEDDED SYSTEMS; SUPER-COMPUTER.

Clocks. Another important measure of a microcomputer's computing power is its clock rate. The

frequency of the clock signal determines how fast the computer can process information. Commercial microprocessors operate in the range from 25 MHz to 1 GHz and more. Depending upon its architecture, a microprocessor can execute one or more instructions per clock cycle. Consequently, a computation rate of hundreds of millions of instructions per second can be achieved.

Data bus. Previously, communications between the microprocessor and other parts of the microcomputer system occurred over the memory bus. This is because the input/output operations also made use of the memory bus's address and data lines. In this arrangement, the memory bus is commonly called the system bus. The system bus thus provides the routes that collect data from input devices and transfer the data to the semiconductor memory. The CPU then operates on the data, and the system bus transfers the results to an output device. However, in many high-performance microcomputer designs, such as servers, the architecture separates the two functions—accessing memory and device input/output—on two separate buses. This arrangement effectively doubles the data paths through which information can flow through the system, boosting throughput. It also allows the memory subsystem to operate at a higher clock speed than the input/output bus, which further improves performance. The input/output bus is typically based on the peripheral component interconnect (PCI) standard.

In some high-end designs, certain microcomputer subsystems contain their own "local" memory and microprocessors. These intelligent subsystems can handle certain tasks independently of the CPU, using programs stored in the local memory. They use the system bus only when transferring data between the device's local memory and the memory subsystem.

Data storage. Data are stored on a microcomputer in various ways: RAM chips, ferromagnetic hard disk drives, and read-only memory (ROM) such as CD-ROMs. Data bits are stored as electronic "on" or "off" states in the array of transistors that make up the semiconductor chip. RAM is used for the temporary storage of program code and volatile data because of its high access speed (typically in nanoseconds). Programs and data are saved on persistent storage media such as a hard drive. The bits that constitute a program's instructions are encoded as magnetic field changes on the ferromagnetic material. These devices have slower access speeds (measured in milliseconds) but have the advantage of preserving the data when the microcomputer is turned off. *See* COMPACT DISK; COMPUTER STORAGE TECHNOLOGY; MAGNETIC RECORDING; OPTICAL RECORDING; SEMICONDUCTOR MEMORIES.

Input and output. Microcomputers interact with users in a variety of ways. They accept data from input devices such as a keyboard, a mouse, touch screen, microphone, scanner, or video camera. In each case, the data collected by these devices are translated via hardware and software into a form usable by the microcomputer. For example, when

keys are pressed on a keyboard, a controller determines which keys were pressed and converts these actions into data bytes (a byte being a group of bits, usually eight, that encode a character). The microcomputer's operating system converts the bytes into characters, and passes them to a program. The program, perhaps a word processor, then displays the characters on the screen. Talking into a microphone triggers speech recognition software, which parses the tones into specific commands. The software then relays these commands to the operating system, which executes the appropriate actions. A scanner can convert an image into an array of bytes that a drawing program can manipulate. Or an optical character recognition (OCR) program might interpret the patterns of bytes, converting them into text. An example of an advanced input device is a data glove, which senses the position of the fingers on a user's hand. Microcomputer software uses this information to allow a user to control the orientation, position, and rate of objects displayed in a simulation program. When an input device such as the data glove, combined with an output device such as 3-D goggles, enables users to immerse themselves in a simulation program, this type of interaction is called virtual reality. *See* CHARACTER RECOGNITION; COMPUTER PERIPHERAL DEVICES; SPEECH RECOGNITION; VIRTUAL REALITY.

The graphic output produced by the microcomputer is limited only by the resolution of the bit-mapped graphics display hardware. Early display resolutions were either 320 by 240 pixels, or 640 by 480 pixels (where a pixel represents one element on the display). Today, standard screen resolutions range from 800 by 600 pixels to 1024 by 768 pixels. Specialized graphics applications can require displays of even higher resolutions. Most graphics subsystems use an 8- or 24-bit value for each pixel that represents its color. (An 8-bit display has $2^8 = 256$ colors or shades of gray; a 24-bit display has $2^{24} = 16,777,216$ colors.) Operating systems use these bit-mapped capabilities to present text in any typeface, color, and size. The graphics capability can be used in simulations, in what-you-see-is-what-you-get (WYSIWYG) editors, and in displaying pictures. By presenting a stream of images on a screen rapidly, the graphics subsystem can even present full-motion video. Such multimedia technology allows the microcomputer to combine sound, graphics, and text to enhance the effectiveness of communications. Multimedia-capable microcomputers are part of the reason that users from all walks of life can easily work with information on the Internet. *See* COMPUTER GRAPHICS; ELECTRONIC DISPLAY; HUMAN-COMPUTER INTERACTION; MULTIMEDIA TECHNOLOGY; VOICE RESPONSE.

Networks and electronic services. Attaching computers to each other via computer networks not only multiplies the information available to a microcomputer but also adds the dimension of communications to its capabilities. The incorporation of the network subsystem as standard—rather than optional—hardware has made this task even easier. Network services, such as electronic mail, and dial-

up services, such as online systems or Internet service providers, enable users to work remotely from the office and correspond with coworkers or relatives around the globe. Such services have played a major role in the increasing utility and popularity of microcomputers. *See* ELECTRONIC MAIL; LOCAL-AREA NETWORKS; WIDE-AREA NETWORKS.

Portable computers. The rapid advance of electronic technology has made it possible to build extremely small, light, yet powerful microcomputers. Low-power microprocessors, flat-panel displays, and miniature hard drives have played a critical role. Many laptops weigh from just less than 2 lb (1 kg) up to 7 lb (3 kg). Yet they have RAM, hard-disk capacity, and a display resolution equal to that of stationary desktop computers.

In addition, some portable computers serve as hand-held, fixed-purpose devices. An example is a digital camera that runs an operating system, captures image data, and displays it on a viewfinder's LCD screen in real time. When a picture is taken, the microcomputer applies color corrections to the image, then compresses and stores it on memory cards or a miniature hard drive. All of this is accomplished by an integral microcomputer that is inexpensive and operates on battery power. *See* CAMERA.

Software. Microcomputer software is the logic that makes microcomputers useful. Software consists of programs, which are sets of instructions that direct the microcomputer through a sequence of tasks. A startup program in the microcomputer's ROM initializes all of the devices, loads the operating system software, and starts it. All microcomputers use an operating system that provides basic services such as keyboard input, simple file operations, and the starting or termination of programs. While the operating system used to be one of the major distinctions between personal computers and workstations, today's personal computer operating systems also offer advanced services such as multitasking, networking, and virtual memory. All microcomputers exploit the use of bit-mapped graphics displays to support windowing operating systems. This capability enables the display to act as a series of overlapping windows, each showing the content of a single application program. Windowing operating systems allow an application to be written independently of the actual display size. This lets users allocate screen space as they see fit for multiple applications running simultaneously. *See* OPERATING SYSTEM; SOFTWARE.

Applications. The widespread availability of microcomputers has led to a proliferation of uses. Businesses use microcomputers for payroll, personnel, and other types of management information systems as well as for word processing, electronic mail, presentation formatting, electronic publishing, spreadsheets, project planning, financial projections, scheduling, personal information management, and many other areas. *See* WORD PROCESSING.

The low cost of microcomputers allows them to serve as controllers for scientific instruments. They can collect and process data from the instrument as well as store, analyze, and display the captured

data. Microcomputers are powerful enough to do sophisticated analysis and even maintain databases of previous measurements. The higher-performance microcomputers or workstations are widely used in science and engineering. In science, they can simulate prototype experimental procedures, and predict results that can be compared with the results of actual experiments. Microcomputers have had such a dramatic impact in this area that computation methods based on microcomputers are now considered a "third leg" of science, complementing both theory and experiment. In engineering, microcomputers support design and evaluation of products through computer-aided design (synthesis), evaluation (simulation), and databases (reuse of designs, and the collection of problem reports). In fact, many cars and airplanes are designed and thoroughly tested in simulations before a working prototype is ever built, at tremendous savings in cost and improved safety. See COMPUTER-AIDED DESIGN AND MANUFACTURING; DATABASE MANAGEMENT SYSTEM; SIMULATION.

Microcomputers are natural tools for both education and recreation. They are used in schools as a complement to classroom instruction. Software programs can provide specific details in depth, then bring that detail to life interactively. Furthermore, microcomputers can provide personalized responses, customizing lessons by allowing the student to explore where there is a greater interest, and to progress at an individualized pace. The Internet allows students to broaden their educational horizons by providing access to sources of information not available locally, and to converse with other students across the country or around the globe.

Finally, a wide variety of computer games, ranging from flight simulators to war games, fantasy, and problem solving, are available. As the computational capabilities of microcomputers have increased, these games have increased in sophistication and realism. See AIRCRAFT TESTING; DIGITAL COMPUTER; VIDEO GAMES. Tom Thompson

Bibliography. P. Freiberger and M. Swaine, *Fire in the Valley: The Making of the Personal Computer*, McGraw-Hill, 1999; H.-P. Messmer, *The Indispensable PC Hardware Book*, 3d ed., Addison-Wesley, 2000; M. Predko, *PC Ph.D.: Inside PC Interfacing*, McGraw-Hill, 1999; M. Rosenthal, *Build Your Own PC*, 2d ed., McGraw-Hill, 2000; S. Veit, *Stan Veit's History of the Personal Computer*, Worldcomm Press, 1993; R. White, *How Computers Work*, 5th ed., MacMillan Computer, 1999.

Microdialysis sampling

An approach for sampling the extracellular space of essentially any tissue or fluid compartment in the body. Continuous sampling can be performed for long periods with minimal perturbation to the experimental animal. Microdialysis provides a route for sampling the extracellular fluid without removing fluid, and administering compounds without adding fluid. The resulting sample is clean and amenable to

direct analysis. Microdialysis was initially developed to study neurochemical processes in the brain. The success of this technique in the study of neurotransmitter release has led to the development of microdialysis techniques for general pharmacokinetic and drug distribution studies.

Fundamentals. Microdialysis sampling is performed by implanting a short length of hollow-fiber dialysis membrane at the site of interest. The fiber is slowly perfused with a sampling solution (the perfusate) having an ionic composition and pH that closely matches the extracellular fluid of the tissue being sampled. Low-molecular-weight compounds in the extracellular fluid diffuse into the fiber and are swept to a collection vial for subsequent analysis. The system is analogous to an artificial blood vessel that can deliver compounds and remove the resulting metabolites. A microdialysis system for awake animals consists of a precision microinfusion pump connected to the microdialysis probe through a low-volume liquid swivel. The outlet of the microdialysis probe is then brought back through a second channel of the liquid swivel to a fraction collector. The experimental animal is housed in a containment system so that the tubing does not become entangled as the animal moves. An alternative to the use of a liquid swivel is to place the animal containment system on a turntable that automatically rotates to keep the animal stationary with respect to the microdialysis equipment (Fig. 1). See DIALYSIS; MEMBRANE SEPARATIONS.

Process. Microdialysis is a diffusion-controlled process. The perfusion rate through the probe is generally in the range of 0.5 to 5.0 ml/min. At this flow rate, there is no net flow of liquid across the dialysis membrane. The driving force for mass transport is the concentration gradient between the extracellular fluid and the fluid in the probe. If the concentration of a compound is higher in the extracellular fluid, some fraction will diffuse into the probe; this is termed a recovery experiment (Fig. 2a). Conversely, if the concentration of a compound is higher in the probe, some fraction will diffuse out of it; this is termed a delivery experiment (Fig. 2b). See DIFFUSION; TRANSPORT PROCESSES.

At typical perfusion rates, equilibrium is not established across the microdialysis membrane. The concentration of analyte (compound sampled) collected in the dialysate (the perfusate after it has gone through the microdialysis probe) is some fraction of the actual concentration in the extracellular fluid. The relationship between the fraction recovered and the actual concentration is termed the extraction efficiency and is defined as the ratio of the concentration of analyte in the dialysate to the true concentration of analyte in the tissue of interest. The extraction efficiency depends on characteristics of the probe, the sample matrix, and the analyte. Probe characteristics that affect recovery include the type of membrane, the length of membrane, the geometry of the probe, and the perfusion flow rate. The extraction efficiency is influenced by matrix characteristics, such as temperature, tortuosity, and effective

volume. The size, hydrophobicity, charge, and diffusion coefficient of the analyte affect recovery. In addition, metabolic and active transport processes affect extraction efficiency. Usually, the extraction efficiency of a compound is independent of its concentration. Under normal microdialysis sampling conditions, these parameters remain constant; and although equilibrium is not established, a steady state is rapidly achieved.

Probe design. The major design consideration for a microdialysis probe is to provide a short length of hollow-fiber dialysis membrane, through which a solution can be pumped, in a geometry that can be implanted into an animal with minimal tissue damage. Microdialysis membranes are made of cellulose, cellulose acetate, polyacrylonitrile (PAN), and polycarbonate ether. These hydrophilic membranes provide little difference in chemical selectivity. Membranes with nominal molecular weight cutoffs between 5000 and 75,000 daltons (unified atomic mass units, u) are available, although membranes in the range of 10,000 to 20,000 Da are most commonly used. The dialysis efficiency decreases dramatically as the molecular weight of a compound increases. Compounds with molecular weights much above 1000 Da typically have very low efficiencies and are not usually appropriate for microdialysis sampling. A variety of probe designs have been described that provide advantages for use in specific tissues. These designs can be divided into two classes: cannula-style (small-tube) probes with parallel inlet and outlet flow; and linear-style probes in which the inlet, membrane, and outlet are in series (Fig. 3).

The most common design is the concentric cannula type that consists of an inner and outer length of stainless steel tubing (Fig. 3a). The inner cannula extends beyond the outer cannula and is covered with the dialysis membrane. This design is most appropriate for stereotaxic implantation in the brain. Its rigid nature provides good mechanical stability and allows for precise placement in the brain. The probe can be cemented to the skull to prohibit movement of the probe when the animal moves. Alternatively, a guide cannula can be implanted and glued in place. A concentric cannula microdialysis probe can then be implanted through the guide cannula. This provides the ability to remove the probe and precisely reinsert it or use another probe later.

Microdialysis probes constructed from rigid materials are not appropriate for implantation in peripheral tissue. Unlike probes implanted in the brain, where the skull provides an excellent site to secure the probe and protect it from movement, peripherally implanted probes must be capable of moving as the animal moves without being damaged or causing damage. A modification of the concentric cannula design in which side-by-side pieces of fused-silica are used has been found to be useful for intravenous implantation. In this design, one piece of fused silica extends beyond the other and is covered with the dialysis membrane (Fig. 3b). This flexible design is still needlelike and thus can be inserted into blood vessels. However, the fused silica bends as the animal

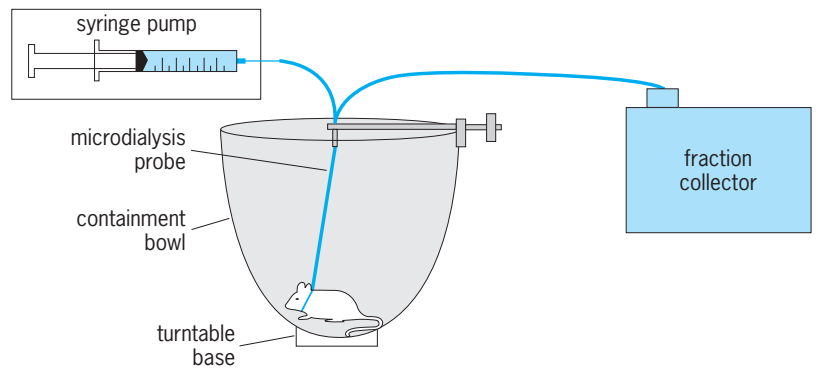


Fig. 1. Microdialysis sampling system for awake animals.

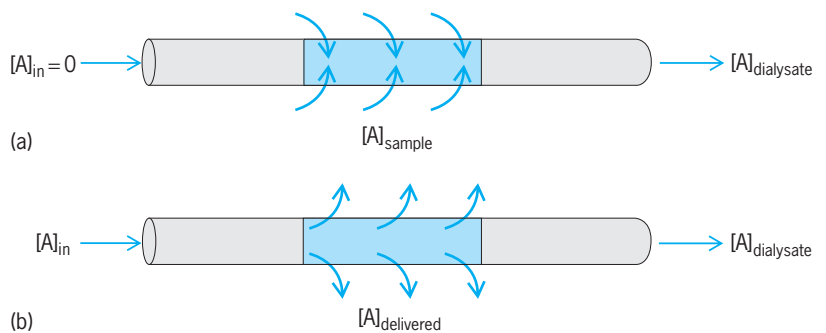


Fig. 2. Mass-transport-driven microdialysis process. (a) Recovery. (b) Delivery.

moves so that neither the dialysis membrane nor the blood vessel is punctured.

The most useful probe design for implantation in peripheral tissue is of linear geometry. Several variations of this design are in use, but the general concept is that the hollow-fiber dialysis membrane is connected to small-bore tubing on both ends

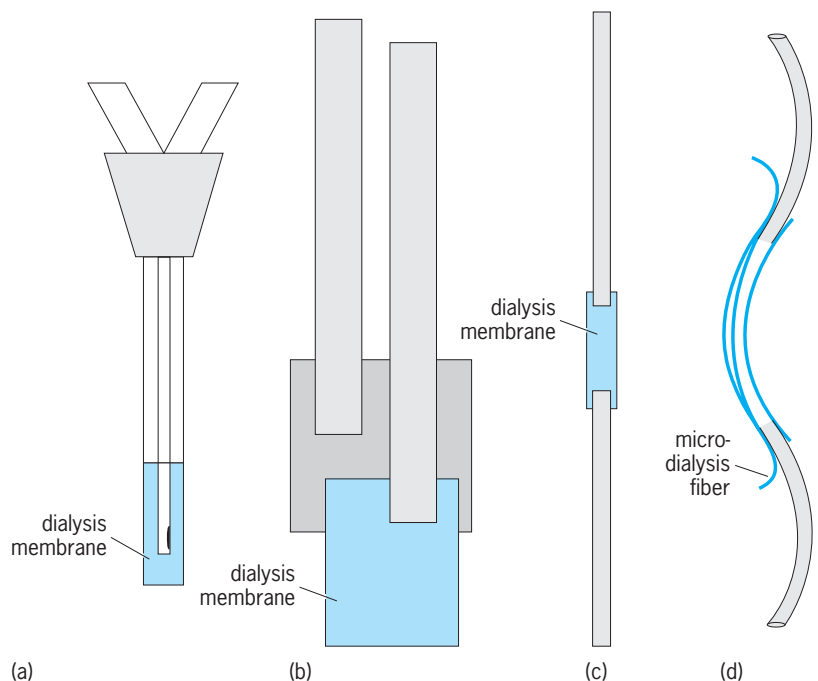


Fig. 3. Microdialysis probes. (a) Rigid concentric cannula probe. (b) Flexible fused-silica probe. (c) Linear probe. (d) Flow-through or shunt probe.

(Fig. 3c). One end is used as the inlet and the other as the outlet. A variety of tubing types can be used for the inlet and outlet, including Teflon, polyetheretherketone (PEEK), polyethylene (PE), and fused silica. Implantation is accomplished simply by pulling the probe through the tissue like a thread. Often a longer piece of the dialysis fiber than desired for sampling is used because the fiber is more flexible than the fused silica. In this case, most of the fiber is coated with an impermeable silicone resin, leaving a small uncoated window for sampling.

A fourth type of microdialysis probe is termed the flow-through or shunt design. The probe is constructed by inserting the microdialysis fiber inside a length of polyethylene tubing (Fig. 3d). The microdialysis fiber is perfused with the sampling solution while the fluid sample flows through the polyethylene tubing. This design is useful for sampling flowing fluids from sites that are too small for implantation of a flexible microdialysis probe. The polyethylene tubing acts as a shunt to bring the biological fluid past the microdialysis membrane. This probe design has been most successfully used to sample bile, but application to blood sampling from arteries or smaller veins may also be possible.

Applications. To date, the greatest use of microdialysis sampling has been in the neurosciences. Microdialysis probes can be implanted in specific brain regions of conscious animals in order to correlate neurochemical activity with behavior. Most studies have focused on determining dopamine or the other monoamine neurotransmitters. Microdialysis sampling has also proven to be a powerful technique for studying excitatory amino acids, such as glutamate, aspartate, and GABA (γ -aminobutyric acid) in the brain. Recently, the use of microdialysis to sample neuropeptides has been explored. An exciting application of microdialysis sampling has been in studying abnormal brain function in humans. Data on the neurochemical processes occurring prior to, during, and after an epileptic seizure have been obtained using microdialysis in humans. See AMINO ACIDS; NEUROBIOLOGY.

The use of microdialysis in the pharmaceutical sciences is growing rapidly. Because microdialysis provides continuous sampling without disruption of biological barriers, the technique is particularly well suited for studying the bioavailability of pharmaceutical compounds. Microdialysis probes have been implanted in the skin of experimental animals and humans to determine the transdermal delivery of drugs from ointments. Delivery of anticancer drugs to tumors has been studied using microdialysis. Alternatively, microdialysis may prove a useful technique for delivering toxic drugs, such as anticancer agents, to specific sites without systemic involvement. While continuous sampling from tissues is possible only by microdialysis, sampling from the blood has been the most common approach taken to pharmacokinetic studies using microdialysis. See DRUG DELIVERY SYSTEMS; PHARMACOLOGY.

Microdialysis sampling has also been used to study the metabolism of compounds in vivo. Metabolic or-

gans such as the liver and kidneys have been studied by microdialysis sampling. By also sampling the bile by microdialysis, complete metabolic profiles can be obtained from a single experimental animal. This approach dramatically decreases the number of experimental animals needed to assess the metabolism of a new drug.

Craig E. Lunte

Bibliography. W. F. Elmquist and R. J. Sawchuk, Application of microdialysis in pharmacokinetic studies, *Pharm. Res.*, 14:267-288, 1997; D. K. Hansen et al., Pharmacokinetics and metabolism studies using microdialysis sampling, *J. Pharm. Sci.*, 2000; C. E. Lunte, Microdialysis and target organ exposure, in D. E. Johnson (ed.), *Drug Toxicodynamics*, Marcel Dekker, 1999; T. E. Robinson and J. B. Justice, Jr. (eds.), *Microdialysis in the Neurosciences*, Elsevier, 1991.

Micro-electro-mechanical systems (MEMS)

Systems that couple micromechanisms with microelectronics. Such systems are also referred to as microsystems, and the coupling of micromechanisms with microelectronics is also termed micromechatronics. Micromechanics refers to the design and fabrication of micromechanisms that predominantly involve mechanical components with submillimeter dimensions and corresponding tolerances of the order of 1 micrometer or less. The types of systems encompassed by MEMS represent the need for transducers that act between signal and information processing functions, on the one hand, and the mechanical world, on the other. This coupling of a number of engineering areas leads to a highly interdisciplinary field that is commensurately impacting nearly all branches of science and technology in fields such as biology and medicine, telecommunications, automotive engineering, and defense. Ultimately, realization of a "smart" MEMS may be desired for certain applications whereby information processing tasks are integrated with transduction tasks, yielding a device that can autonomously sense and accordingly react to the environment. See TRANSDUCER.

Motivating factors behind MEMS include greater independence from packaging shape constraints due to decreased device size, with typical complete packaged devices occupying volumes of less than 0.05 cm^3 (0.003 in.^3). In addition, the advantages of repeatable manufacturing processes as well as economic advantages can follow from batch fabrication schemes such as those used in integrated circuit processing, which has formed the basis for MEMS fabrication. Many technical and manufacturing trade-offs, however, come into play in deciding whether an integrated approach is beneficial. In some cases, the device design with the greatest utility is based on a hybrid approach, where mechanical processing and electronic processing are separated until a final packaging step. See INTEGRATED CIRCUITS.

Two broad categories of devices follow from the transduction need addressed by MEMS: the input transducer or microsensors, and the output

transducer or microactuator. The specifications for these two categories of devices entail unique fabrication methods that have evolved since about 1970 and have resulted in correspondingly unique process technology and manufacturing methods.

Attributes. An ideal sensor extracts information from the environment without perturbing the environment. Thus, minimal energy exchange is desired. Fast response and high sensitivity with minimal power requirements are also important sensor attributes. All of these qualities are fundamental to microsensors, and the degree to which they may be exploited is a direct consequence of mechanical scaling behavior. The progress in integrated pressure sensors provides an example of how it is possible to capitalize on this behavior. Early micromachined pressure transducers with 25- μm -thick membranes and up to 1 cm^2 area consumed 5×10^{-3} watt of power, using a piezoresistive transducer that was sensitive to forces of barely 10^{-2} newton. Reduction of the deflection membrane thickness to 1 μm resulted in a linear size reduction by a factor of nearly 100 and a corresponding increase in force sensitivity to 10^{-6} N, with slightly reduced piezoresistive power dissipation. The next advance involved changing the transduction mechanism from a piezoresistive element to a micromachined vacuum-encapsulated resonant microbeam, typically 200 μm long, 40 μm wide, and 1 μm thick, that optically sensed force via a shift in resonant frequency. An improvement in force sensitivity to better than 10^{-10} N with an input power requirement of 10^{-13} watt was the result. These improvements may be further extended through sensing concepts based on tunneling and atomic force microscopes, and ultimately through magnetic resonance force microscopes with 60-nanometer-thick resonant beams. These devices have resolved forces of 10^{-18} N (equivalent to masses of tens of attograms), and a thermal noise-limited sensitivity of 10^{-19} N is anticipated. See BROWNIAN MOVEMENT; MICROSENSOR; PRESSURE TRANSDUCER; SCANNING TUNNELING MICROSCOPE.

Actuators and corresponding linkage mechanisms are three-dimensional devices that operate via changes in stored energy. The goal is a certain force-versus-displacement character that determines the required volume necessary to be devoted to energy storage, thereby implying a three-dimensional structure. Direct scaling of macroscopic counterparts suggests potential problems for micromechanisms in behavior that is affected by changes in the surface-to-volume ratio. These issues include substantially reduced inertial effects that occur with size reduction, and the relatively greater influence of friction, surface tension, and air damping, which all become more significant relative to mass, particularly in comparison to a 1-horsepower (750-watt) motor or engine, for example. Microactuators and mechanisms based on planar batch integrated circuit fabrication techniques are restricted to essentially two-dimensional elements, a requirement that challenges micro-mechanism design. High-aspect-ratio micromachining (HARM) has

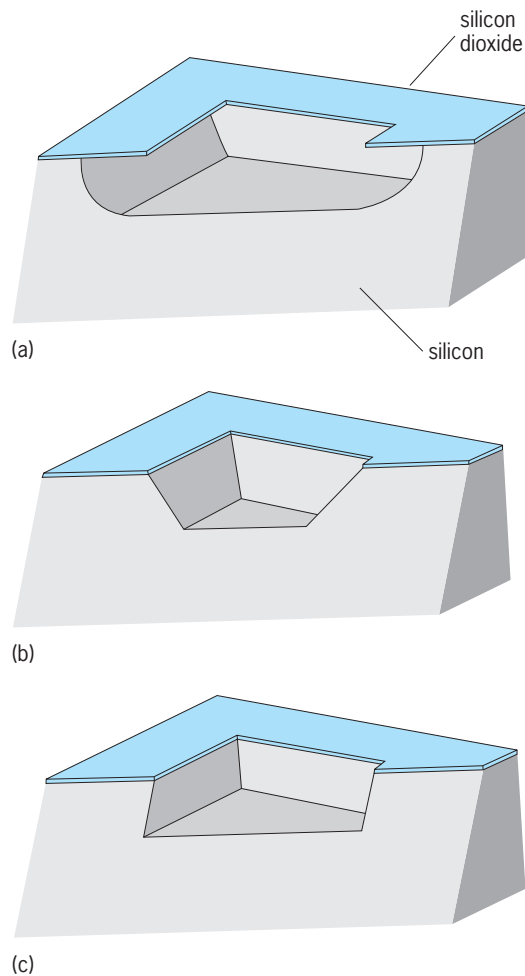


Fig. 1. Silicon etching processes. (a) Isotropic etching. (b) Anisotropic crystal etching. (c) Vertical anisotropic etching.

been developed to alleviate the two-dimensional constraint. High-aspect-ratio processing also aids in the manufacture of inertial sensors, for example, in defining inertial reference masses for applications such as seismic sensing. See SEISMOGRAPHIC INSTRUMENTATION.

Microfabrication technology. The development of process tools and materials for MEMS is the pivotal enabler for integration success. A material is chosen and developed for its mechanical attributes and patterned with a process amenable to co-electronic fabrication. Two basic approaches to patterning a material are used. Subtractive techniques pattern via removal of unwanted material, while additive techniques make use of temporary complementary molds within which the resulting structure conforms. In either case, the goal is precision microstructural definition. Both approaches use a mask to transfer a pattern to the desired material. For batch processes, this step typically occurs via photolithography and may itself entail several steps. The basic process is to apply a photoresist, a light-sensitive material, and use a photomask to selectively expose the photoresist in the desired pattern. A solvent chemically develops the photoresist-patterned

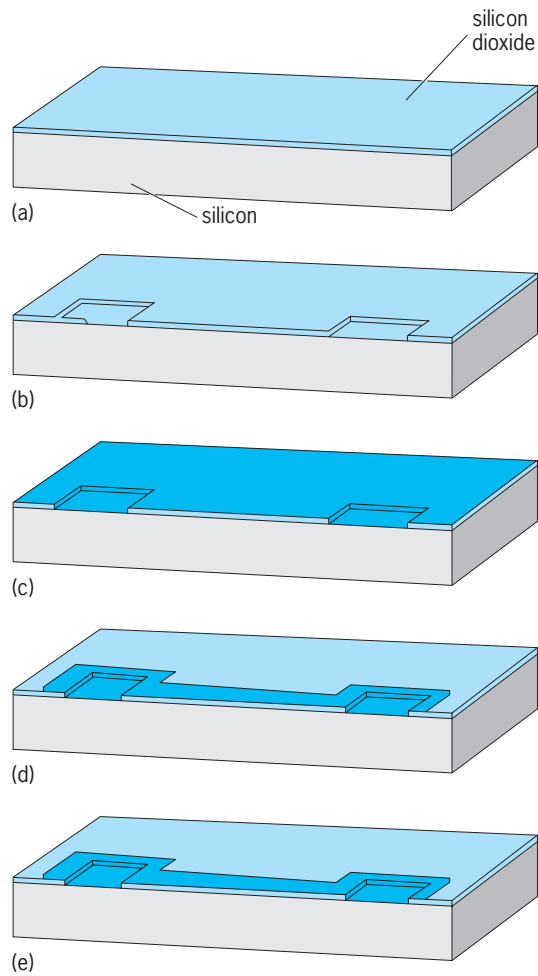


Fig. 2. Basic surface micromachining process. (a) Thermal oxidation of silicon. (b) Silicon dioxide patterning. (c) Polysilicon deposition. (d) Polysilicon patterning. (e) Release etching of silicon dioxide to render the mechanical structure.

image, which then may be used as a mask for further processing.

Subtractive processing is accomplished via chemical etching. Wet etching occurs in the liquid phase, and dry etching or gas-phase etching may occur in a vapor phase or plasma. A key issue in subtractive patterning is selectivity, which is defined as the ratio

of the etch rate of the material that is to be removed to the etch rate of the masking material or any other resident material that the designer wants to be substantially unaffected. See PLASMA (PHYSICS).

Bulk micromachining. An omnipresent material in MEMS due to its use as a semiconductor for microelectronics fabrication is single-crystal silicon. In terms of mechanical stability, silicon is also an outstanding material possessing an exceptionally linear mechanical response as well as resistance to the aging effects prevalent in metals and plastics. Single-crystal compound semiconductor materials, such as gallium arsenide, are also of interest due to their use in photonics and microwave circuitry, but are not as widely developed as silicon technology. In either case, these substrate materials are attractive for use in MEMS if precision mechanical structures can be constructed from them in a fashion that does not interfere with their electronic function.

A primary microfabrication technology that has been used for most commercial devices is bulk micromachining, which is the process of removing, or etching, substrate material. The important aspect of precision bulk micromachining is etch directionality. The two limiting cases are isotropic, or directionally insensitive, and anisotropic, or directionally dependent, up to the point of being unidirectional. All cases may be illustrated with single-crystal silicon (Fig. 1). In many instances, the photoresist possesses insufficient integrity as an etch mask, and an intermediate layer of material is required to be patterned and used as a mask layer. A submicrometer-thick layer of silicon dioxide or silicon nitride, for example, commonly serves as this mask layer.

Isotropic silicon etching may be carried out through liquid chemistry. For example, a mixture of hydrofluoric acid and nitric acid may be applied, whereby the silicon is oxidized by the nitric acid and the hydrofluoric acid converts the resulting silica to a soluble silicon fluoride compound. A nearly directionally independent etch rate results that may be influenced by agitation (Fig. 1a). Alternatively, a plasma with a gas such as sulfur hexafluoride may be used to realize dry silicon isotropic etching, and non-plasma-vapor techniques also exist with xenon difluoride vapor.

Many crystalline materials such as quartz and silicon may be etched preferentially along certain crystal orientations. This attribute may be exploited to chemically machine geometries defined by crystal planes. In the case of silicon, the etch rates in the $\langle 100 \rangle$ and $\langle 110 \rangle$ crystal directions can be several hundred times greater than in the $\langle 111 \rangle$ direction. This type of liquid etching may be accomplished with various alkaline compounds such as potassium hydroxide (KOH) and tetra-methyl ammonium hydroxide (TMAH). The results for a $[100]$ oriented silicon substrate are sidewalls angled at 54.7° with respect to the substrate surface, representing the $[111]$ crystal planes (Fig. 1b). A variety of additional orientations are possible with the restriction that crystalline-based anisotropic etching is limited to faceted structures. Nevertheless, anisotropic

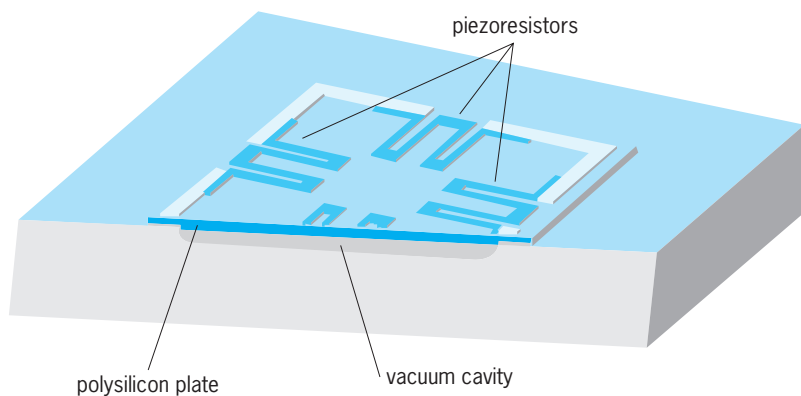


Fig. 3. Surface-micromachined pressure transducer.

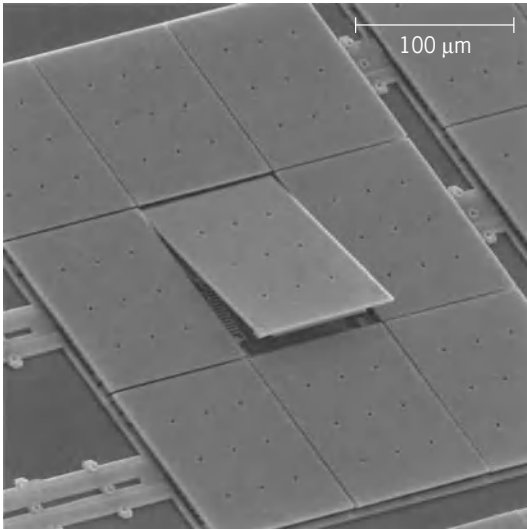


Fig. 4. Surface-micromachined movable mirror array. (E. Garcia, Sandia National Laboratories)

etching in conjunction with microelectronic processing has yielded a tremendous number of useful micromachined devices, including disposable ink-jet print heads, pressure transducers, accelerometers, and microelectrode arrays.

Extending the bulk micromachining tool further, anisotropic silicon plasma etching sequences have been devised to allow an arbitrary planar geometry to be nearly perfectly transferred into vertically defined prismatic cuts (Fig. 1c). This capability yields great flexibility in defining mechanical elements, with the drawback of a substantially increased tool cost over liquid and vapor-phase chemical etching. An extension of bulk micromachining to yield further structure flexibility is afforded through wafer-to-wafer bonding techniques. Two or more wafers containing etched geometry are aligned and bonded using direct fusion bonding, anodic bonding (with the assistance of a high voltage), or an intermediate adhesive layer. These techniques have resulted in new structures, including precisely defined cavities for pressure transducers and microchannels for microfluidic handling, as well as a batch approach to packaging, a particularly troublesome and expensive part of interfacing mechanical behavior with electronics. See ELECTRONIC PACKAGING.

Surface micromachining. An alternative processing approach to bulk microfabrication was driven by the desire to reduce the fraction of the substrate area that had to be devoted to the mechanical components, thereby allowing a larger number of device dies per wafer. The approach, termed surface micromachining (SMM), realizes mechanical structures by depositing and patterning mechanical material layers in conjunction with sacrificial spacer material layers (Fig. 2). In contrast to bulk micromachining, where a substrate may be between 250 and 750 μm thick, these deposited films, realized through deposition processes such as sputtering, evaporation, and chemical vapor deposition (CVD), have thicknesses of the order of 1 μm (micrometer). See CRYSTAL GROWTH;

SEMICONDUCTOR HETEROSTRUCTURES; SPUTTERING.

A substrate material, such as silicon, is used as the basis material for these depositions. The first step in the basic process is to deposit or grow a material to be used as a sacrificial material, such as a silicon dioxide layer (Fig. 2a). The sacrificial layer is patterned to open anchor regions for the structural layer (Fig. 2b) that is subsequently deposited (Fig. 2c). Polysilicon is commonly used for structural layers in surface micromachining, along with other materials that include silicon nitride, silicon carbide, and diamondlike carbon. The structural layer is patterned (Fig. 2d), and the sequence of sacrificial and structural depositions and patterning may then be repeated to enable multilayer geometry. The consequence of this repeated process, however, is an increase in topological variation that ultimately poses problems for patterning. Thus, planarization steps may be inserted to maintain a reasonably flat surface on which to continue to perform microlithography. Ultimately, the last step in silicon micromachining is a release etch of the sacrificial material to render the mechanical

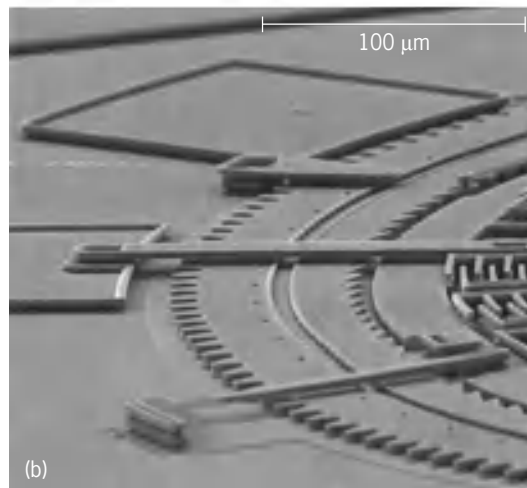
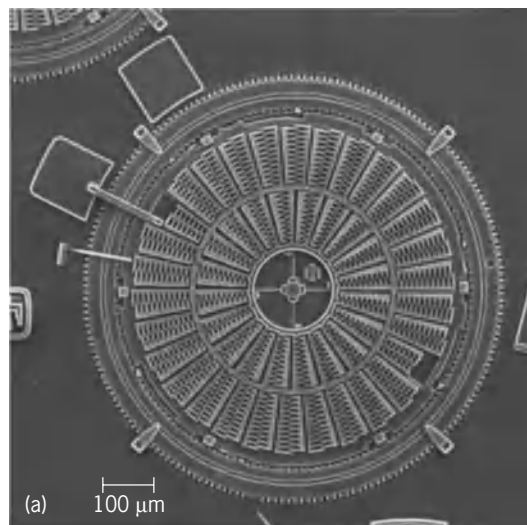


Fig. 5. Torsional ratcheting actuator fabricated by surface micromachining. (a) Overview. (b) Close-up. (J. Jakubczala, Sandia National Laboratories)

structure. In Fig. 2, this structure is a clamped-clamped beam that may be used, for example, in a resonant strain gage (Fig. 2e).

Design. The successful use of any microfabrication process is predicated on the ability to control the mechanical properties of the deposited layers and implement these properties in appropriate design models. Material properties of deposited materials not only differ substantially from their bulk counterparts but also vary depending on deposition conditions. Diagnostic structures must therefore be prepared to measure mechanical properties such as yield strength, bulk modulus, and internal strain. Mechanical test microstructures that may be microfabricated with the same process as the actual device of interest have been developed and are used to measure these properties “on chip.” The MEMS designer is then faced with the task of identifying appropriate models that identify all physical behavior occurring in the microsystem. This task can be daunting due to the interconnectivity between the many different energy transfer mechanisms, but may be reduced by computer-aided design tools that are being tailored for MEMS design problems. See COMPUTER-AIDED DESIGN AND MANUFACTURING.

The path to consolidating micromachining sequences with microelectronics and components from other fields such as photonics is termed process integration. Careful ordering of the process sequence is required to mediate the effects of high temperatures and different chemistries present in disparate and sometimes incompatible microelectronics and micromechanical processes.

Applications. A highly successful device that is fabricated with both bulk and surface micromachining

is the integrated pressure transducer. One surface-micromachined pressure transducer, for example, can measure absolute pressure ranges as high as 70 megapascals (10,000 lb/in.²) and as low as a few pounds per square inch (Fig. 3). The process sequence uses surface micromachining techniques to form a polysilicon-plate-covered cavity that is initially open. The open cavity is then reactively vacuum-sealed with a silicon dioxide, silicon nitride, or polysilicon deposition. After the cavity is closed off, the gas remaining in the cavity continues to react, depositing a solid and thereby self-pumping a vacuum in the cavity. The resulting vacuum reference enables measurement of absolute pressures. Plate deflection is monitored with carefully located polysilicon piezoresistors. Application areas include air pressure sensing in automobile engines, environmental monitoring, and blood pressure sensing. Similar processing has resulted in the integration of surface-micromachined polysilicon inertial reference proof masses with microelectronic processing, yielding single-chip force-feedback accelerometers capable of measurement ranges from a few *g* to hundreds of *g*, where *g* is the acceleration of gravity. Bulk-micromachined single-crystal silicon versions exist that are able to measure several hundred thousand *g*, attesting to the robustness of microsensors. See ACCELEROMETER.

The use of surface micromachining technology to implement microactuators has resulted in steerable micromirror arrays with as many as 1024×768 pixels on a chip. These arrays have revolutionized digital display technology. One device in this category is constructed from four levels of polysilicon and contains $100\text{-}\mu\text{m}$ -square mirrors, separated by $1\ \mu\text{m}$, that may be tilted up to 10° (Fig. 4). Further electrostatic microactuator designs are possible and may be extremely intricate, such as a torsional ratcheting actuator fabricated with five polysilicon levels (Fig. 5). These types of devices are suited for a variety of micropositioning applications. Processing based on deep-x-ray lithography has been used to produce precision magnetic microactuators. One such microactuator directly switches a single-mode optical fiber in a 1×2 switch configuration (Fig. 6). A prime obstacle in manufacturing integrated devices of this sort is that the packaging required to accommodate the interface of the mechanical world with a centralized information network is unique and requires novel design and testing approaches to be economically feasible.

Todd R. Christenson

Bibliography. G. T. A. Kovacs, *Micromachined Transducers Sourcebook*, McGraw-Hill, Boston, 1998; M. Madou, *Fundamentals of Microfabrication*, CRC Press, Boca Raton, FL, 1997; N. Maluf, *An Introduction to Microelectromechanical Systems Engineering*, Artech House, 1999; P. Rai-Choudhury (ed.), *Handbook of Microlithography, Micromachining, and Microfabrication*, SPIE Optical Engineering Press, Bellingham, WA, 1997; W. Trimmer, *Micromechanics and MEMS: Classic and Seminal Papers to 1990*, IEEE Press, 1997.

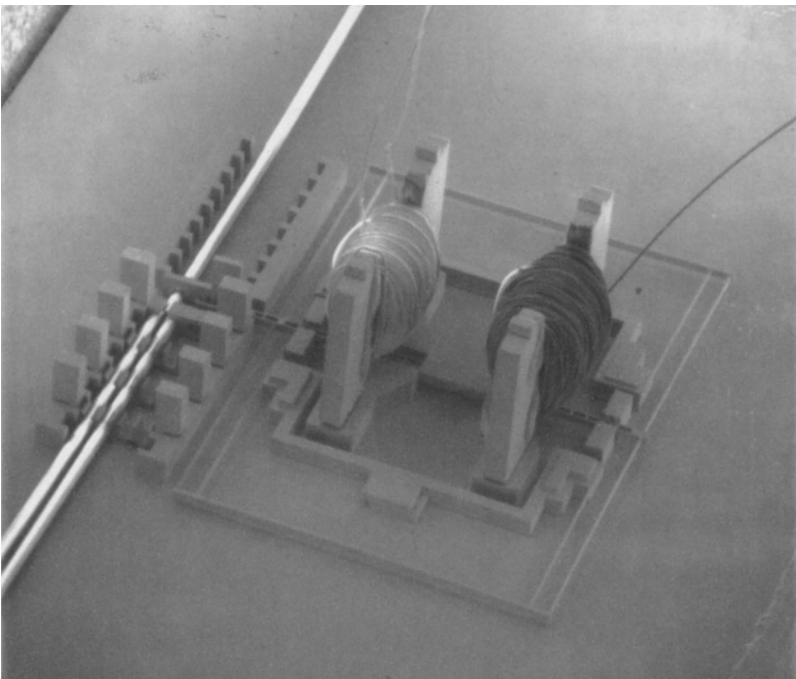


Fig. 6. Magnetic 1×2 optical fiber switch fabricated by deep-x-ray lithography. Total device size is approximately $4\ \text{mm} \times 4\ \text{mm}$. (Henry Guckel, University of Wisconsin)

Microfluidics

Technology that involves manipulating fluids in structures in which at least one linear dimension is less than a millimeter. Currently, the smallest dimension of common microfluidic elements is 10–500 micrometers. The most mature and commercially successful microfluidic systems are the print heads found in common inkjet printers. More recent efforts have been aimed at developing microfluidic systems for performing chemical pro-

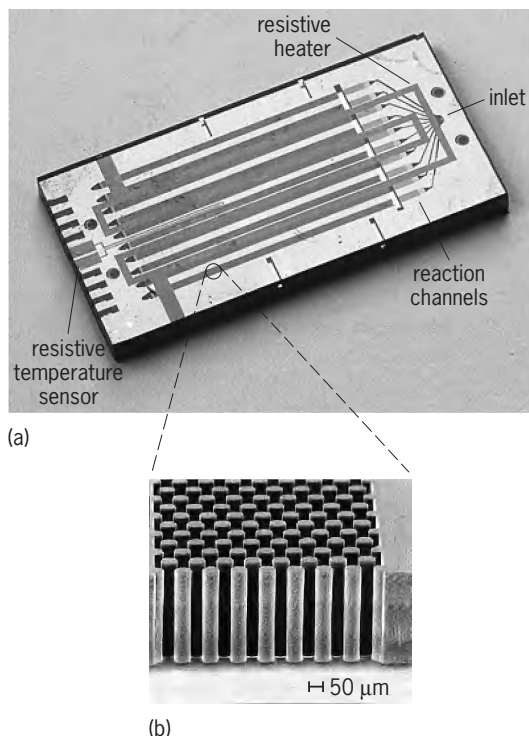


Fig. 1. Microfluidic chip for performing multiphase (liquid/gas) reactions with heterogeneous catalysis. (a) Photograph of chip. This system was used for the hydrogenation of alkenes to form alkanes. The channel structures are in silicon. The heater and temperature sensor are thin metal films (*photo by Felice Frankel*). (b) Scanning electron micrograph showing some of the micropillars that are fabricated inside each reaction channel. These pillars provide a large surface area on which to present catalytic materials in the reactors (*from M. W. Losey et al., Design and fabrication of microfluidic devices for multiphase mixing and reaction, J. Microelectr. Sys., 11(6):709–717, 2002; © 2002 IEEE*)

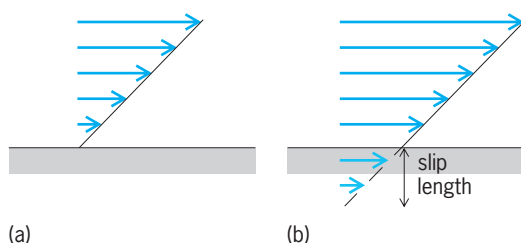


Fig. 2. Flow profiles at a solid boundary (a) with no slip and (b) with slip.

cesses. The principal target applications of these chemical systems are as portable analyzers of chemicals and biomolecules [proteins and deoxyribonucleic acid (DNA)] for biomedical applications and for defense against chemical and biological weapons. Some chemical syntheses have also been performed in microfluidic devices (**Fig. 1**). Microfluidic technologies have only recently begun to be commercialized. The **table** lists applications of microfluidic systems.

Flow characteristics. Experimental and theoretical attention has been paid to the possibility that fluid behavior in microchannels deviates from that found in macroscopic flows. Except in unusual cases, such deviations are neither predicted nor observed; for flows of simple liquids, such as water and organic solvents, the same general equations (Navier-Stokes equations) that govern macroscopic flows, with the no-slip condition at solid boundaries (meaning that the fluid is stationary relative to the solid at the boundary; **Fig. 2a**), are valid on scales above tens of nanometers. One notable exception is the observation of slip of liquids over nonwetting surfaces, that is, surfaces on which the liquid beads up as water does on Teflon[®] (polytetrafluoroethylene). A proposed mechanism for this slip is that a layer of nanoscopic vapor bubbles covers the surface and acts as a lubricating layer of low viscosity between the liquid and the solid. The largest slip lengths that have been observed are about 1–2 μm (**Fig. 2b**). Thus, slip will lead to deviations from expected behavior only in flows of nonwetting liquids through channels with cross-sectional dimensions of less than 10 μm . In flows of gases (less commonly used than liquids), slip has long been predicted and has been observed in microchannels when the mean free path of molecules in the gas is comparable to the

Applications of microfluidic systems

| Application | Advantage of microscopic scale |
|--|---|
| Inkjet printing | High resolution, high operation speed |
| Analytical chemistry | Portability, high operation speed, small sample volume (<1 μL) |
| Protein and DNA analysis for proteomics and genomics | |
| High-throughput screening of chemicals (pharmaceuticals) | |
| Portable chemical detection | |
| Chemical synthesis | Small sample volume, high operation speed, fine control of reaction parameters, rapid transfer of heat and mass |
| In-situ synthesis of toxic compounds | |
| Combinatorial chemistry | |
| Analysis of chemical kinetics | |
| Fuel cell | Portability |
| Cooling for microelectronics | Portability, rapid heat transfer |
| Cell culture/cell sorting | Control of individual cells |

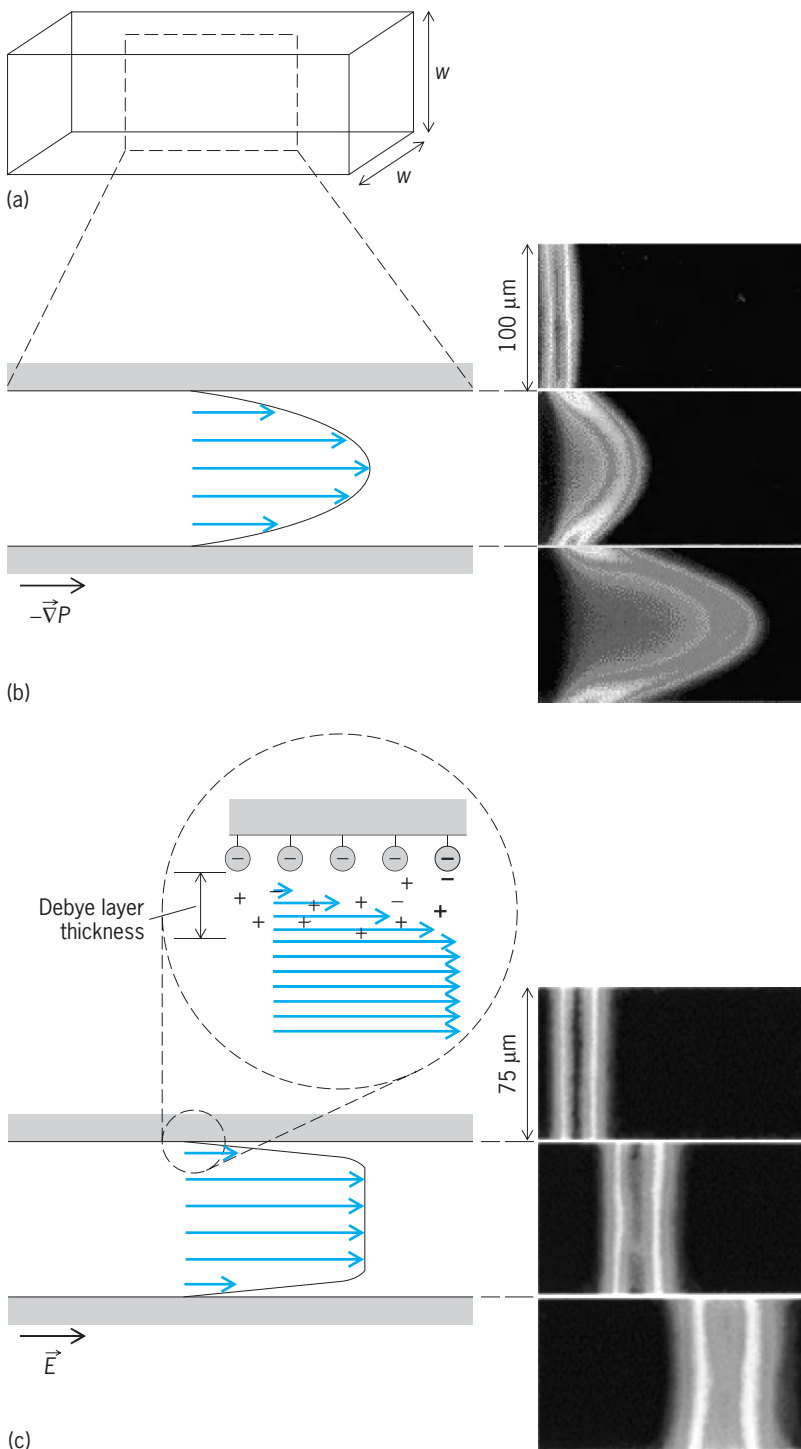


Fig. 3. Flows in microchannels. (a) Diagram of a section of a microchannel with a square cross section. (b) Flow profile of a pressure-driven flow and (c) an electroosmotic flow in a channel. The two series of fluorescent micrographs show the evolution of a band of fluorescent dye in the respective flows. The blurring of the interface between fluorescent and nonfluorescent regions is due to molecular diffusion. (After P. H. Paul, M. G. Garguilo, and D. J. Rakenstraw, *Imaging of pressure and electrokinetically driven flows through open capillaries*, *Analyt. Chem.*, 70:2459–2467, 1998)

dimension of the channel. See BOUNDARY-LAYER FLOW; FLUID-FLOW PRINCIPLES; NAVIER-STOKES EQUATION.

While the basic governing equations are the same, there are several general features that distinguish flows in microstructures from flows in common macroscopic systems (such as water faucets and coffee

cups). As the dimension of a flow decreases, the importance of forces that act on the volume of the fluid, such as inertia and gravity, diminishes relative to that of forces that act at the surfaces, such as viscous friction and surface tension. The ratio of inertial to viscous forces is expressed as the Reynolds number of the flow, $Re = \rho v w / \eta$, where ρ (measured in kg/m^3) is the density of the fluid, v (m/s) is the velocity, w (m) is the characteristic dimension of the flow (\sim volume/surface area of container; Fig. 3a), and η [$\text{kg}/(\text{m} \cdot \text{s})$] is the dynamic viscosity of the fluid. In microchannels, $Re < 100$, so flows are laminar rather than turbulent (flows in channels are typically laminar for $Re < 2000$). Coflowing streams in a laminar flow intermix only by diffusion because no spontaneous eddies carry mass and momentum between them. This characteristic of laminar flows allows spatial control of solute within the flow, but it hinders rapid mixing. See LAMINAR FLOW; REYNOLDS NUMBER; TURBULENT FLOW.

Fabrication. Most methods of fabricating microfluidic systems are adaptations of techniques for fabricating microelectronic structures; photolithography is at the core of these methods. These techniques lead to a flat, chiplike format in which microchannels are constrained in planar layers on a flat substrate (Fig. 1). This format facilitates integration of electronics and optics, but it restricts the geometries that are accessible for the design of fluidic elements. The channel structures are etched into a flat surface of silicon or glass, or molded into a soft or hard plastic sheet; the channels are closed by sealing the structured material to another flat surface. See INTEGRATED CIRCUITS; MICRO-ELECTRO-MECHANICAL SYSTEMS (MEMS); MICROLITHOGRAPHY.

Flow characterization. Simple characterization of flows in microfluidic devices can be achieved by measuring global variables such as the applied pressure difference (Pa) across a channel and the volumetric flow rate (m^3/s) through the channel. Electronic sensors of pressure and flow speed have been integrated into microchannels to provide local measurements of the flow. Fluorescence microscopy is commonly used to perform characterization of these flows with micrometer-scale resolution. Micro particle image velocimetry (μPIV) maps the streamlines in a flow by following the trajectories of submicrometer fluorescent tracer beads in sequences of micrographs. The evolution of streams in microflows can be followed through a microscope by coinjecting streams of fluorescent solution with streams of clear solution. In steady flows, the three-dimensional evolution of fluorescent streams can be imaged through a confocal microscope (Fig. 4). See CONFOCAL MICROSCOPY; FLUORESCENCE MICROSCOPY.

Pumping fluids. The two most common methods for driving flows in microchannels are with externally applied pressure gradients and with externally applied electric fields. Capillary, acoustic, and magnetic forces have also been used to move fluids in microchannels.

Pressure gradients are often generated by macroscopic pumps that are connected via tubes to the

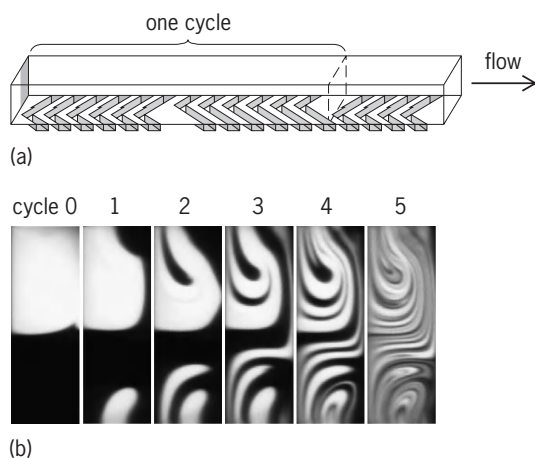


Fig. 4. Passive micromixer. (a) Section of a microchannel with grooves in the form of asymmetric herringbones on the bottom wall; the grooves lead to mixing in pressure-driven flows through the channel. One mixing cycle is made up of two regions of asymmetric herringbones; the direction of the asymmetry switches from one region to the next. (b) Confocal fluorescence micrographs of the vertical cross sections of flow through such a channel. The evolution of two coflowing streams (one of fluorescent liquid and one of clear liquid) in the mixer is seen: The number of folds approximately doubles with each cycle. The blurring of the interface between fluorescent and nonfluorescent regions is due to molecular diffusion. (After A. D. Stroock et al., *Chaotic mixer for microchannels*, *Science*, 295:647–651, 2002)

inlets of the microfluidic chip. Integrated microdiaphragm and microperistaltic pumps have also been developed. In pressure-driven flow in channels, the flow speed is given roughly by (1), where ∇P is the

$$v \sim -w^2 \nabla P / \eta \quad (\text{m/s}) \quad (1)$$

applied pressure gradient (Pa/m). To maintain a given flow speed, the required pressure gradient grows rapidly as the channel shrinks. In pressure-driven flows, the flow speed varies from zero at the wall to its maximum value in the middle of the channel (Fig. 3b). This variation leads to dispersion (spreading) of solute along the direction of the flow. This dispersion is unfavorable for transporting narrow bands of solute such as in a chemical separation.

Electric fields are applied by placing electrodes in the inlet and outlet of a microchannel. In electrolyte solutions (such as salty water), flows are generated by the interaction of the applied electric field with ions that accumulate in a thin (less than $1 \mu\text{m}$ thick) layer of fluid, the Debye layer, adjacent to channel walls; these are called electroosmotic flows (Fig. 3c). Fluid in the Debye layer moves due to the electrical body force and entrains the remainder of the fluid. The flow speed in the bulk is given roughly by (2), where μ_{co} ($\text{m}^2/\text{s} \times \text{V}$) is the electroosmotic

$$v_{\text{co}} \sim \mu_{\text{co}} E \quad (2)$$

mobility that depends on the surface charge density, the concentration of ions in the liquid, and the viscosity of the liquid, and E (V/m) is the magnitude of the electric field. The magnitude of the electroos-

mot mobility is given roughly by (3). In uniformly

$$|\mu_{\text{co}}| \sim 1 \mu\text{m}/(\text{s} \cdot \text{V}/\text{m}) \quad (3)$$

charged channels, the flow speed v_{co} is constant over the cross section except in the Debye layer. This property allows bands of a single type of solute to be transported along the channel with little dispersion. Capillary electrophoresis is a useful tool for chemical analysis that exploits this behavior and is performed in microfluidic devices. The flow speed v_{co} is also independent of the dimension of the channel down to the thickness of the Debye layer. Electroosmotic pumping is therefore appropriate in channels down to submicrometer dimensions. Disadvantages of electroosmotic pumping include the requirement of high voltages ($V \sim 1 \text{ kV}$), sensitivity to the chemical characteristics of the walls, and the tendency for species to separate electrophoretically in the applied field. (This tendency is an advantage for analysis but a disadvantage for transporting general chemical mixtures.) See ELECTROPHORESIS.

Mixing. In microchannels of common dimensions ($500 \mu\text{m} > w > 10 \mu\text{m}$), the motion of solute by diffusion across the laminar flow is often slow relative to the flow speed along the channel. In this situation, mixing is often the slow step in a chemical process, unless transverse flows that stir the fluid are purposely induced. (In turbulent flow, these transverse flows occur spontaneously.) An efficient stirring flow stretches and folds the fluid such that the interface between unmixed regions grows exponentially in time; this type of flow is called chaotic (Fig. 4). Active mixers create transverse motion in the principal flow with local, oscillatory forces generated with bubbles, applied electric or magnetic fields, or flows in cross-channels. These mixers have the potential to be very efficient but require sophisticated controls. Passive mixers use fixed geometrical features in the channel to induce transverse components in the principal flow (Fig. 4). These designs are simple to operate but may not achieve full mixing as quickly as active designs.

Challenges. With the exception of inkjet print heads and, perhaps, on-chip capillary electrophoresis, microfluidic technology is still in its infancy. No consensus has been reached on even the most basic procedures and designs. Optimal designs for pumps, valves, injectors, and so forth must still be invented and characterized, and ideal materials and methods of microfabrication must be developed. Future applications of microfluidic devices must also be conceived of and explored. For example, microfluidic systems could act as implants to deliver drugs and monitor physiological parameters, or as active materials that control environmental parameters in clothing and buildings.

Nanofluidic systems. These systems, in which the characteristic size is less than a micrometer, are already being developed. In this regime, pressure may not be a useful driving force, whereas electroosmosis likely will. Diffusion will often transport solute molecules as rapidly as the flow does; mix-

ing will not be a problem, but careful delivery of solute will. Whereas the design principles for microfluidics were largely borrowed from macroscopic chemical systems (pumps, channels, and so on), the best guides to designing nanofluidic systems may be biological systems such as cells: transport will be controlled on the molecular rather than the hydrodynamic level, with molecular motors carrying chemical building blocks along definite paths between distinct chemical environments. See FLUID MECHANICS.

Abraham D. Stroock

Bibliography. C.-M. Ho and Y.-C. Tai, Micro-electro-mechanical-systems and fluid flows, *Annu. Rev. Fluid Mech.*, 30:579–612, 1998; M. W. Losey et al., Design and fabrication of microfluidic devices for multiphase mixing and reaction, *J. Microelectr. Sys.*, 11(6):709–717, 2002; P. H. Paul, M. G. Garguilo, and D. J. Rakestraw, Imaging of pressure and electrokinetically driven flows through open capillaries, *Analyt. Chem.*, 70:2459–2467, 1998; A. D. Stroock et al., Chaotic mixer for microchannels, *Science*, 295:647–651, 2002; D. C. Tretheway and C. D. Meinhart, Apparent fluid slip at hydrophobic channel walls, *Phys. Fluids*, 14:L9–L12, 2002; G. M. Whitesides and A. D. Stroock, Flexible methods for microfluidics, *Phys. Today*, 54(6):42–48, June 2001.

Microlithography

The formation of small three-dimensional relief images on the surface of a substrate for subsequent transfer of this pattern into the substrate itself, as used in such applications as semiconductor fabrication. The fabrication of an integrated circuit (IC) requires a variety of physical and chemical processes performed on a semiconductor (for example, silicon) substrate. In general, the various processes used to make an IC fall into three categories: film deposition, patterning, and semiconductor doping. Films of both conductors (such as polysilicon, aluminum, and more recently copper) and insulators (various forms of silicon dioxide, silicon nitride, and others) are used to connect and isolate transistors and their components. Selective doping of various regions of silicon allows the conductivity of the silicon to be changed with the application of voltage. By creating structures of these various components, millions of transistors can be built and wired together to form the complex circuitry of a modern microelectronic device. Fundamental to all of these processes is lithography. See INTEGRATED CIRCUITS; SEMICONDUCTOR.

The word lithography, from Greek, literally means writing on stones. In the case of semiconductor lithography, our stones are silicon wafers and we are writing our patterns in a light-sensitive polymer called a photoresist. To build the complex structures that make up a transistor and the many wires that connect the millions of transistors of a circuit, lithography and etch pattern transfer steps are repeated at least 10 times, but more typically are done 20–30 times in order to make one cir-

cuit. Each pattern being printed on the wafer is aligned to the previously formed patterns, and slowly the conductors, insulators, and selectively doped regions are built up to form the final device. See POLYMER; TRANSISTOR.

Optical microlithography is a photographic process by which a photoresist is exposed and developed to form three-dimensional images on the substrate. The general sequence of processing steps for a typical optical lithography process is as follows: substrate preparation, photoresist spin coat, prebake, exposure, postexposure bake, development, and postbake. The resist is removed in the final operation of the lithographic process, after the pattern has been transferred into the underlying layer. This sequence is shown in Fig. 1, and is generally performed on several tools linked together into a contiguous unit called a lithographic cluster.

Substrate preparation. Substrate preparation is intended to improve the adhesion of the photoresist material to the substrate. This is accomplished by one or more of the following processes: substrate cleaning to remove contamination, dehydration bake to remove water, and addition of an adhesion promoter. One common type of contaminant, adsorbed water, is removed most readily by a high-temperature process called a dehydration bake. A typical dehydration bake does not completely remove water from the surface of silica substrates (including silicon, polysilicon, silicon oxide, and silicon nitride). Surface silicon atoms bond strongly with a monolayer of water forming silanol groups (SiOH). Bake temperatures in excess of 600°C (1112°F) are required to remove this final layer of water. Since this approach is impractical, the preferred method of removing this silanol is by chemical means. Adhesion promoters

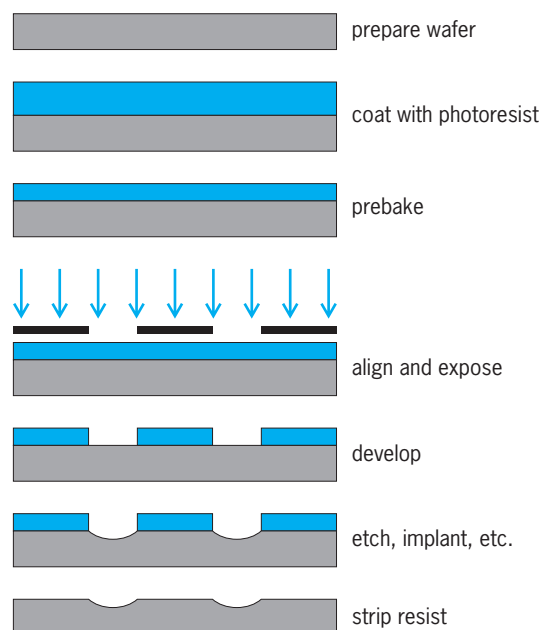


Fig. 1. Example of a typical sequence of lithographic processing steps (with no postexposure bake in this case), illustrated for a positive resist.

are used to react chemically with surface silanol and replace the —OH group with an organic functional group which, unlike the hydroxyl group, offers good adhesion to the photoresist. Silanes are often used for this purpose, the most common being hexamethyldisilazane (HMDS).

Photoresist coating. A thin, uniform coating of photoresist at a specific, well-controlled thickness is accomplished by the seemingly simple process of spin coating. The photoresist, rendered into a liquid by dissolving the solid components in a solvent, is poured onto the wafer and then spun on a turntable at a high speed to produce the desired film. Stringent requirements for thickness control and uniformity and low defect density call for particular attention to be paid to this process, where a large number of parameters can have significant impact on these properties. There is the choice between static dispense (wafer stationary while resist is dispensed) or dynamic dispense (wafer spinning while resist is dispensed), spin speeds and times, and accelerations to each of the spin speeds. Also, the volume of the resist dispensed and properties of the resist (such as viscosity, percent solids, and solvent composition) and the substrate (substrate material and topography) play an important role in the resist thickness uniformity. At the end of this cycle, a solvent-rich film of photoresist covers the wafer, which is ready for the postapply bake. Although theory exists to describe the spin coating process rheologically, in practical terms the variation of photoresist thickness and uniformity with the process parameters must be determined experimentally. A photoresist spin-speed curve is an essential tool for setting the spin speed to obtain the desired resist thickness. The final resist thickness varies as one over the square root of the spin speed and is roughly proportional to the liquid photoresist viscosity.

Postapply bake. After coating, the resulting resist film will contain 20–40% by weight solvent. The postapply bake process, also known as a softbake or a prebake, involves drying the photoresist by removing this excess solvent. There are four major effects of removing the solvent from a photoresist film: (1) film thickness is reduced, (2) postexposure bake and development properties are changed, (3) adhesion is improved, and (4) the film becomes less tacky and thus less susceptible to particulate contamination. Typical prebake processes leave between 3 and 8% residual solvent in the resist film, a sufficiently low amount to keep the film stable during subsequent lithographic processing.

Unfortunately, there are other consequences of baking most photoresists. At temperatures greater than about 70°C (160°F), the photosensitive component of a typical resist mixture, called the photoactive compound, may begin to decompose. Thus, one must search for the optimum prebake conditions which will maximize the benefits of the solvent evaporation and minimize the detriments of resist decomposition. For photoresists known as chemically amplified resists, residual solvent can significantly influence diffusion and reaction properties during

the postexposure bake, necessitating careful control over the postapply bake process. Fortunately, these resists do not suffer from significant decomposition of the light-sensitive components during prebake.

Although the use of convection ovens for prebaking the photoresist was once quite common, currently the most popular bake method is the hot plate. The wafer is brought into either intimate vacuum contact with or close proximity to a hot, high-mass metal plate. Due to the high thermal conductivity of silicon, the photoresist is heated to near-hot-plate temperature quickly (in about 5 s for hard contact or in about 20 s for proximity baking). The greatest advantage of this method is an order-of-magnitude decrease in the required bake time (to about 1 min) over convection ovens and the improved uniformity of the bake. In general, proximity baking is preferred to reduce the possibility of particle generation caused by contact with the back side of the wafer.

Alignment and exposure. The basic principle behind the operation of a photoresist is a change in solubility of the resist in developer upon exposure to light (or other type of exposing radiation). For the case of the standard diazonaphthoquinone positive photoresist, the photoactive compound, which is not soluble in the aqueous base developer, is converted to a carboxylic acid upon exposure to ultraviolet (UV) light in the range 350–450 nanometers. The carboxylic acid product is very soluble in the basic developer. Thus, a spatial variation in light energy incident on the photoresist will cause a spatial variation in solubility of the resist in the developer.

Contact and proximity lithography are the simplest methods of exposing a photoresist through a master pattern called a photomask (**Fig. 2**). Contact lithography offers high resolution (down to about the wavelength of the radiation), but practical problems such as mask damage and resulting low yield make this process unusable in most production environments. Proximity printing reduces mask damage by keeping the mask a set distance (for example, 20 micrometers) above the wafer. Unfortunately, the resolution limit is increased to greater than 2–4 μm , making proximity printing insufficient for today's technology. By far the most common method of exposure is projection printing.

Projection lithography derives its name from the fact that an image of the mask is projected onto the wafer. Projection lithography became a viable alternative to contact/proximity printing in the

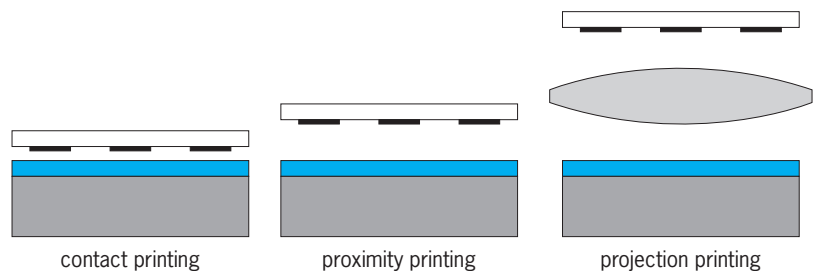


Fig. 2. Lithographic printing in semiconductor manufacturing has evolved from contact printing (in the early 1960s) to projection printing (from the mid 1970s to today).

mid-1970s, when the advent of computer-aided lens design and improved optical materials allowed the production of lens elements of sufficient quality to meet the requirements of the semiconductor industry. These lenses have become so perfect that lens defects, called aberrations, play only a small role in determining the quality of the image. Such an optical system is said to be diffraction-limited, since for the most part diffraction effects and not lens aberrations determine the shape of the image. See ABERRATION (OPTICS); DIFFRACTION; LENS (OPTICS); OPTICAL MATERIALS.

There are two major classes of projection lithography tools: scanning and step-and-repeat systems. Scanning projection printing employs reflective optics (that is, mirrors rather than lenses) to project a slit of light from the mask onto the wafer as the mask and wafer are moved simultaneously by the slit. The exposure dose is determined by the intensity of the light, the slit width, and the speed at which the wafer is scanned. These early scanning systems, which use polychromatic light from a mercury arc lamp, are 1:1 (that is, the mask and image sizes are equal). Step-and-repeat cameras (steppers for short) expose the wafer one rectangular section (called the image field) at a time and can be 1:1 or reduction. These systems employ refractive optics (that is, lenses) and are usually quasimonochromatic. Both types of systems (Fig. 3) are capable of high-resolution imaging, although reduction imaging is required for the highest resolutions. See MIRROR OPTICS.

Scanners replaced proximity printing by the mid-1970s for device geometries below 4–5 μm . By the early 1980s, steppers began to dominate as device designs pushed to 2 μm and below. Steppers continued to dominate lithographic patterning throughout the 1990s as minimum feature sizes reached the 250-nm levels. However, by the early 1990s a hybrid step-and-scan approach was introduced. The step-and-scan approach uses a fraction of a normal stepper field (for example, 25 \times 8 mm), then scans this field in one

direction to expose the entire 4 \times reduction mask. The wafer is then stepped to a new location and the scan is repeated. The smaller imaging field simplifies the design and manufacture of the lens, but at the expense of a more complicated reticle and wafer stage. Step-and-scan technology is the technology of choice today for below-250-nm manufacturing.

Resolution, the smallest feature that can be printed with adequate control, has two basic limits: the smallest image that can be projected onto the wafer, and the resolving capability of the photoresist to make use of that image. For the projected imaging, resolution is determined by the wavelength of the imaging light (λ) and the numerical aperture (NA) of the projection lens according to the Rayleigh criterion below. Lithography systems have progressed from

$$R \propto \frac{\lambda}{\text{NA}}$$

blue wavelengths (436 nm) to UV (365 nm) to deep-UV (248 nm) to today's mainstream high-resolution wavelength of 193 nm. In the meantime, projection tool numerical apertures have risen from 0.16 for the first scanners to amazingly high 0.85 NA systems today, producing features well under 100 nm in size. See OPTICAL IMAGE; RESOLVING POWER (OPTICS).

Before the exposure of the photoresist with an image of the mask can begin, this image must be aligned with the previously defined patterns on the wafer. This alignment and the resulting overlay of the two or more lithographic patterns are critical since tighter overlay control means circuit features can be packed closer together. In the drive toward more functionality per chip, closer packing of devices through better alignment and overlay is nearly as critical as smaller devices through higher resolution.

Another important aspect of photoresist exposure is the standing-wave effect. Monochromatic light, when projected onto a wafer, strikes the photoresist surface over a range of angles, approximating plane waves. This light travels down through the photoresist and, if the substrate is reflective, is reflected back up through the resist. The incoming and reflected light interfere to form a standing-wave pattern of high and low light intensity at different depths in the photoresist. This pattern is replicated in the photoresist, causing ridges in the sidewalls of the resist feature (Fig. 4). As pattern dimensions become smaller, these ridges can significantly affect the quality of the feature. The interference that causes standing waves also results in a phenomenon called swing curves, the sinusoidal variation in linewidth with changing resist thickness. These detrimental effects are best cured by coating the substrate with a thin absorbing layer, called a bottom antireflective coating (BARC), that can reduce the reflectivity seen by the photoresist to less than 1%.

Postexposure bake. One method of reducing the standing-wave effect is called the postexposure bake (PEB). Although there is still some debate as to the mechanism, it is believed that the high temperatures used (100–130°C or 212–266°F) cause diffusion of

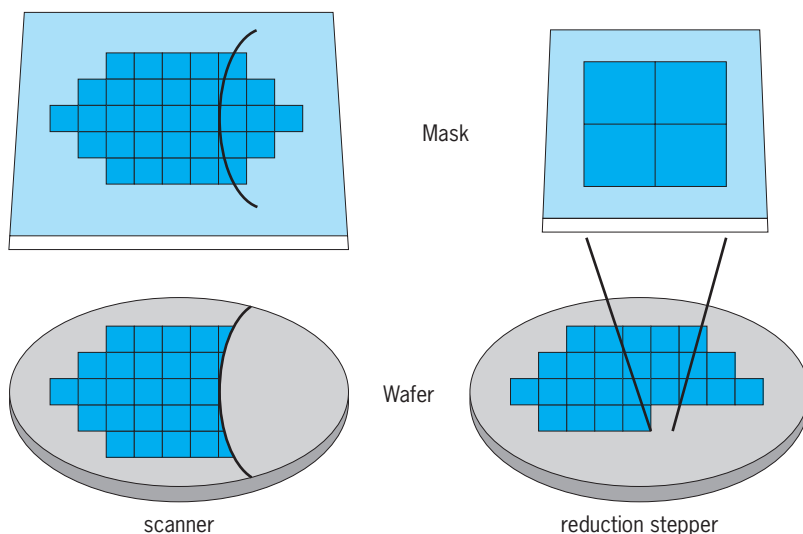


Fig. 3. Scanners and steppers use different techniques for exposing a large wafer with a small image field.

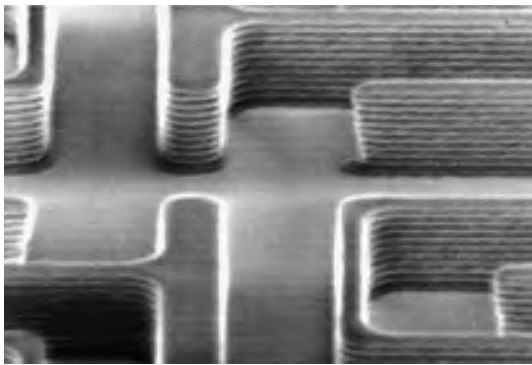


Fig. 4. Photoresist pattern on a silicon substrate showing prominent standing waves.

the exposed photosensitive material, thus smoothing out the standing-wave ridges. It has also been observed that the rate of diffusion is dependent on the prebake conditions since the presence of solvent enhances diffusion during a PEB; that is, a low-temperature prebake results in greater diffusion for a given PEB temperature. For a conventional resist, the main importance of the PEB is diffusion to remove standing waves. For a type of photoresist, called chemically amplified resists, the PEB is an essential part of the chemical reactions that create a solubility differential between exposed and unexposed parts of the resist. For these resists, exposure generates a small amount of a strong acid which does not itself change the solubility of the resist. During the postexposure bake, this photogenerated acid catalyzes a reaction which changes the solubility of the polymer in the resist. Since the photogenerated acid is not consumed in this reaction, it continues to cause more solubility changing events and thus “amplifies” the effects of the exposure. Control of the PEB is extremely critical for chemically amplified resists.

Development. Once exposed, the photoresist must be developed. Most commonly used photoresists use aqueous bases as developers. In particular, tetramethyl ammonium hydroxide (TMAH) is used almost universally at a concentration of 0.26 *N*. Development is undoubtedly one of the most critical steps in the photoresist process. The characteristics of the resist-developer interactions determine to a large extent the shape of the photoresist profile and, more importantly, the control of the sizes of the features being printed.

The method of applying developer to the photoresist is important in controlling the development uniformity and process latitude. Spin development processes, in which developer is poured onto the rotating wafer, use equipment similar to that used for spin coating. The wafer is also rinsed and dried while still spinning. Spray development has been shown to have good results using developers specifically formulated for this method. Using a process identical to spin development, the developer is sprayed on the wafer with a nozzle that produces a fine mist (Fig. 5). This technique reduces the amount of developer used and gives more uniform developer coverage. Another in-line development strategy is called pud-

dle development. Again using developers specifically formulated for this process, the developer is poured onto a stationary wafer which is then allowed to sit motionless for the duration of the development time. The wafer is then spin-rinsed and dried. Note that all these processes can be performed in the same piece of equipment with only minor modifications, and combinations of these techniques are frequently used.

Pattern transfer. After the small patterns have been lithographically printed in photoresist, these patterns must be transferred into the substrate. There are three basic pattern transfer approaches: subtractive transfer (etching), additive transfer (selective deposition), and impurity doping (ion implantation). Etching is the most common pattern transfer approach. A uniform layer of the material to be patterned is deposited on the substrate. Lithography is then performed such that the areas to be etched are left unprotected (uncovered) by the photoresist. Etching is performed using wet chemicals (such as acids) or more commonly in a dry plasma environment. The photoresist “resists” the etching and protects the material covered by the resist. When the etching is complete, the resist is stripped, leaving the desired pattern etched into the deposited layer. Additive processes are used whenever workable etching processes are not available, for example, for copper interconnects. Here, the lithographic pattern is used to open areas where the new layer is to be grown (by electroplating, in the case of copper). Stripping of the resist then leaves the new material in a negative version of the patterned photoresist. Doping involves the addition of controlled amounts of contaminants that change the conductive properties of a semiconductor. Ion implantation uses a beam of dopant ions accelerated at the photoresist-patterned substrate. The resist blocks the ions, but the areas that are not covered by the resist are embedded with ions, creating the selectively doped regions that make up the electrical heart of the transistors. See ELECTROPLATING OF METALS; ION IMPLANTATION; PLASMA (PHYSICS); TRANSISTOR.

Limits of optical microlithography. To date, optical lithography has been the technology of choice for IC manufacturing. The resolution limit as described

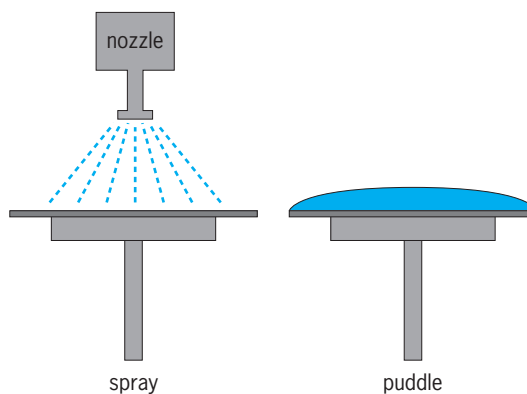


Fig. 5. Different developer application techniques that are commonly used.

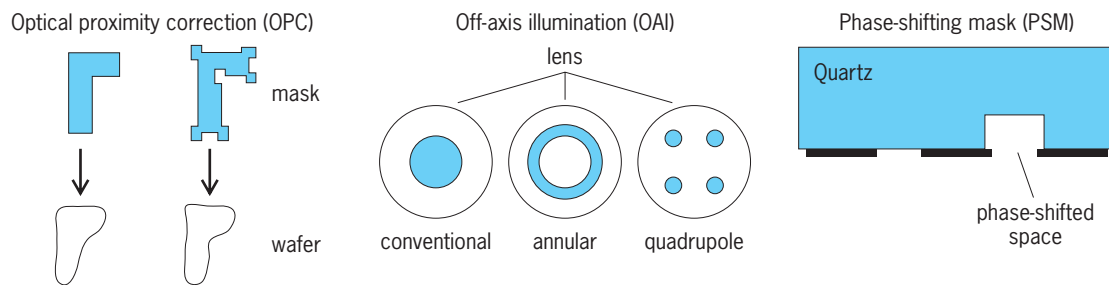


Fig. 6. Examples of several types of resolution enhancement technologies.

earlier in the equation outlines the obvious challenges. Lower wavelengths require expensive, unproven materials (such as superpure fused silica and fluoride salts). Higher numerical apertures result in increasing aberrations, which can only be reduced by more complicated designs and more exacting lens manufacturing processes.

To make the situation more difficult, any improvement in resolution is always accompanied by a decrease in the depth of focus (DOF). According to the Rayleigh criterion, the DOF for small features should decrease as the feature size squared. In reality, empirical results have shown the DOF to decrease as about the feature size to the first power (due to improvements in photoresists and other factors). Today's 100-nm features typically have a DOF of a few tenths of a micrometer. An important requirement of any improvement in the practical resolution of an imaging system is the ability to live within the confines of this reduced DOF. Improvements in wafer and mask flatness, autofocus and autoleveling systems, and wafer planarization by chemical-mechanical polishing of the wafer are examples of how the industry is coping with reduced DOF.

Attempts to simultaneously improve the resolution and DOF by optical means (sometimes called optical "tricks") include optimizing the mask pattern shape (called optical proximity correction, OPC), optimizing the angles of light illuminating the mask (called off-axis illumination, OAI), adding phase information to the mask in addition to intensity information (called phase shifting masks, PSM), and controlling the polarization of the illumination. Collectively, these optical approaches are known as resolution enhancement technologies (RET) [Fig. 6]. OPC predistorts the mask, taking into account the nonlinear nature of the imaging process, to make the final wafer pattern more like the design. OAI improves resolution and depth of focus for small dense features (though often at the expense of larger or more isolated features). PSM adds phase information to the mask, using the property that two beams of light that are 180° out of phase will cancel, creating darkness. Like OAI, PSM can improve resolution and depth of focus, but for a wider range of feature sizes and types. Often several or all of these RET approaches are used together.

The limits of optical lithography may be further pushed with immersion lithography. By replacing

the air between the lens and the substrate with a higher index fluid, it is possible to design and build a lens with a numerical aperture greater than 1.0. In fact, numerical apertures up to 1.3 may be possible. These hyper NAs in conjunction with aggressive resolution enhancement technologies should enable 193-nm lithography to extend to resolutions of 45 nm and below.

Chris A. Mack

Bibliography. R. Dammel, *Diazonaphthoquinone-based Resists*, SPIE Tutorial Texts, vol. TT 11, Bellingham, WA, 1993; P. Rai-Choudhury (ed.), *Handbook of Microlithography, Micromachining, and Microfabrication*, vol. 1: *Microlithography*, pp. 597-680, SPIE Press, Bellingham, WA, 1997; J. R. Sheats and B. W. Smith (eds.), *Microlithography Science and Technology*, pp. 109-170, Marcel Dekker, New York, 1998.

Micromanipulation

The technique or practice involving manipulation of objects too small to be easily seen with the unaided eye. When a microscope is used to allow the operator to visually guide the microtools used in the manipulation, the technique is called micrurgy. When an object that is not extremely small needs to be positioned with extreme precision, as when aiming a laser to make a surgical lesion, the technique is called micropositioning.

To the unaided eye, the hand of a skilled person can appear completely steady and thereby seems able to reliably guide a fine-pointed probe onto the thinnest visible line. However, observed under a microscope, a probe held in even the steadiest of trained hands can be seen to drift about constantly within the width of this line, with a frequency around 8-12 Hz, and the rate of movement is magnified in the same proportion as the distance is magnified. Under high magnification, the hand appears capable of only large, fast movements and cannot guide a probe onto a line that is so thin as to be visible only with a microscope.

To manipulate microscopic objects, it is necessary to use a tool called a micromanipulator, which allows relatively coarse hand movements to execute proportionately slower and smaller movements of a probe or microtool. Micromanipulators are essential to much of modern technology. They are used

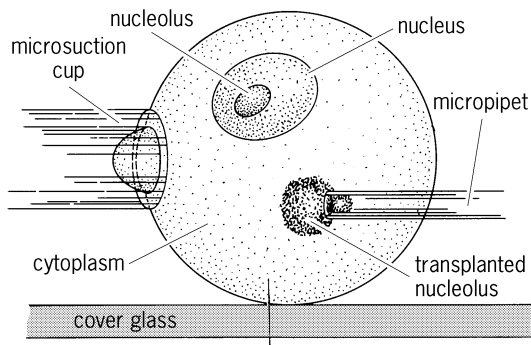


Fig. 1. Diagram of a cell undergoing transplantation of a nucleolus. The microsuction cup is used to hold the cell in place. The micropipet, positioned inside the cell cytoplasm with a micropositioner, holds the nucleolus. A microinjector is used to control both the microsuction cup and the micropipet.

to probe and test integrated circuit chips, to align fiber-optic communication cables, as well as to accomplish techniques in biological research (Fig. 1).

The simplest and most common micromanipulators consist of three orthogonal mechanical slides (Fig. 2), which can be dovetail or ball- or roller-bearing slides and whose position is controlled by a threaded screw or a rack and pinion. Such devices are relatively inexpensive and adequate for magnifications up to 150 power. However, just as the imperfections of the human hand are revealed by magnification, the imperfections of simple machines are revealed by magnifications greater than 150 power.

Imperfections of simple mechanical micromanipulators. A mechanical object that is pushed along a surface tends to move backward slightly when the force used to advance it is removed. This phenomenon, called backlash, poses a major problem in high-magnification micromanipulation. A probe just

inside a cell or one that touches a tiny circuit element may move back out of contact when advance is stopped. To reduce backlash, the relatively coarse micromanipulators usually have one fine drive mounted on top, which is used to make the final advance in the probe direction. These drives have precision-machined and -fitted threaded drives (calibrated in micrometers), as well as a spring to maintain pressure of the slide against the drive. They are adequate to 250-power magnification; for greater precision a more sophisticated micromanipulator is required.

A lag between the beginning of a hand movement and the beginning of a probe movement is called lost motion. Lost motion occurs when a threaded or rack-and-pinion drive is reversed in direction, since there is never a microscopically perfect fit between the nut and the threads, or the rack and pinion gear.

Stiction occurs because the coefficient of static friction is greater than the coefficient of sliding friction. For this reason, greater pressure is required to start a slide moving than to maintain constant motion or even to continue constant acceleration. Stiction results in a tendency to overshoot when a very small initial movement is desired.

A micromanipulator with three orthogonal axes can reach any point in a cubic range of movement, but it must be guided to it by successive approximation on each axis, one knob at a time, which is a slow, painstaking process.

Axial movements of the slide must be accompanied by a certain amount of random lateral deviation since no slide surface is perfectly flat and smooth. Dovetail slides are durable and resistant to damage, but the slide lubrication can attract dirt, which worsens lateral deviation at points on the slide. Ball and roller bearings have relatively less backlash and stiction and tend to force dirt out of their way, but they have less load capacity. Also, the slide track can easily be permanently marred at the point of contact by shock forces such as tipping over on the table. Marring can produce permanent lateral deviation.

Since temperature causes materials to expand or contract, there is inevitably some drift of position of the probe tip with even small temperature change. Drift is negligible in mechanical micromanipulators in a room environment, and when it occurs with other designs it is detrimental only in applications that require long-term holding of a position, as in recording cell intracellular potentials.

Normal movements of the hand that holds the drive knob (the initial reason for employing a micromanipulator) can still create some hand-generated vibration of the probe tip. Such movements can be reduced by adding mass to the micromanipulator, mounting the drive knobs near the base or solid attachment point, or by mechanical isolation of the control mechanism from the probe-carrying mechanism.

Types of micromanipulators. For each of the above imperfections, a micromanipulator has been designed that reduces or virtually eliminates that

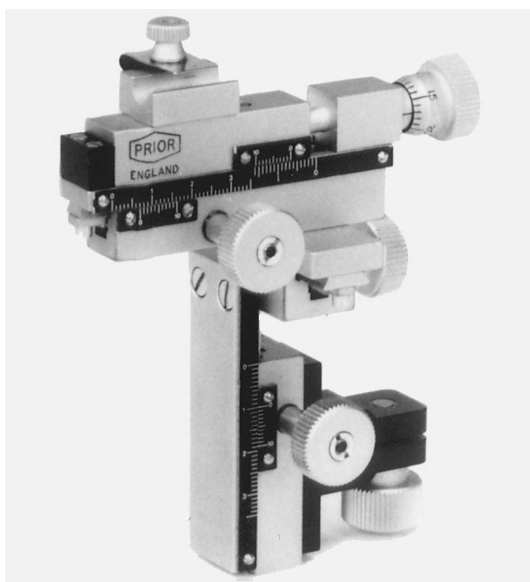


Fig. 2. Simple mechanical micromanipulator, with dovetail slides and rack-and-pinion drives. A fine micrometer drive is mounted on top.

limitation. However, a micromanipulator that is the best choice for all applications has yet to be designed.

High-magnification micromanipulators. Hundreds of different designs of micromanipulators have been published or have been commercially produced since the mid-nineteenth century, when such devices began to appear in the literature. Two or three interesting new designs are introduced each year, but only about 12–20 basic concepts are manufactured commercially.

One popular design, first described by T. H. Huxley, replaces the mechanical slides with flexure strips, that is, short ribbons of stiff spring steel that are flexible in only one dimension. These are made to flex by levers that have a 10:1 mechanical reduction and that are operated by three precision micrometers. Movement resolution of 0.1 micrometer is achieved, and position can be read from the micrometer scales. The nearly negligible backlash, stiction, and lost motion of the micrometer are all further reduced in a ratio of 10:1 by the levers. Lateral deviation is zero, as there are no slide surfaces, but the movement range is limited to a tenth of the range of the micrometer. A coarse micromanipulator is therefore usually mounted on top of this micromanipulator for quick initial positioning. Successive approximation is worsened, however, because the probe tip moves around an arc that follows the bend of the flexure strip. The substantial mass of a Huxley-type micromanipulator eliminates hand vibration but occupies a great deal of space.

Several applications are made significantly more efficient by single-lever joystick control of microtool movements, which eliminates the problem of successive approximation. The pneumatic designs of the Cailloux and DeFonbrune manipulators have such controllers mounted separately from the probe actuator, which is linked by pneumatic hoses. Remote controllers eliminate hand-generated vibration of the probe. Vibration of the hand on the joystick produces movement, but it is reduced in the same proportion as the intentional movements, so that they remain negligible as in normal macroscopic movements. The probe actuators are backlash-free in both pneumatic designs. The Cailloux design is unique in that the probe follows hand movement in three dimensions. In other joystick designs, the probe follows the hand in two dimensions while rotation of the stick or some other movement controls the vertical axis. When the load changes, as when attempting to penetrate a cell wall, the air in the cylinders compresses and then abruptly shoves the probe forward. Pneumatic micromanipulators are best for sorting cells or for other applications where the load is relatively constant.

The Leitz micromanipulator employs an eccentrically mounted sphere with a joystick to rotate the sphere in either direction, like a two-dimensional cam. The slides that move the probe have pushers that are held against the sphere surface by springs. The Leitz thus has direct mechanical linkage between the joystick and the probe and gives a very smooth positive movement, but it suffers from back-

lash. The designs of the Leitz and the DeFonbrune micromanipulators have a significant advantage in that the ratio of movement reduction is adjustable, allowing the same manipulator to have adequate range at low magnification and adequate resolving power at high magnification.

Another micromanipulator that features an adjustable ratio of movement reduction uses four sets of flexure strips in a square called a cloister stage. If one corner is anchored, the opposite corner can move only in a rectilinear, not curvilinear, fashion. The movement is controlled by a joystick. This rather large, complex, and expensive micromanipulator achieves near freedom from mechanical flaws in its movements.

Hydraulic micromanipulators are like the pneumatic instruments but employ an incompressible oil instead of air. Although they are capable of very smooth, forceful movements, they are most subject to drift of all micromanipulator designs, and the magnitude of drift is sufficient to inhibit their use for certain applications.

Electrically powered micromanipulators. There has been increasing use of electrically powered micromanipulators. The three types are direct-current-motorized, stepper-motorized, and piezoelectric.

Direct-current-motor-driven micromanipulators are relatively inexpensive compared to other sophisticated models, and they offer several advantages. They may be operated by remote control with a joystick or single-step push button, which allows for precise control of pulse length. It is easy to adjust the ratio of movement reduction electrically. Lost-motion correction may be built into the controller. They are also relatively light and space-efficient, but at highest magnification the slight vibration of the motors begins to interfere with visualization, making them best suited to medium magnifications.

Micromanipulators with stepper-motor drives provide the ultimate in precise small movements. These are the only micromanipulators capable of fast, direct movements to a specific set of coordinates, and can be directed by a computer over a pre-planned route. Movements are precise enough to move around a surface and return to the original set of coordinates with a high degree of accuracy. Stepper-motor-drive micromanipulators can also be joystick- or button-controlled. However, stepper motors broadcast electronic noise when in operation and require large, expensive, high-voltage controllers. Thus, they are not easily adapted to a situation in which high-impedance amplifiers are employed, such as in recording cell-membrane potentials.

Specialized piezoelectric micromanipulators, called cell penetrators, provide movements over a very short range of 1–10 μm . They move very abruptly but are excellent for puncturing small and hard-to-penetrate cell membranes. However, these must be mounted on another form of micromanipulator to bring the membrane within range of the 1- μm punch. See ELECTRON MICROSCOPE; MICROSCOPE; OPTICAL MICROSCOPE. Charles W. Scouten

Micrometeorite

A submillimeter extraterrestrial particle that has survived entry into the atmosphere without melting. Meteoroids are natural interplanetary objects that orbit the Sun, and they range in size from small dust grains to objects that are miles (kilometers) in diameter. Particles below 0.04 in. (1 mm) in diameter are considered micrometeoroids, and the micrometeoroids that enter the atmosphere without melting are called micrometeorites. Meteoroids of all sizes enter the atmosphere with velocities in excess of the Earth's escape velocity of 7.0 mi/s (11.2 km/s), and all but the smallest ones are heated sufficiently by air friction to produce at least partial melting. Micrometeorites survive entry without severe heating because they are small and they totally decelerate from cosmic velocity at high altitudes near 55 mi (90 km). In the thin air at such altitudes the power generated by frictional heating is low enough to be radiated away without a particle reaching its melting point, typically about 2400°F (1300°C) for common meteoritic samples. Larger objects penetrate deeper into the atmosphere before slowing down, and are melted and partially vaporized by friction with the comparatively dense air. Most of the mass of extraterrestrial matter that annually collides with the Earth is in the micrometeoroid size range, a total of about 10^4 tons (10^7 kg), but only a small fraction survives as micrometeorites. Usually only the particles smaller than 0.1 mm survive as true unmelted micrometeorites, although the survival of an individual micrometeorite depends on entry velocity, angle of entry, melting point, and density as well as size. The flux of micrometeorites falling onto the Earth's surface is approximately 1 per square meter per day (0.1 per square foot per day) for particles with diameters of at least 10 micrometers and approximately 1 per square meter per year (0.1 per square foot per year) for particles with diameters of at least 100 μm . *See* METEORITE.

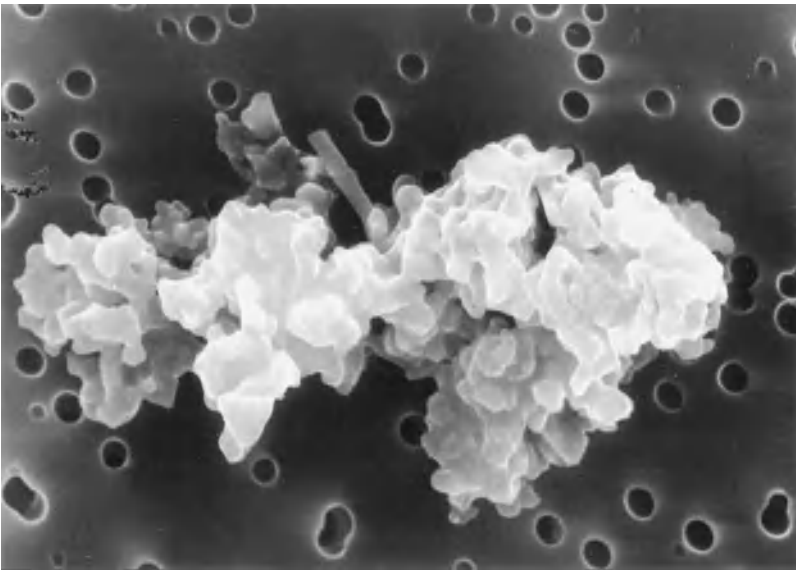
Sources. Micrometeorites are of particular interest because they are samples of comets and asteroids, small primitive bodies that have survived without major change since the earliest history of the solar system. Some of these particles are generated by collisions in the asteroid belt, while others are released from comets when these bodies approach the Sun and ice volatilization releases dust grains and propels them into space. Once released from a parent comet or asteroid, particles survive only for a few thousand to a hundred thousand years, depending on size, before they are either destroyed or collide with a planet. Particles are destroyed either when they collide with other particles or when they spiral into the Sun because of the Poynting-Robertson drag, an effect of sunlight that causes the orbits of small particles to decay. During exposure in space, the small particles accumulate large amounts of helium implanted by the solar wind, and they also are riddled with radiation damage tracks produced by solar cosmic rays, high-energy particles acceler-

ated from solar flares. *See* ASTEROID; COMET; COSMIC RAYS; SOLAR WIND; SUN.

Although it is certain that micrometeorites have both cometary and asteroidal sources, the relative importance of the two sources is not well understood. The dust released from comets can be seen directly as sunlight reflected from the comet tail, and in some cases infrared observations show paths of relatively large particles in the wake of the comet. Most meteors or shooting stars in the night sky are millimeter-sized pieces of comet dust entering the atmosphere at high velocity. Debris from asteroid collisions is observable in the infrared as bands that stretch completely across the sky. Dust from both comets and asteroids can also be seen with the naked eye as the zodiacal light, a glow preceding sunrise and following sunset that is caused by sunlight reflecting off micrometeoroids in interplanetary space. *See* INTERPLANETARY MATTER; METEOR; ZODIACAL LIGHT.

Collection. The collection and laboratory analysis of micrometeorites provide an important source of information on the nature of materials in comets and asteroids. This work complements research on larger conventional meteorites because the dust particles are probably a more representative sampling of early solar-system materials than are the conventional meteorites. Most and possibly all conventional meteorites are believed to be samples of a small number of asteroids from the inner regions of the asteroid belt. These samples are strongly biased because they have to be strong to survive atmospheric entry without fragmenting into dust and they are perturbed to orbits that intersect that of the Earth's only after rare gravitational perturbations, whereas dust particles do not have to be strong to survive atmospheric entry and light-pressure effects allow dust to diffuse through the solar system. Thus, collected micrometeorites include fragile cometary samples and samples of a diverse set of asteroids.

Most micrometeorites are collected in the stratosphere with aircraft such as the U2, which is capable of flying at an altitude of 12 mi (20 km) where terrestrial particles as large as 10 μm are rare. The spatial density of micrometeorites in the stratosphere is exceedingly low, but it is a million times larger than in space because of the low atmospheric fall speeds of micrometeorites relative to their original velocity in space. Micrometeorites are collected from the stratosphere by direct impact onto sticky plates that are extended from aircraft wings into the ambient airstream. With a collection surface area of 5 in.² (30 cm²), the collection rate is one 10- μm particle per hour of flight time. After a cumulative exposure of many hours, the plates are returned to a clean room where the microscopic particles are picked off with needles and placed onto mounts where they can be studied by electron microscopes (see *illus.*), mass spectrometers, and other instruments. Because of limitations imposed by flux and interference by terrestrial particles, the collection of micrometeorites in the stratosphere is usually limited to the size range from 2 to 100 μm in diameter. Most particles larger than this limit melt to form



A 10-mm-long micrometeorite collected in the stratosphere with a U2 aircraft. The image was taken with a scanning electron microscope. The holes on the mounting substrate are unrelated to the particle.

cosmic spherules during atmospheric entry and are not true micrometeorites. A small fraction of particles up to 1 mm in diameter do manage to survive as giant micrometeorites. The flux of these larger particles is too low to collect in the air, but they can be collected from a few locations on the Earth's surface where they accumulate without being severely diluted with terrestrial particles. The best location on Earth for collecting particles of this type is ultrapure polar ice that forms in regions that are distant from rock outcrops or other sources of submillimeter particles. See COSMIC SPHERULES.

Properties. Typical micrometeorites are small black particulates. Some are composed of a few relatively large mineral grains, but most are aggregates of large numbers of submicrometer mineral grains, plus glass and carbonaceous matter. This latter group is often called chondritic micrometeorites because their elemental composition matches that of chondritic meteorites. The composition matches that of the Sun for condensable elements such as magnesium, iron, silicon, aluminum, sulfur, and sodium. Particles dominated by a small number of mineral grains have elemental compositions similar to that of the largest constituent grain. Most of the chondritic particles have similar elemental compositions, but they vary significantly in mineralogical composition. The two most common mineralogical groups are dominated respectively by hydrous minerals such as serpentine and smectite and anhydrous minerals such as olivine, pyroxene, and iron sulfide. Some of the hydrous particles are similar to carbonaceous chondrite meteorites and are thought to have asteroidal origins. They show evidence of moderate aqueous alteration, which is most likely to have occurred inside a moderately warm asteroidal parent body. Comets are smaller, cold bodies that are close to the Sun only for small fractions of their lifetimes. In comets, the damp and moderately warm

internal environments required for aqueous alteration must be very rare. The anhydrous particles are often very porous, show no evidence of aqueous alteration, and are likely to have cometary origins. These particles are unlike any material found in conventional meteorites, and apparently they are samples of meteoroid types that are too fragile to survive atmospheric entry as bodies larger than the size of dust. These particles are being investigated with a broad range of laboratory instruments to determine elemental, chemical, mineralogical, and isotopic compositions and provide clues on the nature and origin of the materials that formed comets and asteroids in the early solar system. Detection of regions inside the particles approximately $1\ \mu\text{m}$ in diameter that are highly enriched in deuterium provide evidence that micrometeorites may also contain records of interstellar materials that predate the origin of the Sun and planets. See COSMOCHEMISTRY; ELEMENTS, COSMIC ABUNDANCE OF; MINERAL; SOLAR SYSTEM. Donald E. Brownlee

Bibliography. D. E. Brownlee, Cosmic dust: Collection and research, *Annu. Rev. Earth Planet. Sci.*, 13:147-173, 1985; J. F. Kerridge and M. S. Matthews (eds.), *Meteorites and the Early Solar System*, 1988; I. D. R. Mackinnon and F. J. M. Rietmeijer, Mineralogy of chondritic interplanetary dust particles, *Rev. Geophys.*, 25:1527-1553, 1987; S. A. Sandford, The collection of extraterrestrial dust particles, *Fundamentals of Cosmic Physics*, 12:1-73, 1987.

Micrometeorology

The study of small-scale meteorological processes associated with the interaction of the atmosphere and the Earth's surface. The lower boundary condition for the atmosphere and the upper boundary condition for the underlying soil or water are determined by interactions occurring in the lowest atmospheric layers. Momentum, heat, water vapor, various gases, and particulate matter are transported vertically by turbulence in the atmospheric boundary layer and thus establish the environment of plants and animals at the surface. These exchanges are important in supplying energy and water vapor to the atmosphere, which ultimately determine large-scale weather and climate patterns. Micrometeorology also includes the study of how air pollutants are diffused and transported within the boundary layer and the deposition of pollutants at the surface.

In many situations, atmospheric motions having time scales between 15 min and 1 h are quite weak. This represents a spectral gap that provides justification for distinguishing micrometeorology from other areas of meteorology. Micrometeorology studies phenomena with time scales shorter than the spectral gap (time scales less than 15 min to 1 h and horizontal length scales less than 2-10 km or 1-6 mi). Some phenomena studied by micrometeorology are dust devils, mirages, dew and frost formation, evaporation, and cloud streets. See AIR POLLUTION; ATMOSPHERE; MESOMETEOROLOGY.

Atmospheric stability. The behavior of the lowest part of the atmosphere depends strongly on stability. In an adiabatic process, there is no loss or gain of heat, no mixing of an air parcel with its surroundings, and no condensation or evaporation of water. An important adiabatic process involves expansion (or contraction) of air as the exterior air pressure decreases (or increases), accompanying rising (or descending) motion of an air parcel with temperature decrease (or increase) by 9.8°C for each kilometer (29.4°F for each mile) of ascent (or descent). A decrease of 9.8°C per kilometer is called the adiabatic lapse rate. Many atmospheric motions are approximately adiabatic. *See* ADIABATIC PROCESS.

For any layer in which the air temperature does not decrease as rapidly as the adiabatic lapse rate, an air parcel that is moved adiabatically upward or downward experiences a buoyant force that tends to cause the air to return to its original level. In this case, buoyancy tends to reduce turbulent mixing, and the atmosphere is said to be stable. An inversion is a very stable situation in which temperature actually increases with height. Whenever the temperature decreases more rapidly with height than the adiabatic lapse rate, any small adiabatic displacement of an air parcel results in a buoyant force that causes the air to move still farther away from its starting level. This is known as a superadiabatic lapse rate. This rate is unstable, meaning that the air will rapidly overturn within this layer. A layer in which temperature decreases at a rate equal to the adiabatic lapse rate is said to be neutral. An air parcel that is adiabatically moved upward or downward in a neutral layer will have no buoyant force acting on it. The neutral case is intermediate between the stable and unstable cases. *See* AIR TEMPERATURE; BUOYANCY.

Atmospheric turbulence. The airflow near the Earth's surface is nearly always turbulent. Friction causes the wind to be zero at the Earth's surface regardless of the wind speed higher up. The difference in wind speed near the ground and the air above is one manifestation of wind shear. The large wind shear near the ground is one of the main sources of energy for turbulent motions. Away from the surface, buoyancy increases the turbulent energy during unstable situations and decreases the turbulent energy during stable situations. Turbulent flows are highly chaotic, vary rapidly in time and space, and contain motions of many different sizes. Some appreciation for the complexities of turbulent motion can be gained by watching the rapid changes and the many sizes of motion of smoke flow. Even in carefully controlled laboratory studies, a turbulent flow is never exactly the same during any two repetitions of the same experiment. Because it is not possible to predict the exact details of turbulent flow, statistical methods are used. Wind, temperature, humidity, and other properties of the flow are written as an average part plus a fluctuation from the average. This separation, known as Reynolds decomposition, can be used to produce equations that govern turbulent flows. The rates at which quantities such as momentum and heat are transported by turbulent fluctua-

tions are known as Reynolds fluxes or simply fluxes. *See* AERONAUTICAL METEOROLOGY; REYNOLDS NUMBER; TURBULENT FLOW.

Whenever the surface is warmer than the overlying air, an unstable, superadiabatic layer forms near the surface, and buoyant convection produces an upward heat flux that tends to mix the air vertically and enhances turbulent mixing. For surfaces that are cooler than the air, the surface layer is stable, and buoyant forces reduce the amount of vertical mixing. The degree of stability of the surface layer is quantified by the use of a dimensionless group known as the Richardson number. *See* CONVECTIVE INSTABILITY; DIMENSIONLESS GROUPS; DYNAMIC INSTABILITY; LAMINAR FLOW.

Plant canopy. When plants are growing on the surface, the layer from the ground to the plant top is known as the plant canopy. Within this layer, airflow is largely determined by the geometry of the plants. Wind speed varies with height within the plant canopy for several types of plants. The wind speed within the canopy is less than above the canopy due to the frictional drag of the plant leaves and stems or trunks and branches. The isolated conifer stand has a layer of maximum wind speed at about 0.1 times the height of the canopy. This type of forest has dense needles near the top of the canopy that produce large amounts of drag and slow the flow. The lowest part of the canopy is more open, mostly consisting of trunks that offer less resistance to the flow and allow greater wind speed. The balance between the pressure gradient force and the frictional drag on the trunks determines the flow in this lowest layer. The flow in this layer tends to be perpendicular to the isobars, whereas the flow above the canopy usually crosses the isobars at an angle less than 40° . *See* GEOSTROPHIC WIND; ISOBAR (METEOROLOGY).

The plant canopy absorbs solar radiation. This energy is used for photosynthesis in the leaves and heats the leaves and the surrounding air. Dense canopies shade the lower part of the canopy layer. Leaves of deciduous trees flutter in the wind, allowing sunlight to reach leaves lower in the canopy than would otherwise be possible.

Water vapor exchange between the plants and the air within the canopy is known as evapotranspiration. Rates of evapotranspiration depend on the photosynthetic and respiration rates of the plants and upon the amount of water within them. Much research is being conducted to determine and test techniques that use satellites for remote sensing of the state of surface plants, the evapotranspiration rate, and the surface heat flux. *See* BIOMETEOROLOGY; REMOTE SENSING.

The exchange of momentum, heat, water vapor, and carbon dioxide between the plant canopy and the overlying air is not steady, but occurs in short episodes. Gusts of air from above the canopy called incursions bring high-momentum air into the canopy. Slow-moving air from the canopy is ejected into the faster air above. These incursions and ejections account for 50–60% of the transfer but occur during only about 5% of the time.

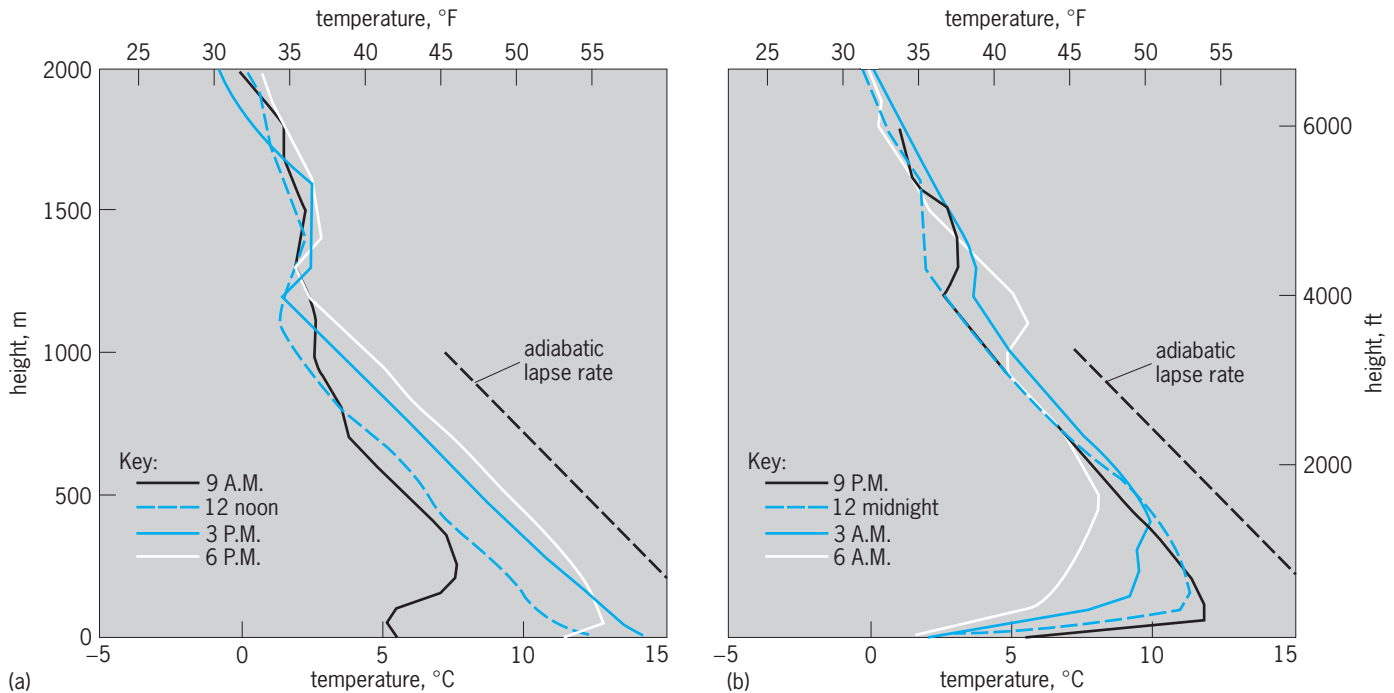


Fig. 1. Soundings of temperature versus height for (a) day time and (b) nighttime obtained by radiosonde measurements during the Wangara experiment near Hay, New South Wales, Australia, in 1967. Soundings are shown for 9 A.M. through 9 P.M. on day 33 of the experiment and 12 midnight through 6 A.M. on day 34. (After R. B. Stull, *An Introduction to Boundary Layer Meteorology*, Kluwer Academic, 1988)

Surface layer. The surface layer extends from above the canopy to a height of about 10 m (30 ft). Within this layer the transfer rates of momentum, heat, and water vapor are nearly equal to the values they have at the top of the canopy (the surface fluxes of momentum, heat, and vapor). Diabatic profile equations (Businger-Dyer equations) are used frequently to relate the surface fluxes to the changes of average wind speed, temperature, and humidity with height within the surface layer. When average wind speed, temperature, and humidity are measured at one or more heights, these equations can be used to compute surface flux values. See GAS DYNAMICS.

Boundary layer. The atmospheric boundary layer extends upward from the surface. It is the entire layer that is directly influenced by the presence of the Earth's surface. It includes the plant canopy and the surface layer and those higher layers in which turbulent mixing takes place. In stable situations the boundary layer may be only a few tens of meters thick. Under convective conditions it is typically 1 km (0.6 mi) or more thick.

The atmospheric boundary layer undergoes distinctive variations during the course of a day (diurnal cycle; Fig. 1). As the Sun warms the surface, convection transfers heat upward into the air, increasing the air temperature and mixing the air vertically. Warming and deepening of the boundary layer continues throughout the day as long as the radiation absorbed by the surface from the Sun exceeds the losses by long-wave radiation. In regions of high atmospheric pressure, the deepening of the boundary layer is retarded by subsidence.

A typical sounding of air temperature in the middle of the day may show an unstable superadiabatic layer near the surface (Fig. 2). In such a situation, when surface heating is strong, the temperature at the surface can be several degrees Celsius higher than the air above. This strong temperature gradient can cause mirages. In the middle of the mixed layer the rate of temperature change is close to the adiabatic lapse rate, indicating that rapid mixing is taking place. The boundary layer is capped by an inversion layer that limits the extent of vertical mixing and acts like a lid for the convection. During highly convective conditions (Fig. 2), substances are mixed rapidly throughout the boundary layer. These conditions produce

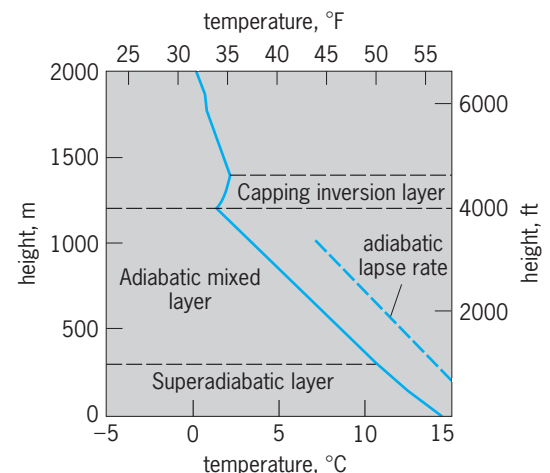


Fig. 2. The 3 P.M. sounding from Fig. 1a showing the layers of the daytime convective boundary layer.

rapid dispersion of atmospheric pollutants throughout the boundary layer. Prediction of the height of the inversion is one step in making air-quality forecasts. *See* CONVECTION (HEAT); MIRAGE.

Late in the afternoon, solar heating decreases, the surface heat source shuts off, and convection stops. During the night the surface cools because of infrared radiative heat loss, and a layer of cool air forms near the surface, giving a surface inversion (6 P.M. sounding of Fig. 1). During the night the air near the surface continues to cool (9 P.M. through 6 A.M. soundings of Fig. 1*b*). The surface inversion is stable and tends to suppress turbulence. When the wind speed is low, there is little wind shear and little turbulent mixing. If surface cooling is very strong, the surface layer may become so stable that the layer is no longer turbulent. On clear nights when radiative cooling is strong, the surface can reach low temperatures, and a thin layer of cold air can form near the surface. These conditions favor the formation of dew, frost, and possibly crop-damaging hard freezes. When wind speeds are greater, there is more turbulence in the surface layer, warm air is mixed downward toward the surface, and temperature decline in the surface layer during the night is less extreme. Crops can be protected from frost damage by smudge pots, orchard heaters, and fans. These are used to create turbulent mixing, which warms the air near the surface and thus mitigates crop damage. Spraying water on crops releases latent heat and can prevent frost damage for freezes lasting a few days or less. However, the weight of ice can cause damage if spraying is attempted for longer freezes. *See* AGRICULTURAL METEOROLOGY; DEW; FROST; WIND.

Surface energy budget. Complex interactions take place near the surface of the Earth over land as heat and moisture are exchanged between the soil, the plants, the air within the plant canopy, and the overlying air. During the day, solar radiation and downwelling infrared radiation are absorbed by the soil and by plants. The resulting heat energy is lost by emission of infrared radiation; or is conducted downward into the soil, increasing soil temperature; or provides the latent heat for evaporation of water from the soil into the air and evapotranspiration of water from plants into the air. The remainder of the heat energy causes an increase in the temperature of the plants, and the air within the plant canopy is carried upward by the turbulent heat flux to warm the air above. At night there is no absorption of solar energy, and infrared radiation usually produces a net loss of heat energy from the surface, accompanied by surface cooling, conduction of heat upward from the soil, and turbulent transport of heat downward from the atmosphere. Latent heat associated with evaporation and evapotranspiration often continues to cool the surface during the night; however, dew or frost formation, when present, releases heat and tends to slow the cooling of the surface.

Surface plant cover has several important effects on micrometeorology. The density of the plant canopy, its color, and the size and shape of leaves and stems affect the absorption of solar radiation

and the emission of infrared radiation. The plants actively transport moisture upward from the soil. Plants open or close their stomata in an attempt to optimize their own well being. Plants need open stomata to obtain carbon dioxide for photosynthesis. When stomata are open, moisture readily escapes from plants. At night and during times when plants are water-stressed, they close their stomata to reduce water loss. The behavior of plants is also dependent on the stage in the plant life cycle. Modern computer programs to simulate surface micrometeorology contain many equations intended to model the effects of plants. *See* HEAT BALANCE, TERRESTRIAL ATMOSPHERIC; INSOLATION; TERRESTRIAL RADIATION.

Ocean-atmosphere interactions. Over the two-thirds of the Earth that is covered by ocean, the boundary layer has a structure somewhat different from that occurring over land. When wind speeds are lower than 3 or 4 m·s⁻¹ (7 or 9 mi/h), the water surface is aerodynamically smooth, and viscosity and diffusion are the main mechanisms transferring momentum and heat through the first few millimeters above the surface. The transfer of momentum from the air to the water results in generation of waves on the ocean surface. At greater wind speeds, these waves make the surface aerodynamically rough and increase the frictional drag on the air. The transfer of momentum from the air to the water also can produce ocean currents and oceanic coastal upwelling. *See* DIFFUSION; OCEAN CIRCULATION; UPWELLING; VISCOSITY.

The ocean has a much larger heat capacity than does soil, and therefore does not change temperature as rapidly as land does. As a result, the marine atmospheric boundary layer does not exhibit much diurnal variation, and it is seldom as stable or as unstable as the boundary layer over land. An important exception is that the marine boundary layer can be extremely unstable during outbreaks of cold air, when cold, dry air from over land passes over warm coastal waters. During these episodes, large amounts of heat and moisture enter the air from the ocean. Because of this transfer of heat and moisture, the Atlantic Ocean off eastern North America and the Pacific Ocean off eastern Asia are regions in which many storms are born and grow rapidly. *See* CYCLONE; MARITIME METEOROLOGY; WIND STRESS.

Other applications. Much of the early understanding of micrometeorology was obtained by studying conditions in large, flat, uniform areas that are relatively simple situations. Micrometeorologists have turned their attention to more complex situations that represent conditions over more of the Earth's surface. The micrometeorology of complex terrain, that is, hills and mountains, is important for air pollution in many towns and cities and for visibility in national parks and for locating wind generators. Another interest is the study of micrometeorology in areas of widely varied surface conditions. For instance, several different crops, dry unirrigated lands, lakes, and rivers may be located near one another. In these cases it is important to understand how the micrometeorology associated with each of these surfaces interacts to produce the overall heat and

moisture fluxes of the region so that these areas can be correctly included in weather and climate forecast computer programs. See CLIMATOLOGY; MOUNTAIN METEOROLOGY; WEATHER FORECASTING AND PREDICTION.

Research methods. Microscale meteorological features are too small to be observed by the standard national and international weather observing network. Generally, micrometeorological phenomena must be studied during specific experiments by using specially designed instruments. Instruments used to study turbulent fluxes must be able to respond to very rapid fluctuations. Special cup anemometers are made from very light materials, and high-quality bearings are used to minimize drag. Other anemometers use the speed of sound waves or measure the temperature of heated wires to measure wind. Tiny thermometers are used, so that time constants are short. Instruments are usually placed on towers or in aircraft, or are suspended in packages from tethered balloons. Instruments have been developed that can measure turbulence remotely. Wind speed and boundary-layer convection can be measured with Doppler radar, lidar devices using lasers, and sodar (sound detection and ranging) using sound waves. See ATMOSPHERIC ACOUSTICS; LIDAR; METEOROLOGICAL INSTRUMENTATION; METEOROLOGICAL RADAR.

Theoretical studies of atmospheric turbulence make extensive use of high-speed digital computers. Because atmospheric turbulent flows have high Reynolds numbers, they have a wide range of sizes and time scales of motion. Even with the fastest modern computers, detailed prediction of only 1 h of boundary-layer turbulence based on the exact equations governing flow would require a time equal to many times the age of the universe to complete. Faced with this obstacle, micrometeorologists have developed approximate techniques based on theoretical and empirical relationships. Higher-order closure techniques use equations that approximate some of the statistics of the flow in terms of simpler quantities and thus obtain a set of equations that can be solved in a reasonable period of time. These models have had good success at reproducing observed behavior of the atmospheric boundary layer.

Large-eddy simulation techniques compute the behavior of the largest features in the flow and use approximations for the smallest features. Even on powerful computers, these techniques require large amounts of time and are thus expensive to use, but they produce very accurate results and are therefore valuable research tools.

Despite the intricate nature of turbulent flow, the rich variety of phenomena involved in micrometeorology, and the detailed computer models available, scientists who make computer weather forecasts are forced to use greatly simplified equations for the boundary layer. These simulate the aspects of the boundary layer that most affect the large-scale weather patterns, including more detailed boundary-layer equations that would be too expen-

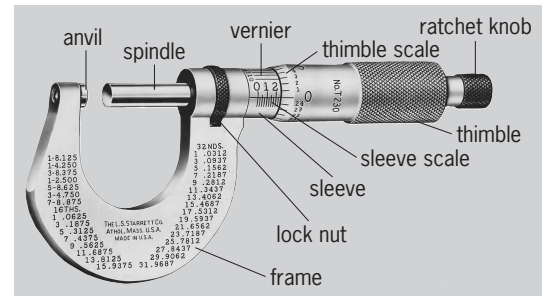
sive and time-consuming for routine forecasts. See BOUNDARY-LAYER FLOW; CLIMATE MODELING; COMPUTER; METEOROLOGY; SUPERCOMPUTER.

Steven A. Stage

Bibliography. S. P. S. Arya, *Introduction to Micrometeorology*, 1988; D. A. de Vries and N. H. Afgan (eds.), *Heat and Mass Transfer in the Biosphere*, 1975; D. A. Haugen (ed.), *Workshop on Micrometeorology*, 1973; T. R. Oke, *Boundary Layer Climates*, 2d ed., 1988; R. B. Stull, *An Introduction to Boundary Layer Meteorology*, 1988.

Micrometer

A precision instrument used to measure small distances and angles. A common use is on a machinist's caliper, as in the **illustration**. See CALIPER.



Machinist's outside caliper with micrometer reading 0.250 in. (L. S. Starrett Co.)

The spindle of the caliper is an accurately machined screw, which is rotated by the thimble or the ratchet knob until the object to be measured is in contact with both spindle and anvil. The ratchet slips after correct pressure is applied, ensuring consistent, accurate gaging. The number 1 on the sleeve represents 0.1 in.; the smallest divisions are 0.025 in. The thimble makes one complete turn for each 0.025 in. on the sleeve, and the 25 divisions on the thimble allow reading to the nearest 0.001 in. A vernier scale allows accurate reading to 0.0001 in. See VERNIER.

Frank H. Rockett

Micro-opto-electro-mechanical systems (MOEMS)

A class of microsystems that combine the functions of optical, mechanical, and electronic components in a single, very small package or assembly. MOEMS devices can vary in size from several micrometers to several millimeters. MOEMS may be thought of as an extension of micro-electro-mechanical systems (MEMS) technology by the provision of some optical functionality. This optical functionality may be in the form of moving optical surfaces such as mirrors or gratings, the integration of guided-wave optics into the device, or the incorporation of optical emitters or detectors into the system. The term may be confused with micro-opto-mechanical systems (MOMS), which more properly

refers to microsystems that do not include electronic functions at the microsystem location. MOEMS is a rapidly growing area of research and commercial development with great potential to impact daily life. The basic concept is the miniaturization of combined optical, mechanical, and electronic functions into an integrated assembly, or monolithically integrated substrate, through the use of micromachining processes derived from those used by the microelectronics industry. These processes, utilizing microlithography and various etch (subtractive) or deposition (additive) steps on a planar substrate, enable the production of extremely precise shapes, structures, and patterns in various materials. *See* INTEGRATED CIRCUITS; INTEGRATED OPTICS; MICRO-ELECTRO-MECHANICAL SYSTEMS (MEMS); MICRO-OPTO-MECHANICAL SYSTEMS (MOMS).

Capabilities. The microsystems realized by these techniques can have many unique capabilities. The miniaturization that is realized is useful in itself, allowing the systems to be utilized as sensors or actuators in environments that were not previously accessible, including inside living organisms, in hand-held instruments, or in small spacecraft. The miniaturization also allows for high-speed operation of the system, as the operating speed of mechanical systems is related to their inertial and frictional properties as well as the actuating forces. Optomechanical systems have been historically constrained in this area because of the mass required for stable optical elements and the extremely precise alignment requirements of most opto-mechanical systems, which limits the forces that can be tolerated for rapid motion. In the more integrated forms of MOEMS, the systems are prealigned by the precise fabrication processes, eliminating one of the more expensive aspects of assembling conventional optical systems. The miniaturization along with the scalability of microfabrication processes allows the development of massively parallel opto-mechanical systems, with millions of moving parts, that would not be possible in conventional technologies. MOEMS can incorporate detection and drive electronics in close proximity to provide improvements in signal-to-noise ratio for sensors and simplified interfaces for actuated systems. Ultimately, these electronics may be monolithically integrated in some technologies. Because of the production volumes achievable with micromachining techniques, MOEMS are potentially much less expensive than their conventional counterparts.

Sensors and actuators. Applications of MOEMS devices typically involve sensing of their environment, or the use of actuators to perform useful work. The sensing mechanisms can be any of those applicable from MEMS technology, with the added capability to utilize optical processes to enhance sensitivity or provide optical communication of data. The actuation mechanisms provided by MEMS technology are limited in many applications by the small forces that can be applied. MOEMS applications often involve the control of light energy, which is an ideal role for actuators that cannot deal with large forces, masses, or energies. The actuators can provide lin-

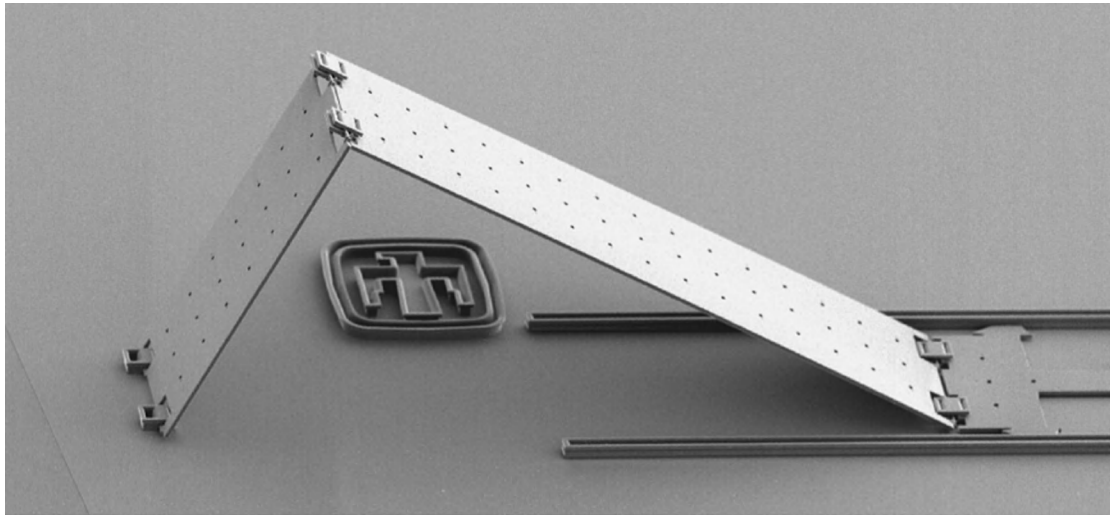
ear or rotary motion. MOEMS systems can also include microfluidics, in which liquid solutions can be transported through an optical detection system for analysis or other processes. *See* MICROSENSOR.

Fabrication. The fabrication processes used for MOEMS are the same as for other MEMS technologies. Most involve one of a few basic approaches, sometimes combined with unique materials or combinations of devices that utilize different processes. The fabrication technologies for MOEMS are based on the use of microlithography as a means of defining microscopic structures in large quantities and great precision. Microlithography produces a two-dimensional pattern on the surface of the substrate in a resist material that is coated on the substrate surface. The resist will protect some areas from subsequent process steps or act as a barrier that can be removed later as a sacrificial material. The microlithographic processes are then combined with various micromachining processes to define the mechanical structure. The micromachining processes are generally additive, in which material is deposited in sequence to build up a structure on a substrate, or subtractive, in which deposited films on the substrate or the substrate itself are removed to form the structure.

Surface micromachining in thin-film layers on silicon or other substrates is the most advanced process in terms of miniaturization, integration, and industrial support, as it uses the processes and tools of the silicon integrated circuit industry. The limitations of this process are due to the thin structures that are produced and the weak actuator forces that can be implemented by the thin-film structures. Surface micromachining with compound semiconductors (primarily from group III and group V, that is, columns 13 and 15 of the periodic table), such as gallium arsenide and indium phosphide, allows for light emission in the microsystem, in the form of semiconductor lasers or light-emitting diodes. These materials can also provide photodetectors in wavelength bands where silicon is not sensitive.

Bulk micromachining involves mostly subtractive processes that sculpt the silicon or other substrates to produce simpler, more robust structures that may incorporate thin-film membranes on the silicon surface as well. Variations of this process use deep reactive ion etching to form these structures with more versatility than can be attained by wet etch processes. These processes may be supplemented by wafer bonding to join structures together that have been processed on separate substrates originally.

Another process available is LIGA (a German acronym for *Lithographie, Galvanoformung, und Abformtechnik*), involving the combination of deep x-ray lithography with electroforming of metals and (sometimes) injection molding, to produce mechanical parts that are larger and more robust than the products of semiconductor fabrication processes. The disadvantage of LIGA is its lower level of integration, with more assembly required for functioning systems. Polymer replication processes, including casting, embossing, and injection molding, are



Scanning electron microscope image of a simple MOEMS tilting mirror, actuated by a MEMS rack-and-pinion drive that pushes the hinged mirror up from the right side. The structure is made by surface micromachining in polysilicon. The mirror is approximately $100\ \mu\text{m}$ wide. (Sandia National Laboratories)

also applicable to some MOEMS fabrication needs. In all these cases, the use of micromachining, which can be performed in high volume, can produce extremely small and precise devices at low cost per device.

Integration and packaging problems. Some of the major problems that are still unsolved for MOEMS are integration difficulties and packaging challenges.

Although some MOEMS can be manufactured as a monolithically integrated device, in a single material system and process, the combination of different functions in a MOEMS sometimes calls for dissimilar materials and devices that cannot be integrated into a single material and processing line. In those cases, MOEMS are fabricated using hybrid assembly or heterogeneous integration techniques that allow the final system to include parts fabricated by different processes and from different materials to be precisely located on a common substrate and share electrical and optical signals. These assembly processes add expense and complexity in comparison to monolithically integrated systems.

The other major challenge to the development of MOEMS is the packaging of the devices. The mechanical actuators are often very sensitive to contamination and must be protected from the environment. The sealed packages that are required must usually also allow for optical access into and out of the package. In addition, optical systems are often complex three-dimensional structures and MOEMS are usually two-dimensional planar structures. The package is often required to incorporate and hold in position additional optical elements in the system, including optical fibers, lenses, and detectors. This requirement can add great complexity to the packaging. See ELECTRONIC PACKAGING.

Applications. The most sophisticated MOEMS device in commercial use is the Digital Micromirror Device (DMD), manufactured by Texas Instruments. This device consists of 2 million individual tilting micromirrors (see *illus.*), each $16\ \mu\text{m} \times 16\ \mu\text{m}$ in

size. Each mirror is attached to a torsion beam suspension that allows it to tilt in one axis. The entire array of mirrors is fabricated by postprocessing of a static random access memory chip that is fabricated in conventional silicon complementary metal-oxide-semiconductor (CMOS) technology. Surface micromachining techniques are used to deposit metal films and sacrificial polymer layers to build the mirrors on top of the memory chip. The cells of the memory chip hold electrical charges, depending on the data stored, that actuate the mirrors above by electrostatic forces. This system is an excellent example of a scalable process and architecture and of monolithic integration. The DMD is used primarily in video projection systems. See OPTICAL PROJECTION SYSTEMS.

A new area of application of MOEMS that may have great impact on everyday life is in optical-fiber switching systems. As the need for high-bandwidth data transmission grows with increasing use of the Internet as a means of communication and commerce, fiber-optic communications are increasingly used and will eventually link to homes and offices directly. A necessary component of these optical-fiber networks is a switching system that can independently switch each of the channels carried in the fiber. Conventional opto-mechanical devices are large and slow, and other nonmechanical approaches have not been successful. The MOEMS optical switch, scalable to hundreds or thousands of channels, is under rapid development by many companies and may enable the development of direct access to very high bandwidth communications. See OPTICAL COMMUNICATIONS; OPTICAL FIBERS.

Mial E. Warren

Bibliography. M. J. Madou, *Fundamentals of Microfabrication*, CRC Press, Boca Raton, FL, 1997; M. E. Motamedi and R. Goering (eds.), *Miniaturized Systems with Micro-Optics and MEMS*, *Proc. SPIE*, vol. 3878, 1999; *Proceedings of the 3d International Conference on Micro Opto Electro Mechanical Systems: MOEMS '99*, Mainz, 1999; J. M. Younse,

Mirrors on a chip, *IEEE Spectrum*, pp. 27–31, November 1993.

Micro-opto-mechanical systems (MOMS)

Miniaturized opto-mechanical devices or assemblies that are typically formed using micromachining techniques that borrow heavily from the microelectronics industry. The term may be used to distinguish devices and microsystems that combine optical and mechanical functions without the use of internal electronic devices or signals. Systems that use electronic devices as part of the microsystem may be referred to as MOEMS (micro-opto-electro-mechanical systems). In some cases, these terms may be used synonymously. A related area is MEMS (micro-electro-mechanical systems), in which electronic and mechanical functions are combined in a miniature device or system, but not necessarily implementing optical functions. The progress of MOMS technology has been greatly enabled by the simultaneous development of microelectronics and optical fiber-based telecommunications technology. These two technologies have resulted in a large manufacturing and scientific base from which MOMS has been developed. The commonly used MOMS fabrication processes are the same as are used for variations of the MEMS fabrication processes with some additional materials choices, due to the need to be compatible with optical functions. *See* INTEGRATED CIRCUITS; MICRO-ELECTRO-MECHANICAL SYSTEMS (MEMS); MICRO-OPTO-ELECTRO-MECHANICAL SYSTEMS (MOEMS); OPTICAL COMMUNICATIONS.

Advantages. Although similar in concept to MOEMS technologies, MOMS has unique advantages for some applications. The use of only optical energy and signals gives MOMS an inherent immunity to electromagnetic interference (EMI) that is important for applications in electrically noisy or high-voltage environments. The absence of semiconductor electronic devices greatly increases the high-temperature tolerance of the system. MOMS devices can be designed to work immersed in liquids, which is of great importance for chemical sensing and biomedical applications. The fact that the power and signal sources can be remotely provided via an optical fiber, allowing the sensor to be passive, is of great utility and reduces the impact of a MOMS sensor on its local environment. MOMS can be used safely in flammable and explosive environments, making them uniquely valuable in the petrochemical industry.

Sensors and actuators. A micromechanical system is able to function either as a sensor or as an actuator. In sensor design, the goal is to measure a physical parameter with minimum impact on that value from the operation of the sensor and with low sensitivity to other parameters that may be changing. An actuator performs some mechanical work that will either change the local environment or control or enable energy or information to be transported or controlled.

MOMS sensors. MOMS applications have been primarily in sensors, utilizing a number of physical phenomena. MOMS sensors consist of some movable mechanical element that can be detected or measured by optical means. The motion may be coupled to temperature, electrical and magnetic fields, acoustic energy, acceleration, chemical forces on surfaces, and so forth. Many sensors can be made sensitive to various physical or chemical phenomena by addition of special surfaces or coatings to the mechanical element.

The state or position of the mechanical structure can be detected optically by various means. Many sensors use intensity modulation of a light beam. This may be by mechanical interruption of a beam path or by more subtle means, such as defocus or deflection of light to be coupled into an optical fiber. Many sensors use phase modulation as the sensing means. This requires the use of an interferometer to convert phase into a detectable amplitude modulation. The interferometer may be one of several types. Fabry-Perot devices can utilize free-space propagation of beams normal to the mechanical surfaces to detect motion or position of the surfaces. Mach-Zehnder and other designs that require splitting and recombination of light beams are often implemented as guided-wave devices, in which the light is confined in planar thin-film structures. Interferometric detection can be very sensitive, but places additional requirements on the light source and detection electronics. A third mechanism is wavelength, in which a broad spectral source can be spectrally dispersed or filtered. Polarization is also an optical parameter that can be used in some cases. *See* DISPERSION (RADIATION); INTERFEROMETRY; POLARIZATION OF WAVES.

Since many MOMS sensor applications use optical fibers as the sources of power and means of data retrieval, MOMS sensors are often discussed in the context of optical-fiber sensors. MOMS can be distinguished from the more mature field of fiber-optic sensing by their use of miniaturized or integrated mechanical structures, formed by micromachining techniques, as the sensor. A MOMS sensor is extrinsic to the optical fiber, as opposed to intrinsic optical-fiber sensors, in which the fiber is the sensing device. *See* FIBER-OPTIC SENSOR; MICROSENSOR; OPTICAL FIBERS.

MOMS actuators. MOMS actuator technology is much less developed. The actuation mechanisms that are available to allow light to do mechanical work are few. Photothermal mechanisms allow the conversion of light to heat in order to power a thermomechanical actuator. Examples of this process include expansion of a working fluid or deforming a bimetallic strip through differential thermal expansion. Photoexcitation of semiconductor structures can be used to generate electrostatic and surface charge effects or cause photostriction in piezoelectric materials to impart mechanical force. Light can also produce direct radiation pressure, but the forces are small relative to the light intensity. A system can be powered by light through photovoltaic conversion, but then would

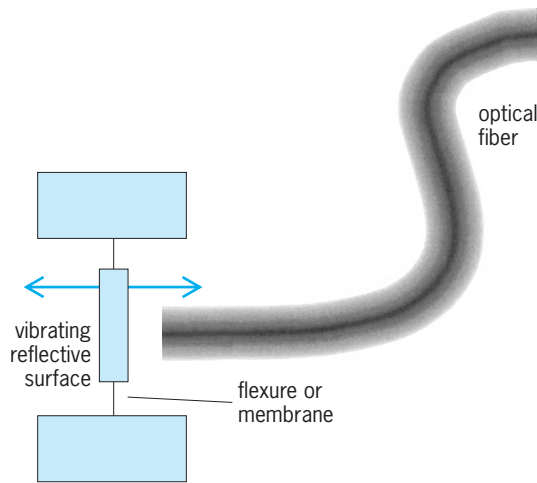


Fig. 1. A simplified MOMS sensor using a reflecting surface on a flexible mount or membrane and the end face of an optical fiber to form an optical interferometer that can sense vibration. The vibration of the flexible membrane allows the reflecting surface to move, changing the resonance wavelength of the interferometer, modulating the intensity of the light that is reflected back into the interferometer. At the other end of the interferometer is a light source and a detector.

not be considered purely opto-mechanical and classified as a MOEMS device. See PHOTOVOLTAIC EFFECT; PIEZOELECTRICITY; RADIATION PRESSURE.

Applications. Some examples of MOMS technology include optical pressure transducers—microphones or hydrophones that have a thin mechanical membrane that is one surface in a Fabry-Perot interferometer formed by the reflection from the membrane surface and the reflection from the end of the fiber. (A similar arrangement for sensing vibration is shown in Fig. 1.) Other versions have a planar optical waveguide on the surface of a sensitive membrane that is one arm of a two-beam Mach-Zehnder interferometer. Another example is an accelerometer in which a small mass is suspended from flexure attachments to the substrate. Optical fibers are positioned with a small gap in which the moving mass can interrupt the transfer of light from one fiber to another to modulate the light intensity transmitted through the fibers. One of the most well developed MOMS applications is optical sensing

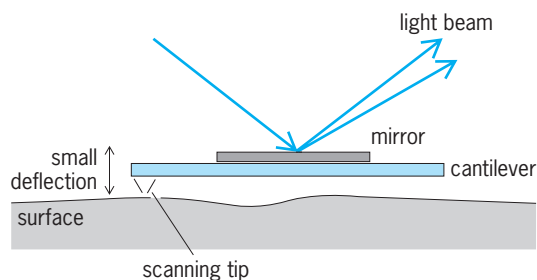


Fig. 2. A simplified representation of how a light beam is used to sense minute deflections of a micromachined cantilever in an atomic force microscope. The sharp tip is scanned very close to a surface. The interatomic forces on the tip deflect the cantilever. The light beam reflected off the top of the cantilever amplifies the motion. This MOMS device can resolve individual atoms on the surface.

of the position of small cantilevers used in scanning tip microscopy processes such as atomic force microscopy. In some versions of these extremely sensitive instruments, a micromachined cantilever or balance mechanism is subject to forces from a sharpened tip on the end of the lever that interacts with the interatomic forces of a surface. The forces bend or tilt the mechanism with displacements of fractions of a nanometer. The minute bending or tilt can be measured optically by angular change of reflected light from the surface of the cantilever beam without perturbing its position (Fig. 2). See ACCELEROMETER; MICROPHONE; PRESSURE TRANSDUCER; SCANNING TUNNELING MICROSCOPE.

Prospects. Micro-opto-mechanical systems are still in their infancy. As optical fiber communications becomes more pervasive, use of MOMS can be expected to increase as a means of optical sensing and in some cases actuation, for situations in which it provides unique value. Some challenges have to be addressed to further improve the performance, cost, and utility of MOMS. Integrating optical and micromechanical functions, which often require different materials and processes, is difficult and limits the sophistication and complexity that can be achieved with MOMS technology. The packaging of MOMS requires input and output coupling of optical signals into a package that often needs to be in intimate contact with its environment for the sensor to function, while still protecting the system. See ELECTRONIC PACKAGING. Mial E. Warren

Bibliography. A. J. Jacobs-Cook, MEMS versus MOMS from a systems point of view, *J. Micromech. Microeng.*, 6:148-156, 1996; M. J. Madou, *Fundamentals of Microfabrication*, CRC Press, Boca Raton, FL, 1997; G. Meyer and N. M. Amer, Novel optical approach to atomic force microscopy, *Appl. Phys. Lett.*, 53:1045-1047, 1988; A. Wang (ed.), Harsh Environment Sensors II, *Proc. SPIE*, vol. 3852, 1999.

Micropaleontology

A branch of paleontology dealing with the fossilized microscopic organic remains (microfossils) of the geologic past, their structure, biology, phylogenetic relations, and distribution in space and time. The study of these microfossils has become an independent scientific field largely because of the following: (1) The size of these fossils requires special methods for collection and examination. (2) Their abundance in geologic formations makes it possible to analyze their spatial distribution and the rates of morphological changes during the course of evolution by means of statistical methods which can be used only under exceptional circumstances in the study of larger fossils. (3) Microfossils have become indispensable tools in certain branches of applied geology, especially in the exploration for oil-bearing strata, because countless numbers of these minute fossils may be obtained from small pieces of subsurface rock recovered from drill holes. (4) The diversity of microfossils, their

wide spatial distribution in varied environments, and their distinctive steps in evolution and the ease of studying them have contributed to make micropaleontology one of the most actively studied branches of the earth sciences. Many organisms are restricted to certain environments best suited to their life activities. The occurrence of certain organisms in sediments, therefore, can in turn provide a clue to a particular environment that existed in the past. Micropaleontologists can obtain important information about depositional environments, such as their ancient environment, water depth, temperature, current systems, water mass distribution, proximity of shore lines and, of course, these organisms are used to date ages of enclosing sediments.

The material subjected to micropaleontological studies forms a spectrum from primitive plants to advanced vertebrates. The only prerequisite for organisms to become the subject of micropaleontological studies is their possession of resistant skeletal components ensuring their preservation in sedimentary strata as fossilized remains even after biological, chemical, or mechanical processes have destroyed the organisms' soft parts.

Most major groups of organisms incorporate, besides organic compounds, hard resistant materials that serve for structural support or protection. The more common substances found among the microfossils are calcium carbonate, silicon dioxide (or silica), calcium phosphate in the form of the mineral apatite (typical of bones and teeth), sporonine (principal constituent of pollen and spore walls), and various complex organic compounds.

Micropaleontological objects representing the plant kingdom (**Fig. 1**) range from primitive unicellular forms such as calcareous nannoplankton, diatoms, charophytes, silicoflagellates, chrysomonads, dinoflagellates and their cysts, tintinnids, the enigmatic chitinozoans, to portions of advanced plants such as pollen, spores, and seeds. *See FOSSIL SEEDS AND FRUITS; PALYNOLOGY.*

Micropaleontological objects representing the animal kingdom (**Fig. 2**) include almost entire organisms of the unicellular Radiolaria and foraminiferans; carapaces or shells of the minute crustacean Ostracoda; embryonic or immature shells of larger organisms such as bryozoans, corals, or mollusks; individual skeletal elements of advanced organisms, including sponges, corals, sea cucumbers, echinoderms, worms, and fishes; and toothlike skeletal elements of animals of unknown taxonomic affinity known as conodonts, which have been assigned variously to the worms, gastropods, cephalopods, and fishes.

Occurrence and abundance. Microfossils occur in most types of marine and nonmarine sediments and sedimentary rocks, particularly in limy shales, limestones, black shales, siltstones, and fine-grained sandstones. They are usually disseminated through the rock in quantities ranging from a few to several hundred fossils per gram of rock. Certain rock types, however, are composed almost entirely of microfossils. Diatomite is a siliceous earth made up of the

bivalved frustules of diatoms; chalk is a soft variety of limestone composed of the calcareous tests of foraminiferans and coccoliths; certain limestones of the Pennsylvanian and Permian systems are essentially composed of fusulinid Foraminiferida; and the Gizeh limestone of the Eocene of Egypt, with which the pyramids were built, is made up almost entirely of the shells of the large foraminiferid *Nummulites*. Many reef limestones in the sedimentary column are built by lime-secreting algae, corals, bryozoans, and stromatoporids. The dense siliceous rock known as chert, in many instances, was formed from accumulations of the shells of diatoms and radiolarians, while plant spores and pollen are believed to be the essential particles composing the rock known as cannel coal. Deep-sea environments harbor special types of sediments entirely made up of shells of microorganisms. These include *Globigerina* ooze, consisting largely of calcareous nannofossils and tests of planktonic foraminifers, and siliceous ooze, of which chief constituents are radiolarians and diatoms. *Globigerina* ooze covers nearly one-half the area of the entire world ocean floor, and siliceous ooze blankets nearly 15%.

Collection and preservation. Surface exposures of sedimentary rock are sampled for micropaleontological analysis by collecting about a pint (a few hundred cubic centimeters) of fresh unweathered rock at the outcrop. Choice of vertical sampling interval depends upon the total thickness of rock exposed, the degree of change in the sedimentary characteristics within the vertical section, and the rate of accumulation of sediments.

Selection of samples containing microfossils larger than 0.02 in. (0.5 mm) may be greatly facilitated by inspecting the sample with a 10× or 15× hand lens. The presence of smaller microfossils cannot be detected in the field without the use of a microscope.

Subsurface samples of fossiliferous rocks are obtained from drill cuttings and from cores obtained in the course of drilling wells for oil, gas, or water. They consist of small chips or rock fragments washed from the borehole during drilling operations, or of cylindrical cores of the rock strata obtained by the use of a special drilling bit designed for such a purpose.

Laboratory treatment to free the microfossils from the enclosing rock matrix varies greatly, depending upon the hardness, texture, and mineral composition of the rock and upon the size and chemical composition of the microfossils. Since most microfossils are quite fragile, the sample is given a minimum of mechanical and chemical treatment to reduce the rock to individual particles and to liberate the microfossils.

Soft shales, siltstones, and chalks are usually soaked and gently boiled in water containing a deflocculant such as sodium hexametaphosphate. This treatment reduces the sample to a thick mud, which is flushed through a series of graduated screen sieves to remove the finest particles. The various-sized fractions are dried and placed in envelopes or vials to await inspection with the microscope. Tougher and more resistant rocks are broken down by first heating

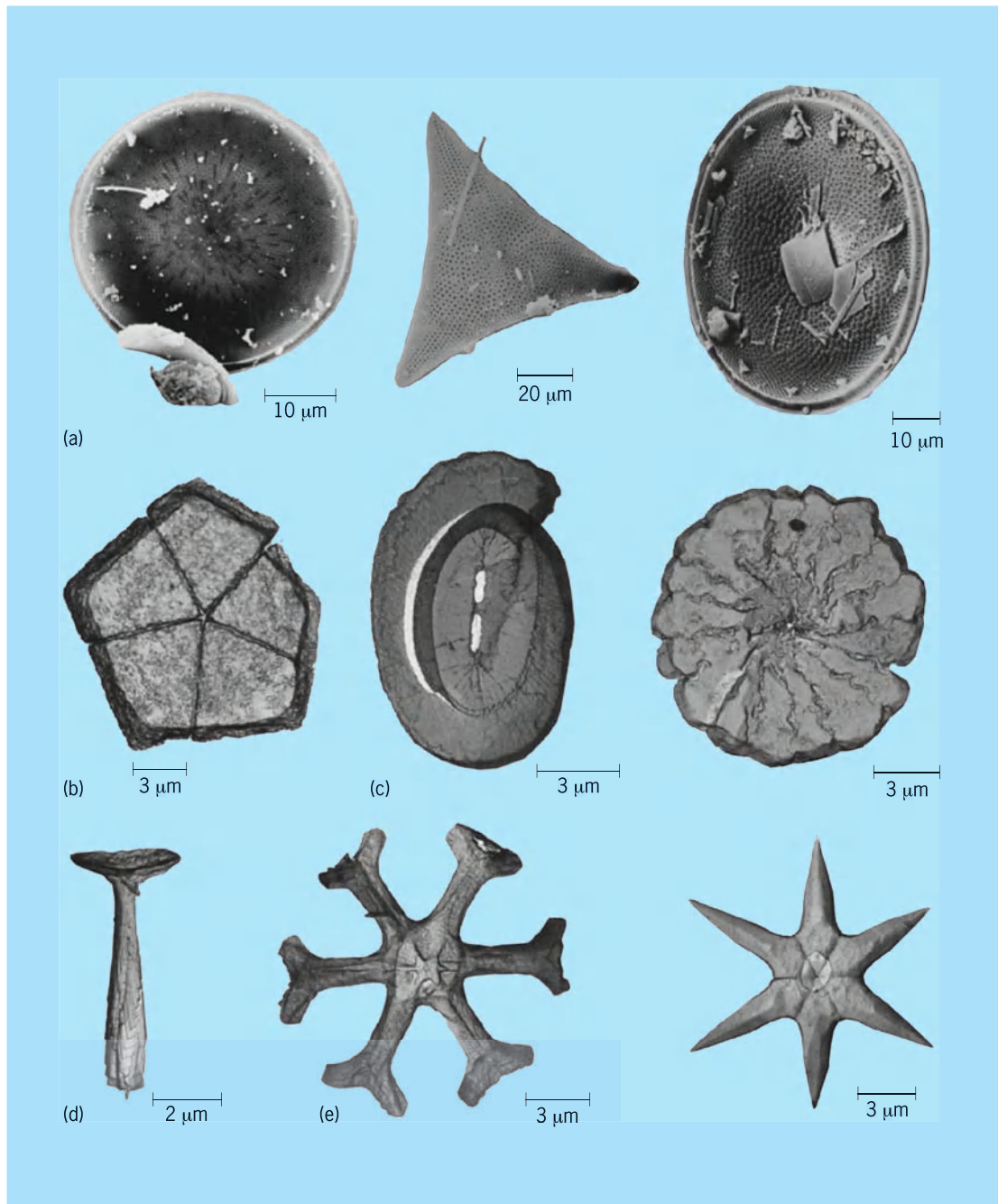


Fig. 1. Representative microfossils of the plant kingdom. (a) Diatoms. (b–e) Examples of calcareous nannoplankton: (b) pentalith, (c) placoliths, (d) rhabdolith, and (e) asteroliths.

samples (crushed to a chestnut size) in an oven at a temperature of more than 212°F (100°C) and then soaking them with either a saturated solution of sodium sulfate or kerosine extract. Samples treated with sodium sulfate should be left at room temperature for a few days until the chemical crystallizes, breaking up the rocks. The kerosine-soaked rocks also give a better result when they are left at room temperature for a few days and then boiled in water with a deflocculant.

Limestones and other carbonate rocks are digested with dilute hydrochloric acid or glacial acetic acid to dissolve the carbonate minerals, leaving a residue

of insoluble mineral grains and such noncalcareous microfossils as siliceous sponge spicules, conodonts, diatoms, radiolarians, and plant microfossils.

Disaggregation of coal and mudstone to liberate spores, pollen, and other very small microfossils requires special treatment. This includes digestion of the rock sample with hydrofluoric acid and Schulze's solution (a mixture of nitric acid and potassium chlorate), concentration of the fine-grained residues by centrifuging, suspension of smears of the residues in glycerin jelly, selective staining, and mounting of the stained material on glass microslides for microscopic examination. Both the Schulze's solution and

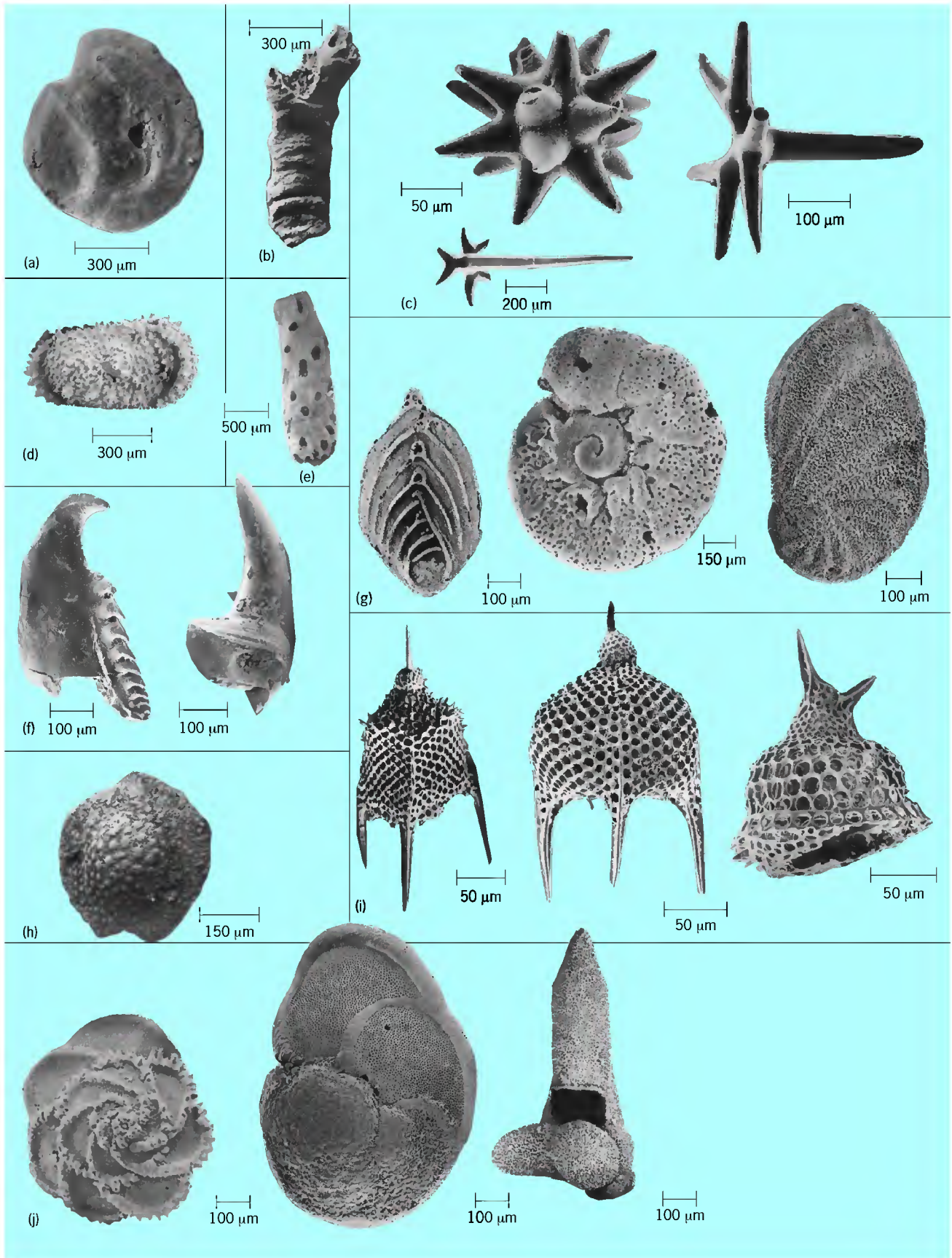


Fig. 2. Microfossils of the animal kingdom. (a) Vertebrate remains: fish ear bone. (b) Alcyonarian coral. (c) Sponge spicules. (d) Ostracod. (e) Bryozoans. (f) Scolecodonts. (g) Smaller benthonic foraminiferans. (h) Larger benthonic foraminiferans. (i) Radiolarians. (j) Planktonic foraminiferans.

hydrofluoric acid are highly toxic and corrosive, and extra care must be used when these chemicals are handled.

Microfossils larger than 0.04 in. (0.5 mm) are separated from the dried and size-sorted residues by scattering the grains on a small black tray and inspecting them on the stage of a wide-field stereoscopic binocular microscope at magnifications of 10 or 15 diameters. The individual fossils are selected from the residue by picking them up on the moistened and pointed tip of a small red sable watercolor brush (size 000); transferred to a special micropaleontological slide made of cardboard, which has a holed-out and blacked-in mounting area; and coated with a water-soluble adhesive such as gum tragacanth.

Smaller microfossils are studied in smears or suspensions of fine residues mounted on glass micro slides and protected with cover slips. Depending on the type of study, both the light microscope at magnifications of 100–1500 and the electron microscope at magnifications of 600–25,000 are used to examine microfossils.

Certain large fossils, such as bryozoans, lime-secreting algae, colonial corals, stromatoporids, and large Foraminiferida, are prepared for examination by grinding down an oriented fragment of the large fossil to a thin slice through which light can be transmitted after it is mounted on a glass microslide with a cover slip. Such sections are studied at magnifications of 30–500.

Applied micropaleontology. In the historical development of micropaleontology, researches on various microfossils have progressed at a markedly different pace, reflecting the interests of micropaleontologists working at various periods of time.

Beginning around the 1930s, great impetus was given to the study of Foraminiferida, Ostracoda, and pollen and spores. Micropaleontologists employed by oil companies have found these minute fossils, which could be easily secured from well cuttings and cores, often in great abundance, useful for correlating oil-bearing strata.

Research interest in these areas surged to such an extent during the 1940s that micropaleontologists found it necessary to create ever finer subdivisions of micropaleontology, each of which would be devoted to one particular group of microfossils. For instance, the term palynology, defined as the study of both modern and fossil pollen and other spores and their dispersal, and applications thereof, was introduced in 1944. Probably less well known, a few other terms such as foraminiferology and ostracodology also came into existence to represent the field of research exclusively devoted to a certain group of microorganisms.

The great expansion of oceanographic research during the 1960s gave another advantageous turn in the course of micropaleontological research, inviting earnest investigations of minute plants and animals inhabiting the surface waters of oceans. In contrast to the sessile organisms (termed the benthos) whose habitats are ocean floors, these organisms maintain the state of suspension throughout their life-span and

are known collectively as plankton. The free suspensive mode of life of plankton in oceanic waters favors their wide geographic distribution by ocean currents, and fossilized representatives of various planktonic microorganisms have become very important tools in establishing mutual time relationships of marine strata in distant parts of the world.

In contrast to the fast-changing continental environments which were subjected to such forces as advancing glaciers, orogenic upheavals, and inundation by the sea, earth scientists would generally agree that the deep-sea environment is one of the most stable on the Earth. For millions of years, an incessant “snowfall” of tiny particles has been settling over the ocean floor. In deep ocean basins, far removed from the continents, the most significant contributors of these particles are the innumerable planktonic microorganisms, which pass their life in the sunlit surface waters, building up their skeletons or shells from the silica or calcium carbonate present in the water. After the plankton dies, these hard parts, freed from organic matter through bacterial action, sink to the bottom and build up the unique deep-sea sediments known as organic oozes. The oozes, rich in calcium carbonate and largely consisting of the shells of calcareous nannoplankton and foraminiferans, dominate in lower latitudes, while organisms such as diatoms and radiolarians form layers of siliceous oozes in the far north and south, especially around the icebound Antarctic continent.

At any particular location on the deep-ocean floor, if the successively accumulated layers of the organic oozes for millions of years have not been disturbed by slumping or denudation by bottom currents, a core of sediment plugged out of the sea bottom would reveal the succession of floras and faunas that inhabited the oceans of the geologic past. No plant or animal has existed for all geologic time. Each species evolved from some ancestor with a definite time of beginning and, if it is extinct today, also a time of ending. Each extinct species therefore may provide three distinct time levels which can be used for comparing time relationship of marine sediments: (1) the time before it evolved; (2) the time during which it existed; and (3) the time since it became extinct. The evolution of organisms through time thus provides a means by which discrete units of time represented by the material accumulation of sediments can be recognized. Dispersal by ocean currents of these planktonic organisms to distant parts of the world further enables a comparison of the established time units between regions, particularly with the marine sections in Europe on which the presently recognized geologic systems were defined. The long, continuous record of biota in the deep-sea basins was intensively studied in the 1960s because the marine record preserved in the continents is frequently punctuated by breaks in sedimentation.

Planktonic biota in deep-sea basins also play a major role in understanding oceanographic conditions which existed in the geologic past. When marine planktons build their shells from minerals dissolved in the water, they incorporate within their

shells oxygen and carbon isotopes in the proportion which existed at the time of their growth. Proportions of various oxygen and carbon isotopes in oceanic waters are controlled by physicochemical parameters of the water, so that by measuring the isotopic compositions incorporated in microfossils, scientists, in turn, can estimate such parameters as temperature and salinity of past oceans.

When magnetic iron oxide minerals slowly settle out of a fluid, they align their magnetic polarity with any magnetic field that existed during the past. This produces a preferential magnetization in sediments that contain the minerals; this property, which is measurable in the laboratory, is termed remnant magnetism. Study of remnant magnetism in rocks of different ages all over the world indicates that the Earth's North and South poles have abruptly switched positions many times in the geologic past. Since each of these paleomagnetic reversals simultaneously affected the field for the entire Earth, reversal records identified in sediments can be compared with the similar record in continental rocks which can be dated precisely by radiometric age determinations. This permits micropaleontologists to compare rates of microfossil evolution against actual time lines for the first time.

Study of deep-sea sediments represents a unique but active field of research, where micropaleontologists, geochemists, and geophysicists can cooperate in elucidating the interaction of environmental changes with the course of organic evolution. See INDEX FOSSIL. Tsunemasa Saito

teristics are also of interest, that is, the voltage output as a function of the direction of incidence for constant sound pressure. See DIRECTIVITY; SOUND; SOUND PRESSURE; TRANSDUCER.

In addition to directional characteristics, some other important characteristics of microphones include open-circuit sensitivity, equivalent noise level, dynamic range, and vibration sensitivity.

Open-circuit sensitivity is defined as the ratio of open-circuit output voltage and sound pressure. The pressure sensitivity refers to the actual pressure acting upon the diaphragm of the microphone, while the free-field sensitivity refers to the pressure that existed in the sound field before insertion of the microphone. Pressure sensitivity and free-field sensitivity are equal at low frequencies. Sensitivities are measured in volts/pascal (V/Pa).

Equivalent noise level is equal to the level of a sound pressure which generates an output voltage of the microphone corresponding to its inherent A-weighted noise voltage. It is measured in dB(A).

Dynamic range is defined as the range of sound pressure levels in decibels (dB) extending from the equivalent noise level to the level where the nonlinear distortion reaches 3%.

Vibration sensitivity is defined as the ratio of the output voltage of the microphone as a result of acceleration of its case to the magnitude of the acceleration. Vibration sensitivities are measured in volts/ g , where g is the acceleration of the Earth's gravity, or in volts/(m/s^2).

Some aspects of the acoustic performance of a few microphones are given in the table. Only microphones of current usage have been considered.

Microphone

An electroacoustic device containing a transducer which is actuated by sound waves and delivers electric signals proportional to the sound pressure. Microphones are usually classified with respect to the transducer principle used. Their directional charac-

Microphone Types

The various microphone types differ by the transducers used for converting the acoustic into electric signals. Most commonly used are electrostatic, piezoelectric, dynamic, and magnetic transducers. These transducer principles are reversible, that is,

Acoustic characteristics of microphones

| Microphone type | Frequency range, Hz | Sensitivity, mV/Pa | Equivalent noise level, dB(A) | Directional characteristics |
|---|---------------------|--------------------|-------------------------------|-----------------------------|
| Condenser | | | | |
| Measuring microphone, dc-biased, 1-in. (25-mm) diameter | 3-8000 | 50 | 10 | Omnidirectional |
| Measuring microphone, dc-biased, 0.12-in. (3-mm) diameter | 7-140,000 | 1.0 | 55 | Omnidirectional |
| Measuring microphone, electret, 0.5-in. (12-mm) diameter | 7-18,000 | 50 | 14 | Omnidirectional |
| Low-noise measuring microphone, dc-biased, 1-in. (25-mm) diameter | 7-10,000 | 100 | -2.5 | Omnidirectional |
| Studio microphone, RF-biased | 12-20,000 | 25 | 10 | Omnidirectional |
| Lavalier microphone, electret | 100-20,000 | 8 | 30 | Cardioid |
| Studio shotgun microphone, RF-biased | 50-20,000 | 40 | 9 | Club-shaped |
| Miniature microphone, electret, 0.16 × 0.12 × 0.08 in. (4 × 3 × 2 mm) | 100-5000 | 10 | 25 | Cardioid |
| Boundary microphone, electret | 20-20,000 | 20 | 27 | Hemisphere |
| Silicon (MEMS) microphone | 100-10,000 | 8 | 35 | Omnidirectional |
| Piezoelectric | | | | |
| Telephone transmitter, unimorph | 300-3400 | 4 | | Omnidirectional |
| Dynamic | | | | |
| Studio microphone | 30-20,000 | 1.3 | 30 | Omnidirectional |
| Directional studio | 30-20,000 | 1.8 | 30 | Superdirectional |

the microphones can also be used as sound generators. Very recently, micromachining techniques were introduced into transducer production for fabricating some electrostatic and piezoelectric microphones in silicon. These sensors are referred to as MEMS (microelectromechanical systems) or silicon microphones. Another new type is the optical microphone which converts acoustic signals into modulations of light waves; such an acoustic sensor has an optical rather than an electrical output. The classical carbon microphones, while used extensively in telephony in the past, have been almost completely replaced by other types of microphones. *See* LOUD-SPEAKER.

Electrostatic (condenser) microphones. These consist of a fixed electrode (the backplate), a movable electrode (the diaphragm), and an air gap between the electrodes. To decrease the acoustic stiffness of the airgap, which is generally about 20 to 30 micrometers (0.8 to 1.2 mils) thick, the backplate is often perforated with holes connecting the air gap to a larger air cavity. The diaphragm is a thin [typically 4 to 6 μm (0.16 to 0.24 mil) nickel, Duralumin, or metalized plastic (for example, Mylar) foil under mechanical tension. The resonance frequency of the microphone is determined by the mass of the diaphragm and by the restoring forces, consisting of the tension of the diaphragm and the stiffnesses of the diaphragm and of the air gap. Below resonance, the system is controlled by the restoring forces and has a frequency-independent sensitivity. To achieve such a frequency-independent sensitivity, the resonance frequency is therefore placed at the upper end of the audio-frequency range. The two electrodes form a capacitor whose capacitance is very small, typically between 2 and 100 picofarads. The electrical impedance of the transducer is therefore relatively large. *See* CAPACITANCE; CAPACITOR; ELECTRICAL IMPEDANCE; RESONANCE (ACOUSTICS AND MECHANICS).

Impinging sound waves produce a motion of the diaphragm and a corresponding variation of the capacitance. This variation can be utilized in various ways to generate electrical output signals. Accordingly, three types of condenser microphones are customary, namely direct-current-biased systems, electret-biased systems, and radio-frequency (RF)-biased systems.

In the dc-biased system (Fig. 1a), a polarizing voltage of about 200 V (or 20 V in portable systems) is applied between the microphone electrodes. Electrical output signals are thus generated by the diaphragm deflections. These are fed into a cathode follower with an input impedance R of typically 10^8 ohms or more. For frequencies larger than $1/RC$, where C is the microphone capacitance, the voltage delivered to the cathode follower is proportional to the sound pressure. To avoid capacitive loading of the transducer by stray capacitances and to eliminate electric pickup of RF signals, the cathode follower is directly built into the microphone case.

In the electret-biased system (Fig. 1b), a permanently charged dielectric (electret) is inserted be-

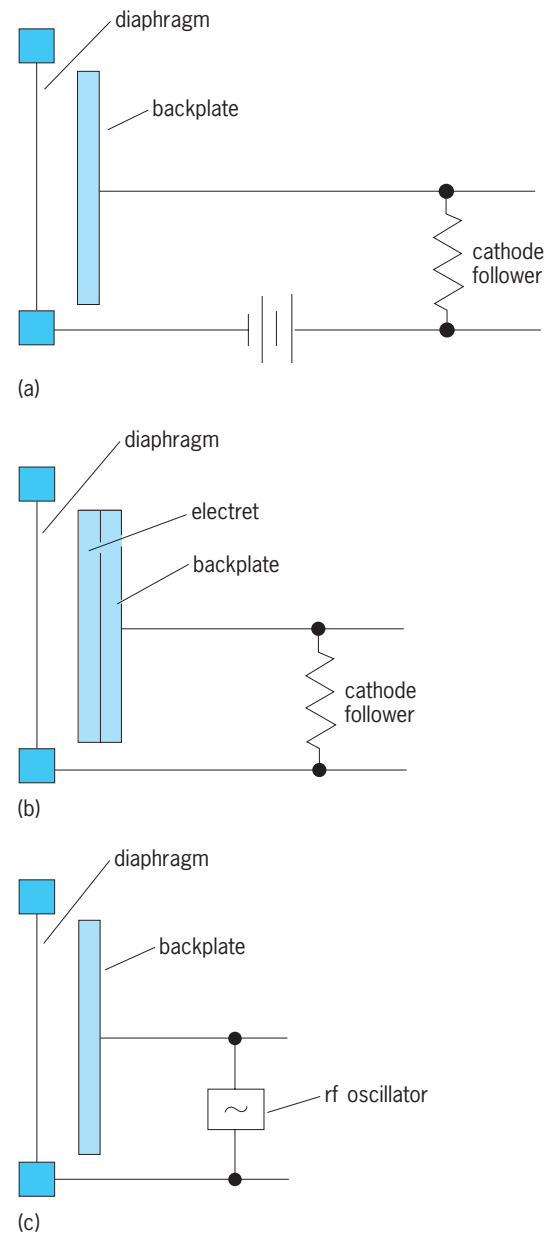


Fig. 1. Electrostatic transducers: (a) dc-biased system, (b) electret-biased system, (c) RF-biased system.

tween the two electrodes of the transducer. The electret, usually made of a 12 to 25 μm (0.5 to 1.0 mil) polymer film, such as poly(tetrafluoroethylene), either is cemented onto the backplate or is metalized on one side and used as the diaphragm of the microphone. The amount of charge is chosen so that it corresponds to a bias voltage of about 200 V. As in dc-biased systems, the electric output signal of the transducer is fed into a cathode follower. *See* ELECTRET TRANSDUCER.

In the RF-biased system (Fig. 1c), the capacitance of the transducer determines the frequency of oscillation of an RF circuit, usually about 10 MHz. Motion of the diaphragm produces a frequency-modulated signal. On demodulation, a low-frequency signal corresponding to the sound pressure is obtained. Unlike dc-biased systems, the RF-biased transducer

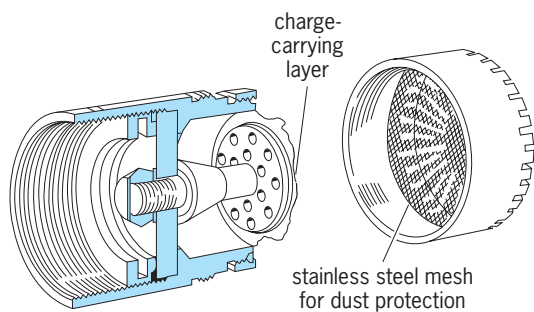


Fig. 2. Measuring condenser microphone. (Brüel & Kjaer)

has no low-frequency cutoff. See FREQUENCY MODULATION.

Condenser microphones are renowned for their excellent acoustic qualities such as flat frequency response, high sensitivity, large dynamic range, and small vibration sensitivity. Also important is their suitability for miniaturization, with the smallest units having dimensions of only about $0.12 \times 0.12 \times 0.08$ in. ($3 \times 3 \times 2$ mm). They can be designed as precision instruments (Fig. 2) and as such are widely used in measurement and in high-fidelity sound production. Other applications, particularly of electret microphones, are in telephones, cassette recorders, camcorders, hearing aids, and toys. Electret microphones are the most widely used microphones with worldwide annual production of about 2 billion units (2006). See HEARING AID; MAGNETIC RECORDING; SOUND RECORDING; TELEPHONE.

Piezoelectric microphones. These consist of a material having piezoelectric properties. A deformation of the material leads to the generation of a voltage which corresponds to the deformation. Piezoelectric materials can be crystals, polycrystalline ceramics, or semicrystalline polymers. The best-known piezoelectric crystals are quartz and ammonium dihydrogen phosphate (ADP). Representative of polycrystalline ceramics are lead zirconate titanate (PZT) and barium titanate, which are initially electrostrictive; they have to be poled, that is, exposed to a high electric field at elevated temperatures, to become piezoelectric. An example of a semicrystalline polymer is poly(vinylidene fluoride) [PVDF]. It is also made piezoelectric by poling. See ELECTRET; ELECTROSTRICTION; PIEZOELECTRICITY.

Piezoelectric microphones for audio and near-ultrasonic frequencies are made from these materials in the form of flexure-mode transducers, such as bimorphs, unimorphs, or curved benders.

The bimorph consists of two piezoelectric disks, usually made of PZT, which are cemented together. Either the disks are poled in the same direction with the outputs connected in parallel (Fig. 3a), or they are poled in opposite directions with the outputs connected in series (Fig. 3b). Since bending of the structures shortens one disk and lengthens the other, the voltage signals from the two disks add in both cases. The unimorph (Fig. 3c) consists of a single piezoelectric disk cemented to a thin metal plate. It has the advantage of being less fragile than the

bimorph. Bimorphs and unimorphs are made in rectangular, square, or circular shapes.

To be operational, these elements have to be mounted in some way. Frequently applied is edge mounting along the entire edge or, for rectangular disks, cantilever mounting at one of the short sides of the rectangle. The transducers are then either directly actuated by the sound waves or actuated by means of a membrane connected by a stylus to some properly chosen point of the system.

The single-element bender (Fig. 3d) consists of a curved piezoelectric foil, usually PVDF, which is clamped around part of its edge. Preferred geometries are cylindrical sections or domes. If actuated by a sound wave, a voltage signal due to the piezoelectric effect.

In piezoelectric microphones, the resonance frequency is placed at the upper end of the audio range. At much lower frequencies, the system is stiffness-controlled. Since the output voltage is proportional to the displacement, a constant sensitivity results in this range. Because the resonance is very prominent, it has to be damped to ensure relatively constant sensitivity around this frequency.

Well-designed piezoelectric microphones have acceptable quality. A drawback is the relatively high vibration sensitivity. Directly actuated unimorph microphones are still in occasional use in telephones in some countries. Unimorphs and bimorphs are also employed in the near-ultrasonic range at frequencies up to about 100 kHz.

Dynamic microphones. These consist of a conductor located in the gap of a permanent magnet. Motion of the conductor produces a voltage proportional to its velocity. Depending on the kind of conductor used, moving coil and ribbon microphones are distinguished.

In the moving-coil microphone (Fig. 4) the coil, often referred to as voice coil, is connected to a diaphragm actuated by the sound waves. Motion of the coil induces a voltage proportional to its velocity. To obtain a frequency-independent sensitivity, the coil must respond to the sound pressure with frequency-independent velocity. This is accomplished by

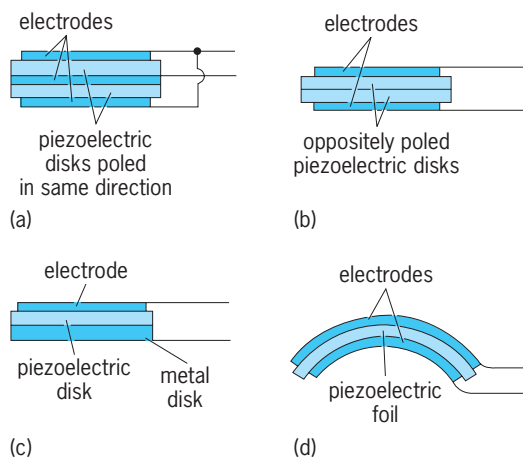


Fig. 3. Piezoelectric transducers: (a) parallel bimorph, (b) series bimorph, (c) unimorph, (d) curved bender.

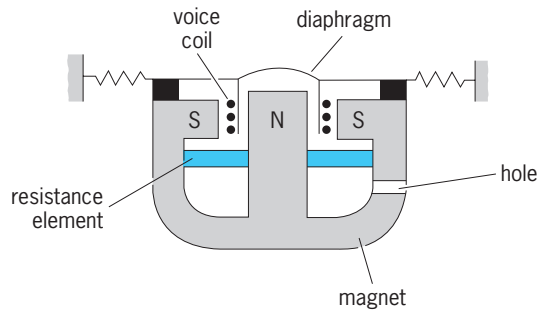


Fig. 4. Dynamic microphone of the moving-coil type.

resistance-controlling the system: the acoustical resistance is made larger in magnitude than the acoustical reactance due to the mass of the diaphragm and coil and due to the compliance of the suspension. A silk cloth or a piece of felt placed behind the voice coil is used for this purpose. The major resonance frequency of the system is usually placed in the range of 500 to 1000 Hz. To equalize the sensitivity in the lower and upper parts of the audio range, other resonances of the microphone are used. In modern moving-coil microphones, the diaphragm is made of a plastic film, for example, 15- μm (0.6-mil) Mylar. The impedance of the voice coil is typically 200 to 1000 ohms. See ACOUSTIC IMPEDANCE.

In the ribbon microphone, a metallic ribbon is placed in a magnetic field. As in the moving-coil microphone, the ribbon has to be terminated into an acoustical resistance that is large compared to the reactance of the system if a frequency-independent sensitivity is desired.

Dynamic microphones are relatively complicated systems. If well designed, they are of good quality. Drawbacks are the difficulties encountered in miniaturization and the relatively high vibration sensitivity. Moving-coil microphones are still widely used in high-fidelity, radio, television, and concert applications. In many other areas they have been replaced by electret microphones. Ribbon micro-

phones were very popular in the past but are now rarely employed.

Magnetic microphones. These consist of a diaphragm connected to an armature which, when vibrating, varies the reluctance in a magnetic field. The variation in reluctance leads to a variation in the magnetic flux through a surrounding coil and therefore to an induced voltage. This voltage is proportional to the velocity of the armature. To obtain a frequency-independent sensitivity, the velocity of the armature in response to the sound pressure must be independent of frequency. As in dynamic microphones, this is accomplished by resistance-controlling the system, for example, by placing an acoustic resistance behind the diaphragm.

Magnetic microphones are relatively complicated and have poor frequency response and high vibration sensitivity. While never extensively used, they have now disappeared completely. However, the magnetic principle is still used in telephone receivers and in earphones employed in hearing aids. See EARPHONES.

Silicon (MEMS) microphones. Microelectronic processing methods allow fabrication of batch-processed, high-performance sensors. Among these methods are doping, deposition, oxidation, lithography, and, as key technologies, various etching processes. The last are usually subdivided into dry and wet methods. The dry processes, such as plasma, sputter, ion, and laser etching, allow one to fabricate holes with walls perpendicular to the surface of the substrate. The wet methods, subdivided into isotropic and anisotropic processes, usually result in curved walls or walls extending under a certain angle to the surface. An important variant of the etching processes is sacrificial-layer etching where a buried layer of material is etched out through holes in the covering layer. The etching technologies enable fabrication of membranes, holes, pits, and recesses as required in acoustic sensors. The methods in their entirety, if used to make three-dimensional mechanical structures, are referred to as micromachining, and the resulting devices are called microelectromechanical systems, or MEMS.

Acoustic MEMS sensors or silicon microphones utilizing the condenser and the piezoelectric principles have been built since the mid-1980s. In addition, new concepts of transducer design, such as the modulation of the drain current of a field-effect transistor or the modulation of light propagation in an optical waveguide by the sound waves, have been realized in silicon. Silicon microphones based on the condenser principle are now commercially available. See FIBER-OPTIC SENSOR; TRANSISTOR.

Silicon condenser microphones can be designed as two-chip or one-chip devices. A typical two-chip transducer (Fig. 5) consists of a membrane chip and a back electrode chip where the two chips, which are bonded together, are separated by a thin air layer. Since bonding of two chips complicates the design, modern silicon condenser microphones consist of only a single chip.

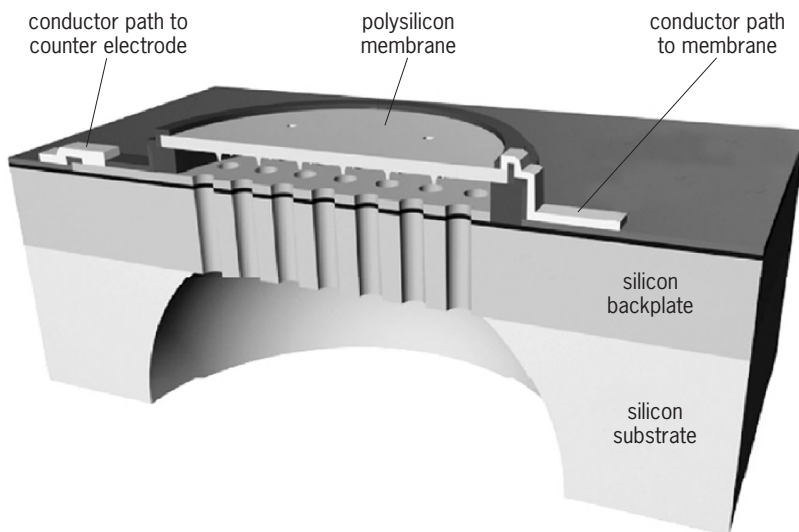


Fig. 5. One-chip silicon condenser microphone consisting of a 1-mm polysilicon membrane, an approximately 2- μm air gap, and a highly perforated silicon back plate.

The cross section of a typical one-chip condenser microphone is shown in Fig. 5. The sensor is fabricated by first growing an epitaxial and doped silicon layer which is highly perforated and which later serves as the backplate. Next, a sacrificial layer of silicon dioxide is deposited whose thickness determines the height of the air gap. On top of this, a polysilicon layer is placed and metalized on its upper side. Thereafter, the wafer backside is wet-etched and the sacrificial oxide is removed through the perforation holes.

The electroacoustic properties of such microphones depend strongly on the geometric, mechanical, and electrical characteristics. For typical microphones with membrane dimensions of $1 \times 1 \text{ mm}^2$ and membrane thicknesses of 0.2 to $0.4 \text{ }\mu\text{m}$, the resonance frequency is in the near-ultrasonic range, and the sensitivity in the audio-frequency range is about 10 mV/Pa for a bias voltage of 3 V (see table).

Compared to conventional condenser microphones, silicon microphones are extremely shock-resistant and have, due to the small mass density of the membrane, a very low vibration sensitivity. The MEMS microphones are also, compared to electret microphones, much less sensitive to exposure to high temperatures. They can be permanently exposed to temperatures of 100°C and are resistant to temperatures up to 260°C for short exposure times. Because they withstand the heat prevailing during standard reflow soldering processes and because of their small height, such microphones can be inserted as surface-mounted devices (SMDs) in circuit boards on automated production lines. Some new silicon SMD microphones possess an integrated AD converter and thus deliver a digital output signal. Silicon microphones are now in commercial use in cellular phones, laptops, PDAs, digital cameras, MP3 recorders, and other devices. Although introduced rather recently, these microphones are already produced in quantities of several 100 million annually.

Piezoelectric silicon microphones are always one-chip designs (Fig. 6). They consist of a diaphragm, usually an approximately $1\text{-}\mu\text{m}$ -thick silicon nitride layer, which carries the lower electrode. Onto this, a piezoelectric layer of about equal thickness is deposited. It consists of either a ceramic such as lead zirconate titanate, or a polymer such as polyvinylidene fluoride. Another metal layer forms the upper (second) electrode. Microphones of this kind show somewhat higher noise levels than the condenser types.

Silicon microphones have several advantages as compared to conventional microphones. They can be made considerably smaller with membrane areas of only about 1 mm^2 , as opposed to about 5 mm^2 for the smallest conventional transducers. They also have very low vibration sensitivity due to the use of thin diaphragms. They are thus not susceptible to pickup originating from vibrations due to walking or other motions or caused by vibration sources such as motors in cassette recorders or camcorders. Furthermore, they can be produced together with proper signal-processing electronics on the same chip with

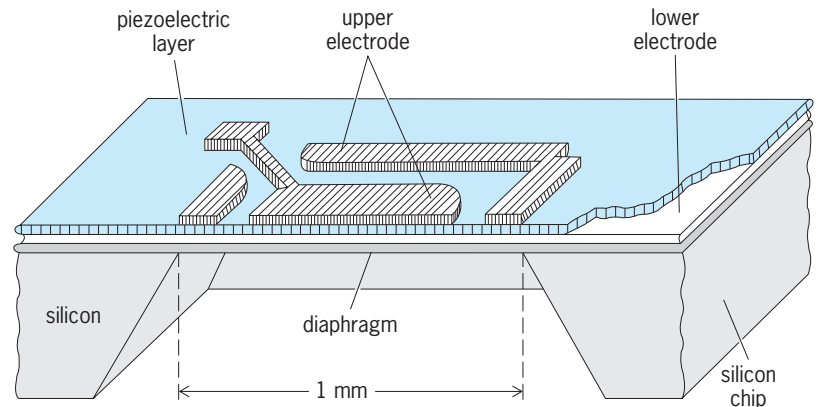


Fig. 6. Piezoelectric microphone consisting of a single silicon chip with a piezoelectric layer integrated into the diaphragm.

the same semiconductor methods. Finally, they can be made inexpensively through batch-processing techniques.

Directional Characteristics

The microphones described above are inherently nondirectional (omnidirectional) as long as they are small compared to the wavelength and if the microphone case shields the rear side of the diaphragm from the sound waves. This means that they pick up sound from all directions with equal sensitivity. They are therefore referred to as pressure microphones since pressure is a scalar and not a vector quantity.

In many applications, however, microphones with other-than-omnidirectional characteristics are desired. Such microphones [have a reduced response for all but certain directions and thus are capable of suppressing unwanted sounds. This results in] an improvement of the signal-to-noise ratio [for the desired acoustic signals.] A great variety of directivity patterns can be achieved with microphones based on either the gradient or the dimensional scheme. Combinations of both principles are also useful.

Gradient scheme. The directivity of a first-order gradient microphone depends on the pressure difference between two microphone elements separated by a distance that is small compared to the wavelength. Second-order gradients are formed by subtracting the outputs of two displaced first-order gradients. Analysis shows that the directivity of such systems is always independent of frequency and transducer dimensions as long as the above condition about the separation of the elements holds. Three common directivity patterns achievable with first-order gradients are shown in Fig. 7.

A simple gradient microphone with the bidirectional sensitivity pattern shown in Fig. 7a is obtained by exposing the diaphragm of a transducer to the sound field from both sides. Such microphones, because of the large sensitivity in the backward direction, are infrequently used.

Of great practical use, however, are the superdirectional and unidirectional microphones based on the gradient principle whose directivity patterns are

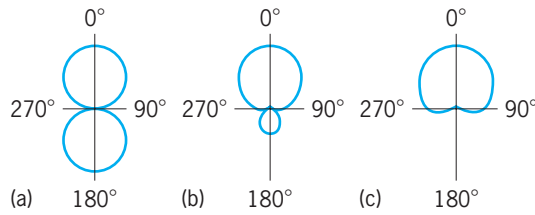


Fig. 7. Polar directivity patterns: (a) bidirectional, (b) superdirectional (supercardioid), (c) unidirectional (cardioid).

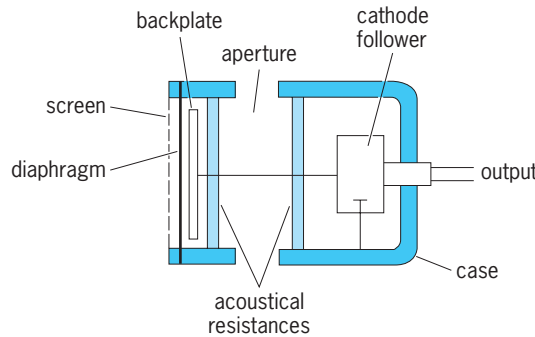


Fig. 8. Sectional view of a unidirectional microphone employing a condenser transducer.

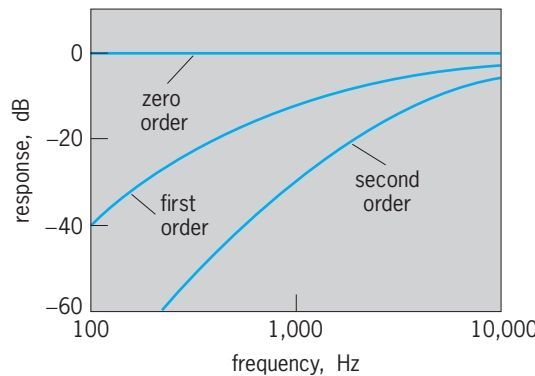


Fig. 9. Frequency-response characteristics of zero-, first-, and second-order gradient microphones, assuming a frequency-independent characteristic of the zero-order system.

shown in Fig. 7b and c. Such transducers are also made by exposing the diaphragm to the sound field from both sides, but they require an additional delay of the sound waves traveling to the back of the diaphragm. A unidirectional microphone employing a condenser transducer is schematically depicted in Fig. 8. An acoustical resistance is located behind the diaphragm to provide resistance control of the vibrating system. The back of the diaphragm is coupled to the acoustic inertance of the ports in the side of the case and the acoustic resistance and acoustic capacitance of the air volume in the case. The phase shift introduced by the acoustic network consisting of the inertance, resistance, and capacitance corresponds to the distance from the ports to the front of the diaphragm. As a result, for sound arriving from the back, the forces on the front and back of the diaphragm are almost equal in phase and amplitude, and the response is low. The phase shift is at a maxi-

imum for sound arriving in a forward direction, with resultant maximum sensitivity. Thus, a unidirectional directivity pattern is obtained.

As compared to the sensitivity of a pressure (zero-order gradient) microphone, the sensitivity of a first-order gradient microphone to plane waves from a far sound source decreases with decreasing frequency. This is shown in Fig. 9 where a flat response for the zero-order gradient is assumed. An even greater sensitivity loss toward low frequencies is found in second-order gradients. However, all gradients show equal sensitivity to the spherical waves of a close sound source. The first- and second-order gradients can thus be used as “close-talking” microphones with frequency-independent sensitivity, having the advantage of suppressing the low-frequency components from far-away noise sources.

Dimensional scheme. The dimensional-directivity scheme depends on interference of signals with phase differences which are exclusively due to the geometry of the system. For the microphone to be directive, the dimensions of the transducer have to be comparable to the wavelength. As opposed to gradient-type microphones, the directivity of such microphones is dependent on the ratio of linear dimensions to wavelength, as shown in Fig. 10 for a line microphone, that is, a microphone picking up sound along a line. As this ratio increases, the directivity changes from a cardioid to a club-shaped pattern. Other examples of dimensional microphones are transducer arrays and reflectors. See INTERFERENCE OF WAVES.

An implementation of the dimensional scheme in the form of a line microphone consists of a long tube which has a large number of sound inlets along its side. An omnidirectional microphone is placed at the end of the tube (Fig. 11). For incidence along the axis of the tube in the direction toward the

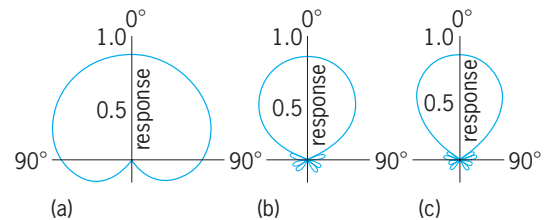


Fig. 10. Directivity patterns for a line microphone at different frequencies. The response maximum is arbitrarily chosen as unity. (a) Length of line = $\lambda/2$, where λ = wavelength (low frequencies). (b) Length of line = 2λ (intermediate frequencies). (c) Length of line = 8λ (high frequencies).

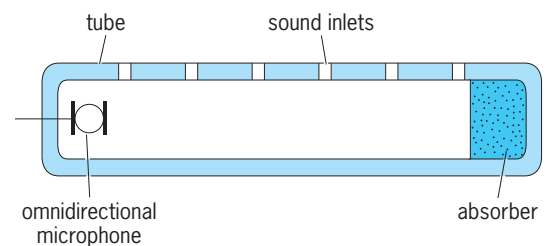


Fig. 11. Schematic section of a line microphone.

omnidirectional microphone, the sound waves entering the tube through any of the holes are all in phase, the interference is constructive, and maximum sensitivity results. For incidence at other angles, phase differences between the sound waves entering the tube through one hole and sound waves entering through another hole appear, and the interference is fully or partially destructive. The directivity patterns are those of Fig. 10. Systems of this kind, referred to as shotgun microphones, are in wide use.

The boundary microphone also uses the dimensional scheme. Such a microphone consists of an omnidirectional transducer set into a disk. At high frequencies, where the wavelength is smaller than the lateral dimension of the disk, sound waves arriving from the front direction experience a 6-dB boost, while sound waves arriving from the rear direction are attenuated. This results in an approximately hemispherical directional (characteristics) of the boundary microphone.

The dimensional scheme is often combined with the gradient principle to improve the directivity of cardioid or supercardioid microphones at higher frequencies. For example, a cardioid capsule can be used in a boundary microphone with the effect of sharpening the unidirectional characteristic.

Sensitivity Calibration

A number of methods may be employed for determining the sensitivity of microphones. The pistonphone and the reciprocity calibration are widely used.

A pistonphone is used to calibrate pressure-type microphones (Fig. 12). The small piston is driven by a crank and thus produces a pressure that can be exactly calculated from the geometry of the system. The pistonphone method is useful for calibrating microphones in the low-frequency range. The upper limit is governed by the permissible speed of the mechanical system, which corresponds to approximately 300 Hz. Pressure and free-field sensitivities do not differ in this range. The accuracy of the pistonphone-calibration method is ± 0.2 dB.

The reciprocity method allows sensitivity measurements in the entire audio-frequency range and beyond, and can be adapted to yield either the pressure or the free-field sensitivity of microphones. In both cases, an auxiliary reversible transducer S_1 and an auxiliary loudspeaker S_2 are required in addition

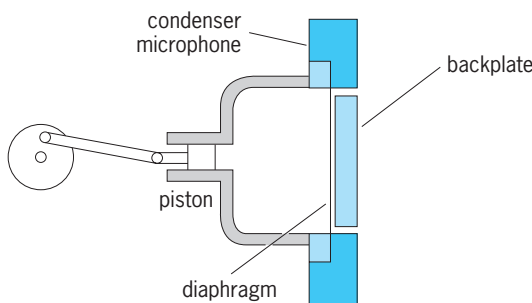


Fig. 12. Sectional view of a pistonphone.

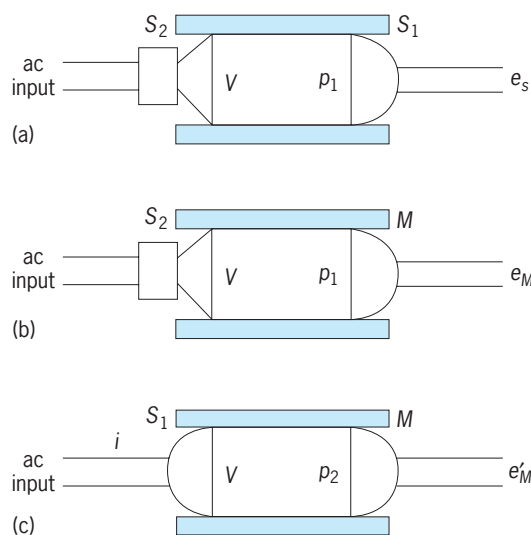


Fig. 13. The three experiments of the reciprocity calibration for obtaining the pressure sensitivity of a microphone. (a) Open-circuit voltage e_s of the reversible microphone loudspeaker S_1 , when used as a microphone and actuated by a sound pressure p_1 . (b) Open-circuit voltage e_M of the microphone M to be calibrated, when actuated by a sound pressure p_1 . (c) Open-circuit voltage e'_M of the microphone M to be calibrated, when actuated by a sound pressure p_2 produced by the reversible microphone loudspeaker S_1 used as a loudspeaker with a current input i in a coupling volume V .

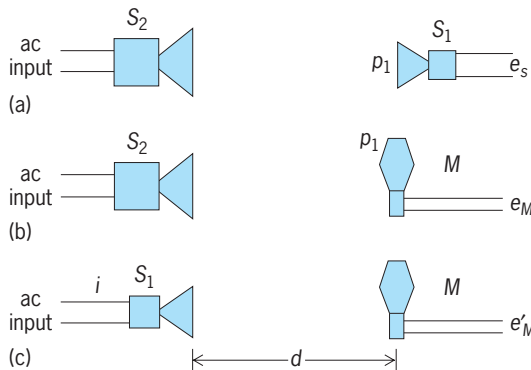


Fig. 14. The three experiments of the reciprocity calibration for obtaining the free-field sensitivity of a microphone. (a) Open-circuit voltage e_s of the reversible microphone loudspeaker S_1 when used as a microphone and actuated by a sound pressure p_1 . (b) Open-circuit voltage e_M of the microphone M to be calibrated, when actuated by a sound pressure p_1 . (c) Open-circuit voltage e'_M of the microphone M to be calibrated, when actuated by a sound pressure produced by the reversible microphone loudspeaker S_1 used as a loudspeaker with a current input i and a spatial separation d .

to the microphone M to be calibrated. A number of experiments have to be performed to determine the sensitivity of M .

The procedure for the pressure-response calibration is schematically shown in Fig. 13. In Fig. 13a, an alternating current is fed to the loudspeaker S_2 . A sound pressure p_1 is produced in the volume V . It generates an open-circuit voltage e_s of S_1 . In the next experiment (Fig. 13b), the same current is fed to the loudspeaker S_2 , resulting in an open-circuit voltage e_M of the microphone M . In the third experiment, a current i is fed to S_1 which is now used as a

loudspeaker. This produces the output voltage e'_M of M . The sensitivity K_M of the microphone M is then given by Eq. (1), where $\omega = 2\pi f$, f is the frequency,

$$K_M = \sqrt{\frac{\omega V e_M e'_M}{\rho_0 c^2 i e_s}} \quad (1)$$

ρ_0 is the density of air, and c is the sound velocity.

The free-field calibration by means of the reciprocity procedure is shown in Fig. 14. In this case, the sensitivity K_M of the microphone M is given by Eq. (2), where e_s , e_M , e'_M , and i are obtained from the

$$K_M = \sqrt{\frac{2d\lambda e_M e'_M}{r_A i e_s}} \quad (2)$$

experiments in Fig. 14, d is the distance as shown in the figure, λ is the wavelength of the sound wave, and r_A is the characteristic impedance of air.

The reciprocity calibration is a very accurate method since it depends solely on measurements of electrical quantities and length dimensions. Its accuracy is estimated to be ± 0.02 dB. Gerhard M. Sessler

Bibliography. J. Eargle, *The Microphone Book*, 2d ed., Focal Press-Elsevier, 2004; G. Elko, Microphone arrays, in T. Rossing (ed.), *Springer Handbook of Acoustics*, Springer, 2006; E. L. Hixson and I. J. Busch-Vishniac, Transducer principles, pp. 1375-1388, and I. J. Busch-Vishniac and E. L. Hixson, Types of microphones, pp. 1411-1422, in M. J. Crocker (ed.), *Handbook of Acoustics*, Wiley, 1998; R. Lerch, Sensors for measuring sound, in W. Göpel, J. Hesse, and J. N. Zemel (eds.), *Sensors: A Comprehensive Survey*, vol. 7, Wiley-VCH, 1993; G. M. Sessler, Silicon microphones, *J. Audio Eng. Soc.*, 44:16-22, 1996; G. S. K. Wong and T. F. W. Embleton (eds.), *AIP Handbook of Condenser Microphones*, AIP Press, 1995; G. Wong, Microphones and their calibration, in T. Rossing (ed.), *Springer Handbook of Acoustics*, Springer, 2006.

Microprocessor

A device that integrates the functions of the central processing unit (CPU) of a computer onto one semiconductor chip or integrated circuit (IC). In essence, the microprocessor contains the core elements of a computer system, its computation and control engine. Only a power supply, memory, peripheral interface ICs, and peripherals (typically input/output and storage devices) need be added to build a complete computer system. See COMPUTER PERIPHERAL DEVICES.

Internal architecture. A microprocessor consists of multiple internal function units. A basic design has an arithmetic logic unit (ALU), a control unit, a memory interface, an interrupt or exception controller, and an internal cache. More sophisticated microprocessors might also contain extra units that assist in floating-point math calculations, program branching, or vector processing (see *illus.*).

The ALU performs all basic computational operations: arithmetic, logical, and comparisons.

The control unit orchestrates the operation of the other units. It fetches instructions from the on-chip cache, decodes them, and then executes them. Each instruction has the control unit direct the other function units through a sequence of steps that carry out the instruction's intent. The execution path taken by the control unit can depend upon status bits produced by the arithmetic logic unit or the floating-point unit (FPU) after the instruction sequence completes. This capability implements conditional execution control flow, which is a critical element for general-purpose computation. See BIT.

The memory interface enables the microprocessor to maintain two-way communication with off-chip semiconductor memory, which stores programs and data. This interface typically supports memory reads and writes in blocks of words (the number of bits that the processor operates on at one time). The block size facilitates burst data transfers to and from the chip's internal cache. See SEMICONDUCTOR MEMORIES.

The interrupt or exception controller enables the microprocessor to respond to requests from the external environment or to error conditions by allowing interruptions of the ongoing operation. An interrupt might be an external peripheral requesting service, while an exception typically consists of a floating-point math error or an unrecognized instruction. The interrupt controller can prioritize and selectively handle these interrupts.

The internal cache is an on-chip memory storage area that holds recently used data values or instruction sequences that are likely to be used again in the near future. Since this information is already on-chip, it can be accessed rapidly, thereby accelerating the computation rate. Items not in the cache can take several or more extra operations to access, which significantly degrades the computation rate. Software writers often organize a program's code and data structures so that the most frequently used elements often occupy the cache, thus maintaining a high level of computational throughput. See COMPUTER ARCHITECTURE; COMPUTER STORAGE TECHNOLOGY.

Computational power. The computation power of microprocessors can be roughly characterized by their word size and computational rate (operations per second). The earliest microprocessors, introduced in the mid-1970s, were 4- and 8-bit machines operating at a clock rate of 1 MHz. They could process a few hundred thousand operations per second. By the start of the twenty first century, microprocessors had evolved into 32- and 64-bit machines operating at clock speeds of 500 MHz to 1 GHz, and capable of processing rates from 100 million to nearly 2 billion operations per second. See DIGITAL COMPUTER; MICROCOMPUTER.

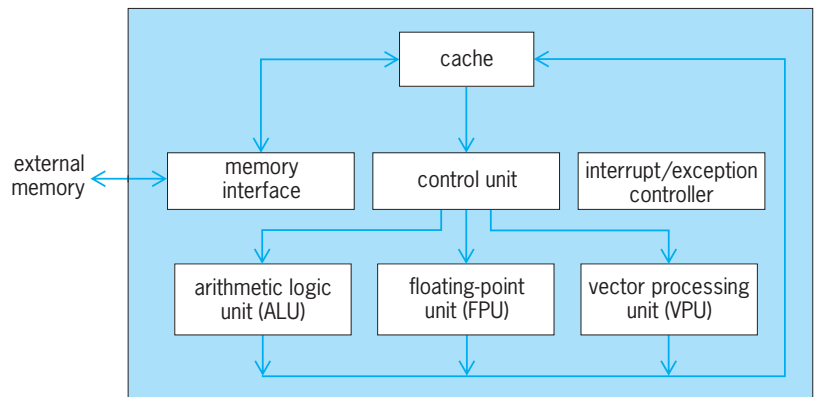
The rapid growth of the microprocessor's computing power, especially in the 1990s, was due to two factors. First, the rapid advance of semiconductor fabrication technology reduced the dimensions of the transistor (the basic building block of a digital

computer) from hundreds of micrometers to fractions of a micrometer. This allows more transistors to be packed onto the chip. The smaller transistors can also switch faster, resulting in faster computations. Today's microprocessors often contain 10 million transistors, with some having more than 20 million transistors. Over half of the transistors in these microprocessors are part of the cache unit, which indicates how critical this unit is in maintaining computational throughput. *See* INTEGRATED CIRCUITS; TRANSISTOR.

Improvements to the microprocessor's architectural design is the second contributor to its improved performance. Such improvements include simplified instruction sets and sophisticated function units that operate in parallel. The latter ability enables the microprocessor to support multiple issue, a technique where several instructions start at the same time. This is done by the control unit issuing instructions to the other autonomous function units. Function units also support pipelining, where an instruction's control sequence is divided into distinct stages (for example, decode, data fetch, execute, data write). Much like an auto assembly line, each stage of the pipeline contains a partially processed instruction. On each tick of the processor clock, one instruction exits the last pipeline stage, other instructions move to subsequent processing stages, and another instruction enters the pipeline. The net result is that the microprocessor is capable of executing one instruction per clock cycle. The combination of multiple instruction issue and instruction pipelining effectively multiplies the amount of work that the microprocessor accomplishes during a clock period. *See* CONCURRENT PROCESSING.

Another trend improving microprocessor performance is the integration of more functions onto the chip. For example, the floating-point unit used to be a separate processor, but it is now part of a microprocessor's complement of function units. This accelerates floating-point math calculations since the floating-point data and instructions now execute on-chip rather than on an external device. Some microprocessors have gone further in this area by integrating either peripheral integrated circuits or vector processing units onto the chip.

RISC versus CISC. The design of instruction sets (the commands that produce basic work when executed by the microprocessor) often influences the design of the microprocessor itself. Instruction sets—and as a consequence, the microprocessor architecture—fall into two camps: reduced instruction set computers (RISC) and complex instruction set computers (CISC). Because of the limits of early computer technology, most computers were by necessity RISC machines. Since most of the software was written in assembly language (that is, a programming language that represented the program's intent in actual machine instructions), there was a drive to build instruction sets of greater sophistication and complexity. These new CISC instruction sets made assembly language programming easier, but they also made it difficult to build high-speed computer hard-



A microprocessor consists of multiple independent function units. The memory interface fetches instructions from, and writes data to, external memory. The control unit issues one or more instructions to other function units. These units process the instructions in parallel to boost performance.

ware. First, CISC instructions were harder to decode (a block move instruction might have many different parameters that the control unit had to recognize). In addition, since CISC instructions involved long and complex operation sequences, they incurred a major cost by requiring more complicated logic to implement. Second, such instructions were also difficult to interrupt or abort if an exception occurred. Finally, such instructions usually carried many data dependencies that made it more difficult to support advanced architectural techniques such as pipelining and multiple issue. By returning to a RISC design, much faster computers can be built. In fact, an enhancement in performance by a factor of 2 to 3 has been attributed to this simple organizational change. To achieve these efficiencies, most of the RISC microprocessor's function units must be kept as busy as possible. This requires optimizing compilers that can translate a program's high-level source code and then reorder the resulting low-level instructions in such a way as to ensure the high throughput. *See* COMPUTER PROGRAMMING; PROGRAMMING LANGUAGES.

Microprocessor families. There are several widely known microprocessor families. Like all computer families, microprocessor families are organized around software compatibility. Each new member of the family must be capable of executing the programs that run on its progenitors, preserving the substantial investment in software.

The most widely used microprocessor family, the x86, is a CISC architecture. This sacrifices some performance, but offers compatibility with software whose roots can be traced to the early 1980s. (However, many current x86 microprocessors actually decode the CISC instructions into RISC-like "micro operations" so that they can take advantage of pipelining and multiple issue techniques.)

The x86 family consists of Intel's 8080, 8085, 8086, 80286, 80386, 80486, the Pentium, Pentium II, and Pentium III processors. Other members of this family include Advanced Micro Device's K6 and Athalon processors. In this family, the Intel 80386 was the first 32-bit microprocessor. The 80486 integrated the floating-point unit on-chip, and later x86 family

processors used on-chip caches and multiple instruction issue to further boost performance.

The best representative of RISC processors is the PowerPC family, a joint design effort by Motorola and IBM. It consists of the PowerPC 601, 603, 603e, 604, 604e, 750 (also known as G3), and 7400 (also known as G4). The first PowerPC-based desktop computer appeared in 1994. The PowerPC is a 32-bit microprocessor, and features on-chip caches and an integrated floating-point unit. The G4 processor integrates a vector processing unit on-chip, which is valuable for three-dimensional graphics and scientific applications. Although the clock speed of the PowerPC family currently lags behind that of the x86 family, its RISC architecture, combined with instruction pipelining and multiple instruction issue, places its performance on par with that of the fastest x86 processors. *See* COMPUTER GRAPHICS.

Applications. In general, microprocessors can easily and effectively replace custom, large-scale integrated circuits in a variety of applications beyond basic computers. More importantly, the functions and features of a device—such as a network switch—can be readily improved or modified by simply installing new software into the device, rather than replacing custom hardware.

Microprocessors are found in virtually every consumer product that requires electric power, such as microwave ovens, automobiles, video recorders, cellular telephones, digital cameras, and hand-held computers. High-performance microprocessors implement the servers that store and distribute Web content, such as streaming audio and video, desktop computers, and the high-speed network switches that constitute the Web's infrastructure. More modest-powered microprocessors are at the heart of notebook computers and electronic games. Low-power microprocessors provide the control and flow logic of hand-held devices, digital cameras, cellular and cordless phones, pagers, and the diagnostic and pollution control of automobile engines. *See* INTERNET; VIDEO GAMES; WIDE-AREA NETWORKS; WORLD WIDE WEB.

Tom Thompson

Bibliography. K. Diefendorff, *The Microprocessor Report*, August 2, 1999; T. Halfhill, *BYTE*, April 1995; K. Short, *Microprocessors and Programmed Logic*, 2d ed., 1987; M. Slater, *A Guide to RISC Microprocessors*, 1992; M. Slater, *Microprocessor-Based Design*, 1989; J. Stewart, *Microprocessor Systems*, 1990; D. Tabak, *Advanced Microprocessors*, 2d ed., 1994.

Micropygoida

An order of regular echinoids belonging to the Diadematacea, established for the two recognized species of *Micropyga*. They have an aulodont lantern with grooved teeth. Test plating is imbricate, and ambulacra are composed of trigeminate compound plates in which upper and lower elements are reduced to demiplates. Pore pairs are biserially arranged in ambulacral columns, and there are unique umbrellalike

aboral tube feet. *Micropyga* is a deep-water echinoid, found between 480 and 4290 ft (150 and 1340 m) depth in the Indo-West Pacific. There are no fossils that can be placed in this taxon with certainty, but it is possible that the Upper Jurassic *Pedinothuria* may belong here. *See* ECHINODERMATA. Andrew B. Smith

Bibliography. M. Jensen, Morphology and classification of Euechinoidea Bronn, 1860, a cladistic analysis, *Vidensk. Meddr dansk naturb. Foren.*, 143:7-99, 1981.

Microradiography

The process of producing enlarged images of the interior of thin, usually small specimens by penetration of low-energy (0.1–10 keV) x-rays. The magnification can be obtained geometrically during the exposure by subsequent enlargement of the initial image by optical or electronic means, or by a combination of both processes. As with radiography on other size scales, microradiography shows the spatial distribution of mass and elemental composition of the sample. If a pulsed (flash) source or time-gated detector is employed, radiographs also provide stop-motion details of fast-changing objects. Microradiography has numerous applications in biology, material science, the characterization of fabricated microstructures, and assessment of plasma-driven compression of thermonuclear fuel.

Microradiography is largely synonymous with x-ray microscopy, both techniques being concerned with producing enlarged images of opaque objects by using x-rays. However, x-ray microscopy of defects in crystals can be performed by using diffraction, rather than simple absorption, in a technique termed x-ray topography. Microradiography by x-ray absorption is complementary, or related to a variety of other techniques for characterization of microstructures. *See* X-RAY DIFFRACTION; X-RAY MICROSCOPE.

Principles. Radiography on all size scales is based on Beer's law, which relates the x-ray intensity incident on (I_0) and transmitted by (I) an object of density ρ and thickness x , as given by Eq. (1). The

$$\frac{I}{I_0} = \exp\left(-\frac{\mu}{\rho}\rho x\right) \quad (1)$$

mass absorption coefficient μ/ρ (cm²/g) and the linear absorption coefficient μ (cm⁻¹) depend on the x-ray wavelength and absorbing material.

Constituents of cells such as protein and water have x-ray absorption lengths (the inverse of the linear absorption coefficient) that vary rapidly with energy in the soft x-ray region (**Fig. 1**) and correspond to small penetration ranges (approximately 1–10 micrometers). Radiography with submicrometer spatial resolution requires thin samples in order to avoid blurring and superposition of images, with attendant loss of resolution and contrast, in addition to obtaining the requisite penetration. Usually, microradiographic samples are less than 10 μm thick and 10–1000 μm in lateral dimension.

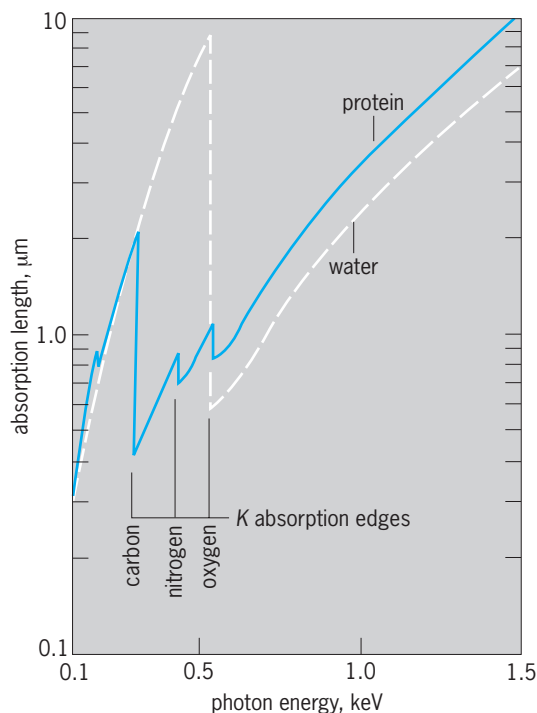


Fig. 1. Soft x-ray absorption lengths (inverse of the linear absorption coefficient) for protein typical of cellular materials and for water. K-electron absorption edge jumps are indicated for carbon, nitrogen, and oxygen. The large difference in absorption lengths between the carbon and oxygen edges provides good contrast for live biological specimens. (After J. C. Solem and G. F. Chapline, *X-ray biomicroholography*, *Opt. Eng.*, 23: 193–203, 1983)

The very large steps in absorption length which occur at absorption edges (Fig. 1) are highly significant because they provide chemical specificity and contrast between different materials. These thickness (x) and chemical (μ) sensitivities can be quantified by differentiating Beer's law, as in Eq. (2). A

$$\frac{dI}{I} = -xd\mu - \mu dx \quad (2)$$

relative intensity change of 1% or greater is needed for reliable measurements. Hence, for constant μ , a change in thickness $dx = 10^{-2}/\mu$ is required. In the soft x-ray region, μ is typically of the order of 10^4 cm^{-1} , so a thickness change dx of approximately 10 nanometers is marginally resolvable. For constant thickness x , the required change in linear absorption coefficient is $d\mu = 10^{-2}/x$. Given typical values of $x = 10^{-4} \text{ cm}$, a change $d\mu = 100 \text{ cm}^{-1}$ is near the chemical resolution limit. Since μ is given by Eq. (3),

$$\mu = \sum_i W_i \mu_i \quad (3)$$

where W_i is the weight fraction of the i th component of a homogeneous material and μ_i is the corresponding linear absorption coefficient, concentration variations dW_i of 1% weight are barely resolvable in binary samples of $\mu_1 \gg \mu_2$ or vice versa. The technique of detecting small chemical variations by monochromatic microradiography is termed x-ray absorption microanalysis.

Equipment. There are two necessary components of each radiographic exposure, in addition to the object under study: a source of x-rays, and a means of detecting the transmitted x-radiation. Sources are sometimes augmented by x-ray optics to improve the efficiency of radiation collection or to provide monochromatic radiation. Optics are also employed to form images of specimens. Detection systems sometimes include intensifiers in order to shorten exposure times or to permit work with weaker sources.

Sources. There are three classes of sources employed for microradiography: electron-impact devices, storage ring sources of synchrotron radiation, and multimillion-degree plasmas. The first microradiography measurements, and most studies since then, have been done with electron-impact sources. These contain an electron-emitting cathode and an anode in a high vacuum, between which a high voltage is applied. Such devices vary from sealed x-ray tubes to components mounted within special-purpose vacuum systems. Electron-impact sources emit both line and continuum x-radiation. They range in size from about $1 \mu\text{m}$ to 1 cm. Although usually run continuously, some electron-impact sources can be operated in a pulsed mode. See X-RAY TUBE.

High-energy (of the order of giga-electronvolts) electrons orbiting within an evacuated toroid, called a storage ring, produce intense continuum radiation which is tightly collimated and polarized. Such synchrotron radiation is emitted in nanosecond pulses at MHz rates, and can be 1000 or more times as intense as the x-rays from electron-impact sources. The source cross section is typically $0.2 \times 2 \text{ mm}$. A monochromator between the storage ring and sample allows radiation to be tuned over absorption edges of elements in the sample. The use of synchrotron radiation for microradiography began in 1977 and has rapidly expanded. See SYNCHROTRON RADIATION.

Very high-temperature plasmas can be heated by high-power lasers or electrical discharges. They are typically $10 \mu\text{m}$ to 1 mm in size. Multimillion-degree plasmas emit uncollimated line and continuum spectra predominantly in the soft x-ray region, below a few kiloelectronvolts, with pulse lengths of 0.1–100 ns. Stop-motion microradiographs were first made with plasma x-radiation in 1980. Fresnel zone plate x-ray lenses are sometimes used to gather (condense) synchrotron radiation, that is, to image the source, onto a microscopic sample. See PLASMA PHYSICS.

Optics. Devices which monochromatize or focus x-rays incident upon or transmitted by specimens are employed in some microradiographic techniques. The dispersive elements in x-ray monochromators can be reflection or transmission gratings (manufactured) or crystals (natural or grown), depending on the x-ray wavelength. Focusing elements can also be of transmission or reflection character. Fresnel zone plates consist of equal-area alternating open and x-ray opaque annuli. They form soft x-ray images by interference of transmitted radiation. Since Fresnel zone

plates bring different wavelengths to focus at different axial positions, they can also serve as monochromators. Gratings and Fresnel zone plates are often made by using optical or electron beam patterning techniques and materials-processing methods developed for microcircuit production. Multilayer-coated optics, which serve simultaneously to monochromatize and focus x-rays, are produced by highly controlled vapor deposition onto substrates figured by diamond turning, followed by superpolishing. Grazing-incidence focusing optics can also be employed for microradiography. They are also produced by modern techniques for production of precision optics. See DIFFRACTION; INTEGRATED CIRCUITS; X-RAY OPTICS.

Detectors. Passive detectors which absorb x-rays and retain the image are most commonly used for microradiography. Photographic film, which provides resolution down to around $0.1\ \mu\text{m}$, is usually employed. Images recorded on film usually are magnified and observed optically. See PHOTOGRAPHIC MATERIALS; PHOTOGRAPHY.

Grainless polymeric media, called photoresists, offer resolution down to $0.01\ \mu\text{m}$. Absorption of x-radiation alters their solubility in specific solvents. In positive resists, absorption produces polymer chain scissions, reducing the molecular weight and increasing solubility. The reverse is true for negative resists. Development (dissolution) in a solvent matched to the resist results in an image which consists of surface relief in the photoresist. The images are magnified and transferred to film with an electron microscope.

Active electronic detectors of a wide variety can be employed for microradiography. Two-dimensional, "area" detectors employing multiple x-ray collecting anodes, or arrays of small solid-state detectors, or television-style readouts are available. Image converters transduce images from x-rays to electrons or visible light. They sometimes serve as image intensifiers also, reducing required source intensity. The images collected with area detectors can be stored digitally or they can be recorded on film. If a microradiographic image is formed by the two-dimensional raster scanning technique, in which a very fine x-ray beam is moved sequentially over the sample, then electronic detectors which do not offer spatial resolution are employed. See CHARGE-COUPLED DEVICES; LIGHT AMPLIFIER.

Techniques. X-ray micrographs may be obtained by four geometrically different techniques. Two of them, the contact and the projection methods, have been in use for decades. However, they have undergone significant development because of the availability of bright synchrotron radiation and plasma radiation sources. The other two techniques, the true imaging and the scanning methods, require the bright sources. Work on these methods goes back only to the mid-1970s, with major developmental efforts underway.

Contact. In this approach, the specimen either touches or is very close to the recording medium (Fig. 2a). Resolution in the one-to-one image is

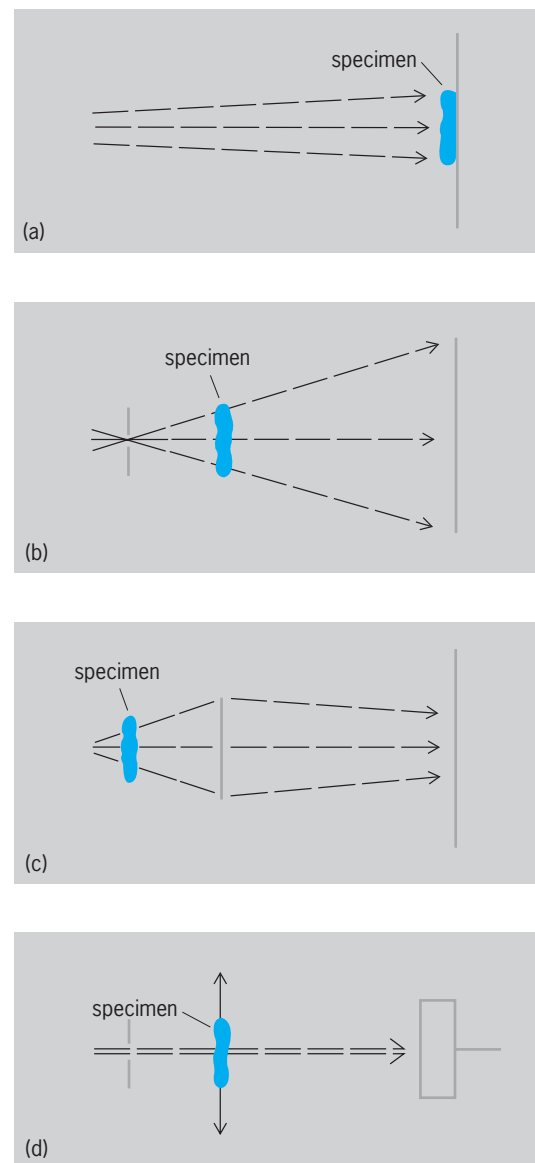


Fig. 2. Schematics of techniques for x-ray microradiography. X-ray paths are indicated by broken lines. Alternative configurations for some techniques are discussed in the text. (a) Contact. (b) Projection. (c) Imaging. (d) Scanning.

determined primarily by the detector, although penumbral effects can contribute to image blurring for extended sources and relatively thick samples. Photographic film is most commonly employed for recording contact microradiographs, with magnified images being produced from the original record by optical means. However, the film grain size of $0.1\ \mu\text{m}$ or greater sometimes provides an unacceptable resolution. In such cases, for example, imaging of live biological materials, it is necessary to use x-ray-sensitive grainless photoresist media. After coating the developed resist with a thin layer to ensure electrical conductivity, the images are read and enlarged with a scanning electron microscope for a thick resist or substrate, or a transmission electron microscope if the resist is thin and on a very thin substrate. Contact x-ray microradiographs have been made with

diverse specimens. Photoresist images of biological specimens exhibit features as fine as 10 nm. Time resolution of about 100 ns has been demonstrated with contact microradiography.

Projection. As with the contact method, the projection approach to microradiography depends on simple geometric shadow casting to form the image. However, in projection, the initial recording is a magnified image of the specimen. This recording is usually on film, although electronic image converters are also employed. The scale of the image depends on the ratio of the source-to-detector and source-to-specimen distances. The source for projection microradiography can be either an intrinsically small x-ray emitter or a larger source limited by a pinhole (Fig. 2*b*). If an extended source and pinhole are employed, the pinhole can be placed between the sample and recording medium. This is tantamount to taking an x-ray pinhole-camera photograph of a back-lighted specimen.

The spatial resolution in projection is usually determined by the effective source diameter. Micrometer-sized focused electron-impact sources have long been used to provide striking projection pictures of the interior of complex biological specimens, notably insects. Laser-heated plasma sources of soft x-rays have been used to take static or dynamic microradiographs. Pinholes several micrometers in diameter have been employed to obtain flash microradiographs of laser-driven targets with time resolution of 100 picoseconds, and subpicosecond time resolutions are foreseen. Although the projection method has not been pushed below about 1 μm spatial resolution, it has the potential for resolving 100 nm.

Imaging. Ordinary refractive transmission lenses do not form images in the x-ray region. However, both transmission and reflection x-ray optics are available for x-ray imaging. Fresnel zone plates, as discussed above, with annuli small enough to provide resolution below 100 nm are small and inefficient. Hence, bright synchrotron sources are required for imaging microradiography. Two Fresnel zone plates are employed: the first monochromatizes and condenses x-rays onto the specimen, and the second then forms an image of the sample magnified several hundred times (Fig. 2*c*).

Film has usually been used to record imaged microradiographs with zone plates in a few seconds with synchrotron radiation. Image intensifiers will reduce exposure times to less than 100 ms. Flash microradiographs formed with submicrosecond pulsed plasma x-radiation are also possible. Imaging techniques may eventually yield spatial resolution near 10 nm.

Scanning. The above methods require that the entire sample be exposed to x-rays during image formation in relatively insensitive, high-resolution recording media. This constant exposure produces radiation damage in sensitive specimens, such as biological specimens, if prolonged exposures are made with steady sources. Fast, pulsed sources provide records before the sample is altered by x-ray absorption. A more attractive approach is to

use sensitive, photon-counting techniques. However, high-resolution area detectors with the ability to sense individual soft x-ray photons are not available.

Large x-ray detectors, such as gas proportional or scintillation counters, are sensitive and convenient. With them, spatial resolution is achieved by raster-scanning the specimen through a fine x-ray beam while counting transmitted x-ray photons. X-rays strike, and damage, any spot on the sample only when its transmission is being measured. Live cells can be held at lower temperatures to immobilize them during a complete scan. The resolution-defining beam can be simply formed by pinhole aperturing of a larger x-ray beam (Fig. 2*d*). It can also be formed by use of two Fresnel zone plates, one for condensing and another for focusing, as in the imaging method. Reflective x-ray optics, coated with multilayer films which diffract x-rays as do crystals, can also be employed to form submicrometer, monochromatic spots for scanning microscopy.

Spatial resolution of 300 nm has been demonstrated with the scanning technique, and resolutions in the 10–50-nm range are anticipated. Because of the technique's sequential nature, scanning microradiographs require at least 100 s to record, although each picture element (pixel) may require only a few milliseconds to record. Brighter synchrotron radiation sources may reduce the time to record a complete scanning microradiograph to below 1 s. An attractive feature of scanning microradiography concerns the immediate digital nature of the data. Image magnification and analysis is readily performed with modern computer techniques.

Applications. Microradiography was developed in order to observe fine details in the interior of opaque natural (for example, biological) and manufactured (for example, metallurgical) samples. Optical microscopy has long been available to observe the surfaces of opaque specimens, or the interior of clear samples, with spatial resolutions approaching 200 nm. Early interest in microradiography waned for two reasons: weak sources required long exposure times, and electron beam techniques which provide submicrometer details were developed. The availability of bright soft x-ray sources, especially synchrotron radiation, has led to a resurgence of interest in microradiography, especially for cellular samples. Sub-100-nm resolution of live specimens is needed to understand biological structures with sizes between those observable with light microscopes (down to about 200 nm) and the molecular level (about 1 nm) probed by diffraction from ordered arrays of molecules. Microradiography of inorganic natural, manufactured, and dynamic structure remains of interest. See ELECTRON MICROSCOPE; OPTICAL MICROSCOPE.

Biology. Microradiographic techniques which would yield images of the ultrastructure of live cells with resolution near 10 nm are under development. Such techniques have immense potential as a research tool and, perhaps, even for clinical medicine. As shown in Fig. 1, radiation at energies in the range

between the carbon *K* absorption edge (0.284 keV) and the oxygen *K* edge (0.532 keV) provides excellent contrast between organic material and the water needed to sustain life in the specimen.

Two approaches to microradiography of cells have been undertaken. In one, contact radiographs are made of cells placed on photoresist. A very thin x-ray transmitting window [for example, 100 nm of silicon nitride (Si_3N_4)] retains a layer of water about $1\ \mu\text{m}$ thick around the cells. Specimens for the initial demonstration were blood platelets. The microradiographs were recorded in a resist copolymer of poly(methyl methacrylate) and methacrylic acid. The exposure was made with a 100-ns flash of soft x-rays from a discharge-heated plasma, and a scanning electron micrograph was made of the developed resist (Fig. 3). Structures as fine as 10 nm are visible. Details not visible in exposures of dried platelets are observed. In particular, new structures near the base of pseudopods extending from the platelets are seen for the first time.

Scanning microradiography of biological materials is also under development. The use of zone plates to condense and focus radiation to a small spot has been demonstrated in a scanning microscope. Figure 4 is a digitized (10^4 -pixel) image of an algal cell taken with synchrotron radiation in about 1 h with 300-nm pixels. The relatively long exposure was due to the use of monochromatized 0.387-keV radiation. When scanning microradiography with 30-nm resolution is possible, it is expected to yield images exhibiting new structures.

Natural inorganic structures. Interest in microradiography of inorganic materials is not as great as for live biological materials. Inorganic materials can be coated with a conductor and examined by electron microscopes offering resolution to 1 nm. However, microradiography has utility for specimens thicker than 100 nm, especially if chemical information is desired.

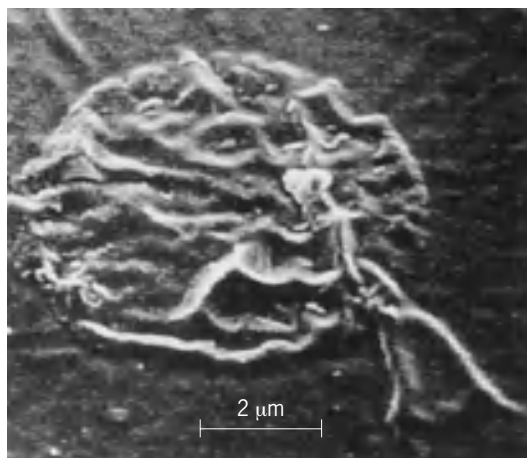


Fig. 3. Scanning electron micrograph of the contact microradiograph of a live human blood platelet recorded in a photoresist with a 100-ns burst of soft x-rays from a plasma heated by an electrical discharge. (From R. Feder et al., *Flash x-ray microscopy*, *Science*, 227: 63–64, 1985)

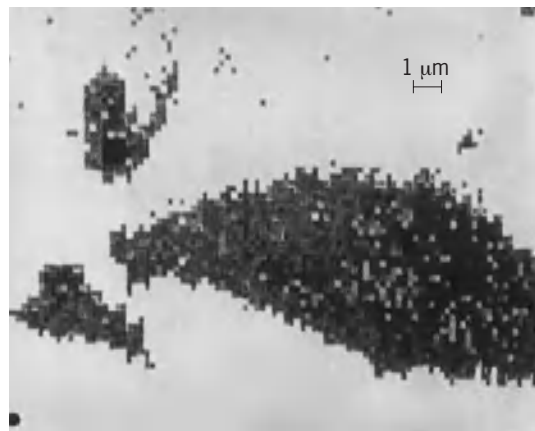


Fig. 4. Digital presentation of a scanning x-ray micrograph of a wet *Phaeodactylum* algal cell taken with synchrotron radiation. (From H. Rarback et al., *Recent results from the Stony Brook scanning microscope*, in G. Schmahl and D. Rudolph, eds., *X-ray Microscopy*, pp. 203–216, Springer-Verlag, 1984)

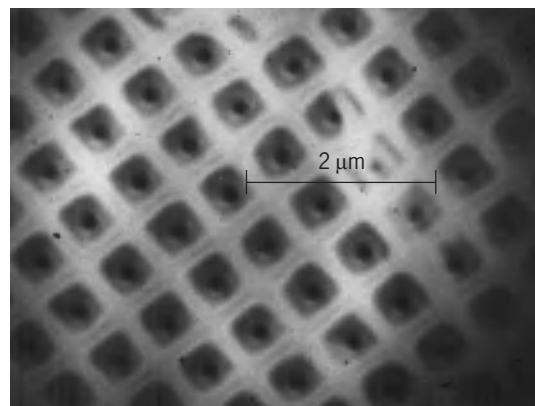


Fig. 5. Soft x-ray image of a diatom recorded on film with synchrotron radiation. (From G. Schmahl et al., *Zone plates for x-ray microscopy*, in G. Schmahl and D. Rudolph, eds., *X-ray Microscopy*, pp. 63–74, Springer-Verlag, 1984)

Diatoms are calciferous structures with submicrometer detail produced by sea life. They serve as convenient test objects for developmental microradiography techniques. Figure 5 is an x-ray image of a diatom produced by passing 0.276-keV synchrotron radiation through a pair of Fresnel zone plates. The second plate yielded a $230\times$ magnified image on film in a 3-s exposure. Details as fine as 50 nm can be seen in such images. Microradiographs of wet biological specimens have been taken with the imaging technique used to produce Fig. 5.

Artificial microstructures. A variety of techniques, especially those used in microelectronics production, are available to produce structures with detail on the micrometer scale. As with natural inorganic materials, scanning electron microscopy is usually the tool of choice for examining artificial microstructures. However, again, there is some need for examination of the interior of artificial microstructures which microradiography can fulfill.

Microballoons with diameters of 0.1 to 1 mm, filled with deuterium and tritium, are of interest in

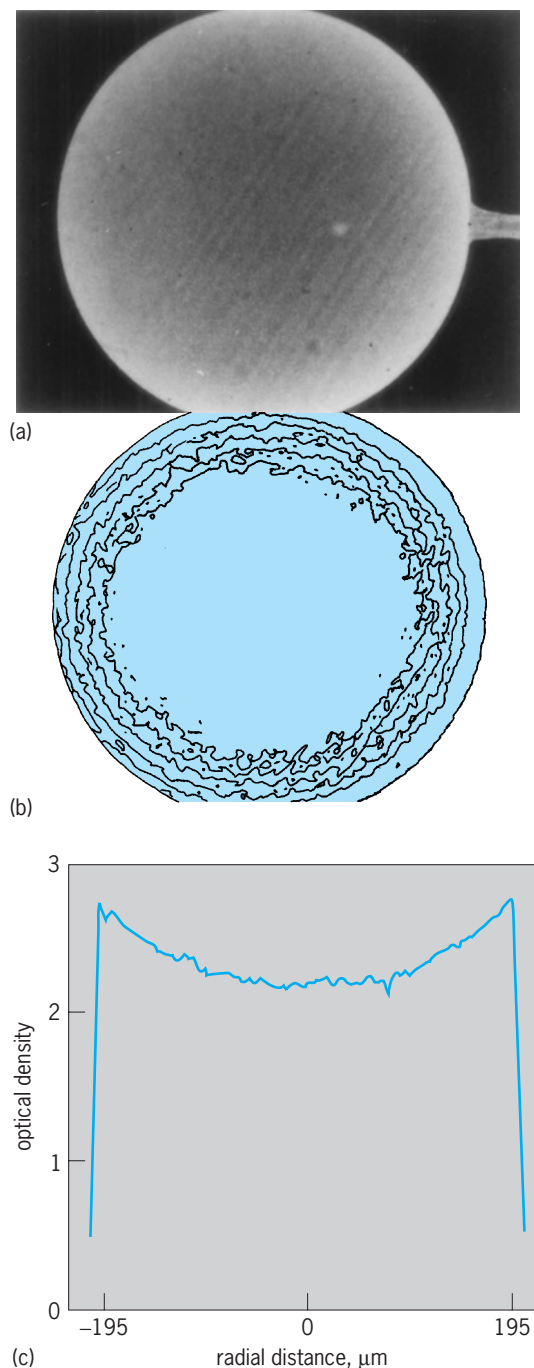


Fig. 6. Examination of a glass microballoon with 390- μm outer diameter. (a) Contact x-ray microradiograph of microballoon (mounting stalk on right). (b) Two-dimensional contour density representation. (c) One-dimensional density scan. (From H. Kim and M. D. Wittman, *X-ray microradiography of inertial fusion targets*, *J. Vac. Sci. Technol.*, 3:1262–1265, 1985)

fusion research. Exterior irradiation of such small spheres (pellets) with very high-power, nanosecond-pulsed lasers blows off the pellet's surface. In response, the pellet and its nuclear fuel are crushed to a small diameter. This raises the fuel density and temperature to conditions required for ignition of thermonuclear burn, as in the hydrogen

bomb but on a very much smaller scale. Uniform compression is needed to attain ignition, requiring uniform laser irradiation of pellets with uniform wall thickness. X-ray microradiography is employed, usually with steady electron-impact sources, for preshot examination and selection of microballoons. However, plasma radiation has also been used to inspect microspheres. **Figure 6** shows a 250 \times magnification of the radiograph of a pellet with a diameter near 400 μm , together with a two-dimensional contour density representation and a one-dimensional density scan. The contact image was recorded on film with 9-keV radiation in a 1-ns pulse from a laser-heated plasma. This rapid exposure, typical of laser-plasma x-ray sources, precludes vibration-induced blurring which may be present when a steady x-ray source is employed. See NUCLEAR FUSION.

Dynamic exposures. Very short x-ray pulses are needed to record rapidly changing structures, for example, a laser-driven microballoon. X-rays are employed because optical radiation will not penetrate the plasma which surrounds and drives the pellet inward. Projection microradiography with a pinhole between the microballoon being studied and film is employed. A separate laser beam which can be arbitrarily delayed relative to those striking the pellet provides a stroboscopic x-ray backlighting source. **Figure 7** shows a sequence of flash microradiographs of imploding pellets with initial diameter of 150 μm and final diameters during stagnation (prior to blowup) of about 50 μm . These microradiographs were taken with x-ray pulses of 100 picoseconds, to achieve very high simultaneous spatial resolution (about 10 μm) and time resolution (100 ps).

Related techniques. Microradiography is only one of many techniques useful for structural or chemical characterization of samples.

Electron microscopy, as discussed above, can provide spatial resolution better than 1 nm for the surface of thick samples (measured in the scanning mode) or for thin (<100 nm) samples imaged or scanned in transmission. Scanning and transmission electron microscopy is used to produce enlarged images of contact microradiographs recorded in photoresists (Fig. 3). Emitted x-ray intensity produced when a focused beam of energetic electrons is scanned over a sample can be used to provide maps of elemental distribution. This method, termed electron probe microanalysis, is a form of x-ray microscopy. However, it does not involve x-ray transmission through the sample and, hence, is not microradiography.

X-ray lithography, a method of pattern replication developed for microcircuit and microstructure production, is closely related to contact x-ray microradiography. In lithography, artificial patterns called masks are reproduced rather than natural specimens such as cells. The bright sources developed for fast commercial replication of a lithographic mask, notably dense multimillion-degree plasmas, are now applied to contact microradiography.

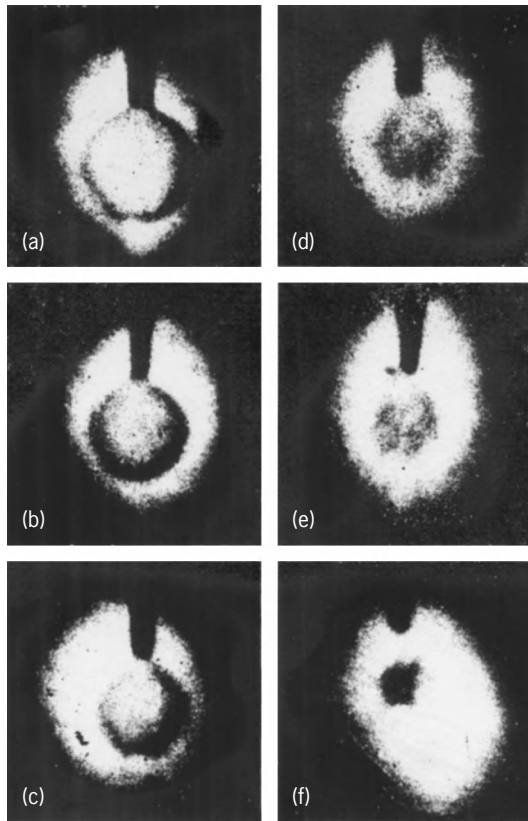


Fig. 7. Projection (pinhole) photographs of laser-irradiated microballoons recorded on film with 100-ps flashes of soft x-rays from oval-shaped laser-heated backlighting plasmas, at time delays between the 100-ps implosion-driving laser pulse and the x-ray strobe pulses of (a) 200 ps, (b) 350 ps, (c) 400 ps, (d) 450 ps, (e) 500 ps, and (f) 1000 ps. (From C. Yamanaka, *Progress of inertial confinement research in Japan*, in C. Yamanaka, ed., *Advances in Inertial Confinement Systems*, pp. 43–56, Institute of Laser Engineering, Osaka, 1980)

Holography ordinarily involves recording the interference pattern between an optical reference wave and the same wavelength scattered from the surface of an object. Holograms can also be formed by interference of reference and transmitted waves of penetrating x-radiation. Reconstruction of x-ray holograms with longer-wavelength (optical) radiation can provide enlarged, three-dimensional images of the object. X-ray holography requires sources with

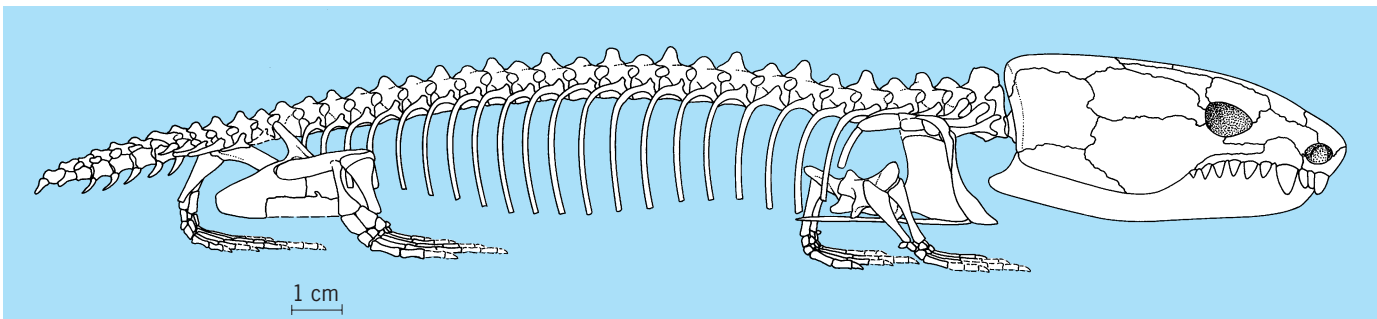
adequate temporal (longitudinal) and spatial (transverse) coherence to produce the requisite interference. X-ray lasers are under development, but they do not yet have wavelengths short enough to penetrate samples, nor do they exhibit adequate coherence. Bright synchrotron x-radiation sources can be monochromatized (for temporal coherence) and apertured (for spatial coherence) to provide intensity adequate for development of x-ray holographic techniques. X-ray holography promises to be an important technique, distinct from but closely related in its methodology and goals, to conventional microradiography. See HOLOGRAPHY.

David J. Nagel

Bibliography. P. C. Cheng and G. Jan, *X-ray Microscopy*, 1987; D. F. Parsons (ed.), *Ultrasoft x-ray microscopy: Its application to biological and physical sciences*, *Annals of the New York Academy of Sciences*, vol. 342, 1980; G. Schmahl and D. Rudolph (eds.), *X-ray Microscopy*, 1984; K. Shinohara and K. Yada (eds.), *X-ray Microscopy in Biology and Medicine*, 1990.

Microsauria

A diverse order of small, extinct amphibians, known only from the Pennsylvanian and lower Permian of North America and Europe. They range from obligatorily aquatic, perennibranchiate genera with lateral-line canal grooves, to fully terrestrial lizardlike forms. Several families include long-bodied, possibly burrowing, species. Limbs are always retained, and the tail is never specialized as a swimming organ. Microsaurians are recognized by the possession of a broad, strap-shaped occipital condyle, and no more than a single bone in the temporal series. The trunk vertebrae are spool-shaped. Some species have trunk intercentra, and most show caudal hemal arches. The number of presacral vertebrae ranges from 19 to 45. Limb and skull proportions are extremely variable, as is the dentition (see *illus.*). The specific origin of the group remains unknown. They are customarily allied with lysorophids, aistopods, and nectridians in the subclass Lepospondyli, but all the shared, derived characteristics of these groups may be correlated with small body size and do not necessarily



Pantylus, a microsauroid from the lower Permian of Texas. (After R. L. Carroll, *The postcranial skeleton of the Permian microsauroid Pantylus*, *Can. J. Zool.*, 46(6):1175–1192, 1968)

indicate close relationship. Although intermediate forms are not known, microsaurians appear to be the most probable group of Paleozoic amphibians from which two modern amphibian orders, the apodans and the salamanders, have evolved. See AMPHIBIA; LEPOSPONDYLI.

Robert L. Carroll

Bibliography. R. L. Carroll and P. Gaskill, The order Microsauria, *Mem. Amer. Phil. Soc.*, 126:1–211, 1978.

Microscope

An instrument used to obtain an enlarged image of a small object. The image may be seen, photographed, or sensed by photocells or other receivers, depending upon the nature of the image and the use to be made of the information of the image. Microscopes are classified as simple or compound according to the kind of radiation used to form the image, and the use for which they are designed.

Simple microscope. A simple microscope, hand lens, or magnifier usually is a round piece of transparent material, ground thinner at the edge than at the center, which can form an enlarged image of a small object. Commonly, simple microscopes are double convex or planoconvex lenses, or systems of lenses acting together to form the image (Fig. 1). The lens can be mounted in a simple holder, in a folding case for hand use, or with a support which has a mechanical focusing mechanism, stage, and mirror to make a dissecting microscope. See LENS (OPTICS).

Compound microscope. The compound microscope utilizes two lenses or lens systems. One lens system forms an enlarged image of the object and the second magnifies the image formed by the first. The total magnification is then the product of the magnifications of both lens systems. Theoretically, several simple microscopes could be used in line, each to magnify the image of the one before it. Practically, the losses from aberrations, reflections, and other defects limit the compound microscope to two such systems (Fig. 2).

The typical compound microscope consists of a stand, a stage to hold the specimen, a movable body-tube containing the two lens systems, and mechanical controls for easy movement of the body and the specimen. The lens system nearest the specimen is called the objective; the one nearest the eye is called

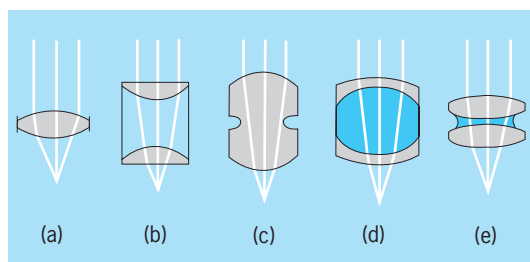


Fig. 1. Some common types of magnifier. (a) Double convex. (b) Doublet. (c) Coddington. (d) Hastings triplet. (e) Achromat. (After F. A. Jenkins and H. E. White, *Fundamentals of Optics*, 3d ed., McGraw-Hill, 1957)

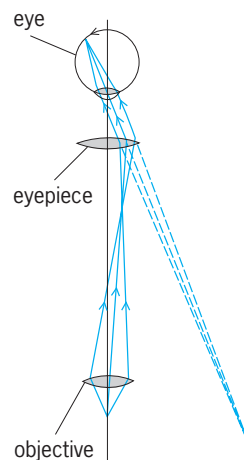


Fig. 2. Diagram of compound microscope. (After F. A. Jenkins and H. E. White, *Fundamentals of Optics*, 3d ed., McGraw-Hill, 1957)

the eyepiece or ocular. A mirror is placed under the stage to reflect light into the instrument when the illumination is not built into the stand. For objectives of higher numerical aperture than 0.4, a condenser is provided under the stage to increase the illumination of the specimen. Various optical and mechanical attachments may be added to facilitate the analysis of the information in the doubly enlarged image.

Special compound microscopes have two image-forming systems to give an enlarged image of an image. These instruments utilize electrons, x-rays, sound, or other forms of radiation for image formation, and electromagnetic or electrostatic fields or mirrors to form the enlarged images. Because these images are not visible, photography, television, and special receivers must be used to record and analyze the image. Special microscopes are usually named according to the kind of radiation used, such as electron, x-ray, ion, and ultrasonic.

The compound microscopes which employ a lens system are essentially similar in their working principles. Basically, they are modifications of the ordinary laboratory microscope (bright-field microscope) for specialized or specific purposes. Following are descriptions of some of the different types.

Light or photon microscopes utilize light of wavelengths from 380 to 760 nanometers for image formation. Such microscopes include the laboratory or bright-field microscope, and modifications of it such as the capillary, centrifuge, chemical, comparison, crystallographic, dark-field, dissecting, fluorescence, integrating, interference, inverted microprojection, museum, nuclear track, petrographic, phase, phosphorescence, and profile microscopes.

Reflecting microscopes utilize a mirror rather than a lens system. The infrared microscope uses radiation of wavelengths greater than 700 nm, and the ultraviolet employs light of 180–400 nm. The ultraviolet microscope requires reflecting optics or special quartz and crystal objectives. Color-translating microscopes employ three different wavelengths of light to reveal details produced by ultraviolet or other nonvisible radiation.

In the electron, proton, x-ray, and beta-particle microscopes the image is usually recorded on a fluorescent screen or is photographed.

Mechanical vibrations, generated into an elastic system, provide the basis for the ultrasonic microscope employed for locating foreign bodies or for the analysis of reflecting surfaces. *See* ELECTRON MICROSCOPE; FLUORESCENCE MICROSCOPE; INTERFERENCE MICROSCOPE; OPTICAL MICROSCOPE; PHASE-CONTRAST MICROSCOPE; POLARIZED LIGHT MICROSCOPE; REFLECTING MICROSCOPE; SCANNING TUNNELING MICROSCOPE; X-RAY MICROSCOPE.

Oscar W. Richards

Microsensor

A very small sensor with physical dimensions in the submicrometer to millimeter range. A sensor is a device that converts a nonelectrical physical or chemical quantity, such as pressure, acceleration, temperature, or gas concentration, into an electrical signal. Sensors are an essential element in many measurement, process, and control systems, with countless applications in the automotive, aerospace, biomedical, telecommunications, environmental, agricultural, and other industries. The stimulus to miniaturize sensors lies in the enormous cost benefits that are gained by using semiconductor processing technology, and in the fact that microsensors are generally able to offer a better sensitivity, accuracy, dynamic range, and reliability, as well as lower power consumption, than their larger counterparts.

Technology and materials. Microsensors are typically batch fabricated from silicon wafers by using standard semiconductor process technologies in combination with specially developed processes. This technology, known as micromachining, allows hundreds of complex microsensors (or microstructures) to be produced on a single silicon wafer, resulting in a very low unit cost. This process technology is accurate and repeatable. A wide variety of passive and active materials are used to make microsensors. Single-crystal silicon is commonly used as a passive substrate material, although other substrate materials, such as gallium arsenide, quartz, and silicon carbide, are used for more specialized microsensor applications. These base materials are used in combination with a variety of thin films, including polysilicon, oxides, nitrides, and metals. It is often necessary to deposit an active material which is essential to the operation of the microsensor. Examples include zinc oxide, which is commonly used as the active material in piezoelectric sensors, and conducting polymers, which are commonly used in chemical microsensors. *See* INTEGRATED CIRCUITS; SEMICONDUCTOR.

Engineering a microsensor (sometimes called microengineering) requires the use of a number of micromachining processes, together with an understanding of micromechanics and microelectronics. The term bulk micromachining is often used for a set of processes, such as anisotropic etching, laser ablation, silicon-wafer bonding, and deep reactive

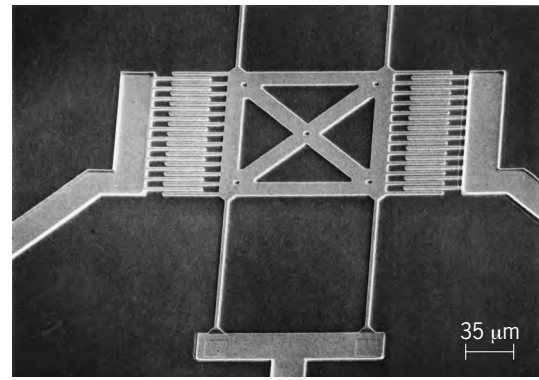


Fig. 1. Scanning electron micrograph of a polysilicon resonant structure made with a surface micromachining technique. (From C. J. Welham et al., *A laterally driven micromachined resonant pressure sensor*, *Sensors and Actuators A*, 52:86-91, 1996)

ion etching, that enable the three-dimensional sculpting of single-crystal silicon to make a small structure, such as a diaphragm for a pressure sensor or a pressure switch, or a suspended mass structure for an accelerometer and flow sensor. Surface micromachining refers to a set of processes based upon deposition, patterning, and selective etching of thin films to form a free-standing microstructure on the surface of a silicon wafer. **Figure 1** shows a surface-micromachined microflexural resonator that has been integrated onto a thin silicon diaphragm to form a pressure sensor. The resonator is electrostatically driven into a lateral resonant mode of oscillation (around 50 kHz) by using a comb capacitor, and its motion is sensed by using another comb capacitor. The deflection of the diaphragm under an applied pressure stretches the microresonator, thus changing its spring rate and its fundamental resonant frequency. *See* PRESSURE TRANSDUCER.

Mechanical microsensors. Mechanical microsensors form perhaps the largest family of microsensors because of their widespread availability. Microsensors have been produced to measure a wide range of mechanical measurands, including force, pressure, displacement, acceleration, rotation, and mass flow. Force sensors generally use a sensing element that converts the applied force into the deformation of the elastic element. **Figure 2** shows a scanning electron micrograph of a simple microcantilever stylus structure that is used in an atomic force microscope for analyzing surfaces at an atomic scale of resolution. The tip of a flexible force-sensing cantilever stylus (with a tip curvature radius of a few nanometers) is scanned over the sample surface. The forces acting between the tip and sample cause minute deflections of the cantilever, which can be detected optically, capacitively, or with integrated piezoresistors to produce an atomistic surface image. *See* SCANNING TUNNELING MICROSCOPE.

Solid-state microsensors. Some microsensors can be made by using conventional bipolar or metal oxide semiconductor (MOS) technology, and a variety of radiation sensors are available that can detect visible and near-infrared radiation by using, for

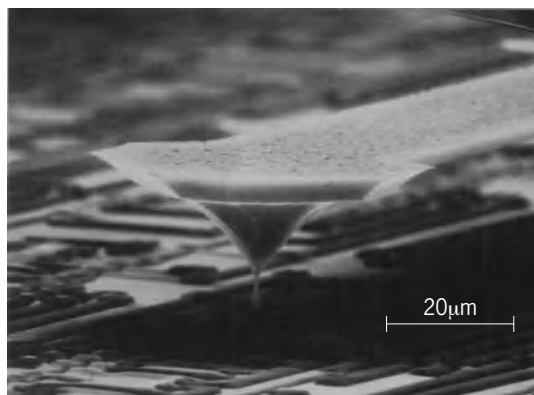


Fig. 2. Micrograph of a silicon microcantilever above an integrated circuit. (From J. Brugger et al., *Silicon cantilevers and tips for scanning force microscopy, Sensors and Actuators A*, 34:193–200, 1992)

example, silicon *pn* diodes, avalanche photodiodes, and pyroelectric devices. Digital thermometers can be engineered by using a thermally sensitive device, such as a resistor, *pn*-junction diode, or transistor. A variety of magnetic microsensors, based upon Hall-plate devices (which are sensors that measure the magnetic field strength through the Hall effect), magnetoresistors, magnetodiodes, and magnetotransistors, are also available. See OPTICAL DETECTORS.

Chemical and biochemical microsensors. Two applications for chemical and biochemical microsensors are environmental monitoring and medicine. Both are relatively undeveloped, but should be of great importance in the future.

Unfortunately, chemical microsensors are the least well developed, for many reasons. For example, it has proved difficult to make a stable sensor that is sensitive to just the one gas of interest. One approach successfully applied to overcome this problem has been to employ an array of nonspecific gas sensors in a microprocessor-based instrument. The microprocessor is able to apply an appropriate pattern-recognition technique to extract the required information from the output of the microsensor array. This approach has opened up an enormous range of

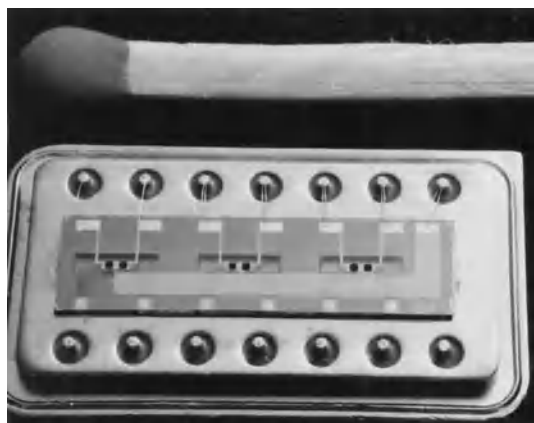


Fig. 3. Micromachined six-element chemical microsensor array in a dual in-line package. (From J. W. Gardner et al., *Integrated sensor array for detecting organic solvents, Sensors and Actuators B*, 26-27:135–139, 1995)

applications for gas and odor measurement in many different areas, such as the automotive, medical, and food-beverage industries and environmental monitoring. In addition, there have been advances in the design and synthesis of new materials such as conducting polymers, which coupled with the above approach make high-quality gas microsensors a distinct possibility in the near future. Microsensor array devices are also able to measure multiple measurands and compensate for interfering variables, such as temperature. For example, one chemical microsensor (Fig. 3) is employed as the sensing element in an electronic-nose instrument. See MICROPROCESSOR.

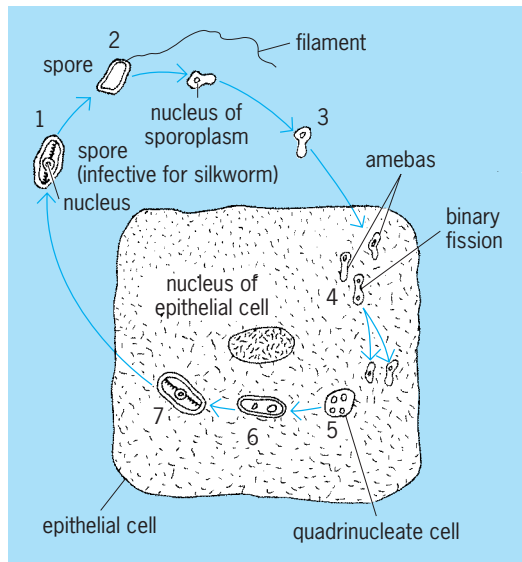
Applications in the medical industry may involve monitoring blood, urine, and breath, which contain a wealth of information about the patient's state of health. Only a few such devices now exist. Examples include a glucose biochemical microsensor and ion-selective field-effect devices used to measure blood pH. The use of microsensors to gather medical diagnostic information is an attractive proposition, and eventually there may even be implanted microsensors to diagnose health problems, using smell-sensitive array devices. See BIOELECTRONICS.

Smart sensors. Since the processes used to fabricate microsensors are similar to the ones used to produce conventional integrated circuits, signal-conditioning electronics such as analog-to-digital converters can be integrated on the same chip, resulting in an integrated microelectromechanical component. Such a smart or intelligent sensor is able either to process information itself or to communicate with an embedded microprocessor. One notable example is a microaccelerometer developed for the automotive industry to control air-bag inflation. This inexpensive monolithic accelerometer comprises a surface-micromachined polysilicon sensor and bipolar-MOS interface circuitry on a single silicon chip. The sensor has been developed for a measurement range of ± 5 g (where g is the acceleration of gravity, 32 ft/s² or 9.8 m/s²), operates on a single 5-V supply, and has a power consumption of 40 mW. As the level of integration increases, a single-chip device incorporating microsensors, signal conditioning, and a microprocessor can be expected. See ACCELEROMETER.

Andrew C. Pike; Chris J. Welham; Julian W. Gardner
Bibliography. J. W. Gardner, *Microsensations*, *IEE Rev.*, pp. 185–188, September 1995; J. W. Gardner, *Microsensors: Principles and Applications*, 1994; E. Kress-Rogers (ed.), *Handbook of Biosensors and Electronic Noses*, 1997; S. Middlehoek and S. A. Audet, *Silicon Sensors*, 1989.

Microsporidea

A class of Cnidosporea characterized by the production of minute spores with a single intrasporal or one or two intracapsular filaments and a single sporoplasm. The spore membrane of these protozoa is usually a single piece. Microsporidians are mainly



Diagrammatic life cycle of *Nosema bombycis*, cause of pébrine, a fatal disease of silkworms. (1) Spore is typical of the order, with a single filament and sporoplasm. Upon ingestion (2) the filament is extruded and (3) the sporoplasm enters an epithelial cell of the gut where it divides many times to form uninucleate amebas which fill the cell. The cell may dissolve. (4) Amebas attack other cells, perhaps almost all in the worm. (5) Eventually parasites form four-nucleate cells, which (6,7) develop into spore. Silkworms are infected by eating mulberry leaves or other food contaminated by feces or body fragments of infected worms.

intracellular parasites of arthropods and fishes. Microsporidia is the only order of this class.

The sporoplasm or amebula, released from the spore in the intestine of the host, passes through the gut wall to reach the site of infection by way of the bloodstream. The amebula enters a cell and becomes a trophozoite, feeding and growing at the expense of the host. Eventually, the trophozoite divides by binary fission, or schizogony. The cells resulting from this asexual division develop into sporonts. A sporont may transform directly into a sporoblast and give rise to a single spore, or it may undergo further nuclear division and produce several sporoblasts, each of which will form a spore. In this order there is no evidence that special cells differentiate in the developing spore to form valves or a capsule as occurs in the Myxosporida and Actinomyxida.

An interesting feature of microsporidiosis is that the parasites may induce an extreme enlargement (hypertrophy) of the cell or nucleus. Thus, certain cells of the stickleback and smelt, infected with *Glugea anomala* and *G. hertwigi*, respectively, often enlarge from a diameter of 8-10 micrometers to one of 5000 micrometers. Such cells are called *Glugea* cysts.

The microsporidian *Nosema bombycis* is a parasite of the silkworm, the larval stage of the insect *Bombyx mori* (see **illus.**). Invasion by the parasite of all the cells of the larva, pupa, and adult causes the lethal disease pébrine which was first studied in detail by L. Pasteur in 1865. See CNIDOSPORA; INSECT PATHOLOGY; LEPIDOPTERA; PROTOZOA.

Ross F. Nigrelli

Microtechnique

The art of preparing objects for examination under the microscope and of preserving objects so prepared. Few objects yield useful information if examined without such preparation, which may involve, in addition to preliminary preservation, hardening, rendering transparent, selective coloration of parts, and cutting into thin slices.

The four types of microscope slide commonly made today are wholemounts, smears, squashes, and sections. The last three methods are merely devices to make thinner, or smaller, objects that are unsuitable for the first method. In all four methods, the objects are permanently preserved in a mounting medium between a glass slide, about 1 mm thick, and a glass cover slip 0.2-1 mm thick. The preliminary steps of fixation and preservation, the practice of staining, and the final mounting media are common to all four types.

Fixation and preservation. Biological specimens may be preserved in any environment, physical or chemical, which inhibits enzyme action and thus prevents either autolysis or microbiological growth. Objects thus simply preserved, however, rarely present to microscopists those details of external and internal anatomy which they desire to study. It is therefore customary to "fix" specimens before preservation so as to retain in them a reasonable facsimile of their appearance when alive. A perfect fixative solution would immobilize highly contractile organisms in a fully expanded condition, and at the same time preserve both the relations of the organs and the protoplasmic structure of the cells within them. The immobilization of highly contractile creatures requires instant denaturing of the proteins, while the preservation of protoplasmic structure is best retained by reasonably slow tanning. Actual fixative solutions, which by convention bear the name of the technician who first published them, are compromises (**Table 1**). Violent denaturants, which coagulate proteins, are mixed with less violent tanning agents which do not coagulate proteins. The best-known coagulants are heat, trinitrophenol (picric acid), chromium trioxide (chromic acid), mercuric chloride, and ethanol. Noncoagulating tanning agents in common use are potassium dichromate, osmium tetroxide (osmic acid), acetic acid, and formaldehyde.

No combination of these is perfect. Compositions should be selected not only for their immobilizing, coagulating, or tanning properties but also in regard to the relative importance of nuclear preservation, cytoplasmic preservation, and, in tissues intended for sectioning, the degree of brittleness produced. In general, very acid coagulants result in conventional images of nuclei and chromosomes, while weakly acid tanning agents display cytoplasmic detail and produce the least brittleness. The weakly alkaline fixatives of the electron microscopist produce images which have not proved acceptable to the optical microscopist.

Free fixative is generally removed from specimens

TABLE 1. Common fixatives

| Mixture | Example | Penetration | Use |
|-----------------------------|---|---|--|
| Osmic-chromic-acetic | Flemming's solution | Poor penetration; tissues should be fixed for about 6 h | Excellent for small animals and small pieces of tissue |
| Chromic-acetic-formaldehyde | Navashin's solution | — | Botanical fixative |
| Picric-acetic-formaldehyde | Bouin's solution; Allen's solution | Rapid penetration | Excellent immobilizer for contractive forms; produces little hardening |
| Dichromate-mercuric-acetic | Zenker's solution | Good penetration; results in brilliant staining | Most widely used histological and pathological fixative |
| Mercuric-acetic | Gilson's solution Carnoy and Lebrun's solution | Excellent Penetration is rapid | Excellent general-purpose fixative Used where penetration of hard objects is required, as seeds |

by washing in water. Trinitrophenol-fixed material must be washed in 70% ethanol to avoid the removal of water-soluble albumen-trinitrophenol complexes. Alcoholic fixatives, such as Carnoy and Lebrun, are washed out in ethanol. After washing, specimens are conventionally preserved in 70% ethanol, although 4% formaldehyde may be equally well used for anything except trinitrophenol-fixed specimens.

Stains and staining. Wholmounts, smears, squashes, and sections are usually stained before being mounted. This process is still justified where it is desired to differentiate chromatically one part of the specimen from another. In many cases, however, this traditional procedure yields no more useful information than the examination of an unstained specimen by phase microscopy.

Dyes used in staining microscopic specimens are acid dyes, which color the nucleus (Table 2), and basic dyes, which color other cellular components. Each class contains dyes which attach directly (direct dyes) and those which attach to an intermediary (mordant) which is either applied before, or in the same solution as, the dye.

Orcein, safranin, methylene blue, and crystal violet are typically direct nuclear dyes, while carmine, hematoxylin, and celestin blue B are commonly used with mordants. Acid dyes will also attach to bacteria and cellulose. The various eosins, orange G, ponceau 2R, light green SF, and methyl blue are basic dyes. These give a general background stain when used as direct dyes, but some may be rendered specific to special cytoplasmic substances when used with

mordants. Thus methyl blue, applied after a solution of phosphotungstic acid, will remain attached only to collagens.

Many basic and acidic dyes form compounds known as neutral dyes which may be used to stain both nucleus and cytoplasm from a single solution. The numerous compounds of various eosins with various thiazins, such as eosin Y-methylene blue, which is used to stain blood films, are in this category.

Basic dyes are commonly applied from simple solutions in water or weak ethanol. Acid dyes, however, usually require compounding into staining solutions, sometimes of considerable complexity and usually of an empirical composition established by tradition. All except orcein, and some oxazines, must be used indirectly; that is, the object is first impregnated with the dye solution and then differentiated, usually with a weak acid, until the dye is removed from the cytoplasm. After this treatment, hematoxylin, which is pink in acid, must be "blued" by treatment with a weak alkali.

Mounting media. A mounting medium should be a preservative. It should have an index of refraction sufficiently high to render objects transparent, and, preferably but not necessarily, it should cause the cover to adhere to the slide. Canada balsam, a natural exudate of *Abies balsamifera*, has all of these properties and is used for most slide preparation. Many proprietary substitutes have been proposed. Glychrogel may be substituted in those cases in which organic solvents would remove cell contents such as fat. Berlese's medium is excellent for small, hard-bodied specimens which are not to be stained. Glycerol may be used for similar purposes, but the edges of the cover slip must be sealed. Fant's medium, applied molten from a heated needle, is satisfactory for sealing glycerol mounts.

Preparation of mounts. Wholmounts, smears, squashes, and sections of tissues are preparations of plant and animal structures for specific purposes. Wholmounts allow the observer to study an entire organism or specific organ structure in some detail. More detailed study of the tissues or cells requires the preparation of sections. Such sectioned material is used in the study of normal and pathological material. Smears are regularly made for bacteriological and blood examinations, while a technique for

TABLE 2. Examples of nuclear stains

| Stain | Use |
|---|--|
| <i>Direct nuclear stain</i> | |
| Lillie's ammonium oxalate crystal violet | Stain bacterial smears |
| LaCour's acetocarmine | Squashed cells for chromosomes |
| Gray's celestin blue B | Sections and small objects |
| <i>Indirect nuclear stain</i> | |
| Grenacher's alcoholic borax carmine | Wholmounts of small plants and animals |
| Ehrlich's acid alum hematoxylin | Plant and animal sectioned tissues |
| Johansen's safranin | Sections of plant and animal tissues |

chromosome study is the employment of squashes, in which the material to be studied is literally flattened and squashed.

Wholemounds. Wholemounts of objects which can be preserved dry, such as seeds, moss archegonia, microfossils, and small beetles, may be attached with mucilage of tragacanth to a piece of black paper cemented to a slide. A low cylinder, punched from a sheet or cut from a tube, will serve to support the cover slip. Cylinder and cover may be cemented in place with any cellulose ester cement.

Small objects which must be rendered transparent but which need not be stained mount excellently in Berlese's medium. Thus, the majority of small arthropods may be placed, without preliminary fixation or preservation, in a drop of this medium and a cover slip applied.

Most wholemounts, however, are prepared as stained specimens in balsam. This involves the following sequence of operations: fixation, preservation, staining, dehydration, clearing, and mounting.

Fixation, in one of the fluids previously discussed, may have to be preceded by narcotization in the case of contractile invertebrates. Experience alone can determine the most successful narcotic to be used. All but the most delicate specimens, such as Rotatoria, medusae, and fresh-water Bryozoa, should be fixed in hot (50–70°F, or 10–21°C) fluids after narcotization. In many cases, hot fixation of *Hydra* and *Amoeba* without narcotization is quite adequate. The fixed, washed specimens are almost invariably preserved in 70% ethanol.

Wholemounts are best stained in a nuclear stain which differentiates internal organs, either by virtue of their density, which delays the extraction of the dye during differentiation, or because of the greater number of nuclei in small-celled organs. In any case, the specimen is taken from alcohol and placed in, for example, Grenacher's carmine overnight. The specimen is then rinsed in 0.1% hydrochloric in 70% ethanol and transferred to a fresh supply of the same reagent in which the process of differentiation may be watched. These transfers, and those subsequently required in the course of dehydration and clearing, are most readily made with the aid of cloth-bottomed tubes. These tubes are easily prepared by cutting the top from a test tube and tying bolting silk over the inverted rim.

Dehydration is necessary since water would prevent penetration of the balsam. The specimen is accordingly transferred from the acid alcohol to 95% ethanol and thence to pure ethanol. A half-inch-long (1.27-cm) leech will require 24 h in each. A fern prothallus will be perfectly dehydrated after 10 min.

Clearing the dehydrated specimen involves substituting some substance for the alcohol which is miscible with balsam. Wholemounts are usually cleared in essential oils or their synthetic equivalents. Clove oil and thyme oil are widely used, but terpineol renders specimens less brittle.

The specimen is now lifted from the clearing agent and placed in a drop of Canada balsam on a slide. A cover is then lowered vertically in place (Fig. 1).

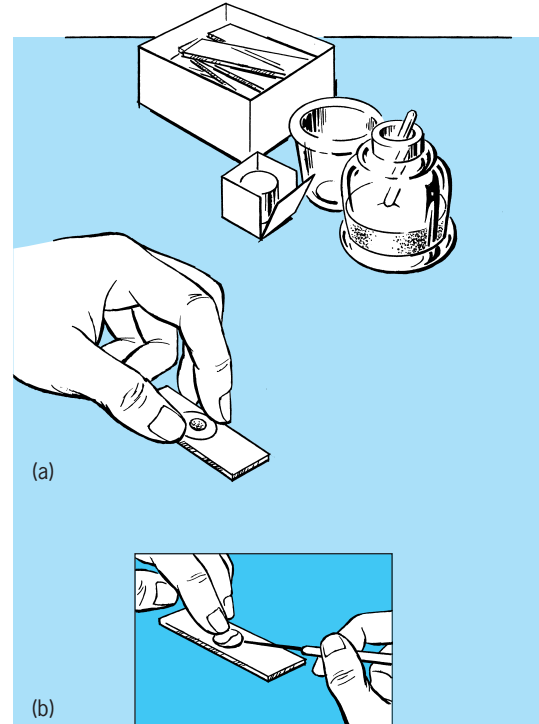


Fig. 1. Materials and method for applying cover slip to balsam wholemount. (a) Correct way; notice that the cover slip is held vertically centered over the drop of balsam. (b) Wrong way; this method draws the object to one side. (After P. Gray, *Handbook of Basic Microtechnique*, 3d ed., McGraw-Hill, 1964)

Smears and squashes. Smears are the most easily prepared microscope slides. A drop of organic fluid, such as blood, is placed on the end of a clean slide; a second slide is touched to it (Fig. 2) and pushed forward to leave a uniform film.

Blood smears of this type are usually stained in a methylene blue–eosinate dissolved in methanol, of which Wright's stain is a typical example. This stain also requires a phosphate buffer with pH of 6.4.

The air-dried smear is flooded with a specific amount of stain and laid on a rack, or across the top of a beaker, for 1 min. Buffer is then added to the stain from a drop bottle in the proportion of two drops of buffer to each drop of stain. After 2 min, the mixture is washed from the slide with distilled water. Blood smears are usually preserved in the dry state.

Squashes are just what their name indicates, and their success is dependent on the condition of the material selected. Anthers of plants, or the testes of insects, may be squashed directly, but root tips or plant ovaries require softening before squashing. Cellulases derived from the snail's stomach are much used for this preliminary.

The salivary glands of *Drosophila* are so widely prepared by this technique that they will serve as an example. Glands are taken from a third instar larva, easily recognized by the sluggish movements with which it crawls on the side of the culture vessel, by pulling them out with the head. This is best performed in a drop of LaCour's acetoorcein. Excess

tissue is rapidly stripped away, a cover slip placed on top, and strong pressure exerted with the forefinger through a sheet of bibulous paper. Inexperienced technicians may find it desirable to dissect out the glands in a drop of saline before transferring them to stain. A successful preparation shows sharply stained chromosomes against a faintly pink background.

These preparations may be rendered permanent. The slide is frozen for about 5 min on a block of dry ice and the cover slip is then gently levered off with a safety razor blade. The squashed material adheres to the slide, which is next placed in a coplin jar containing 70% ethanol before being dehydrated, cleared, and mounted in the same manner as a section.

Sections. All investigations of the structure of cells, or of the relations of cells within organs, require the examination of thin slices known biologically as sections.

Optical microscopy. The useful range of thickness for examination under the optical microscope is from about 25 to 2 micrometers; 10- μ m sections are routine in most histological and pathological work.

Thick sections of crisp or waxy structures, particularly plant stems and roots, may be cut by hand. The razor is drawn across the plate with gentle pressure and the section is then washed into a stender dish. Such sections are subsequently stained and mounted, as though they were wholmounts, by the methods described above.

Most biological specimens, however, do not have the necessary consistency for such treatment but must be impregnated with, and embedded in, a substance which sections well. Paraffin is the most widely used. The preparation of paraffin sections requires the following steps: fixation, preservation, dehydration, clearing, impregnation, embedding, cutting, mounting on slides, dewaxing, rehydrating, staining, differentiating, counterstaining, dehydrating, clearing, and mounting.

Fixation, preservation, and dehydration of small pieces of tissue are the same as for the wholmounts described above. Clearing, however, is in this case preliminary to impregnation with molten wax, and xylol is the reagent of choice. Five-millimeter cubes of animal tissue should have at least 2 h in one change of 70% ethanol, two changes of 95% ethanol, two changes of pure ethanol, and two changes of xylol before being placed in molten wax. Plant tissues, in which the protoplasm tends to shrink from the cell wall, should be passed through the following sequence of reagents: (1) 5% ethanol, (2) 10% ethanol, (3) 18% ethanol, (4) 30% ethanol, (5) 40% ethanol plus 10% *tert*-butanol, (6) 50% ethanol plus 20% *tert*-butanol, (7) 50% ethanol plus 35% *tert*-butanol, (8) 40% ethanol plus 55% *tert*-butanol, (9) 25% ethanol, 75% *tert*-butanol, (10) pure *tert*-butanol. It should be understood that, for example, solution (7) of this series is prepared by mixing 50 parts of ethanol with 35 parts of butanol and then "making up" to 100 parts, by addition of water. Five-millimeter cubes of plant tissue will be adequately dehydrated

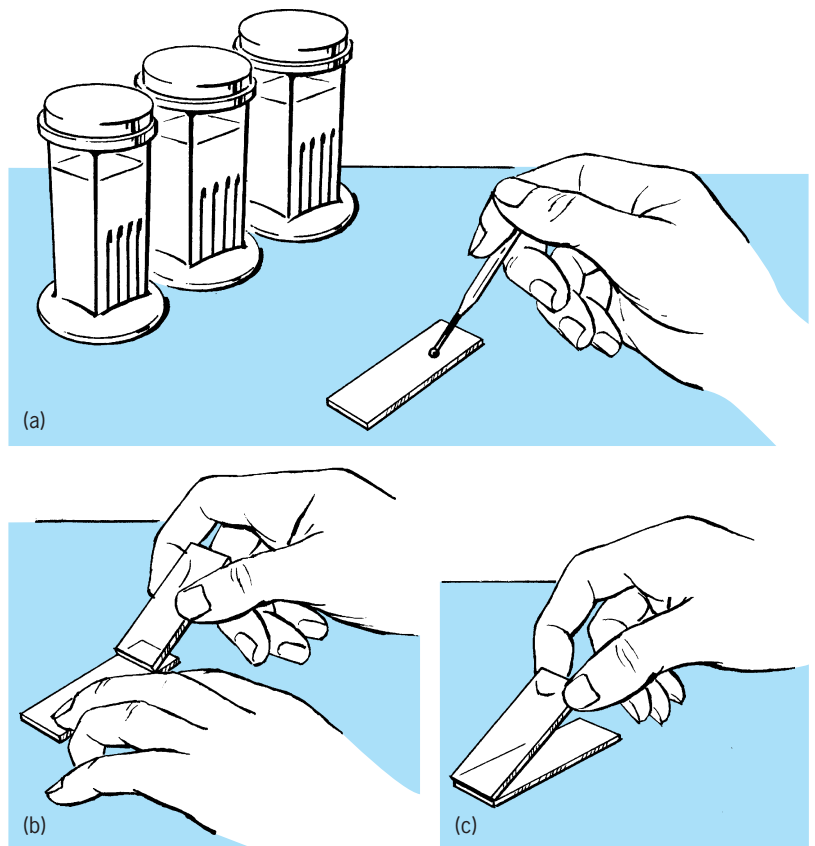


Fig. 2. Making a smear preparation. (a) Place the drop about 1 in. (2.5 cm) from end of slide. (b) Apply a second slide just in front of the drop. (c) Push slide smoothly forward to spread the smear. (After P. Gray, *Handbook of Basic Microtechnique*, 3d ed., McGraw-Hill, 1964)

and cleared after 1 h in each of the solutions except (6), in which the pieces should remain for 6–12 h.

The tissues, whether cleared in xylol or butanol, must now be transferred to molten paraffin, maintained at just above its melting point in a thermostatically controlled oven or bath. Wax of 129–133°F (54–56°C) melting point, held at 136.4°F (58°C), is

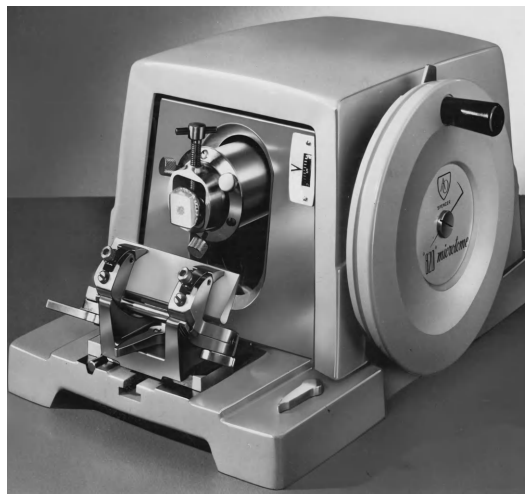


Fig. 3. Rotary microtome produced by the American Optical Co. (From P. Gray, *Handbook of Basic Microtechnique*, 3d ed., McGraw-Hill, 1964)

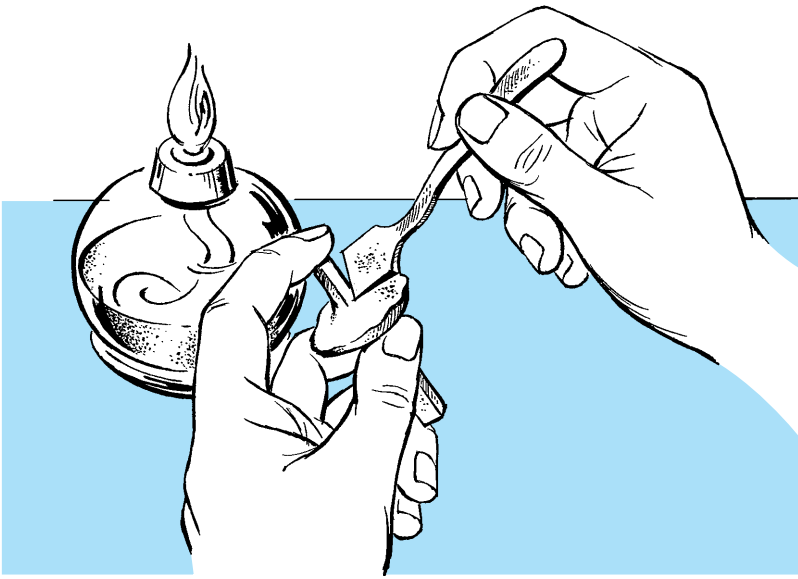


Fig. 4. Method of mounting the wax block on the block holder. (After P. Gray, *Handbook of Basic Microtechnique*, 3d ed., McGraw-Hill, 1964)

conventional. Softer wax, 126–129°F (52–54°C), is difficult to cut but should be used, at 131°F (55°C), for muscular tissues, such as heart and tongue, which tend to harden at 136.4°F (58°C). Harder waxes are rarely used. It is customary to add wax shavings to the tube containing delicate specimens in solvent, to leave this overnight, and then to place the tube in the oven. Pieces of homogeneous tissue like liver may be transferred directly to molten wax.

The impregnated pieces must now be embedded in a wax block. To this end paper boats, or other suitable containers, are filled with molten wax, the specimen is placed in the boat, a heated pipet is used to melt any film which may have formed around the specimen, and the block is then chilled and solidified by partial immersion in water. A satisfactory block is translucent and almost flat on top. A chalky block has either been cooled too slowly or made from solvent-contaminated wax. A block with a deep conical depression in the top has been cooled too fast.

Sections are cut from the block after it has been mounted on a microtome (Fig. 3). The revolution of the handle of this device raises and lowers the block vertically against a knife. The block is advanced horizontally at the top of its rise by the action of a micrometer screw advancing a pressure point along a spring-loaded diagonal plate. The extent of the advance, and thus the thickness of the section, is controlled by varying the moment of engagement of the pawl with the ratchet wheel attached to the micrometer screw.

The block of wax containing the object is removed from the container in which it was cast and trimmed until about equal volumes of specimen and wax remain. It is essential that two of the faces be parallel. The trimmed block is fused (Fig. 4) to the object carrier of the microtome. The carrier is placed in the holder of the microtome with the parallel faces parallel to the knife edge, adjusted until it almost touches the knife, and the handle turned until the ribbon of sections starts to form. The commonest defects are that the ribbon splits because of a nick in the knife edge or that the ribbons roll into cylinders without

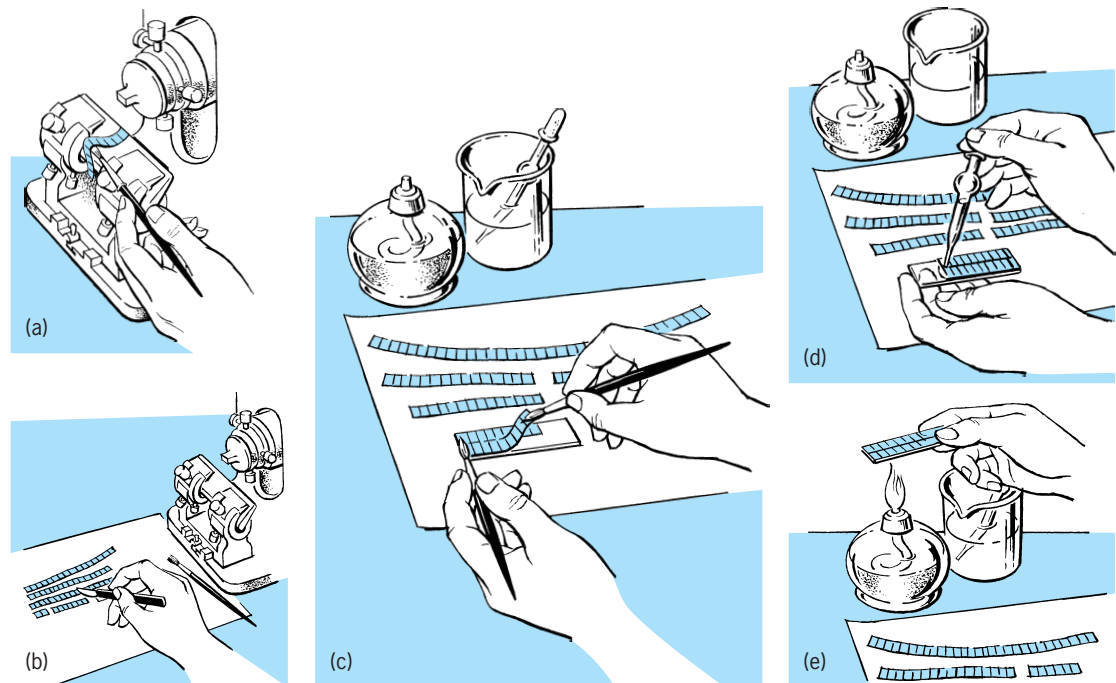


Fig. 5. Basic steps in cutting and mounting paraffin ribbons prior to staining. (a) Starting the ribbon. (b) Cutting the ribbon in lengths. (c) Mounting the dry ribbon. (d) Flooding the ribbons. (e) Warming the flooded ribbons to flatten them. (After P. Gray, *Handbook of Basic Microtechnique*, 3d ed., McGraw-Hill, 1964)

forming sections. The latter defect is due to a faulty relationship between the hardness of the block and the angle at which the knife strikes it. This relation must be established empirically for each block. The block lifting the ribbon from the knife blade sometimes results from a faulty angle and sometimes from a dirty knife.

When enough sections have been cut, the ribbon is divided (**Fig. 5**) into pieces about 2 in. (5 cm) long in preparation for mounting. A clean slide is lightly smeared with a mixture of equal parts of egg albumen and glycerin. The pieces of ribbon are laid on the prepared slide and flooded with water before being gently warmed until flat. The excess water is then drained off and the slide set on a warming table, at about 95°F (35°C), until dry.

The sections are now ready to be stained. To prepare the conventional celestin blue B–eosin histological slide, the following coplin jars (**Fig. 6**) should be set up: (1) xylol, (2) 50% xylol-ethanol, (3) pure ethanol, (4) 95% ethanol, (5) 70% ethanol, (6) water, (7) Gray's celestin blue B, (8) 0.2% ethyl eosin in 95% ethanol, placed behind the 95% ethanol jar. The slide is warmed until the wax melts, dropped into xylol for 1 min and then passed down the series, with about 30 s in each jar, until it is in water (6). It is then transferred to stain (7) for 1–2 min, rinsed in water, and transferred up the series again until it reaches 95% ethanol (4). The slide is now dipped up and down in the eosin solution (8) until the sections are sufficiently yellow—a point on which opinions vary widely. A quick rinse in 95% ethanol (4) precedes transfer to pure ethanol (3) for about 30 s and thence into xylol. The slide is then withdrawn from xylol, an adequate amount of balsam is placed on the surface, and a cover slip is applied.

Electron microscopy. The ultrathin (20–30 nm) sections used in electron microscopy are prepared by essentially similar methods but require different equipment and materials. Animal tissues are commonly fixed in Palade's fixative, a 1% solution of osmium tetroxide buffered to pH 7.4, but plant tissues are more usually fixed in Luft's fixative, which is a 1% solution of potassium permanganate buffered to the same pH. Both solutions differentially deposit electron-scattering materials in the tissue and thus function also as "stains."

The fixed and washed fragments, usually about 0.25-mm cubes, are dehydrated and "cleared" in acetone or, more rarely, ether before being placed in a plastic monomer which is subsequently catalyzed. Neither the hardness nor crystalline structure of wax permits ultrathin sections, but plastics have the necessary colloidal structure and their hardness can be adjusted at will. Mixtures of methyl methacrylate (very hard) and butyl methacrylate (very soft) were preferred in the past, but the heat generated in the course of polymerization has caused their abandonment in favor of epoxy resins, the hardness of which can be modified by the addition of plasticizers.

The "blocks" are cast in standard pharmacist's

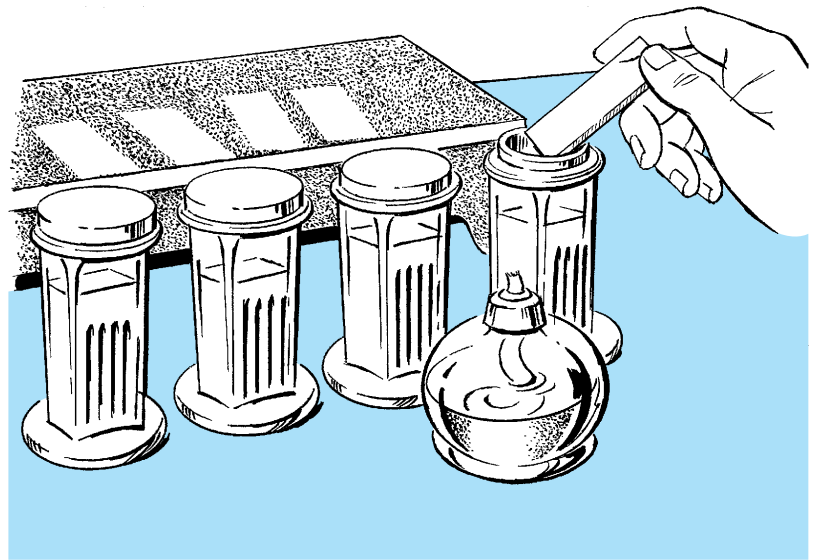


Fig. 6. Coplin jars for starting a slide through the reagent series. (After P. Gray, *Handbook of Basic Microtechnique*, 3d ed., McGraw-Hill, 1964)

gelatin capsules and the end of the block trimmed to a tiny pyramid containing the specimen before mounting in the chuck of the ultramicrotome. This, as the regular microtome, moves the block up and down against a knife edge while advancing the block a given amount between each stroke. The advance is controlled either by a micrometer screw working against a motion-reducing lever or, in most modern instruments, by controlled thermal expansion. The very slight forward motion, coupled with a frequent irregularity of cut, necessitates either withdrawing the block slightly on the upward stroke or giving it a circular motion which avoids the knife on the upward stroke.

The best knife is undoubtedly a freshly broken edge of a 45° wedge of plate glass, but specially ground diamond knives are used for extremely hard materials. The sections, sometimes "stained" in solutions of lead or uranium salts, are collected on wire mesh grids for examination under the electron microscope. See ELECTRON MICROSCOPE.

Finishing and storing slides. Mounts made with balsam should be hardened on a warm plate at about 113°F (45°C) for 3 or 4 days. Surplus exuded balsam may then be wiped off with a rag moistened with xylol. Another day of hardening will prepare the slide for a final cleaning in warm soapy water, after which the permanent label may be attached.

Wholemounds must be, and sections are better, stored in flat trays than in vertical grooves. Slides so stored, in a cool dark place, do not alter appreciably in 50 years.

Peter Gray
Bibliography. G. P. Berlyn and J. P. Mikshe, *Botanical Microtechniques and Cytochemistry*, 1976; W. Burrells, *Microscope Techniques*, 1980; P. Gray, *Handbook of Basic Microtechnique*, 3d ed., 1964; P. Gray, *The Microtomists' Formulary and Guide*, 1954, reprint 1975; J. K. Presnell et al., *Humason's Animal Tissue Techniques*, 5th ed., 1997.

Microwave

Electromagnetic energy with wavelengths in free space ranging roughly from 0.3 to 30 cm (**Fig. 1**). Corresponding frequencies range from 1 to 100 GHz. Frequency and wavelength are related by Eq. (1),

$$f\lambda = c \quad (1)$$

where f is the frequency, λ is the free-space wavelength, and c is the velocity of light in vacuum, approximately 3×10^8 m/s. The range shown is an arbitrary one, especially at the higher frequency end, where wavelengths are in the millimeter and submillimeter range and similar techniques are used. See ELECTROMAGNETIC RADIATION; FREQUENCY (WAVE MOTION); SUBMILLIMETER-WAVE TECHNOLOGY; WAVE MOTION; WAVELENGTH.

The first portions of the microwave spectrum to be developed were around 30 cm and 10 cm. These bands are designated L (for long) and S (for short), respectively. As other bands were developed, other letter designations were added (see **table**).

Characteristic transmission media for microwaves are hollow-pipe waveguides, where the cross-sectional dimensions are of the order of the wavelength and thus are of convenient size. Coaxial transmission lines are also used, however, especially in the lower-frequency bands, and various stripline techniques are used on microwave integrated circuits. Resonant cavities are commonly used as circuit elements, and radiation or reception of the energy is typically by horns, parabolic reflectors, or arrays. See ANTENNA (ELECTROMAGNETISM); CAVITY RESONATOR; COAXIAL CABLE; ELECTROMAGNETIC WAVE TRANSMISSION; TRANSMISSION LINES; WAVEGUIDE.

Generation. For most applications, microwaves are generated in electronic devices that produce oscillations at a frequency in one of the frequency bands (see **table**). The devices may be single-frequency or tunable, and continuous-wave (cw) or pulsed. Vacuum-tube generators include klystrons, magnetrons, and backward-wave oscillators; solid-state generators include tunnel diodes, Gunn diodes, IMPATT diodes, transistor oscillators, masers, and harmonic generators using varactor diodes. The vacuum-tube generators are used to produce higher powers, which can be as much as thousands of kilowatts. Solid-state generators were formerly limited in power to a few watts, but their power capabilities are continually increasing and now may reach hundreds of watts. See GYROTRON; KLYSTRON; MAGNETRON; MICROWAVE SOLID-STATE DEVICES; MICRO-WAVE TUBE; TRAVELING-WAVE TUBE.

For some applications, microwave noise is useful.

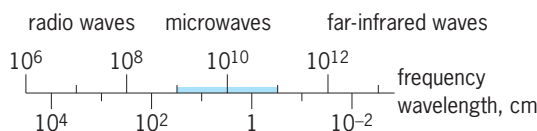


Fig. 1. Portion of the electromagnetic spectrum, with microwave range indicated.

| Microwave frequency bands* | |
|----------------------------|----------------------|
| Microwave band | Frequency range, GHz |
| L | 1.120–1.700 |
| LS | 1.700–2.600 |
| S | 2.600–3.950 |
| C (G) | 3.950–5.850 |
| XN (J, XC) | 5.850–8.200 |
| XB (H, BL) | 7.05–10.00 |
| X | 8.20–12.40 |
| KU (P) | 12.40–18.00 |
| K | 18.00–26.5 |
| V (R, KA) | 26.5–40 |
| Q (V) | 33–50 |
| M (W) | 50–75 |
| E (Y) | 60–90 |
| F (N) | 90–140 |
| G (A) | 140–220 |
| R | 220–325 |

*After T. K. Ishii, *Microwave Engineering*, Harcourt Brace Jovanovich, 1989.

This noise, generated in ordinary fluorescent light bulbs mounted in waveguides, is uniform across a wide band of frequencies, and the power emitted is known accurately and is quite stable with respect to time. Generators of this type are used in measuring the noise performance of receivers and, after amplification to high power, as a jamming signal for radar and communication systems. Microwave noise is also generated by current passed through a solid-state diode. See ELECTRICAL NOISE; ELECTRONIC WARFARE; JAMMING; MICROWAVE NOISE STANDARDS.

Circuit elements. Physical elements which produce specific effects on microwaves are called circuit elements. Microwave circuit elements bear little resemblance to circuit elements at lower frequencies because the shorter wavelength leads to an entirely different method of transmitting microwaves from place to place within a system (**Fig. 2**).

Waveguide. The most common method of microwave transmission within a system is through hollow circular or rectangular metal tubes of uniform cross section called waveguides. The microwave energy is confined within these tubes and guided along them.

In general, microwave energy is conveyed by an electromagnetic field whose components are the electric and magnetic intensities, denoted by E and H respectively. In a waveguide, only certain specific patterns of E and H , known as modes, can exist. The exact nature of the modes is determined by Maxwell's equations and the boundary conditions; the latter consist of the cross-sectional dimensions and the electromagnetic properties of the waveguide. Two types of mode can exist in an ideal hollow waveguide, the transverse electric (TE) and the transverse magnetic (TM) modes. See MAXWELL'S EQUATIONS.

The cross-sectional dimensions of a waveguide determine a frequency and a corresponding free-space wavelength below which no transmission is possible for each mode; this is the cutoff frequency or wavelength. The mode with the lowest cutoff frequency

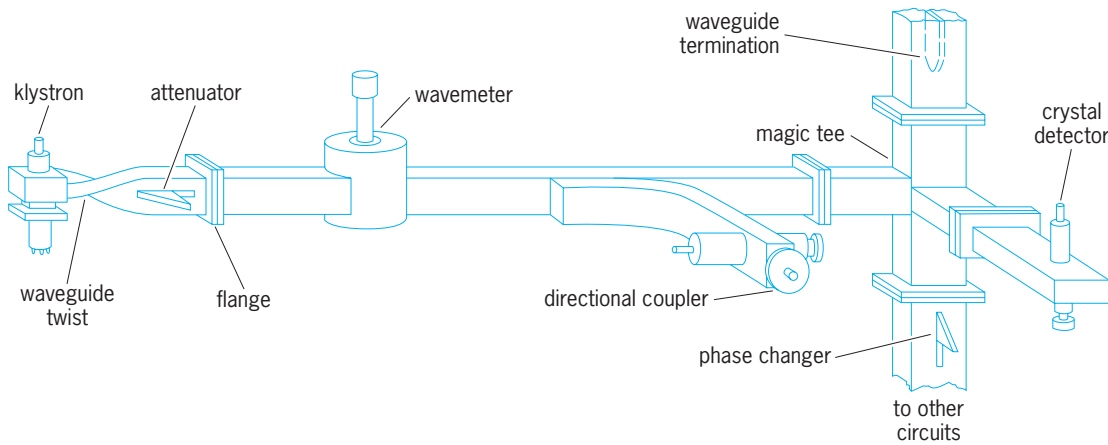


Fig. 2. Sample microwave circuit illustrating some microwave components.

is called the dominant mode; the other modes are called higher-order modes. In most applications, the waveguide dimensions are chosen so that transmission can occur only in the dominant mode, the other modes having cutoff frequencies well above the operating frequency (Fig. 3).

Most waveguides are rectangular in cross section. They are used to carry microwaves between other circuit elements and as integral parts of other circuit elements such as attenuators and phase shifters. The points at which microwaves enter or leave a circuit element are called ports; a length of waveguide has two ports.

Stripline. Various forms of stripline are used in interconnecting components on a dielectric or semiconductor substrate when microwave devices are integrated. One important example is the microstrip,

in which a metallic strip or ribbon is placed on a thin dielectric, which is in turn backed by a conducting ground plane. Another configuration is the coplanar strip, in which two parallel strips, separated by a gap, are placed on the same surface of a dielectric. Still other variations are possible. These may be regarded as special cases of transmission lines, but since fields penetrate both the dielectric and air, they have dispersive properties; that is, different frequency components propagate with different velocities. Striplines have more losses than hollow-pipe waveguides but are generally used over very short distances. By proper configurations of the strip, typical functional transmission-line elements, such as impedance-matching sections, directional couplers, isolators, and filters, can be made.

Microwave filter. Filters are needed in communication or information-processing systems for blocking of high frequencies (low-pass), blocking of low frequencies (high-pass), elimination of undesired bands (band elimination), or passing of desired bands while attenuating others (band pass). All of these may be made for microwaves by adding periodic perturbations, such as posts, irises, diaphragms, or dimensional variations, to the waveguide or other transmission system. Simple designs may be made by cascading quarter-wave sections of different propagation characteristics. More sophisticated designs use transformations from low-frequency, lumped-element designs. See ELECTRIC FILTER; MICROWAVE FILTER.

Attenuator, phase shifter, and termination. A thin sheet of plastic can be used to alter the amplitude or phase of microwaves. If the sheet is coated with powdered carbon with appropriate electrical conductivity and placed in a waveguide with the lossy material parallel to the lines of electrical intensity, it will absorb microwave power. Variable attenuation can be achieved by mechanically inserting more or less of the lossy strip into the path of the wave. See ATTENUATION (ELECTRICITY).

A phase shifter changes the phase of a microwave without changing its amplitude. It can be constructed in the same manner as an attenuator without the lossy material.

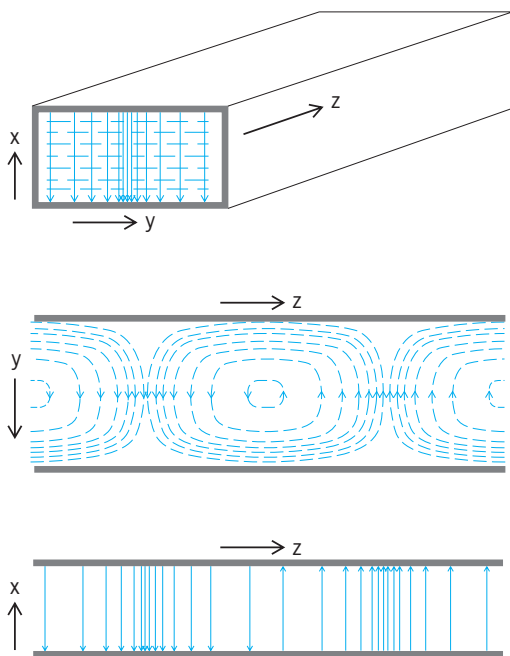


Fig. 3. Instantaneous field pattern for the TE_{01} wave, the dominant TE mode in a rectangular waveguide. The wave propagates in the z direction. Solid lines indicate the electric intensity E , and broken lines the magnetic intensity H .

An attenuator which has a very large loss and is closed at one end is called a termination; it absorbs all the power transmitted into it, reflecting none.

Detector. The most common microwave detector is a silicon diode designed for high frequencies and mounted in a waveguide or a stripline. One terminal of the diode is connected to the waveguide, and the other is connected to a wire post, which is parallel to the direction of the electric field E and brought out of the waveguide through a small hole. The diode rectifies the microwave signal, producing an average current which can be indicated by a direct-current meter connected between the diode terminals. If the microwave signal is modulated in amplitude, the modulation will appear in the output current. See AMPLITUDE MODULATION; SEMICONDUCTOR DIODE.

The bolometer is a detector which absorbs microwave power, causing a temperature increase and a corresponding change in resistance. The bolometer does not respond fast enough to detect high-frequency modulation. It is often used as one arm of a resistance bridge circuit in microwave power meters. See BOLOMETER; MICROWAVE POWER MEASUREMENT.

Antenna. A transmitting antenna takes microwave power from a waveguide and converts it into a plane wave that propagates through space to a distant receiving antenna. Two important characteristics of antennas are efficiency and directivity, efficiency being the ratio of the power delivered into space to the power available in the waveguide. High directivity is accomplished by large antennas which focus the microwave energy in the same way a searchlight focuses a beam of light. In the parabolic antenna, a small waveguide feed horn (Fig. 4) illuminates a large paraboloidal dish which concentrates the energy in a narrow angular beam. In the horn-reflector antenna, a waveguide is expanded into a horn to illuminate a large paraboloidal surface (Fig. 5) which reflects the

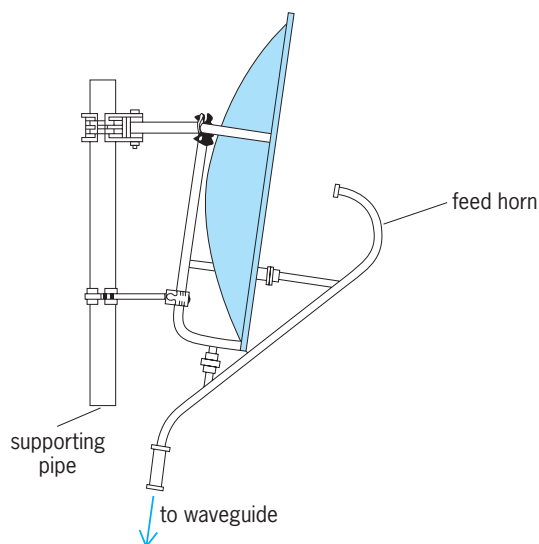


Fig. 4. Parabolic microwave antenna. (After K. L. Dumas and L. G. Sands, *Microwave Systems Planning*, Hayden Book Co., 1967)

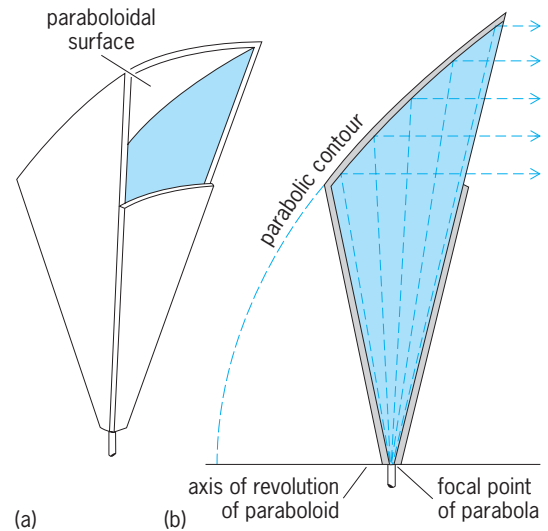


Fig. 5. Horn-reflector antenna. (a) Oblique view. (b) Cross section. (After S. A. Schelkunoff and H. T. Friis, *Antenna Theory and Practice*, John Wiley and Sons, 1952)

microwave energy in a narrow angular beam. Directivity is described by the beam width W of an antenna which is approximately given in radians by Eq. (2),

$$W = \frac{\lambda}{D} \quad (2)$$

where λ is the wavelength and D is the dimension of the antenna aperture. For circular antennas, D is the diameter. For rectangular antennas, D is either dimension, the beamwidth applying in the plane of the dimension. Directivity is also related to the gain of the antenna, which is the ratio of the power received at a distant receiver to the power that would be received if the transmitted power were radiated from an antenna with a uniform spherical pattern. Most microwave applications require highly directive antennas; gains of several thousand are not unusual, corresponding to beamwidths ranging from a fraction of a degree to several degrees. See DIRECTIVITY.

When receiving, an antenna intercepts power from the incident microwave proportional to the area of the antenna aperture. The gain and directivity of an antenna are the same whether it is used to transmit or receive. Efficiency of large antennas ranges from 50% to nearly 100%.

Gyrotator, circulator, and isolator. The gyrotator is a lossless, nonreciprocal, two-port circuit which has 180° more phase shift in one direction than in the other. This principle is used in the broadband microwave circulator. The three-port circulator has the property that all the power into port 1 exits at port 2, all the power into port 2 exits at port 3, and all the power into port 3 exits at port 1 (Fig. 6). The nonreciprocal phase shift is achieved in a magnetic ferrite placed in the waveguide junction and magnetized with an external permanent magnet. An isolator is a circulator with one port terminated, resulting in a circuit which transmits power in one direction and not the other. The input circuit is thus isolated from the output circuit. When the terminated port is internal, the

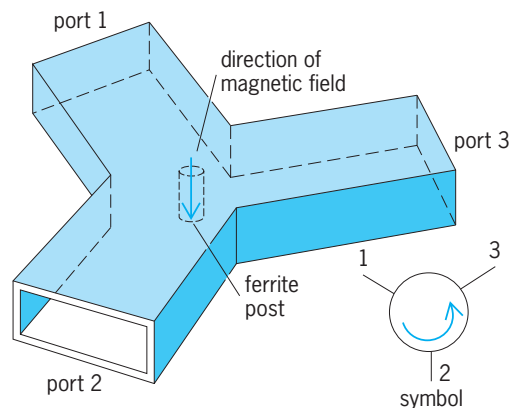


Fig. 6. Microwave circulator. (After G. J. Wheeler, *Introduction to Microwaves*, Prentice-Hall, 1963)

isolator appears to be a two-port element. See GYRATOR.

Varactor. This is a solid-state diode whose capacitance changes with applied voltage. Varactors are used as harmonic generators to obtain microwave power efficiently from lower-frequency sources such as quartz crystal-controlled oscillators in the 10–100-MHz range. Varactors are also used in up-converters and down-converters performing the same functions as similar circuits using resistive diodes, but more efficiently and at the expense of narrower bandwidth. See OSCILLATOR; VARACTOR.

Amplifier. A microwave amplifier converts a low-power input signal to a higher-power output signal while preserving one or more characteristics. A linear amplifier preserves the amplitude, frequency, and phase of the input signal. When a linear amplifier is overloaded, it becomes saturated: the output amplitude tends to remain constant. The envelope of the input signal becomes distorted, and additional frequency components are introduced in this nonlinear regime.

High power output is achieved with klystron and traveling-wave tube (TWT) amplifiers, both of which can be operated in the linear or saturated modes; the traveling-wave tube amplifier has the larger bandwidth, and the klystron has the higher power. Moderate power can be achieved with transistor amplifiers, and these are continually being extended in their frequency range of usefulness. Silicon bipolar transistors dominate at frequencies up to about 4 GHz, with gallium-arsenide metal-semiconductor field-effect transistors (MESFETs) dominant beyond this. In an injection-locked-oscillator amplifier, a high-amplitude oscillator is locked in phase and frequency to a low-amplitude input signal coupled into the oscillator circuit. The output of the oscillator is unaffected by the input signal; therefore only the phase and frequency of this type of amplifier are preserved. For example, the IMPATT diode can be operated as an injection-locked oscillator amplifier in the saturated mode. These are one-port amplifiers, the output power emerging from the input port, and a circulator is used to convert them into two-port amplifiers.

Very low noise levels are achieved in the maser and the parametric amplifier, both of which require power at a single frequency to pump the active element. In the maser, energy is pumped from one atomic level to another, and in the parametric amplifier a varactor is driven by the pumping signal to provide the nonlinear capacitance required for amplification. See AMPLIFIER; MASER; PARAMETRIC AMPLIFIER.

Microwave integrated circuits. Microwave integrated circuits (MICs) are of two types, hybrid MICs and monolithic microwave integrated circuits (MMICs). In the hybrid circuits, some or all of the active and passive devices are added to a dielectric or semiconducting substrate and interconnected by striplines as discussed above. For MMICs, all of the active and passive devices of the functional unit are formed on the substrate, and interconnected by striplines through a variety of microfabrication techniques, including photolithography, epitaxy, ion implantation, etching, diffusion, sputtering, and evaporation. Substrate materials include insulating materials such as alumina or sapphire or semiconducting materials such as silicon or gallium arsenide (GaAs). Conducting materials are gold, silver, copper, or aluminum, and dielectric films are such materials as silicon monoxide (SiO), silicon dioxide (SiO₂), or silicon nitride (Si₃N₄).

MMIC technology promises small size, improved reliability, reproducibility and low cost if produced in volume. MMIC switches using gallium-arsenide field-effect transistor devices have given switching speeds of less than 2 nanoseconds with a bias of only 350 microwatts. An entire X-band microwave receiver including local oscillator, mixer, low-noise amplifier, and intermediate-frequency amplifier has been formed on a single chip of dimension 0.25 × 0.25 in. (6 × 6 mm). See INTEGRATED CIRCUITS.

Polarization. An important property of a microwave is its polarization, the direction of its electric field E with respect to a fixed plane. A microwave transmitted parallel to the Earth is said to be vertically polarized if E is perpendicular to the Earth's surface. Two superimposed waves of the same frequency and magnitude, and in both time and phase quadrature, result in a circularly polarized wave in which the electric field direction rotates either clockwise or counterclockwise. See POLARIZATION OF WAVES.

Circuit elements can be constructed whose effect is polarization-sensitive. For example, closely spaced, vertical wires in a plane normal to the direction of a wave will completely reflect a vertically polarized wave and pass, without modification, a horizontally polarized wave.

Microwave receiver. The most simple microwave receiver is a silicon diode detector followed by an amplifier; amplitude modulation on the input signal is detected directly and amplified to a suitable level. This receiver is insensitive and noisy and cannot be used with frequency or phase modulation. See FREQUENCY MODULATION; MODULATION; PHASE MODULATION.

The first active element in nearly all microwave receivers is a silicon diode operated as a down-converter. In this type of receiver, a strong, continuous-wave local oscillator signal is used to pump the diode over its nonlinear resistance range. In this manner, the local oscillator and the input signal are mixed, shifting the input signal down to an intermediate frequency, which is the difference between the frequencies of the local oscillator signal and the received signal. Intermediate frequencies of a few tens of megahertz are common. Frequency, phase, or amplitude modulation on the received signal appears in the detector output at the intermediate frequency. A bandpass intermediate-frequency amplifier, providing most of the gain of the receiver, follows the detector, after which a demodulator converts the modulation on the intermediate-frequency signal to usable form, for example, an audio or a television signal. *See* RADIO RECEIVER.

If extremely low-noise performance is required, the first active component is a maser or a parametric amplifier. These are not required in most applications.

Microwave transmitter. The main components of a microwave transmitter are a microwave power source, a modulator, and, if necessary, a power amplifier. Among the vacuum-tube power sources, the reflex klystron and the backward-wave oscillator are frequency-modulated, and the magnetron is pulse amplitude-modulated, by a signal on an electrode. In solid-state transmitters, by contrast, the modulator is usually separate from the continuous-wave power source. The modulation can be done directly at microwave frequencies or it can be performed at intermediate frequency and shifted to the microwave frequency in an up-converter, which is very much like a down-converter. A common frequency modulator consists of an intermediate-frequency oscillator tuned by the modulation signal applied to a varactor. *See* PULSE MODULATION.

The power amplifier can be a klystron or traveling-wave tube when higher powers are required. For solid-state transmitters, transistor or injection-locked-oscillator amplifiers are common.

Microwave propagation. In free space, microwaves travel in straight lines as do optical waves. Near the Earth, however, the atmosphere has an index of refraction which normally decreases with distance above the Earth and causes the wave to travel in a circular path which bends slightly toward the Earth. The radius of the circular path is larger than the Earth's radius, so normally a wave traveling above the surface of the Earth and parallel to it will not intercept the Earth. Microwaves are reflected and refracted by objects just as are optical waves. *See* MICROWAVE OPTICS.

Occasionally during the summer, atmospheric conditions cause microwaves transmitted from an antenna to travel to a receiver via two or more paths. These waves interfere at the receiver and may cause large decreases in the received signal amplitude. This phenomenon, called multipath fading, is a serious problem in microwave transmission parallel to

the surface of the Earth. Other atmospheric conditions cause the microwave to bend away from the Earth and miss the receiver. This problem is solved by placing the transmitting and receiving antennas on tall towers; the wave initially heads toward the Earth, bends away before intercepting the Earth, and reaches the receiver if it is high enough. If the towers are not sufficiently high, the wave is intercepted by the Earth, resulting in what is known as earth-bulge fading. When microwaves are directed well above the horizon, neither of these two problems occurs. Thus satellite microwave systems do not suffer from either multipath or earth-bulge fading.

At frequencies above about 10 GHz, rain absorbs microwave energy, resulting in large signal losses. Both satellite and point-to-point microwave systems are seriously affected by rain attenuation.

For most frequencies the attenuation of microwaves by the Earth's atmosphere is very small. There are, however, bands of frequencies for which the loss is higher due to molecular absorption. Attenuation due to water vapor occurs in several bands, the first one appearing at about 30 GHz. A larger attenuation, due to oxygen absorption, occurs at 60 GHz. *See* RADIO-WAVE PROPAGATION.

Applications of microwaves. Areas in which microwave radiation is applied include radar, communications, radiometry, medicine, physics, chemistry, and cooking food.

Radar. This is an acronym for radio detection and ranging. In one form a pulse of electromagnetic energy is transmitted in a narrow beam toward a target, and part of the energy is reflected to the receiver. The time difference between the transmitted and received pulses provides the distance to the target, and the direction of the antenna beam provides the target's direction. Radar is used in military applications, commercial aviation, remote sensing of the atmosphere, and astronomy. The high antenna directivity and the excellent propagation characteristics of microwaves in the atmosphere make this the preferred band for radar applications. *See* RADAR.

Electronic countermeasures. Microwaves are also used in electronic countermeasures to radar. The most direct counter is the broadcast of a high-intensity microwave signal, usually with noise modulation, to mask the echoes from the potential target. More sophisticated systems detect the transmitted radar signal, modify it, and send it back in such a way that false information is given concerning the position and direction of the target. *See* ELECTRONIC WARFARE.

Communications. There is at least 100 times as much frequency space available for communications in the microwave band as in the entire spectrum below microwaves. In addition, the high directivity obtainable at microwave frequencies allows reuse of these frequencies many times in the same area, practice not possible at lower frequencies. The high directivity also makes possible communication to satellites and deep-space probes. *See* RADIO SPECTRUM ALLOCATION.

Satellites may act as relay links between ground stations a great distance apart or as direct

broadcast sources to a distribution of ground receivers. Most satellites operate at frequencies between 4 and 12 GHz, although higher frequency bands are available. Light weight, long life, high efficiency of transmitters, and low noise of receivers are of obvious importance for microwave components for satellite use. *See* COMMUNICATIONS SATELLITE.

Communication between the Earth and deep-space probes is accomplished with microwaves because of the great antenna directivity and receiver sensitivity that can be achieved at microwave frequencies. Microwaves are also used in cellular communications, where systems for mobile communications are well established. Higher-frequency systems with cell radii of 0.6 mi (1 km) or less are promising for personal communications. *See* MOBILE COMMUNICATIONS; SPACE COMMUNICATIONS; SPACE-CRAFT GROUND INSTRUMENTATION.

Radiometry. All objects, including liquids and gases, emit electromagnetic radiation in the form of noise, the amount of the noise being proportional to the absolute temperature of the object. A noise temperature T can be assigned to the object corresponding to the amount of noise radiating from it. The determination of the noise temperatures of selected noise sources, including background radiation from outer space, rain, clouds, stars, and the Earth, is called radiometry. *See* COSMIC BACKGROUND RADIATION; RADIO ASTRONOMY.

A microwave radiometer is a sensitive receiver which measures the noise power received by an antenna; from this measurement, the noise temperature of the source object can be determined. The noise powers measured are so small that receivers with very high gain are required, and the accuracy of the measurements is severely limited by small time variations in the gain. R. H. Dicke solved this problem in 1946 by switching the receiver input between the antenna and a reference noise source at a rate faster than the gain changes in the amplifier. The detected output of the receiver is also switched synchronously at the same rate resulting in two output signals, one due to the noise received by the antenna and the other to the reference noise source. These two signals are subtracted to form an output signal proportional to the difference in the input noise temperatures. The reference noise power can then be adjusted until the output signal is zero, indicating the received noise power is equal to the known reference noise power.

Radiometers are used extensively for remote sensing. Atmospheric temperature, water vapor density over oceans, soil moisture content, and the liquid water content of clouds can be determined from radiometer measurements, some of which are made from satellites. Microwave radiometers are used to study astronomical sources of noise and to observe planets from deep space probes. *See* PASSIVE RADAR; RADIOMETRY; REMOTE SENSING.

Physiological effects. Exposure to microwave power of 100 mW/cm^2 for several minutes can lead to pathophysiological effects in laboratory animals. The microwaves penetrate beneath the skin and heat the

tissue, and tissue destruction can result if the temperature rise is faster than the control mechanisms of the body can handle. It is assumed that human tissue reacts to microwave exposure in the same manner, and safety standards have been proposed. In the United States the maximum recommended exposure is 10 mW/cm^2 , although stricter standards are proposed.

Applications of microwaves in medicine include (1) thermography, the measurement of tissue temperature (cancer causes a temperature rise of about 2°F or 1°C , which can be detected with a microwave radiometer); (2) hyperthermia, microwave heating used in the treatment of cancer and in the treatment of hypothermic subjects; and (3) biomedical imaging, the use of microwaves to study the structure of tissue beneath the skin. *See* RADIOLOGY.

Physics and chemistry. Microwave energy is used in large particle accelerators to accelerate charged particles such as electrons and protons to very high energies and cause them to collide. Knowledge of the structure of matter is also obtained from microwave spectroscopy, which is used to study the frequencies and amplitudes of the microwave resonances of molecules. Microwaves are also used in the study of the crystal structure. *See* MICROWAVE SPECTROSCOPY; PARTICLE ACCELERATOR.

Microwave oven. Microwave energy is absorbed in most foods and has been found to be a source of quick, uniform heating or cooking. Microwave ovens based upon this principle are now widely used. The microwave power, supplied by a magnetron, is at a frequency of 2.45 GHz, with rated power from 350 to 750 W. Since oven size is generally much larger than the 12 cm wavelength, standing waves could form, resulting in heating nonuniformities. This problem is avoided by using either a rotating mechanism that shifts or effectively stirs the modes, or a moving platform to shift the food through the standing-wave pattern.

Safety is, of course, the primary concern. Interlocks are provided on all ovens so that the power is off when the door is open, and a backup system should the primary interlock system fail. Leakage levels of microwave energy from the units must meet strict requirements, and all safety features must be operable for at least 100,000 times.

Industrial heating and drying. Microwaves are also used for the industrial heating of foodstuffs and other materials. The advantages are in the rapidity of heating, the accuracy of control, the possibility of selectivity and localization, the clean environment, and the absence of combustion products. For the food industry, cooking, tempering of frozen foods, and drying of pastas and other foodstuffs are examples. Microwaves are also used in the pharmaceutical, paper, fabric, rubber, and ceramic industries. The primary frequency ranges in these uses are at 0.9 and 2.45 GHz. *See* FOOD ENGINEERING.

Beyond microwaves. Applications of electromagnetic radiation in the millimeter-wave region, which extends from about 100 to about 300 GHz, are much like those in the microwave region. Beyond

this range, and throughout the long infrared wavelengths, suitable power for many of these applications is not available.

At the shorter infrared wavelengths and in the visible optical range, the light-emitting diode (LED) and the laser provide the power, and solid-state detectors provide the necessary efficiency, for communication and other applications. Optical communication (including the infrared) is feasible through the atmosphere but is seriously affected by fog, haze, rain, and snow. The advent of the optical fiber has greatly accelerated the application of these wavelengths to communications. Fibers are made from glass, are as small as a human hair, and can be made so that they have very low loss, a few tenths of decibel per kilometer. Transmission through fibers is unaffected by atmospheric conditions. The communication capacity of an optical-fiber system can be very large: many billions of pulses per second can be transmitted over a single fiber. Microwaves are used in the modulation of optical-fiber systems. Optical-fiber systems have been built for use inside buildings, between cities, and across oceans. See JUNCTION DIODE; LASER; LIGHT-EMITTING DIODE; OPTICAL COMMUNICATIONS; OPTICAL DETECTORS; OPTICAL FIBERS.

Clyde L. Ruthroff; J. R. Whinnery

Bibliography. T. K. Ishii (ed.), *Handbook of Microwave Technology*, 2 vols., 1995; T. K. Ishii, *Microwave Engineering*, 2d ed., 1989; G. M. Kizer, *Microwave Communication*, 1990; Y. Konishi (ed.), *Microwave Integrated Circuits*, 1991; S. Liao, *Microwave Devices and Circuits*, 3d ed., 1996; A. D. Olver, *Microwave and Optical Transmission*, 1992; B. L. Smith and M. H. Carpentier (eds.), *Microwave Engineering Handbook*, vols. 1-3, 1993; J. Thuery, *Microwaves: Industrial, Scientific and Medical Applications*, 1992.

Microwave filter

A two-port component used to provide frequency selectivity in satellite and mobile communications, radar, electronic warfare, metrology, and remote-sensing systems operating at microwave frequencies (1 GHz and above). Four types of filter responses are commonly used in such systems (**Fig. 1**): low-pass, high-pass, band-pass, and band-stop. For example, a low-pass filter allows the passage of signals at frequencies below the cutoff frequency, f_c (the passband), while reflecting signals at higher frequencies (the stopband). While an ideal filter would have zero loss in its passband and infinite attenuation in its stopband, such characteristics can only be approximated for realistic filters. Microwave filters perform the same function as electric filters at lower frequencies, but differ in their implementation because circuit dimensions are on the order of the electrical wavelength at microwave frequencies. Thus, in the microwave regime, distributed circuit elements such as transmission lines must be used in place of the lumped-element inductors and capacitors used at lower frequencies. This can make mi-

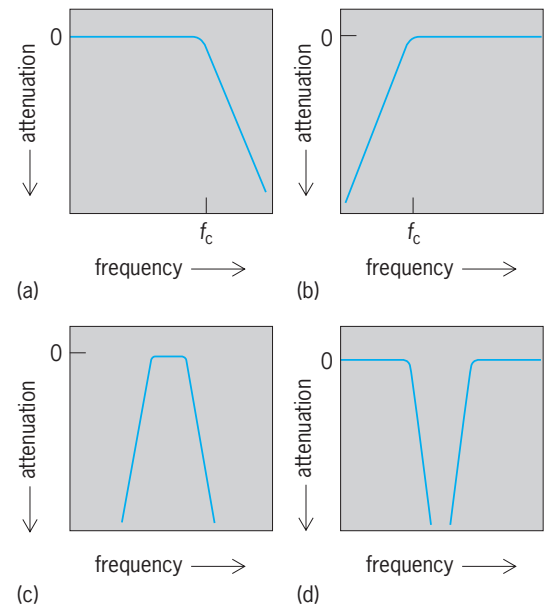


Fig. 1. Common microwave filter attenuation (insertion loss) versus frequency responses. (a) Low-pass filter. (b) High-pass filter. (c) Band-pass filter. (d) Band-stop filter. Attenuation or insertion loss increases down the vertical axis.

crowave filter design more difficult, but it also introduces a variety of useful coupling and transmission effects that are not possible at lower frequencies. In addition, the use of distributed elements often introduces spurious stopband regions of low attenuation at frequencies far above the passband.

Design. The majority of modern microwave filters are designed by using the insertion-loss method, whereby the amplitude response of the filter is approximated by using network synthesis techniques that have been extended to accommodate microwave distributed circuit elements. This method allows great flexibility in specifying filter characteristics, including the passband and stopband amplitude responses, the attenuation rate, and the phase response. Commonly used passband responses include maximally flat (binomial), equal-ripple (Chebyshev), and elliptic responses. Other important microwave-filter performance measures are the insertion loss (the passband attenuation, in decibels), the stopband attenuation rate (in decibels per frequency decade), the group delay (in nanoseconds), and the passband impedance match (the input return loss, in decibels). System considerations may include the power capacity or temperature stability of the filter, as well as weight, size, and cost. The order of a filter is a measure of its complexity, and is the number of capacitors and inductors in its low-pass filter prototype; increasing the order of a filter increases the rate of attenuation in the stopband. See IMPEDANCE MATCHING.

Microwave filter design by the insertion-loss method follows a general four-step procedure: determination of filter specifications, design of a low-pass prototype filter, scaling and transforming the filter, and implementation (conversion of lumped elements to distributed elements). After a realistic set

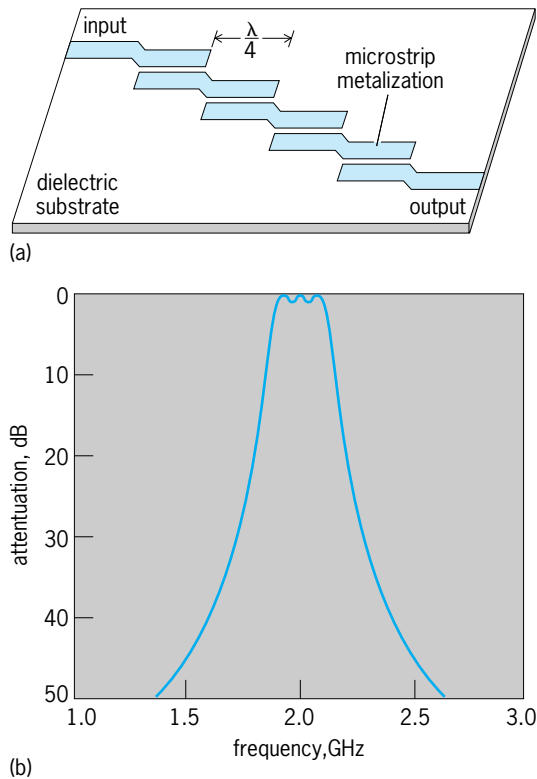


Fig. 2. Third-order (four-section) microwave band-pass filter. (a) Layout of parallel coupled stripline filter. (b) Calculated frequency response of filter, with center frequency of 2 GHz and 0.5-dB equal-ripple passband.

of filter specifications (such as response type, cutoff frequency, impedance, and order) have been determined, a low-pass prototype filter is designed. This is a normalized filter design using lumped inductors and capacitors, and having a unity cutoff frequency and impedance. A wide variety of design tables for low-pass prototype filters of various orders and responses are available. The low-pass prototype can then easily be transformed to a high-pass, band-pass, or band-stop response, and scaled to the desired cutoff or center frequency. The last step, filter implementation, is what makes microwave filter design a unique and challenging process. There are a wide variety of different techniques for constructing microwave filters, usually specific to the type of transmission-line media being used (such as coaxial line, waveguide, microstrip, or coupled line), but some common principles apply. The Richards transform allows lumped-element capacitors to be replaced with open-circuited transmission-line stubs, and lumped-element inductors to be replaced with short-circuited stubs. The Kuroda identities allow the introduction of transmission-line lengths to provide spacing between filter elements. Use of these techniques is called commensurate line filter synthesis, and works well for low-pass and high-pass microwave filters. An additional filter implementation tool is the impedance or admittance inverter, which can be used to transform series resonators to shunt resonators (and vice versa) and to transform internal impedance levels.

Types. Microwave filters are implemented in many different ways depending on considerations such as cost, size and weight, integration with the rest of the system, and electrical performance issues such as insertion loss, phase characteristics, and stop-band attenuation. Waveguide cavity band-pass filters have very low insertion loss, making them preferred for frequency multiplexing in satellite communication systems. Coaxial low-pass filters, made with sections of coaxial line with varying diameters, are compact and inexpensive. Planar filters in microstrip or stripline form (Fig. 2) are important for integration with hybrid or monolithic microwave integrated circuits, and can be made by using transmission-line stubs, transmission-line resonators, or coupled transmission lines. Many compact planar microwave filter designs are made in interdigital, comb-line, or hairpin-line form, consisting of multiple sections of parallel coupled transmission lines. While planar filters are usually more cost-effective than waveguide versions, their insertion loss is usually greater. Dielectric resonators can also be used with planar microwave-circuit technology. Computer-aided design procedures are used in the synthesis of more sophisticated amplitude and phase responses, and active microwave devices (field-effect transistors) are used to provide filters with gain or tunable response characteristics. See COAXIAL CABLE; COMPUTER-AIDED DESIGN AND MANUFACTURING; ELECTRIC FILTER; MICROWAVE; MICROWAVE SOLID-STATE DEVICES; TRANSMISSION LINES; WAVEGUIDE.

David M. Pozar

Bibliography. I. J. Bahl and P. Bhartia, *Microwave Solid State Circuit Design*, 1988; G. L. Matthaei, L. Young, and E. M. T. Jones, *Microwave Filters, Impedance-Matching Networks, and Coupling Structures*, 1980; D. M. Pozar, *Microwave Engineering*, 2d ed., 1997.

Microwave free-field standards

The means for setting up electromagnetic fields of precisely determined intensity at microwave frequencies in unbounded regions of space. Such standards are used to evaluate field probes and antennas for measuring field strength and power density. The standardization of these devices is necessary before they are used for determining the performance of radar and communications systems or for assessing such systems for health and safety risks or electromagnetic compatibility. See ELECTROMAGNETIC RADIATION; MICROWAVE.

Near fields and far fields. An electromagnetic field can be launched into an unbounded region by applying an oscillating voltage or current to a radiating structure or antenna. The energy radiating from an antenna passes through several distinct stages before the final pattern emerges. The true near field comprises induction and static fields which decrease in magnitude as the inverse-squared and inverse-cubed of distance R respectively; they thus become negligibly small a few wavelengths from the antenna and

are rarely significant in practice. The radiated field decreases in intensity as the inverse of distance and thus becomes predominant beyond a few wavelengths. See ANTENNA (ELECTROMAGNETISM).

Radiating antennas carry alternating currents which can be thought of as an array of monochromatic sources spread over the structure according to the Huygens-Fresnel principle, and these sources will give rise to elemental wavefronts which arrive at a point distant from the antenna with phases which impose an interference-type pattern on the wavefront, those at the extremities of the structure having the greatest phase differences. The complex wavefront is obtained by the superposition of these elementary wavelets. The distance at which a 90° phase shift obtains on the elemental wavefronts is known as the Rayleigh distance, and is equal to $D^2/2\lambda$ where D is the maximum aperture dimension and λ the wavelength of the radiation. The region of the field within this distance is the radiated near field, and because of its complex nature it is commonplace free-field measurement practice to use $2D^2/\lambda$ as a minimum working distance for antenna measurements. However, even at this distance the path phase differences are 22.5° and the power density, being proportional to $1/R^2$, is accurate only to 5%. For 1% uncertainty the required separation is at least $10D^2/\lambda$. See HUYGENS' PRINCIPLE; INTERFERENCE OF WAVES; RESOLVING POWER (OPTICS).

Standard antennas. Antennas used in free-field standards are usually either half-wave dipoles or waveguide horns.

Half-wave dipole. The half-wave dipole is a collinear device with a length of approximately one-half of the free-space wavelength of the radiated wave. The radiation pattern is doughnut-shaped and symmetrical about the axis of the dipole. The gain of a simple dipole can be shown to be 2.16 dB.

Waveguide horns. An open-ended waveguide will radiate effectively and, as the aperture dimensions are comparable with the wavelength, its pattern is very broad; an open-ended waveguide has a power gain of about 6 dB. Flaring the aperture decreases the beamwidth and results in a pyramidal horn. Horns can be made with gains ranging from about 8 to about 30 dB, depending on frequency and acceptable size. The gain and pattern characteristics of a pyramidal horn can be calculated to accuracies of about ± 0.2 dB; however, the reflections at the throat and aperture discontinuities significantly influence the gain-frequency characteristic and calibration is usually necessary. See WAVEGUIDE.

Free-field measuring sites. The ideal environment for making measurements on antennas is a large unobstructed volume which is free of reflecting objects and electromagnetically interfering signals—that is, a free-space condition.

Anechoic chambers. A practical solution is to use an anechoic chamber to set up simulated free-space conditions in a bounded environment. Low reflection of electromagnetic signals from the walls of such a chamber is achieved by the use of an electromagnetic-wave absorbent layer covering all of

the reflecting surfaces within the room or chamber, the outer shell of which is a metallic structure to give shielding against interfering signals encroaching on the test region.

Anechoic chambers can be of rectangular or tapered form. For rectangular chambers the width of the chamber is ideally chosen so that the angle of incidence for the reflected ray from the side wall is of the order of 70° or less. The receiving antenna is placed about half the width of the chamber from the back wall so that coupling to the back wall absorbers is minimal. These considerations suggest an optimum length-to-width (and height) ratio of about 3:1. The “tapered” anechoic chamber consists of a pyramidal tapered anechoic section joined to a cubical anechoic section and is an alternative construction suitable for some antenna measurements. The tapered section may be thought of as a large horn antenna terminating in a large waveguide in which a single mode in the form of a plane wave is to be generated.

The size of field strength monitor or antenna which can be measured in an anechoic chamber is determined by the size, and shape, of its quiet zone, which is the volume in which the reflections from the internal surfaces of the chamber are below a specified acceptable signal level compared to the direct radiated value of the field. The measure of the quiet-zone performance, the reflectivity, will vary with position in the zone as well with frequency. For antenna measurements using pseudo-far-field methods the quiet zone needs to be as long as possible with respect to the operating wavelength; this is in order for the separation between receiving and transmitting antennas to be at least equal to several Rayleigh distances at the lowest operating frequency. For normal pattern measurements a reflectivity level of -40 dB in the quiet zone is acceptable; for high-precision gain measurements, with uncertainties of 0.1 dB, -50 dB reflectivity is necessary.

For near-field scanning applications, in which the sampling or detecting probe scans within a few wavelengths of the test antenna, interfering reflections and resonances from the environment which may contaminate the test region can be controlled with a fairly modest application of electromagnetic absorber materials. See ANECHOIC CHAMBER.

Open field sites. Clear open sites can be used at low frequencies, or for high-gain antennas where far-field measurements require such large distances that enclosed or anechoic environments are impractical. The horizontal plane can be readily made free of obstruction, as can the upward vertical plane. However, since electromagnetic waves are strongly reflected by a ground plane, this effect must be accommodated, and one of two courses can be adopted. One is to do measurements as far above the ground as possible. The alternative is to make use of the ground plane by working close to it—if necessary, enhancing its reflection with a metal ground plane or grid—and by making allowance for the reflection in the analysis of performance.

Measurements on antennas. Probably the single most important parameter of a standard antenna, when considered for metrological applications, is the boresight or maximum gain, and a number of methods of determining this parameter have been devised.

Three-antenna method. The basis of this method is a measurement of the transmission between two polarization-matched antennas (1 and 2), which is expressible in terms of the antenna gains by the Friis formula (1), where P_t and P_r are the transmit-

$$P_r = P_t G_1 G_2 \left(\frac{\lambda}{4\pi D} \right)^2 \quad (1)$$

ted and received powers, λ the wavelength, and D the separation. Only the product of the gains can be obtained; however, with three antennas the measured combinations will yield the gain of each antenna uniquely. Small corrections are required due to imperfect matching in the waveguide systems.

Extrapolation method. The extrapolation technique is probably the most accurate method known for determining absolute gain and polarization. The method is one of determining the transmission characteristic between two antennas as the transmission path is increased through about 4 to 10 Rayleigh distances. With good metrology it is possible to get a sufficiently accurate characterization to allow extrapolation to the true far-field range.

Near-field scanning. Antenna metrology has increasingly concentrated on near-field techniques with the objective of improving antenna characterization, particularly at very high (VHF) and ultrahigh (UHF) frequencies where the lack of low-reflectivity far-field ranges has been acute. In this method a probe antenna is used to sample the magnitude and phase, for orthogonal polarizations, of the radiated fields over a well-defined surface, which can be a plane, a cylinder, or a sphere, a few wavelengths from the antenna under test. The objective of the near-field to far-field technique is to mathematically transform the data determined in the near field into the required far-field properties. For a large antenna the number of samples required to define the near field adequately can be large enough to require substantial measurement and computational time. The features of the technique are that complete characteristics can be obtained for any distance, near-field or far-field, for the test antenna and that measurements may be made indoors as close as desired to the antenna in a simple anechoic environment.

Calibration. The essential requirement for the calibration of devices for measuring power flux density or field strength is the creation of a substantially plane wave of known power density which encompasses the effective aperture of the device to be tested. This is effected by launching a known power through an antenna or transverse-electromagnetic (TEM) cell of known characteristics and calculating the field strength or power density from the appropriate equation. The techniques used fall into two groups which are conveniently divided by frequency.

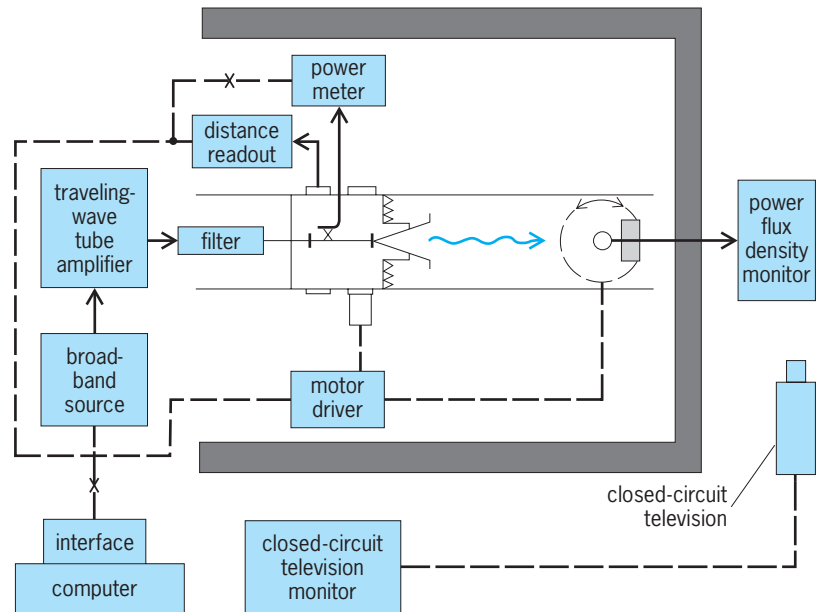


Fig. 1. Microwave free-field calibration system.

Frequencies above 500 MHz. At frequencies above a nominal 500 MHz the conventional technique for calibrating power density or field monitors is to set up a standard field in an anechoic space and position the probe at some prescribed position in front of the waveguide horn.

A practical system for setting up standard fields or power densities involves a small anechoic chamber (Fig. 1) about 15 ft (5 m) long, lined on five sides with electromagnetic absorber. One end is left open for access to the generating and measuring systems. The microwave signal is generated by a broadband signal generator operating into a traveling-wave-tube power amplifier. The output of the latter is filtered to remove spurious harmonics and linked to the monitoring coupler and power meter combination and output (horn) antenna by a low-loss coaxial cable. The instrument under test is mounted on a low-loss dielectric pedestal which is positioned about 3 ft (1 m) from the end wall of the chamber and on the axis of the radiating antenna, which places the test probe in the quiet zone. A computer can be used to control the measurement process and apply corrections. The uncertainties associated with this measurement are about 0.3 dB (equivalent to about 7%). See SIGNAL GENERATOR; TRAVELING-WAVE TUBE.

Frequencies below 500 MHz. The Crawford transverse-electromagnetic transmission cell is commonly used for setting up radio-frequency fields. This consists of a transmission line with a rectangular outer conductor, 1.5–3 ft (0.5–1 m) in size, and an inner conductor in the form of a flat plate or septum (Fig. 2). The input and output ends are tapered so as to form a transition to a conventional coaxial cable. The wave within the cell is essentially plane and can be made to have free-space impedance by proper choice of width and height dimensions. The field distribution within the cell, and thus its uniformity, can be calculated. The electric field within the working area

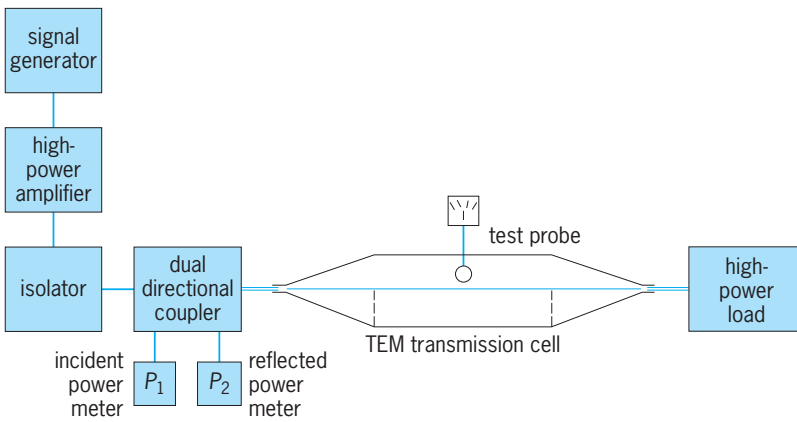


Fig. 2. Guided-wave test facility based on Crawford transverse-electromagnetic (TEM) cell.

is given approximately by dividing the voltage between the plates by the separation of the plates. The voltage is determined by calculation from the power P transmitted through the cell divided by the line impedance Z (equal to 50 ohms). The power density is then expressed by Eq. (2), where Z_0 is the

$$\text{Power density} = \frac{PZ_0}{b^2Z} \quad \text{W/m}^2 \quad (2)$$

impedance of free space (376.7 ohms) and b is the separation of the plates. The uncertainties associated with transverse-electromagnetic cell measurements are of the order of 0.5 to 1 dB. See COAXIAL CABLE.

The upper limit to the operating frequency range of transverse-electromagnetic cells is determined by resonances within the cell; also, it is not good practice to more than half fill the gap between the plates. As a consequence, the larger the test object, the larger must be the cell, and thus the lower its maximum frequency. This problem can be reduced to some extent by placing electromagnetic absorber material in the cell, but this expedient renders the cell suitable for comparison purposes only.

Ralph W. Yell

Bibliography. G. E. Evans, *Antenna Measurement Techniques*, 1990; Institute of Electrical and Electronics Engineers, *IEEE Standard Test Procedures for Antennas*, ANSI IEEE Std. 149-1979, 1979, reaffirmed 1990; J. D. Kraus, *Antennas*, 2d ed., 1988; Y. T. Lo and S. W. Lee (eds.), *The Antenna Handbook*, 4 vols., 1993; A. W. Rudge et al., *The Handbook of Antenna Design*, 2 vols., paper 1986.

Microwave impedance measurement

The determination of parameters, associated with microwave propagation in transmission lines or waveguides, which are generalizations of the impedance concept at lower frequencies and are derived from ratios of complex electric or magnetic field amplitudes.

As an outgrowth of popular usage the term

impedance measurement, at microwave frequencies, has acquired a more general interpretation than the one associated with it at lower frequencies. Throughout the lower portion of the radio-frequency spectrum, impedance is generally defined as the (complex) ratio between voltage and current. Although the low-frequency concepts of voltage and current can be generalized in such a way as to be useful in the microwave region, the almost universal practice is to use the scattering description. Instead of voltage and current, the basic parameters are now the complex electric- or magnetic-field traveling-wave amplitudes in the (assumed) uniform transmission line or waveguide by which the components of interest are interconnected. The coaxial transmission line and rectangular waveguide (where the center conductor is absent) are common examples. See COAXIAL CABLE; TRANSMISSION LINES; WAVEGUIDE.

For a given transmission line or waveguide, there is a preferred band of operating frequencies throughout which the solution of Maxwell's equations is given by a pair of complex numbers a and b which represent the electric or magnetic traveling-wave amplitudes to the left and right (Fig. 1). Let a voltage v and current i be defined by Eqs. (1) and (2), where Z_0

$$v = b + a \quad (1)$$

$$i = \frac{b - a}{Z_0} \quad (2)$$

is a property of the transmission line, called characteristic impedance. These definitions of voltage and current are consistent with the ones in common use at lower frequencies. Obviously, Eqs. (1) and (2) can be solved for a and b in terms of v and i so that the information content is the same in either description; the mode of expression, however, is different. See MAXWELL'S EQUATIONS.

As a counterpart to the lower-frequency definition of impedance (the complex ratio of v to i), a reflection coefficient (which is usually represented by the Greek letter Γ) can now be defined as the complex ratio given by Eq. (3). The measurement

$$\Gamma = \frac{a}{b} \quad (3)$$

of reflection coefficient is usually included if not primarily implied by the term impedance measurement at microwave frequencies. In its more restricted definition, the impedance Z is given by the ratio of

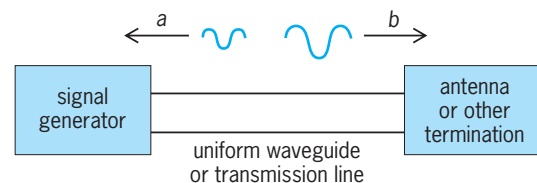


Fig. 1. Microwave system, with complex traveling-wave amplitudes a and b that are generally used to describe the electric and magnetic fields.

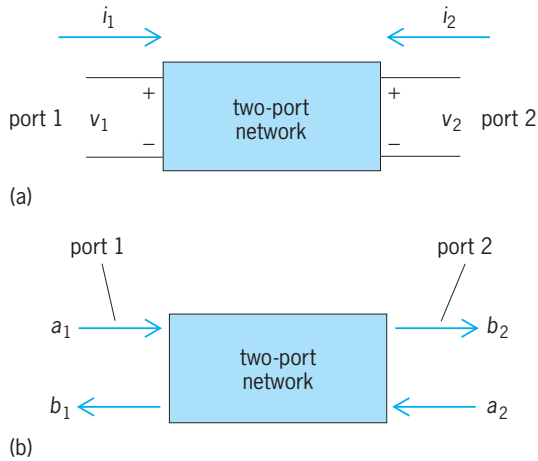


Fig. 2. Two-port network showing (a) the voltages and currents used in y -parameter definitions, and (b) the incident and reflected complex-wave amplitudes used in scattering-parameter definitions.

Eqs. (1) and (2). From here it is easy to show that Eqs. (4) and (5) are valid.

$$\Gamma = \frac{Z - Z_0}{Z + Z_0} \quad (4)$$

$$Z = Z_0 \frac{1 + \Gamma}{1 - \Gamma} \quad (5)$$

Scattering parameters. Linear electrical circuits can be completely characterized by parameters measured at the input and output ports. Several different parameter sets can be used. At low frequencies, b , y , and z parameters are frequently encountered; for example, a two-port network (Fig. 2a) can be characterized by y parameters by using Eqs. (6) and (7),

$$i_1 = y_{11}v_1 + y_{12}v_2 \quad (6)$$

$$i_2 = y_{21}v_1 + y_{22}v_2 \quad (7)$$

where v_1 and v_2 are the voltages across the input and output ports respectively and i_1 and i_2 denote the currents entering the network. See NETWORK THEORY.

To measure b , y , or z parameters, it is necessary to excite the two ports in turn with an open-circuit or short-circuit connected to the other port. At microwavelengths (0.3–30 cm) this measurement is difficult to carry out for several reasons: (1) Lead inductance makes it difficult to get a good short circuit. (2) Stray capacitance makes it difficult to create a good open circuit. (3) Microwave amplifiers sometimes oscillate when their ports are open-circuited or short-circuited. (4) Voltages and currents are rarely measured at microwavelengths, but equipment that yields reflection and transmission coefficients is frequently employed.

For all of these reasons, scattering parameters, rather than b , y , and z parameters, are widely used to characterize microwave networks. Let a_1 and a_2 (Fig. 2b) represent the complex wave amplitudes entering ports 1 and 2 respectively, and let b_1 and b_2

denote the complex wave amplitudes emerging from ports 1 and 2 respectively. The relationships between these complex wave amplitudes are given by Eqs. (8) and (9).

$$b_1 = s_{11}a_1 + s_{12}a_2 \quad (8)$$

$$b_2 = s_{21}a_1 + s_{22}a_2 \quad (9)$$

Interpretation. When a generator is connected to port 1 and a nonreflecting (that is, perfectly matched) termination is connected to port 2, $a_2 = 0$, and from Eq. (8) it follows that Eq. (10) is satisfied. Thus, s_{11}

$$s_{11} = \frac{b_1}{a_1} \quad (10)$$

is the voltage reflection coefficient looking into port 1 when port 2 is perfectly matched. Under the same conditions, it is seen from Eq. (9) that Eq. (11) is

$$s_{21} = \frac{b_2}{a_1} \quad (11)$$

valid. Thus, s_{21} is the ratio of the complex-wave amplitude emerging from port 2 to the complex-wave amplitude incident on port 1 when port 2 is perfectly matched. After interchanging the generator and matched termination, expressions for s_{22} and s_{12} are found in a similar way.

Matrix notation. The scattering coefficients are often written in matrix notation. For a two-port network, the scattering matrix has the general form shown in Eq. (12).

$$[S] = \begin{pmatrix} s_{11} & s_{12} \\ s_{21} & s_{22} \end{pmatrix} \quad (12)$$

For a device with m ports, the scattering matrix has m rows and m columns. For the n th port, some workers let $|a_n|$ and $|b_n|$ represent the root-mean-square values of the incident and reflected voltages; but the expressions for reciprocity and losslessness with unequal port characteristic impedances become more simple if $|a_n|^2$ and $|b_n|^2$ represent the incident and reflected powers respectively. Using this latter convention, the following rules apply:

1. In a linear reciprocal two-port network, Eq. (13) is valid.

$$s_{12} = s_{21} \quad (13)$$

2. In a linear lossless two-port network, Eqs. (14)–(16) are satisfied, where s_{11}^* denotes the complex

$$s_{11}s_{11}^* + s_{21}s_{21}^* = 1 \quad (14)$$

$$s_{12}s_{12}^* + s_{22}s_{22}^* = 1 \quad (15)$$

$$s_{11}s_{12}^* + s_{21}s_{22}^* = 0 \quad (16)$$

conjugate of s_{11} , and so forth. Equations (14)–(16) can be readily derived by equating the input and output powers, using Eqs. (8) and (9) and remembering that $|a_1|^2 = a_1 a_1^*$, and so forth. In mathematical language, S is a symmetric, unitary matrix when the two-port network is linear, reciprocal, and lossless. See MATRIX THEORY.

These rules remain valid when the two ports have different characteristic impedances. There are corresponding rules for multiport networks. The scattering matrices for many two-, three-, and four-port networks are given in standard works and are widely used when analyzing microwave systems.

The scattering matrix for a perfectly matched lossless line of length x is given by Eq. (17), where

$$[S_x] = \begin{pmatrix} 0 & e^{-j\beta x} \\ e^{-j\beta x} & 0 \end{pmatrix} \quad (17)$$

$\beta = 2\pi/\lambda$ is the phase constant of the line and λ denotes the wavelength in it. The rules stated above are all seen to be satisfied by this matrix. Frank L. Warner

Slotted line. Among the most conceptually simple and oldest methods of measuring microwave impedance is the slotted line. In its basic form, a longitudinal slot is cut in the transmission line of interest, which permits the fields to be sampled by means of a movable probe at different positions along the axis of the line. A mechanical transport mechanism is provided so that the depth of penetration or coupling between the probe and internal fields remains fixed as the probe moves along the slot.

At any instant of time, the phase of a or b depends upon the phase of the generator and the propagation time delay between the generator and the probe position where a or b is measured. As the probe is moved along the line, this time delay will be increased for a and decreased for b , or vice versa, depending upon the position of the generator and the direction of the motion. Thus the phase difference between a and b will depend upon the probe position.

In the existing practice, the designs of most probes are such that their response is proportional only to the total field amplitude which is given by $|a + b|$ or $|v|$ [according to Eq. (1)]. Since motion along the line is accompanied by changes in the phase difference between a and b , the magnitude of the sum changes as required by the laws governing the addition of two complex numbers. In particular, if a and b are in phase, the maximum probe response is obtained, and this is proportional to $|a| + |b|$. At the minimum, the response is proportional to $||a| - |b||$. By definition, the ratio between these two responses is the voltage-standing-wave ratio (VSWR)

and is given by Eq. (18). (The reciprocal of this defi-

$$\text{VSWR} = \frac{|a| + |b|}{||a| - |b||} \quad (18)$$

inition is also in use.) The measurement of VSWR also falls within the scope of “impedance measurement” at microwave frequencies. Ordinarily, the directions of wave propagation associated with a and b are chosen such that $|a| \leq |b|$. Equations (3) and (18) may now be combined to obtain Eqs. (19) and (20).

$$\text{VSWR} = \frac{1 + |\Gamma|}{1 - |\Gamma|} \quad (19)$$

$$|\Gamma| = \frac{\text{VSWR} - 1}{\text{VSWR} + 1} \quad (20)$$

The foregoing procedure determines the magnitude of Γ but not its phase. In many applications, however, the interest in the phase of Γ is minimal. The reason is that the generation of energy is far more difficult at microwave than at the lower frequencies. As a consequence, the interface between different components is usually arranged so as to achieve maximum power transfer. As a general rule, this calls for wave propagation in one direction only. Loosely speaking, the magnitude of Γ usually represents an indication of the extent of departure from the system design objectives.

On the other hand, it is possible to obtain the phase by a rather simple extension of the foregoing technique. While the phases of a and b change with respect to line position, their magnitudes and thus $|\Gamma|$ remain constant. At the position of the minimum probe output, the phase of Γ is 180° since this corresponds to a minimum value of the sum $|a + b|$. To find the phase of Γ at any other position in the line, it is only necessary to note that since the phases of a and b move in opposite directions, the phase of Γ changes at twice the rate of either of them and undergoes a total of 360° between two successive minima. By observing the positions of two minima, the phase of Γ at any other location can then be determined provided that its distance from the position of the minimum is known. See PHASE-ANGLE MEASUREMENT.

Reflectometer. In the case of the slotted line, the probe typically consists of a short length of circular rod which projects into the waveguide in a direction perpendicular to its axis. The reflectometer is based on a different type of coupling mechanism (Fig. 3). Here the secondary waveguide is coupled to the primary one by means of a pair of small holes spaced a quarter wavelength apart. On an individual basis, each of these holes is a source of waves in the secondary guide which tend to be of equal amplitude in both directions. When taken as a pair of holes, however, the geometry is such that the waves tend to combine in phase in the forward direction and out of phase (and thus cancel) in the reverse direction. (From a given point in the primary guide to the left of the coupling holes, the distance to a point in the secondary guide to the right of these two holes is the

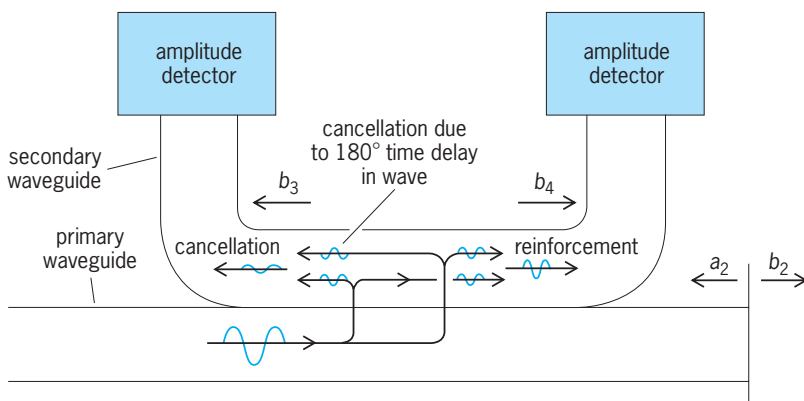


Fig. 3. Directional coupler which forms the basic element of a microwave reflectometer.

same via either of them. But in the reverse direction, that is, to a point in the secondary guide to the left of the two holes, there is a half-wavelength difference in the path length associated with each of the holes. This leads to a reinforcement of the total wave amplitude in one direction and to a cancellation in the other.) See INTERFERENCE OF WAVES.

The foregoing waveguide circuit is aptly called a directional coupler, which is the basic element in a reflectometer. If the wave in the primary guide (Fig. 3) is propagating to the left instead of to the right, the coupled wave in the secondary arm will also be primarily to the left. Let each of the secondary arms be terminated by a suitable detector such as a diode or bolometer. These devices respond to the powers which they absorb and which are proportional to $|a|^2$ and $|b|^2$. The square root of the ratio of the detector responses thus gives $|\Gamma|$. See DIRECTIONAL COUPLER; MICROWAVE POWER MEASUREMENT.

As before, only the magnitude of Γ is obtained, but this time there is no simple extension of the method which yields the phase. If required, the phase can be obtained by one of the techniques described below.

The foregoing methods have the advantage of requiring only the most simple of microwave detectors, usually either a bolometer or diode, but yield only the amplitude (as contrasted with phase) of the detected signals. On the other hand, the accuracy of these methods is limited by the impossibility of constructing either slotted lines or directional couplers which completely satisfy the simple description outlined above. With regard to the reflectometer, techniques based on the use of tuning transformers have been devised which substantially improve its operation at a single frequency, but these methods become time-consuming in a practical application where measurements at multiple frequencies are required.

Network analyzer. The limitations on accuracy and multiple-frequency operation noted above are largely avoided by the network analyzer, which is a multipurpose instrument with applications in the measurement of attenuation, and possibly power, as well as of impedance.

Let a_2 and b_2 (Fig. 3) represent the wave amplitudes in the primary line, while b_3 and b_4 represent those in the secondary or coupled line. In general, b_3 can be expressed by Eq. (21), where A and B are

$$b_3 = Aa_2 + Bb_2 \quad (21)$$

two complex constants which are functions of the scattering parameters of the directional coupler and its associated detectors. The magnitudes $|A|$ and $|B|$ indicate how tightly the secondary line is coupled to the primary line, while their ratio is a measure of the directive properties of the coupling. In a similar way, b_4 can be expressed by Eq. (22), where C

$$b_4 = Ca_2 + Db_2 \quad (22)$$

and D are additional complex constants. For an ideal reflectometer, one has $B = C = 0$, and taking the

ratio of Eqs. (21) and (22) yields Eq. (23), so that if

$$\left| \frac{b_3}{b_4} \right| = \left| \frac{A}{D} \right| \cdot \left| \frac{a_2}{b_2} \right| = \left| \frac{A}{D} \right| \cdot |\Gamma| \quad (23)$$

the detectors measure $|b_3|$ and $|b_4|$ their ratio is proportional to the reflection coefficient. The constant term $|A/D|$ can be easily determined by observing $|b_3/b_4|$ in conjunction with a known value of $|\Gamma|$, usually a short for which $\Gamma = -1$. Thus the system is a reflectometer as described above.

In the network analyzer mode of operation, Eqs. (21) and (22) are combined to yield Eq. (24).

$$\frac{a_2}{b_2} = \Gamma = \frac{\frac{b_3}{b_4} - \frac{B}{D}}{\frac{A}{D} - \frac{Cb_3}{Db_4}} \quad (24)$$

Equation (23) gives Γ in terms of the complex ratio b_3/b_4 , which means that the detection system must provide the phase difference between b_3 and b_4 . In addition, the ratios A/D , B/D , and C/D are also required. These are properties of the coupling network and must be determined at each operating frequency by a suitable calibration procedure. The network analyzer has the advantage that errors due to nonzero values of B and C are in principle eliminated. In addition to the calibration requirement, however, an alternative detection system is required which provides the phase difference between b_3 and b_4 . Ordinarily, this calls for heterodyne conversion to a lower frequency which preserves the phase information and where phase detection techniques are well known. See HETERODYNE PRINCIPLE.

The typical network analyzer (Fig. 4) includes a pair of reflectometers, or test sets, which permits the measurement of a two-port device that has been inserted between the test ports. The analyzer circuit includes a dividing network, which is merely a switch that permits the excitation to be applied to either test set #1 or #2. In the former case, the input reflection is measured by #1, while the transmitted signal is measured by the detectors associated with #2. Given these results, and a similar set with the excitation applied via test set #2, it is possible to obtain the complete set of scattering parameters, and thus the attenuation, of the two-port.

Six-port network analyzer. The six-port network analyzer derives its name from the use of a six-port, rather than four-port, coupling network and the use

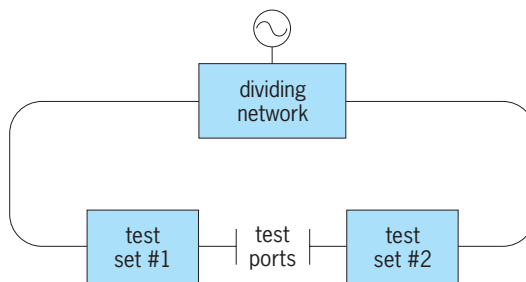


Fig. 4. Basic circuit for a network analyzer.

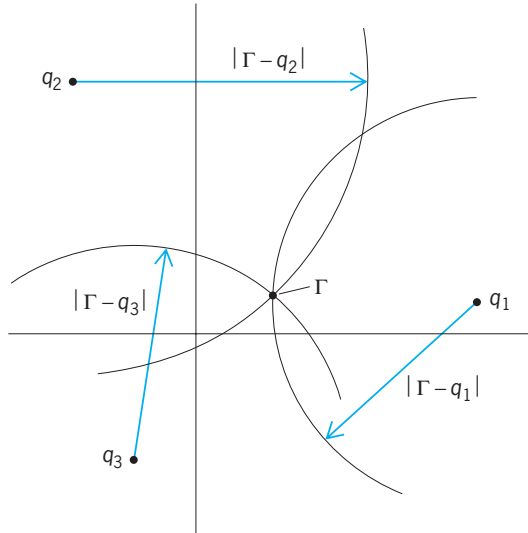


Fig. 5. Geometric construction used with six-port reflectometer to determine the complex reflection coefficient from the intersection of three circles.

of four amplitude detectors rather than a complex-ratio (b_3/b_4) detector. Each detector satisfies Eq. (25),

$$|b_3| = |Aa_2 + Bb_2| \quad (25)$$

similar to Eq. (21), where the absence of phase response is reflected by the use of absolute values.

Equation (25) may be factored to yield Eq. (26), where $q_3 = -(B/A)$.

$$|b_3| = |b_2| \cdot |A| \cdot |\Gamma - q_3| \quad (26)$$

Because of the absolute value sign, it is not possible to solve Eq. (26) for Γ . In the complex plane, $|\Gamma - q_3|$ represents the distance between q_3 and Γ but gives no indication of the direction. The values of Γ which satisfy Eq. (26) lie on a circle whose center is at q_3 and whose radius is determined by $|b_3|$, $|A|$, and $|b_2|$. Here $|b_3|$ is measured by one of the detectors, while $|A|$ and q_3 are parameters of the six-port network.

The term $|b_2|$, however, is an unknown whose value (in common with Γ) depends upon the excitation conditions, and whose determination usually falls within the scope of power rather than impedance measurement. It can be eliminated from consideration in the present context by taking the ratio between two detectors. Here it will prove convenient to assume initially that this second detector is described by Eq. (22) but where $C = 0$. Then,

Eq. (27) is satisfied.

$$\left| \frac{b_3}{b_4} \right| = \left| \frac{A}{D} \right| \cdot |\Gamma - q_3| \quad (27)$$

By use of the remaining two detectors, it is possible to form two more equations similar to Eq. (27), but where the six-port design is such that the values of q_3 are different. Each of these equations fixes the distance between Γ and three arbitrary points (circle centers). It is convenient to visualize the operation as providing the value of Γ through the intersection of three circles (Fig. 5). Here the circle centers, q_1 , q_2 , and q_3 , are parameters of the six-port, while their radii are given by the detector responses. In practice, where the value of C is usually different from zero, the circle centers as well as their radii depend upon Γ , so that the foregoing description is only an approximate one.

The six-port network analyzer has the advantage of simplicity in the detection system, since the requirement for frequency conversion has been eliminated. But, since the detectors tend to be broadband, it is important for the signal source to be free of harmonics. In addition, if bolometric detectors are employed to achieve the best accuracy, a considerable increase in signal power is required.

The test sets in Fig. 4 may also be implemented as six-port reflectometers, which leads to the dual six-port. In this case, however, the dividing network is such as to provide a simultaneous and nominally equal level of excitation to both test sets, but at several values of phase difference between the two.

Multistate reflectometer. Because most of the associated theory is applicable, the multistate reflectometer may be considered a variant of the six-port network analyzer. A typical implementation (Fig. 6) includes a pair of directional couplers, two (rather than four) amplitude detectors, and a moving short in the secondary (or coupled arm) of the coupler on the left. In contrast with the six-port, whose operation requires the ratio of three of the detectors to the fourth, the multistate reflectometer achieves the same result by taking the ratio between the two detectors for three positions of the moving short (whose nominal separation should be a sixth of the wavelength, but are otherwise arbitrary, and must be repeatable). The advantage of the multistate reflectometer is that its operation requires less power, which is an important consideration at the higher microwave frequencies. However, the problems associated with the short adjustment are such as to preclude its operation as a dual six-port.

Calibration methods. As is evident from Eq. (24), the operation of a network analyzer is characterized by the three complex ratios A/D , B/D , and C/D . Before it can be used to measure impedance, these must first be determined by a suitable calibration. Now Eq. (24) may be written as Eq. (28). Thus, obser-

$$\frac{A}{D} \Gamma + \frac{B}{D} - \frac{C}{D} \frac{b_3 \Gamma}{b_4} = \frac{b_3}{b_4} \quad (28)$$

variations of the analyzer response (b_3/b_4) for three

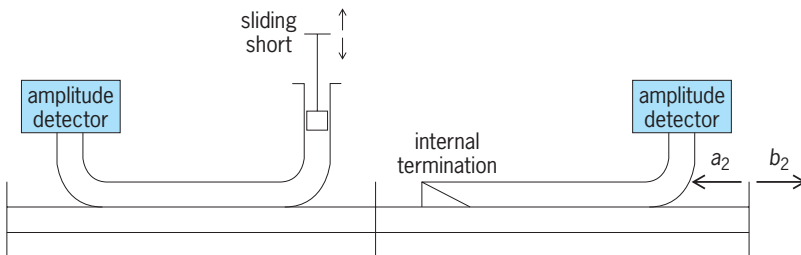


Fig. 6. Basic circuit for a multistate reflectometer.

different terminations for which the values of Γ are known provides a system of three linear equations which may be solved for A/D , B/D , and C/D . In practice, a shorted length of transmission line provides a nominal reflection whose magnitude is unity, while the phase angle may be determined from the line dimensions and the operating frequency. In addition, the response from a nonreflecting ($\Gamma = 0$) termination may be approximated from the response to three or more positions of a weakly reflecting sliding termination. (The latter is a termination of small but unknown reflection which has been designed to slide inside the transmission line or waveguide of interest such that it provides a reflection of constant magnitude but variable phase as its position is changed.) Although this technique may be applied, in turn, to each of the test sets in Fig. 4, the dual configuration permits the use of several alternatives which are based on the thru-reflect-line (TRL) procedure. This procedure calls for observations with the two test ports connected together (thru), then with a strongly reflecting (but otherwise unknown) termination connected, in turn, to each of the test sets (reflect), and finally with a (nominally arbitrary) length of transmission line inserted between the two. Compared with the prior procedure, this technique has the important advantages of eliminating the requirement for terminations of known reflection and sliding terminations, whose implementation is untenable in a line of solid dielectric.

The calibration of the six-port is somewhat more involved. Referring to Eq. (26), the six-port parameters associated with the response, $|b_3|$, are the magnitude $|A|$ and the complex quantity q_3 . A similar observation applies to the remaining three detectors. In all there are four real parameters and four complex ones, or (since a complex number may be reduced to two real ones) a total of twelve real ones. As may be inferred from Eq. (27), however, the measurement of the reflection coefficient (as contrasted with the power) requires only the ratio of three of the magnitudes to the fourth. This simplification leaves eleven parameters to be determined in the six-port calibration. From Fig. 5, it is evident that two of the radii (by which Γ is determined) determine the third to the extent of a choice between two possible values. This means that there is a quadratic relationship among the detector responses which may be characterized by five of these eleven parameters and which may be determined by observing the system response to a suitable collection of terminations whose actual values may remain unknown except as required to assure a well-conditioned solution. This procedure is referred to as a six-to-four-port reduction in that the remaining calibration problem is exactly that which has been described earlier. Moreover, the terminations which are utilized in the initial part of this procedure may include those which will be required for the four-port calibration problem.

Swept-frequency reflectometry. In the above discussion, the problems associated with multiple-frequency operation have been largely ignored. As a practical matter, it is frequently important to know

how the impedance (or reflection coefficient) varies with frequency. One of the more simple methods of achieving this is with the swept-frequency reflectometer. This is nothing more than the reflectometer described above in conjunction with a swept signal source whose frequency varies smoothly from an initial to a final value and then returns quickly to the starting value for a repeat of the sweep. If the reflectometer output is displayed on an oscilloscope, the reflection coefficient magnitude can be obtained as a function of frequency.

In response to increasing frequency range and accuracy requirements, however, this method has been largely replaced by automated versions of the network analyzers described above. The automatic network analyzer (ANA) typically includes a small digital computer for making the necessary computations and for controlling the signal source. It is thus possible to make a rapid succession of measurements at closely spaced frequencies. Moreover, the phase as well as magnitude of the reflection coefficient is usually obtained.

Time-domain reflectometer. In a typical application, the wave amplitude b (Fig. 1) has its source in a generator (to the left) which is delivering energy, via the transmission line, to an antenna or other termination on the right. Provided that the antenna impedance is equal to Z_0 , it is evident from Eqs. (3) and (4) that the amplitude of the wave, a , will equal zero. In a practical application, if a is found to be nonzero, it may be desirable to determine whether the source of this reflected wave is the antenna or possibly some other discontinuity in the transmission line. Although it is possible to answer this question given the results of impedance measurements at a number of different frequencies, a more simple approach, at least conceptually, is provided by the time-domain reflectometer (Fig. 7). In one mode of operation the generator output is a series of short pulses. These will travel down the line and be reflected at discontinuities within the line itself (if they exist) or at the antenna (or other termination) if its impedance is different from Z_0 . Provided that the pulse width is short in comparison with the transit time from the generator to the discontinuity, or

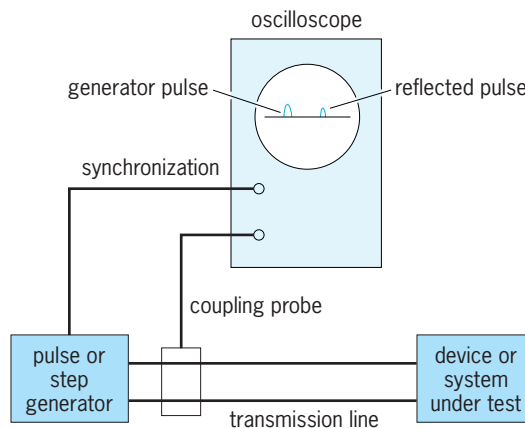


Fig. 7. Basic elements of a time-domain reflectometer.

termination, the oscilloscope responses to the generator pulse and to the reflected pulse will be separated by twice the transit time to the reflection and thus will be displaced from one another along the horizontal time axis. Apart from having confined the transmitted energy to a waveguide rather than free space, the operation is basically the same as radar. Another useful waveform in time-domain reflectometry is the step function. By examining the waveform associated with the reflection, it is possible to learn a great deal about the nature of the discontinuity which produced it. See FUNCTION GENERATOR; PULSE GENERATOR; RADAR.

As an alternative to generating these specialized waveforms, it is also possible to synthesize the time-domain response from measurements at a collection of suitably chosen discrete frequencies. The basis for this is found in Fourier analysis. In particular, any arbitrary repetitive waveform (including, for example, a series of short pulses) can be synthesized by the superposition (or addition) of a (usually infinite) number of harmonically related frequencies of the proper amplitude and phase. (As a practical matter, the synthesis is limited to a finite number of frequencies, and the desired waveform is only approximated.) The system response to the synthesized waveform may be obtained by mathematically combining the system responses to this collection of discrete frequencies, usually with the help of a digital computer. See FOURIER SERIES AND TRANSFORMS.

Locating reflectometer. Another method of determining the position of waveguide discontinuities is provided by the locating reflectometer, which uses a swept-, rather than stepped-, frequency source. This technique is also based upon Fourier analysis, but the Fourier transformation is obtained by analog methods rather than digital methods, and the need for a (digital) computer is eliminated. On the other hand, the associated waveguide circuit tends to be more complicated.

Comparison reflectometer. As noted above, much of the motivation for the network analyzer is to eliminate errors caused by nonzero values of B and C , which in turn are due to imperfections in the directional coupler. An alternative technique for dealing with this problem is provided by the comparison reflectometer. The solution of Eq. (24) for $|b_3/b_4|$ yields Eq. (29). If $|\Gamma|$ is small, the term $C\Gamma$ may ordinarily

$$\left| \frac{b_3}{b_4} \right| = \frac{|A\Gamma + B|}{|C\Gamma + D|} \quad (29)$$

be neglected in comparison with D . On the other hand, $A\Gamma$ and B may be comparable in magnitude so that the nonzero value of B represents a major source of error. In the comparison reflectometer the directional coupler is separated from the item being measured by a long length of waveguide. This has the effect of causing the phase of Γ to vary rapidly with frequency. A small variation in source frequency will thus cause the terms $A\Gamma$ and B to combine in and out of phase, while ordinarily A and B may be assumed to be constant if the frequency excursion

is not too large. The separate contributions of $A\Gamma$ and B to the reflectometer response may then be obtained by techniques similar to those described in conjunction with the slotted line. See MICROWAVE MEASUREMENTS; RADIO-FREQUENCY IMPEDANCE MEASUREMENTS.

Glenn F. Engen

Bibliography. A. E. Bailey (ed.), *Microwave Measurements*, 1985; G. H. Bryant, *Principles of Microwave Measurements*, 1988; G. F. Engen, *Microwave Circuit Theory and Foundations of Microwave Metrology*, 1992; K. C. Gupta, *Microwaves*, 1979; I. Kneppo, *Microwave Measurements by Comparison Method*, 1988; T. Laverghetta, *Modern Microwave Measurements and Techniques*, 1988; P. I. Somlo and J. P. Hunter, *Microwave Impedance Measurement*, 1985; F. L. Warner, *Microwave Attenuation Measurement*, 1977.

Microwave landing system (MLS)

An all-weather aircraft landing-guidance system that operates at microwave frequencies and provides deviations from the landing runway centerline using a time-referenced scanning beam (TRSB) technique. The MLS was standardized in 1988 and approved for use in international civil aviation until at least the year 2020. MLS is used to support low-visibility instrument precision approach and landing operations in North America and Europe. In addition to the fixed-base MLS equipment design, a compact mobile microwave landing system (MMLS) equipment design exists. The instrument landing system (ILS) is also standardized internationally and approved for use indefinitely as countries implement their transition to new technologies. Standards for a third landing system, the Global Navigation Satellite System (GNSS), based primarily on Global Positioning System (GPS) technology, exist. Multimode receivers enable an aircraft to conduct an instrument approach using ILS, MLS, or GNSS. See INSTRUMENT LANDING SYSTEM (ILS); SATELLITE NAVIGATION SYSTEMS.

Design principles. MLS antennas are designed to generate a narrow fan-shaped main beam with side lobes (secondary beams) that are typically 25 dB or more below the main beam level. The antenna pattern design used in conjunction with the TRSB technique makes MLS resistive to guidance errors caused by reflection of the radiated signal by objects in the airport environment (such as hangars, buildings, aircraft, or terrain). A reflection of the main lobe will cause negligible guidance errors if it is separated in angle from the line of sight to the aircraft by two main-lobe widths or more. Since a TRSB technique is used, this separation angle equates to different times of arrival at the airborne receiver for the desired signal and undesired reflections. For any landing guidance system sited near the landing runway stop end, the worst-case lateral reflection geometry assumes a large hangar or other reflector offset 500 ft (150 m) from centerline near the threshold of a 15,000-ft (4500-m) runway. Then, the angular offset from runway centerline is about 2° . For this reflector to cause

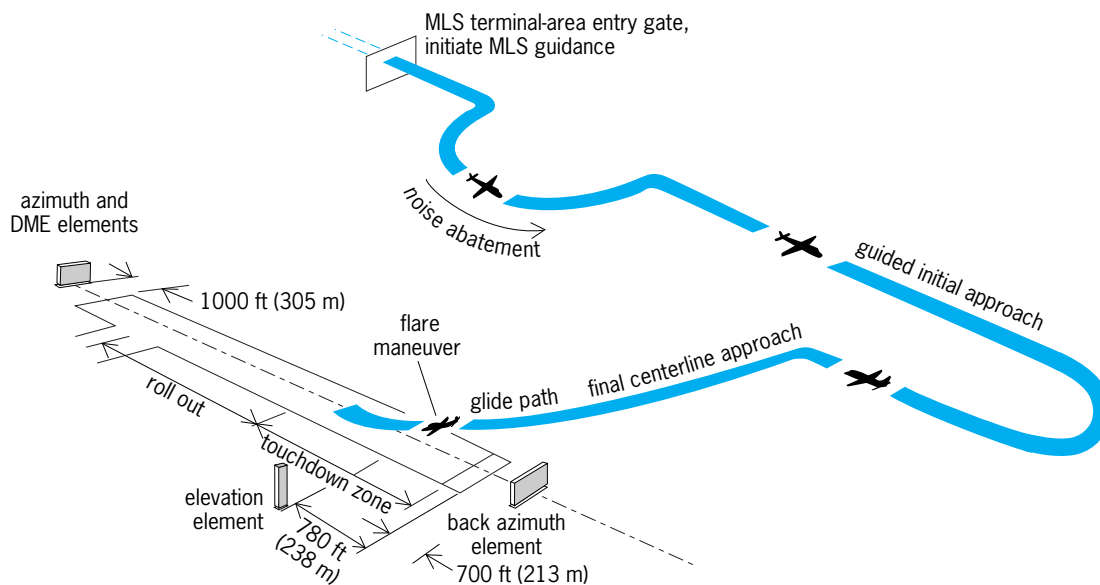


Fig. 1. Landing operation with microwave landing system units located near the runway. The elevation element is sited such that the glide path crosses the threshold above 50 ft (15 m). (After R. J. Kelly and E. F. C. LaBerge, *MLS: A total system approach*, *IEEE AES Mag.*, 5(5):27–39 May 1990)

negligible error, the lateral guidance antenna pattern should have a beamwidth of 1° . In that case, the 2° separation angle approximates two beamwidths. Thus, antennas with beamwidths of 1° keep the lateral guidance errors very small along the centerline approach path. For vertical guidance, the geometry of the reflections from the ground in front of the vertical guidance antenna yields a requirement for antenna pattern beamwidths of $1\text{--}2^\circ$ for a normal approach angle of 3° .

The operating frequencies for MLS lie in a portion of the C band (5030–5091 MHz) designated for use in aeronautical telecommunications. This frequency choice allows a 12-ft (3.6-m) antenna to generate the 1° beamwidth pattern needed to exclude most reflections. The next higher available frequencies (Ku band at 15,000 MHz) would allow the same directivity with an antenna one-third this size (a desirable trade for transportability), but the technological risks increase significantly and the attenuation of signal strength caused by water vapor is severe.

These narrow beams provide accurate vertical guidance down to low heights (8 ft or 2.4 m) above the runway, and thus operations in the lowest-visibility conditions can be supported without limitation from the guidance system. Also, at these microwave frequencies the radio-frequency patterns are well defined at 100–200 ft (30–60 m) from the array, which allows the total radiated signal to be continually validated by field monitors located in protected environments, and provides the high confidence in the signal (high integrity) required for these operations. In order to provide flexible approach paths for a variety of aircraft and rotorcraft, wide lateral and vertical coverage sectors are required. Scanning the narrow microwave beams is an ideal way to generate these coverages, as the numerous reflected signals also generated will be separated, almost

always, from the desired signal by a time interval that equates to an angular separation exceeding two beamwidths and thus will not distort the guidance information.

As with ILS, the MLS equipment is sited near the primary runway, with the azimuth transmitter (lateral guidance) and distance-measuring equipment (DME) transponder located near the runway stop end, and the elevation transmitter (vertical guidance) located alongside the runway near landing threshold (Fig. 1). This type of siting is referred to as a split-site configuration. With this geometry, the approach course and glide path, generated by the ground equipment, are monitored at the landing runway. Also, the aircraft lateral and vertical displacements due to guidance errors become vanishingly small as the runway is approached and the angular guidance converges to its origin. For special-purpose applications, such as providing guidance to a heliport-only facility, all the MLS equipment may be sited at one common location, known as a collocated site configuration.

Unlike ILS, the 50 times higher frequency of the MLS allows generation of narrow beams by relatively small equipment. Because of this 50:1 scale factor, a 1° beamwidth antenna for MLS requires a 12-ft (3.6-m) antenna, while for ILS a 600-ft (180-m) antenna would be required. See ANTENNA (ELECTROMAGNETISM); DISTANCE-MEASURING EQUIPMENT.

Technology. The large coverage volume of MLS is provided by scanning the narrow beams clockwise then counterclockwise for azimuth functions (Fig. 2) and up then down for elevation functions. This scanning is electronically controlled at a precise rate of $20,000^\circ/\text{s}$ and fills a lateral sector of 60° (maximum) on each side of the runway centerline and a vertical sector of 30° (maximum). The angular position of the aircraft is decoded by the

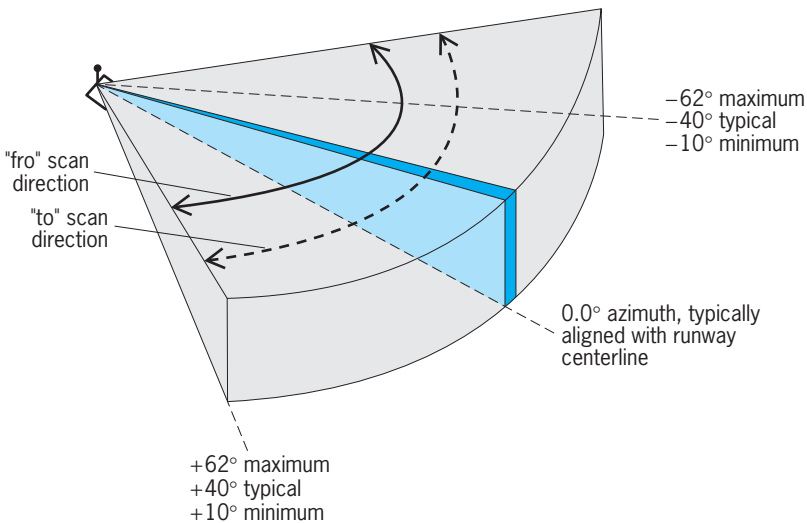


Fig. 2. Azimuth antenna scanning beam and TRSB convention. Maximum, typical, and minimum values of the magnitude of the azimuth at the limits of the coverage volume are indicated. (After M. F. DiBenedetto, *Development of Critical-Area Criteria for Protecting Microwave Landing System Azimuth and Elevation Guidance Signals*, Dissertation, Fritz J. and Dolores H. Russ College of Engineering and Technology, Ohio University, Athens, March 1999)

airborne receiver, which measures the time elapsed between successive passages of azimuth or elevation beams.

The antennas typically used are phased arrays where beam scanning is accomplished by a stored set of commands which, at the appropriate time in the transmission sequence, are directed to variable signal delay devices (phase shifters) associated with each radiating element of the array. These fraction-of-a-wavelength delays are programmed so that the total signal delay increases more rapidly at one end of the array than at the other to cause the radiated phase front to rotate and the pattern to scan in one direction. After a short pause, the program is reversed, and the pattern rotates back to the starting point. The data transmissions use separate fixed-pattern antennas designed to radiate into all parts of the coverage volume simultaneously. Differential phase-shift keying (DPSK) is used to encode the data.

Required navigation performance (RNP). The performance required of an aircraft guidance system may be defined in terms of four parameters: accuracy, integrity, continuity of service, and availability. The criticality of each depends upon the particular flight regime; and for precision approach and landing operations where the aircraft is approaching obstacles (for example, the ground), the emphasis is on guidance signal accuracy and the integrity (believability) of the guidance.

Unlike ILS, the MLS has the same accuracy standard for all categories of operation, which is one result of the reduced environmental effects. Above runway threshold, the lateral error for MLS is limited to ± 20 ft (6 m) and the vertical error is limited to ± 2 ft (0.6 m). The lateral error limits apply throughout the runway region at 8 ft (2.4 m) above the surface, although increased error is allowed from the threshold

out to 20 nautical miles (37 km). The error limits are chosen so that the probability is extremely remote that, with no failures, guidance errors would cause the landing aircraft to stray outside the obstacle-free region near the approach path. The integrity of the guidance is assured by the monitoring system which, for the lowest-visibility operations, is also designed to have an extremely high probability of detecting erroneous guidance and providing an alarm within the required time interval. To meet these very low probabilities, first the ground equipment and the far-field patterns of the radiated MLS signals are continuously monitored, and second the monitor system is verified, end to end, at short intervals. In this manner, the integrity level achieved is controlled by the frequency of the verification.

The MLS receiver provides high integrity by a safety-critical design methodology, including software modules, and by verifying the internal computations at short intervals. For the lowest-visibility applications, the typical aircraft installation may include two or three redundant receivers arranged so that their outputs act as "votes" in an executive monitor.

Mobile microwave landing system (MMLS). The MMLS is a military-grade, all-weather precision instrument approach and landing system. Like MLS, it operates at microwave frequencies and provides deviations from the landing runway centerline using a TRSB technique. The equipment may be sited in either a split-site or a collocated configuration. The MMLS ground equipment is interoperable with both civil and military MLS avionics complying with international standards. The MMLS technology and design principles are very similar to those previously discussed for MLS, with an emphasis on a system design that is compact, mobile, and rapidly deployable. It may be used to support initial deployment of ground forces, for forward-area resupply, for medical evacuation, for rapidly restoring landing service at airports where existing service has been lost due to war or natural disaster, and for other missions not suited for the larger fixed-base civil MLS.

The MMLS equipment is compact and modular in design with a total system weight of less than 600 lb (272 kg). Once on site, the MMLS can be set up and operational in less than 2 hours by a crew of no more than three persons. Achieving the required level of portability mandates the use of antennas that are physically smaller than those used for MLS. This situation results in MMLS antennas with larger antenna pattern beamwidths. Still, antenna beamwidths are typically less than 3.0° for azimuth and less than 2.5° for elevation. Although the coverage and accuracy performance requirements for MMLS are not as stringent as those for MLS, the MMLS performance is well suited for supporting low-visibility approach and landing operations. See AIR NAVIGATION; ELECTRONIC NAVIGATION SYSTEMS.

Douglas B. Vickers; Michael F. DiBenedetto

Bibliography. International Civil Aviation Organization, *International Standards, Recommended*

Practices and Procedures for Air Navigation Services: Aeronautical Telecommunications, Annex 10, vol. I, 1985; M. Kayton and W. Fried (eds.), *Avionics Navigation Systems*, 2d ed., Wiley, New York, 1997; R. J. Kelly and E. F. C. LaBerge, MLS: A total system approach, *IEEE AES Mag.*, 5(5):27-39, May 1990; C. Marton (ed.), *Advances in Electronics and Electron Physics*, vol. 57, 1981.

Microwave measurements

A collection of techniques particularly suited for monitoring of devices and systems where physical size of components varies from a significant fraction of an electromagnetic wavelength to many wavelengths. See MICROWAVE.

Virtually all microwave devices are coupled together with a transmission line having a uniform cross section. The concept of traveling electromagnetic waves on that transmission line is fundamental to the understanding of microwave measurements. See TRANSMISSION LINES.

At any reference plane in a transmission line there are considered to exist two independent traveling electromagnetic waves moving in opposite directions. One is called the forward or incident wave, and the other the reverse or reflected wave. The electromagnetic wave is guided by the transmission line and is composed of electric and magnetic fields with associated electric currents and voltages. Any one of these parameters can be used in considering the traveling waves, but measurements in the early development of microwave technology were made principally on the voltage waves, and this led to the custom of referring only to voltage. One parameter in very common use is the voltage reflection coefficient, Γ , which is related to the incident, V_i , and reflected, V_r , voltage waves by Eq. (1). Often the volt-

$$\Gamma = \frac{V_r}{V_i} \quad (1)$$

age reflection coefficient is referred to as return loss, which is calculated as $-20 \log_{10} |\Gamma|$ and expressed in decibels (dB). For example, if V_r is 1% of V_i , it is considered that 99.99% of the incident energy is lost in the termination relative to the reflected or returned energy, so the return loss would be 40 dB.

Impedance. The voltage reflection coefficient Γ is related to the impedance terminating the transmission line and to the impedance of the line itself. If a wave is launched to travel in only one direction on a uniform reflectionless transmission line of infinite length, there will be no reflected wave. The input impedance of this infinitely long transmission line is defined as its characteristic impedance Z_0 . An arbitrary length of transmission line terminated in an impedance Z_0 will also have an input impedance Z_0 . See ELECTRICAL IMPEDANCE.

If the transmission line is terminated in the arbitrary complex impedance load Z_L , the complex

voltage reflection coefficient Γ_L at the termination is given by Eq. (2).

$$\Gamma_L = \frac{Z_L - Z_0}{Z_L + Z_0} \quad (2)$$

Even when there is no unique expression for Z_L and Z_0 such as in the case of hollow uniconductor waveguides, the voltage reflection coefficient Γ has a value because it is simply a voltage ratio. In general, the measurement of microwave impedance is the measurement of Γ . Both amplitude and phase of Γ can be measured by direct probing of the voltage standing wave set up along a transmission line by the two opposed traveling waves, but this is a slow technique. Directional couplers have been used for many years to perform much faster swept frequency measurement of the magnitude of Γ , and more recently the use of automatic network analyzers under computer control has made possible rapid, accurate measurements of amplitude and phase of Γ over very broad frequency ranges. See DIRECTIONAL COUPLER.

In general, the voltage reflection coefficient Γ measured at the input port of a microwave device derives from a number of reflective discontinuities within the device. The time-domain reflectometer is extremely useful in design work for measuring the positions of the discontinuities and the magnitude of their reflections. A step of voltage with a very short rise time is launched into the input port. Measurement of the time interval between the launching and return of the reflected step gives the position of the discontinuity, and the amplitude of the reflected step gives information on the impedance of the discontinuity. See MICROWAVE IMPEDANCE MEASUREMENT.

Power. A required increase in microwave power is expensive whether it be the output from a laboratory signal generator, the power output from a power amplifier on a satellite, or the cooking energy from a microwave oven. To minimize this expense, absolute power must be measured. Most techniques involve conversion of the microwave energy to heat energy which, in turn, causes a temperature rise in a physical body. This temperature rise is measured and is approximately proportional to the power dissipated. The whole device can be calibrated by reference to low-frequency electrical standards and application of appropriate corrections. See MICROWAVE POWER MEASUREMENT; RADIOMETRY.

The power sensors are simple and can be made to have a very broad frequency response. A power meter can be connected directly to the output of a generator to measure available power P_A , or a directional coupler may be used to permit measurement of a small fraction of the power actually delivered to the load.

Microwave power measurements provide a good illustration of some of the differences between low-frequency and microwave measurement problems. Consider the simple connection of a load to a generator at low frequency (**Fig. 1**). The equivalent circuit of the generator consists of a source of constant

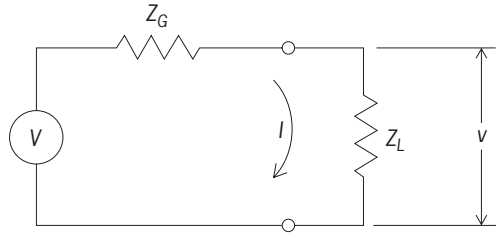


Fig. 1. A load connected to a generator.

voltage V , connected in series with Z_G , the internal impedance of the generator. The load has impedance Z_L .

If the circulating current in Fig. 1 is taken as I and the voltage drop across Z_L as v , Eqs. (3) and (4)

$$I = \frac{V}{Z_G + Z_L} \quad (3)$$

$$v = IZ_L \quad (4)$$

follow from Ohm's law. These are all vector quantities having both a magnitude and a phase angle. See ALTERNATING-CURRENT CIRCUIT THEORY.

The power P_L delivered to Z_L is a scalar quantity and can be calculated from Eq. (5), where I^* is the

$$P_L = \text{Re}\{vI^*\} \quad (5)$$

complex conjugate of the current I and the symbol $\text{Re}\{\}$ stands for the real part of the complex number enclosed inside the braces.

Substituting Eqs. (3) and (4) in Eq. (5) yields Eq. (6) for P_L in terms of the parameters of the circuit.

$$P_L = \text{Re} \left\{ \frac{VZ_LV^*}{(Z_G + Z_L)(Z_G + Z_L)^*} \right\} \quad (6)$$

It can be shown that the maximum power which the generator can deliver is available only when the load has impedance Z_G^* , the complex conjugate of the generator impedance. This power is defined as the available power P_A , a fundamental parameter of the generator, and its value is obtained by substituting Z_G^* for Z_L in Eq. (6) and simplifying to obtain Eq. (7). The actual power delivered to Z_L relative to

$$P_A = \frac{VV^*}{4\text{Re}\{Z_G\}} \quad (7)$$

the available power may be determined by taking the ratio of Eq. (6) divided by Eq. (7) to obtain Eq. (8).

$$\frac{P_L}{P_A} = \frac{4\text{Re}\{Z_G\}\text{Re}\{Z_L\}}{|Z_G + Z_L|^2} \quad (8)$$

The same simple circuit can also be considered at microwave frequencies. The same power ratio can be written as Eq. (9), where the voltage reflection

$$\frac{P_L}{P_A} = \frac{\{1 - |\Gamma_L|^2\}\{1 - |\Gamma_G|^2\}}{|1 - \Gamma_L\Gamma_G|^2} \quad (9)$$

coefficient Γ_G of the generator is given by Eq. (10).

$$\Gamma_G = \frac{Z_G - Z_0}{Z_G + Z_0} \quad (10)$$

The expression in Eq. (9) is called a mismatch factor because it would be equal to 1 if the impedances Z_G and Z_L were equal (matched) to Z_0 . Since usually the magnitudes $|\Gamma_L|$ and $|\Gamma_G|$ are measured, the denominator of the mismatch factor can commonly be determined only to be within certain limits about 1.

If a uniform transmission line is introduced between the generator and load, the voltage on the line can be considered constant at low frequency, but at microwave frequencies both the phase and the amplitude of the voltage will vary along the line. For example, if the forward-traveling voltage wave has value V_f at a certain reference plane on the transmission line, it will have a value $V_f e^{-\gamma x}$ at a distance x from that reference plane measured in the direction of travel. The parameter γ is the propagation constant and has a real and imaginary part so that the factor $e^{-\gamma x}$ modifies both amplitude and phase of V_f to account for the power loss in the transmission line and the finite time taken for the forward wave to travel the distance x .

Scattering coefficients. While the measurement of absolute power is important, there are many more occasions which require the measurement of relative power which is equivalent to the magnitude of voltage ratio and is related to attenuation. Also, there arises frequently the need to measure the relative phase of two voltages. Measurement systems having this capability are referred to as vector network analyzers, and they are used to measure scattering coefficients of multipoint devices. The concept of scattering coefficients is an extension of the voltage reflection coefficient applied to devices having more than one port. The most simple is a two-port device. Its characteristics can be specified completely in terms of a 2×2 scattering matrix (Fig. 2). The incident voltage at the reference plane of each port is defined as a , and the reflected voltage is b . Voltages a and b are related by matrix equation (11), where

$$(b_n) = (S_{nm})(a_m) \quad (11)$$

(S_{nm}) is the scattering matrix of the junction. Writing

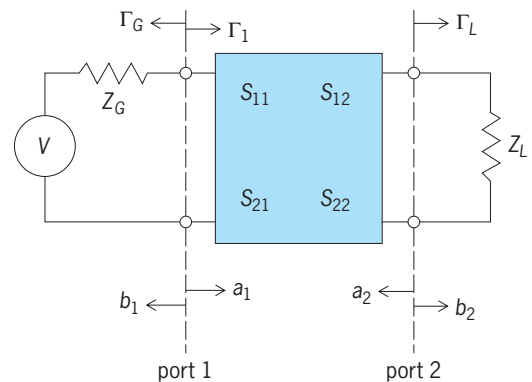


Fig. 2. A two-port inserted between a load and a generator. S_{nm} are the scattering coefficients of the two-port.

Eq. (11) out for a two-port device gives Eqs. (12) and (13). Examination of Eq. (12) shows, for example,

$$b_1 = S_{11}a_1 + S_{12}a_2 \quad (12)$$

$$b_2 = S_{21}a_1 + S_{22}a_2 \quad (13)$$

that S_{11} is the voltage reflection coefficient looking into port 1 if port 2 is terminated with a Z_0 load ($a_2 = 0$). See MATRIX THEORY.

Attenuation. To define attenuation and illustrate the use of the scattering equations, it is useful to consider the case in which the two-port is a microwave attenuator. The insertion loss of the attenuator will first be defined, and the attenuation will follow. Consider a source connected to its load as shown in Fig. 1. The power which will be delivered to Z_L is a function of V , Z_G , and Z_L , and is defined as P_L .

Now if an attenuator is inserted between the source and the load (Fig. 2), the power P_T which will be delivered to Z_L is now a function of V , Z_G , Z_L and the scattering matrix (S_{mm}) of the attenuator. The insertion loss L_I of the attenuator is defined by Eq. (14). Although L_I is a power ratio, independent

$$L_I = 10 \log_{10} \frac{P_L}{P_T} \quad (14)$$

of physical units, it is conveniently expressed as 10 times the logarithm to base 10 and assigned the unit name decibel (dB). See DECIBEL.

It can be shown that the insertion loss is given by Eq. (15), where the reflection coefficient Γ_1 is

$$L_I = 10 \log_{10} \left| \frac{(1 - \Gamma_G \Gamma_1)(1 - S_{22} \Gamma_L)}{S_{21}(1 - \Gamma_G \Gamma_L)} \right|^2 \quad (15)$$

the input reflection coefficient of the attenuator-load combination.

The attenuation A of the attenuator may now be defined, and it is the particular value of insertion loss which the attenuator causes when inserted between a generator and load both having zero reflection coefficients. Therefore the expression for A is Eq. (16),

$$A = L_I \Big|_{\Gamma_G, \Gamma_L=0} = 10 \log_{10} \left| \frac{1}{S_{21}} \right|^2 \quad (16)$$

and the remaining part of the expression in Eq. (15), will be the mismatch factor whose limits may be estimated or whose value may be calculated and applied as a correction if attenuation measurement is being made by using a source and load whose impedances are not matched to the transmission line. See ATTENUATION (ELECTRICITY).

Heterodyne. The heterodyne principle is used for scalar attenuation measurements because of its large dynamic range and for vector network analysis because of its phase coherence. The microwave signal at frequency f_s is mixed with a microwave local oscillator at frequency f_{LO} in a nonlinear mixer. The mixer output signal at frequency $f_s - f_{LO}$ is a faithful amplitude and phase reproduction of the original microwave signal but is at a low, fixed frequency so that

it can be measured simply with low-frequency techniques. One disadvantage of the heterodyne technique at the highest microwave frequencies is its cost. Consequently, significant effort has been expended in development of multiport network analyzers which use several simple power detectors and a computer analysis approach which allows measurement of both relative voltage amplitude and phase with reduced hardware cost. See HETERODYNE PRINCIPLE.

Noise. Microwave noise measurement is important for the communications field and radio astronomy. The noise figure of an amplifier is a measure of the amount of noise added to a signal in the process of amplification. If the amplifier adds a lot of noise, the signal must be correspondingly larger at the input to the amplifier to maintain the minimum signal-to-noise ratio required at the output of the amplifier, for example, to obtain appropriate television picture clarity or insignificant errors in digital communication. If measurement of noise figure shows that one amplifier has a lower noise figure than another, a smaller antenna or a lower transmitter power may be used in a telecommunications link which incorporates the less noisy amplifier.

The measurement of thermal noise at microwave frequencies is essentially the same as low-frequency noise measurement, except that there will be impedance mismatch factors which must be carefully evaluated. The availability of broadband semiconductor noise sources having a stable, high, noise power output has greatly reduced the problems of source impedance mismatch because an impedance-matching attenuator can be inserted between the noise source and the amplifier under test. See ELECTRICAL NOISE; ELECTRICAL NOISE GENERATOR; MICROWAVE NOISE STANDARDS.

Sources, detection, and automation. Microwave sources should be stable in both amplitude and frequency for accurate microwave measurements. The amplitude is stabilized by using an amplitude sensor and a negative-feedback loop. Frequency synthesis which permits automatic frequency adjustment with great precision and stability is now fairly common.

Microwave diodes are very often used as detectors. With their extremely low capacitance, their output can follow rapid changes in the amplitude of the microwave input, which is useful for rapid measurements. Although the diode output response varies from square law at low-level inputs to linear at high-level, careful calibration permits use of the diode over a wide dynamic range. A small penalty in accuracy is paid for the simplicity of the detection hardware. See DIODE; MICROWAVE SOLID-STATE DEVICES.

The need to apply calculated corrections to obtain the best accuracy in microwave measurement has stimulated the adoption of computers and computer-controlled instruments. An additional benefit of this development is that measurement techniques that are superior in accuracy but too tedious to perform manually can now be considered. An example is the multiport network analyzer using power detectors mentioned above.

Field strength. Measurement of electromagnetic radiated field strength at microwave frequencies requires the guided wave techniques and equipment discussed above plus an antenna of known gain. An antenna to be calibrated may be compared with a standard-gain horn by mounting them interchangeably on an elevated turntable and simply receiving the far-field radiation with a similarly elevated receiving antenna. Measurements of this type on outside antenna ranges have been largely superseded by gain determinations in large anechoic chambers because of the improved possibility of control of both the weather and the electromagnetic environment. See MICROWAVE FREE-FIELD STANDARDS.

Uncertainty. As is the norm in all fields of science and engineering, microwave measurements require careful attention to the sources of error that will cause uncertainty. A measurement result is incomplete and, in fact, worthless unless its associated uncertainty is given along with the probability of the measured value falling within the quoted uncertainty range. A cooperative program begun in 1952 involves the national laboratories of many countries in intercomparison of microwave standards to help assure that the microwave measurements made in those countries are meaningful and not in disagreement. See PHYSICAL MEASUREMENT. Richard F. Clark

Bibliography. A. E. Bailey (ed.), *Microwave Measurements*, 1985; G. H. Bryant, *Principles of Microwave Measurements*, 1988; G. F. Engen, *Microwave Circuit Theory and Foundations of Microwave Metrology*, 1992; H. Fukui (ed.), *Low Noise Microwave Transistors and Amplifiers*, 1981; K. C. Gupta, *Microwaves*, 1980; T. S. Laverghetta, *Modern Microwave Measurements and Techniques*, 1988; F. L. Warner, *Microwave Attenuation Measurement*, 1977.

Microwave noise standards

Electrical noise generators which produce calculable noise intensities at microwave frequencies, and which are used to calibrate other noise sources by using comparison methods. A factor limiting the performance of microwave communications and radar systems is receiver sensitivity (signal-to-noise ratio). To measure this, a reference noise source is needed. See RADAR; SIGNAL-TO-NOISE RATIO.

Blackbody radiator. Noise standards are based upon the blackbody or thermal radiator and generate noise power according to Eq. (1), which is derived

$$S = kTP(f) \quad \text{W/Hz} \quad (1)$$

from quantum mechanics. Here k is the Boltzmann constant (1.38×10^{-23} joule/kelvin), T the absolute temperature in kelvins, and $P(f)$ the Planck function given by Eq. (2), where h is the Planck constant

$$P(f) = \frac{hf}{kT} [\exp(hf/kT) - 1]^{-1} \quad (2)$$

(6.62×10^{-34} J · s) and f the frequency in Hz. The

quantity $TP(f)$ is referred to as the noise temperature and is used as a convenient measure of available noise power from a source. In the microwave region and at room temperature $P(f)$ is nearly equal to 1, and thus $S = kT$ W/Hz. See HEAT RADIATION.

The practical realization of a blackbody in the microwave region may best be understood by reference to Kirchhoff's law of radiation, which states that for any body the ratio of emissivity E to absorptivity A is equal to the emissivity E_b of a blackbody; that is, Eq. (3) holds. To realize a blackbody the absorptivity

$$\frac{E}{A} = E_b \quad (3)$$

A is set equal to 1. A microwave absorber with unity absorptivity can be achieved by using a transmission line terminated in its characteristic impedance, or in microwave terminology a matched termination. See TRANSMISSION LINES.

Temperature. The range of sources which require calibration and the desire to obtain low uncertainties dictate that microwave thermal noise standards are required with temperatures both above and below the ambient temperature. Sources have been developed with temperatures in the range from 4 to 1300 K (−452 to 1900°F). The low temperatures are normally achieved by immersion of the matched termination in a cryogenic liquid of which liquid nitrogen (77 K or −321°F) is the most common. Standards for measurement of high-temperature sources have the termination in a heated oven. A transition section supports the temperature gradient from the thermal termination to the ambient temperature output which connects to the measurement system (Fig. 1). The structures are produced in transmission lines with well-established characteristics such as coaxial lines or waveguides. Complications in the design of microwave thermal noise standards arise principally in the transition and output sections. These are lossy elements with nonzero absorptivity and are thus themselves thermal noise generators. Because the absorptivity A is less than 1 for these elements, the noise power which they contribute is proportional to kAT . If the transmission coefficient of the lossy element is represented by Eq. (4), then for

$$\tau = 1 - A \quad (4)$$

a two-port component at a temperature T_A with an input noise temperature T_{in} , the output noise temperature is given by Eq. (5). The transition and output

$$T_{out} = T_{in}\tau + (1 - \tau)T_A \quad (5)$$

sections thus reduce the noise output from the blackbody termination and add some noise themselves. The combined effect is such that the output noise temperature relative to the termination temperature is lower for an above-ambient source and higher for a below-ambient source. See COAXIAL CABLE; WAVEGUIDE.

Radiometers. Measurement of microwave noise involves comparison between an unknown source and a thermal noise source or a secondary standard

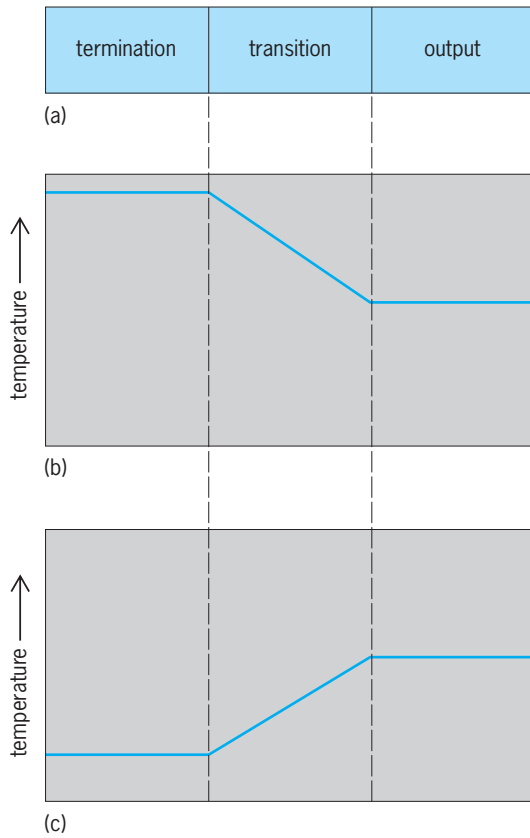


Fig. 1. Microwave noise standard. (a) General form. (b) Temperature variation in a standard with termination at a temperature above ambient. (c) Temperature variation in a cryogenic standard.

which has been compared previously with a thermal source. The measurement systems are classified under the general heading of radiometers, which simply compare the noise powers of the standard and unknown noise sources. The most simple type of radiometer is the total power radiometer (Fig. 2).

The standard and unknown noise sources are connected to the input of the amplifier in turn and the output observed on the power meter. The unknown noise source may be calculated by using Eq. (6),

$$T_x = Y T_s + T_R (Y - 1) \quad (6)$$

where T_x and T_s are the noise temperatures of the unknown and standard sources, and Y is the ratio of the power meter readings when T_x and T_s are connected,

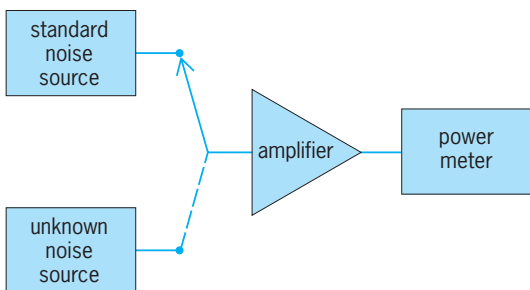


Fig. 2. Total-power radiometer.

respectively. T_R is the input noise temperature of the amplifier system and is determined separately by using two known sources at the system input. The sensitivity of the system is given by Eq. (7), where T_{in}

$$\Delta T = \frac{T_{in}}{(BC)^{1/2}} \quad (7)$$

is the sum of the receiver and source noise temperatures at the input, B is the amplifier bandwidth, and C is the time constant associated with the power meter. ΔT represents the minimum change at the system input which may be detected at the output. The total power radiometer may be used for measurement of above- and below-ambient sources. For the below-ambient (cryogenic) sources the value of T_R must be as low as possible to retain good sensitivity. See MICROWAVE; MICROWAVE MEASUREMENTS; RADIOMETRY.

Malcolm W. Sinclair

Bibliography. A. E. Bailey (ed.), *Microwave Measurement*, 1989; G. H. Bryant, *Principles of Microwave Measurement*, 1988.

Microwave optics

The study of those properties of microwaves which are analogous to the properties of light waves in optics. The fact that microwaves and light waves are both electromagnetic waves, the major difference being that of frequency, already suggests that their properties should be alike in many respects. But the reason microwaves behave more like light waves than, for instance, very low frequency waves for electrical power (50 or 60 Hz) is primarily that the microwave wavelengths are usually comparable to or smaller than the ordinary physical dimensions of objects interacting with the waves.

In his classical experiments to verify Maxwell's theory, H. Hertz first demonstrated the optical properties of damped decimeter waves, such as rectilinear propagation, reflection, refraction, and polarization. It is virtually taken for granted that microwaves inherently possess all these properties, and the language of geometrical or physical optics is freely used wherever allowed by the situation.

Rectilinear propagation. As is the case with light, a beam of microwaves propagates along a straight line in a perfectly homogeneous infinite medium. This phenomenon follows directly from a general solution of the wave equation in which the direction of a wave normal does not change in a homogeneous medium. In the use of radar, a microwave beam is justifiably presumed to travel in a straight line before and after reflection by an object. For microwave communication in cases when two distant stations are not along the line of sight, the waves would be blocked by the Earth's surface. The difficulty is remedied by the use of microwave relay links so that straight-line propagation is maintained in each section. See WAVE EQUATION.

Reflection and refraction. Consider a plane boundary between two semi-infinite media having different physical properties (Fig. 1). If a plane-polarized

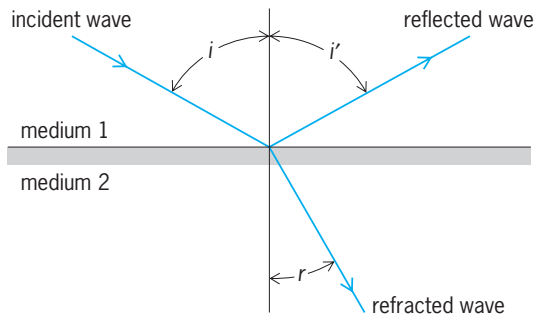


Fig. 1. Reflection and refraction of microwaves at a plane boundary between two insulating media.

microwave is incident from medium 1, the boundary conditions generally require the presence of a reflected wave back to medium 1 and a transmitted (refracted) wave into medium 2. In the case of two insulating media, the familiar relations in optics shown in Eqs. (1) and (2) hold, where the angles i , i' , and r

$$i = i' \quad \text{for reflection} \quad (1)$$

$$\frac{\sin i}{\sin r} = \sqrt{\frac{\epsilon_2}{\epsilon_1}} = N \quad \text{for refraction} \quad (2)$$

are as indicated in Fig. 1, ϵ_1 and ϵ_2 are the dielectric constants of media 1 and 2, respectively, and N is the index of refraction of medium 2 relative to medium 1. The reflected and refracted intensities depend upon whether the incident electric intensity is polarized in the plane of incidence or perpendicular to it. In any case, the well-known Fresnel equations of optics can all be applied to this case. See REFLECTION OF ELECTROMAGNETIC RADIATION; REFRACTION OF WAVES.

With some modification the laws of reflection and refraction can be applied to the propagation of microwaves inside a dielectric-filled metallic waveguide. The usual case is that of a vertical incidence to the plane boundary between two dielectric media perpendicular to the lengthwise direction of the waveguide. The reflection coefficient can be obtained by measuring the standing-wave ratio in the waveguide. Another interesting application is associated with the microwave analog of total internal reflection in optics. It may be seen from Eq. (2) that if $\epsilon_1 > \epsilon_2$ (that is, if the wave is incident from a denser medium), there is a total internal reflection for the wave when $i > \sin^{-1} \sqrt{\epsilon_2/\epsilon_1}$. This means that a properly designed dielectric rod (without metal walls) can serve as a waveguide by totally reflecting the elementary plane waves. Still another case of interest is that of a microwave lens. By using either a natural dielectric of a certain shape or an artificial dielectric consisting of an array of thin metal plates of a certain design, a microwave lens can be constructed which has the required index of refraction. Such lenses have been used as microwave antennas. See ANTENNA (ELECTROMAGNETISM); WAVEGUIDE.

The reflection of a microwave by a conducting plane has all the characteristics of the reflection of a light wave by a metallic mirror. The amplitude of the

reflected wave is practically identical to that of the incident wave, with the angle of reflection equal to the angle of incidence. The wave in the conducting medium does not go much beyond a "skin depth" and is of little practical consequence. Examples of reflection of microwaves by conductors are found in parabolic reflectors used as antennas and in targets for radar beams.

Diffraction. In an analogous manner to light, a microwave undergoes diffraction when it encounters an obstacle or an opening which is comparable to or somewhat smaller than its wavelength. Diffraction problems have been much studied but the results are generally too complicated for a simple description. One case of considerable importance, however, may be cited as an illustration. Let two waveguides be coupled through a small hole in a metallic partition as shown in Fig. 2. The radius of the hole is assumed to be much smaller than $\lambda/2\pi$, where λ is the wavelength. A wave in one waveguide will leak through the hole by diffraction into the other waveguide. While an exact calculation is difficult, a satisfactory treatment can be worked out by regarding the diffraction effect as being equivalent to the presence of an electric and magnetic dipole placed at the position of the hole. The radiation fields of the dipoles supply the necessary wave coupling between the two waveguides. For additional information on microwave diffraction See DIFFRACTION.

Polarization. The polarization of an electromagnetic wave is specified by the direction of the electric and magnetic intensities. For simplicity, consider a plane wave propagating in the z direction. The electric and magnetic intensities are always mutually perpendicular to each other in the xy plane. It is, consequently, only necessary to consider the polarization of the electric intensity alone. If the electric intensity is polarized along one direction, say the x axis, then the wave is said to be plane polarized. If there are components E_x and E_y equal in amplitude but differing in phase by 90° , the wave is circularly polarized

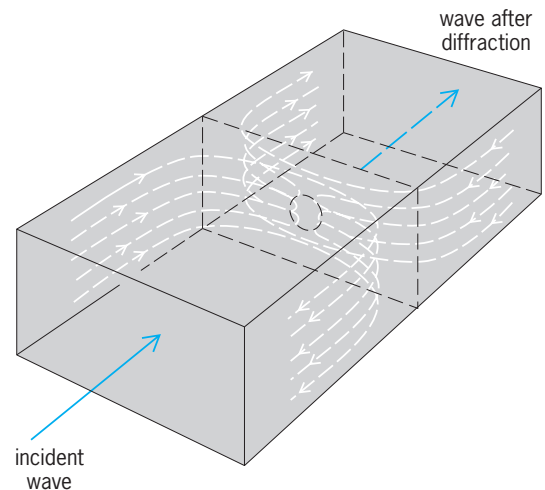


Fig. 2. Diffraction of microwaves through a small aperture between two waveguides. Only magnetic lines of force are shown for the field pattern.

(right-handed for E_y lagging E_x and left-handed for E_y leading E_x). Lastly, if the wave is neither plane nor circularly polarized, it must be elliptically polarized. See POLARIZATION OF WAVES; POLARIZED LIGHT.

The preceding essentially optical description holds true for microwaves in free space or wherever there is a transverse electric and magnetic (TEM) wave. Since a hollow waveguide does not support a TEM wave, the description of the polarization of the wave is much more complicated. However, the general notions expressed here are still valid and useful. For instance, the electric intensity of the dominant mode (TE_{01}) in a rectangular waveguide is plane polarized. This situation is also approximately true for the dominant mode (TE_{11}) in a circular waveguide. A circular waveguide is particularly useful in transforming plane-polarized electric intensity into circularly polarized electric intensity by a technique equivalent to the use of a quarter-wave plate in optics. For this purpose, it suffices to use a thin slab of dielectric material, such as mica, to introduce a 90° phase shift for one of the two equal components of the electric intensity. For information on TEM waves see MICROWAVE.

Faraday effect for microwaves. In optics, the Faraday effect is the rotation of the plane of polarization of a light beam which propagates in a dense transparent medium placed in a magnetic field along the direction of propagation. In microwaves, a similar phenomenon has been discovered and has led to many interesting applications. See FARADAY EFFECT.

Consider, for example, a circular waveguide which contains a slender rod of ferrite (a magnetic material of very low conductivity), as shown in Fig. 3. Place a steady magnetic field along the axial direction. If a wave with vertical electric polarization is incident from the left, then in passing through the ferrite zone its plane of polarization is rotated by an angle θ as indicated in the figure. This Faraday effect can be explained by the action of precessing elementary magnets in the magnetized ferrite upon the phase of the propagating wave. The initial plane-polarized electric intensity is equivalent to two oppositely rotating circularly polarized components. Only one of the components is affected (principally in the phase factor) by the precessing magnets because the latter can only precess in a unique direction corresponding to a given direction of the magnetic field. In the

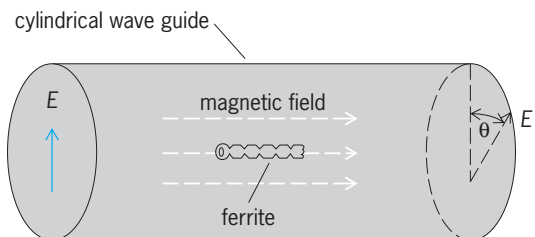


Fig. 3. Rotation of the plane of polarization of a microwave by a ferrite rod in a longitudinal magnetic field. E represents electric intensity.

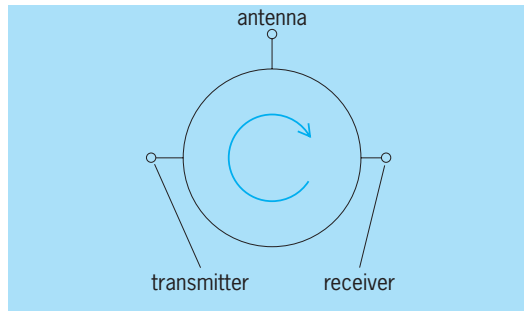


Fig. 4. Circulator formed by one or more gyration elements for nonreciprocal application.

output, the combination of two circularly polarized waves with a relative phase shift is equal to a plane-polarized wave having its polarization rotated by an angle with respect to the initial plane.

If the output wave is sent back to pass by the ferrite from right to left, the plane of polarization of the backward wave rotates another θ degrees in the same direction as the initial rotation. In other words, the polarization of the new output wave on the left is rotated from the initial plane by 2θ , instead of zero degrees as might be expected by the principle of reciprocity. Thus this system constitutes a nonreciprocal circuit element and is sometimes called a gyration or signify the gyrating motion of the elementary magnets. See GYRATION; RECIPROCITY PRINCIPLE.

One of the most important applications of a gyration is found in unidirectional transmission. Suppose a plane-polarized wave is introduced from a rectangular to a circular waveguide containing a gyration and finally to another rectangular waveguide oriented to suit the polarization of the output wave. If the angle of rotation, which defines the output polarization, is adjusted to 45° , any wave from the output which is fed back to the input will be at right angles with the input intensity and hence will not be accepted by the rectangular waveguide. Another important application of gyrations is known as a circulator, a simple example of which is given in Fig. 4. By means of one or more gyrations, a circulator such as the one depicted will allow a wave from the transmitter to go to the antenna but not to the receiver and will let a wave from the antenna go to the receiver but not to the transmitter.

C. K. Jen

Bibliography. S. Cornbleet, *Microwave and Geometric Optics*, 1994; N. Marcuvitz et al., *Waveguide Handbook*, 1986; A. D. Oliver, *Microwave and Optical Transmission*, 1992; D. M. Pozar, *Microwave Engineering*, 2d ed., 1997.

Microwave power measurement

Determination of the rate at which energy is transmitted by microwave propagation. Power is one of the fundamental parameters measured at microwave frequencies. Since the physical dimensions of the microwave system are comparable with the wavelength, the voltage along a conductor is no longer

constant. Further, for energy transmitted inside a hollow, uniconductor waveguide, the voltage difference between two points is not uniquely defined and varies as a function of the path taken to pass from one point to the other. Thus voltage standards, basic to low-frequency electrical measurements, give way to power standards at microwaves. Power meters are connected either temporarily, to measure all the power, in place of the normal termination, or permanently to measure a fixed portion of the power. See MICROWAVE; POWER.

Terminating power meters. A terminating power meter absorbs a power P_M at its one input port. The effective efficiency η_e of the terminating power meter is defined by Eq. (1), where P_I is the power in-

$$\eta_e = \frac{P_I}{P_M} \quad (1)$$

dicated by the power meter. A parameter designated the calibration factor K of a terminating power meter is defined by Eq. (2). In this case the power incident

$$K = \frac{P_I}{\text{power incident}} \quad (2)$$

has a rather special meaning. It is the energy per second flowing in the electromagnetic wave incident on the power meter as measured an infinitesimal distance ahead of the power meter input.

In order to relate the two parameters, η_e and K , consider the terminating power meter connected to a short section of transmission line having characteristic impedance Z_0 which has electromagnetic energy fed into its opposite end (Fig. 1). If the incident electromagnetic wave is assumed to have voltage V at an infinitesimal distance δ in front of the power meter input, the reflected electromagnetic wave at the same plane will have voltage $\Gamma_M V$, where Γ_M is the voltage reflection coefficient of the power meter. See ELECTRICAL IMPEDANCE; MICROWAVE IMPEDANCE MEASUREMENT; REFLECTION AND TRANSMISSION COEFFICIENTS; TRANSMISSION LINES.

Since the energy flow in the traveling waves is proportional to the square of the voltage, assuming an ideal section of transmission line connected to the power meter, it follows that Eq. (3) is valid, where

$$\frac{P_M}{\text{Power incident}} = 1 - |\Gamma_M|^2 \quad (3)$$

$|\Gamma_M|$ is the magnitude of the voltage reflection coefficient. Hence the relation between K and η_e is given by Eq. (4).

$$K = \eta_e(1 - |\Gamma_M|^2) \quad (4)$$

While η_e and K are both useful parameters for a terminating power meter, η_e should be considered the more fundamental since it is independent of the characteristic impedance of the connecting line and of the impedance of the power meter. Furthermore, the two parameters are very simply related, and either one may be determined through knowledge of the other one and an impedance measurement.

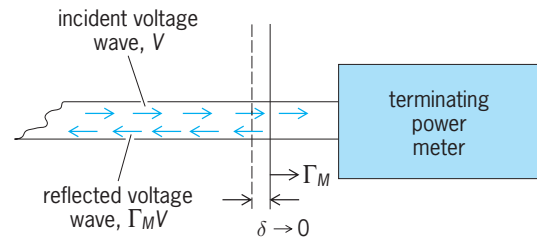


Fig. 1. Microwave power incident on a terminating power meter.

The power P_0 that a microwave source would deliver to a termination with characteristic impedance Z_0 , as measured by the terminating power meter, is given by Eq. (5), where Γ_G is the voltage reflection

$$P_0 = \frac{P_I}{K} |1 - \Gamma_G \Gamma_M|^2 \quad (5)$$

coefficient of the source. Often in the past, the impedance portion of the expression in Eq. (5) was not calculated, but its maximum deviation from 1 was included in the error budget of the measurement. Now, as vector network analyzers with their capability to measure both phase and amplitude of Γ have come into wider usage, it has become more common to do the impedance mismatch calculation to reduce the uncertainty of the measurement.

Two-port power monitor. The use of a two-port power monitor was proposed as a technique for assembling a standard power source. The general form of this type of device consists of a stable three-port network converted to a two-port power monitor by permanent connection of a power meter to one of its ports. An excellent power monitor (Fig. 2) has a terminating power meter connected to the side-arm, port 3, of a three-port directional coupler. The output from port 3 is proportional to the microwave energy traveling in the main line of the directional coupler in the direction from port 1 to port 2. The fourth port of the directional coupler is terminated internally. A source of radio-frequency energy with adjustable amplitude is connected to port 1. Port 2 becomes the output of a standard source with power available and output impedance given by the power meter reading and the characteristics of the three-port. The variable attenuator may be adjusted manually for a desired power meter reading, or, as suggested by the broken line, an electrical output from the power meter may be used to automatically control the attenuator to

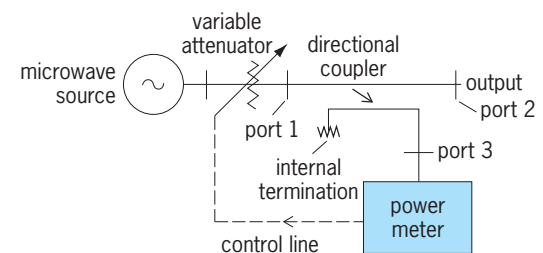


Fig. 2. A standard source of microwave power stabilized by a power meter.

maintain a constant power available from port 2. See DIRECTIONAL COUPLER.

A calibration factor may be assigned to the directional coupler with side-arm power monitor. The calibration factor of interest is K_2 and is defined by Eq. (6), where P_{I_3} is the power indicated by the

$$K_2 = \frac{P_{I_3}}{P_{O_2}} \quad (6)$$

power monitor at port 3 when the power delivered to a load with characteristic impedance Z_0 connected at port 2 is P_{O_2} . Calibration of the coupler-power monitor is thus performed by connection of a power meter having known characteristics to the output port 2 and determination of P_{O_2} for a particular power monitor reading. Use is made of Eq. (5) for this measurement.

The directional coupler-power monitor must now be regarded as a single device and must remain connected together permanently as a calibrated feed-through power monitor. The advantage of this coupler-power monitor is that it can have a very low insertion loss and yet have a very low effective source reflection coefficient. However, its bandwidth is limited by the directional coupler.

For cases in which the low insertion loss is not necessary, it is possible to assemble an extremely broadband feed-through power meter (Fig. 3). The power splitter has no low-frequency cutoff, but three-fourths of the power available from the source is lost. In the particular case where R is 50 ohms, the effective source reflection coefficient has magnitude zero in a 50-ohm transmission line system. This is due to the fact that a monitored T junction behaves as a zero impedance voltage source.

Power sensors. A number of techniques have been used to sense microwave power. These include (1) conversion of the microwave power to heat and sensing the thermal effect on a calorimetric body, (2) measurement of the force caused by microwave radiation pressure, (3) measurement of a microwave Hall effect, (4) measurement of electron acceleration caused by a microwave electric field, and (5) rectification of microwave voltage by using diodes. Rectification of microwave voltage is the most simple and most sensitive technique, but conversion of microwave power to heat is by far the most widely used because of its long-term stability. Two versions of de-

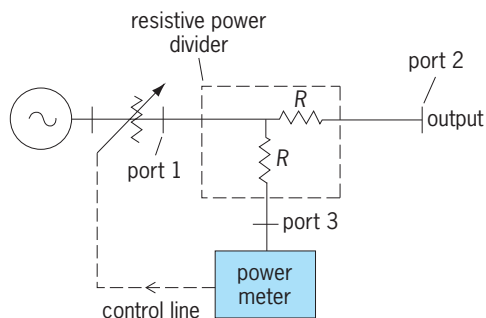


Fig. 3. A standard microwave source having very broad frequency range.

tor, based on conversion to heat, are bolometers and thermocouples. See DIODE; HALL EFFECT; RADIATION PRESSURE; RECTIFIER.

Bolometer mounts. A bolometer is a component whose resistance varies with temperature. Very small bolometers are inside a bolometer mount having a transmission line input. The bolometer is located so that it can simultaneously absorb microwave power and dc power. By using electronic instrumentation, the bolometer resistance is held constant by automatic adjustment of the dc bias power dissipated in the bolometer. When microwave power is fed into the mount, the dc power is automatically reduced to keep the bolometer resistance constant, and the change in dc power required is a measure of the microwave power. The most used type of bolometer is the thermistor with a negative temperature coefficient of resistance. Since this is a power substitution technique, any long-term drift or instability in the thermistor is compensated by an automatic change in dc bias power. See BOLOMETER; THERMISTOR.

Thermocouple detectors. A type of power detector in which the microwave power is dissipated directly on the warm junctions of a thermopile has largely replaced the coaxial thermistor mount for general usage. The electronic instrumentation is simply a low-noise dc amplifier-voltmeter combination to measure the thermopile output. Also, manufacturers have achieved a better impedance match over a broad frequency range with this detector than with a thermistor mount. One complication is the inclusion of a stable radio-frequency reference source into the electronic part of the measurement system to permit checking of the long-term stability of the detector. See RADIOMETRY; THERMOCOUPLE.

Standards. The power sensors and their associated electronics are the standards for measurement of microwave power throughout industry and in the laboratory. The electronic portion can usually be calibrated by reference to dc standards. The efficiency and calibration factor of the microwave power sensors are compared through a hierarchy of laboratories to fundamental standards of microwave power. These fundamental standards are essentially the calorimetric type mentioned above, but are configured so that their efficiency is either calculable or measurable.

In one type of fundamental standard, referred to as a coaxial calorimeter, microwave power is dissipated in a resistor and the temperature rise of the resistor and its mounting is measured by a thermopile. Direct-current power for the same temperature rise is fed into the calorimeter through the same input port as was used for microwave power. The calorimeter is constructed so that the microwave and direct currents in the resistor are as nearly as possible the same. The difference between the microwave loss and the dc loss in the thermally insulating input to the resistor is measured and used to calculate the reduction in efficiency of the calorimeter at high measurement frequencies.

A second type of fundamental standard involves again a calorimetric measurement, but in this case

a thermistor mount is placed in thermal isolation inside an enclosure. The dc bias power causes the thermistor mount to heat up relative to its immediate surroundings, and a thermopile indicates this relative temperature rise. When microwave power is applied, the dc power automatically decreases to maintain constant power dissipation in the thermistor. However, any microwave power not dissipated in the thermistor itself causes an increase in thermopile output which is interpreted as a measure of the inefficiency of the thermistor mount.

Transfer of the accuracy of the power standard to a power sensor under test is a two-step comparison process. For the first step, the power standard is connected to the output of a source (Figs. 1 and 2). Modern microwave sources are quite stable in amplitude so the controlled variable attenuator would not be necessary here. Instead, simultaneous readings of the power indicated by the standard and by the source power meter permits the calibration of the source as a so-called working-standard power source. The second step involves the connection of the power sensor under test to the working-standard power source. Simultaneous power readings and a calculation of impedance mismatch corrections complete the calibration. See ELECTRIC POWER MEASUREMENT; MICROWAVE MEASUREMENTS.

Richard F. Clark

Bibliography. G. H. Bryant, *Principles of Microwave Measurement*, 1988; A. E. Fantom, A. E. Bailey, and A. C. Lynch (eds.), *Radio Frequency and Microwave Power Measurement*, 1990; T. S. Laverghetta, *Handbook of Microwave Testing*, 1981; T. S. Laverghetta, *Modern Microwave Measurements and Techniques*, 1988; T. S. Laverghetta, *Practical Microwaves*, 1996.

Microwave solid-state devices

Semiconductor devices used for the detection, generation, amplification, and control of electromagnetic radiation with wavelengths from 30 cm to 1 mm (frequencies from 1 to 300 GHz). Since 1985, the number and variety of microwave semiconductor devices, used for wireless and satellite communication and optoelectronics, have increased as new techniques, materials, and concepts have been developed and applied. Passive microwave devices, such as *pn* and PIN junctions, Schottky barrier diodes, and varactors, are primarily used for detecting, mixing, modulating, or controlling microwave signals. Step-recovery diodes, transistors, tunnel diodes, and transferred electron devices (TEDs) are active microwave devices that generate power or amplify microwave signals. See MICROWAVE; SEMICONDUCTOR DIODE.

Materials

Typical high-frequency semiconductor materials include silicon (Si), germanium (Ge), and compound semiconductors, such as gallium arsenide (GaAs), indium phosphide (InP), silicon germanium (SiGe), silicon carbide (SiC), and gallium nitride (GaN). In

general, the compound semiconductors work best for high-frequency applications due to their higher electron mobilities. See GALLIUM; GERMANIUM; SEMICONDUCTOR; SILICON.

Silicon microelectronics is a mature technology. Its advantages include good mechanical strength, high thermal conductivity, and a high-quality oxide (SiO₂), allowing for the fabrication of field-effect transistors (FETs). FETs form the basis for complementary metal-oxide semiconductor (CMOS) technology. A major disadvantage of silicon is its low intrinsic resistivity, resulting in very high dielectric losses, which make it almost impossible to transmit microwave signals. To help alleviate the inherent low resistivity of silicon, high-resistivity silicon-on-insulator (SOI) can be used to reduce circuit parasitics (such as parasitic capacitances and resistances). Even with improvements, silicon devices are limited in high-frequency microwave applications, thus requiring inherently faster materials, such as GaAs, InP, or SiGe. The electron mobility and energy bandgap for GaAs and InP are higher than the corresponding values for silicon. A major advantage is that semi-insulating GaAs and InP substrates are easily produced, allowing interconnects to be put over these substrates without introducing large parasitic capacitances. By metalizing the bottom side of the substrate, microstrip lines can be made that integrate the passive microwave circuits with transistors on monolithic microwave integrated circuits (MMICs). The disadvantages of GaAs are low hole mobility, which severely limits its ability to support complementary circuitry, and the inability to form a high-quality oxide on GaAs, a shortcoming that has been partially mitigated by high electron mobility transistors (HEMTs).

Passive Devices

Passive microwave solid-state devices include Schottky-barrier diodes, PIN diodes, and varactor diodes.

PIN and *pn* diodes. A PIN (*p*-type/*i*ntrinsic/*n*-type) diode is a *pn* diode that has an undoped (intrinsic) region between the *p*- and *n*-type regions. The use of an intrinsic region in PIN diodes allows for high-power operation and offers an impedance at microwave frequencies that is controllable by a lower frequency or a direct-current (DC) bias. The PIN diode is one of the most common passive diodes used at microwave frequencies. Under zero and reverse bias the PIN diode has a very high impedance, whereas at moderate forward current its impedance is relatively low. The DC behavior of a PIN diode is basically the same as that of an ordinary *pn* junction. However, at microwave frequencies PIN diodes behave as a frequency-independent conductance that is DC-controlled. The capacitance is mainly determined by the width of the intrinsic region, which is only weakly influenced by the DC bias. The DC control current through a PIN diode can vary the resistance value from over 1 megohm to less than 1 ohm. With reverse or zero bias, the intrinsic region is depleted of carriers and the PIN exhibits very high resistance. When forward bias is applied across the PIN diode, holes from the

p -region and electrons from the n -region are injected into the intrinsic region, which increases the conductivity. High off-resistance and low on-resistance make the PIN diode highly attractive for switching applications. The PIN diode switch can be viewed as a resistor in the forward-biased state and a capacitor in the reverse-biased state. PIN diodes are used to switch lengths of transmission line, providing digital increments of phase in individual transmission paths, each capable of carrying kilowatts of peak power. PIN diodes come in a variety of packages for microstrip and stripline packages, and are used as microwave switches, modulators, attenuators, limiters, phase shifters, protectors, and other signal control circuit elements. See JUNCTION DIODE.

Schottky barrier diodes. A Schottky barrier diode (SBD) consists of a rectifying metal-semiconductor barrier typically formed by deposition of a metal layer on a semiconductor. The SBD functions in a similar manner to the antiquated point contact diode and the slower-response pn -junction diode, and is used for signal mixing and detection. The point contact diode consists of a metal whisker in contact with a semiconductor, forming a rectifying junction. The SBD is more rugged and reliable than the point contact diode. The SBD's main advantage over pn diodes is the absence of minority carriers, which limit the response speed in switching applications and the high-frequency performance in mixing and detection applications. SBDs are zero-bias detectors. Frequencies to 40 GHz are available with silicon SBDs, and GaAs SBDs are used for higher-frequency applications. See SCHOTTKY EFFECT.

Varactor diodes. The variable-reactance (varactor) diode makes use of the change in capacitance of a pn junction or Schottky barrier diode, and is designed to be highly dependent on the applied reverse bias. The capacitance change results from a widening of the depletion layer as the reverse-bias voltage is increased. As variable capacitors, varactor diodes are used in tuned circuits and in voltage-controlled oscillators. For higher-frequency microwave applications, silicon varactors have been replaced with GaAs. Typical applications of varactor diodes are harmonic generation, frequency multiplication, parametric amplification, and electronic tuning. Multipliers are used as local oscillators, low-power transmitters, or transmitter drivers in radar, telemetry, telecommunication, and instrumentation. See VARACTOR.

Active Microwave Devices

Transistors are the most widely used active microwave solid-state devices. At very high microwave frequencies, high-frequency effects limit the usefulness of transistors, and two-terminal negative resistance devices, such as transferred-electron devices, avalanche diodes, and tunnel diodes, are sometimes used.

Microwave transistors. Two main categories of transistors are used for microwave applications: bipolar junction transistors (BJTs) and field-effect transistors (FETs). In order to get useful output power at high frequencies, transistors are designed

to have a higher periphery-to-area ratio using a simple stripe geometry. The area must be reduced without reducing the periphery, as large area means large interelectrode capacitance. For high-frequency applications the goal is to scale down the size of the device. Narrower widths of the elements within the transistor are the key to superior high-frequency performance. See TRANSISTOR.

Microwave bipolar transistors. A BJT consists of three doped regions forming two pn junctions. These regions are the emitter, base, and collector in either an npn or pnp arrangement. Silicon npn BJTs have an upper cutoff frequency of about 25 GHz (varies with manufacturing improvements). The cutoff frequency is defined as the frequency at which the current amplification drops to unity as the frequency is raised. The primary limitations to higher frequency are base and emitter resistance, capacitance, and transit time. Modern silicon npn BJTs are of two designs: the planar design has its regions implanted or diffused under the surface and patterned using photolithography techniques, and the epitaxial design uses a thin, low-conductivity epitaxial layer for part of the collector region (the remainder is highly conductive, n^+ material).

To operate at microwave frequencies, individual transistor dimensions must be reduced to micrometer or submicrometer size. To maintain current and power capability, various forms of internal paralleling on the chip are used. Three of these geometries are interdigitated fingers that form the emitter and base, the overlaying of emitter and base stripes, and the matrix approach. Silicon BJTs are mainly used in the lower microwave ranges. Their power capability is quite good, but in terms of noise they are inferior to GaAs metal semiconductor field-effect transistors (MESFETs) at frequencies above 1 GHz and are mainly used in power amplifiers and oscillators. They may also be used in small-signal microwave amplifiers when noise performance is not critical. See ELECTRICAL NOISE.

Heterojunction bipolar transistors. Heterojunction bipolar transistors (HBTs) have been designed with much higher maximum frequencies than silicon BJTs. HBTs are essentially BJTs that have two or more materials making up the emitter, base, and collector regions (**Fig. 1**). In HBTs, the major goal is to limit the injection of holes into the emitter by using an emitter material with a larger bandgap than the base. The difference in bandgaps manifests itself as a discontinuity in the conduction band or the valence band, or both. For npn HBTs, a discontinuity in the valence band is required. In general, to make high-quality heterojunctions, the two materials should have matching lattice constants. For very thin layers, lattice matching is not absolutely necessary as the thin layer can be strained to accommodate the crystal lattice of the other material. Fortunately, the base of a bipolar transistor is designed to be very thin and thus can be made of a strained layer material. Combinations such as AlGaAs/InGaAs and Si/SiGe are possible. See BAND THEORY OF SOLIDS; ELECTRICAL CONDUCTIVITY OF METALS; HOLE STATES

IN SOLIDS; SEMICONDUCTOR HETEROSTRUCTURES.

Aluminum gallium arsenide (AlGaAs) and GaAs fulfill the lattice-matching requirement. AlGaAs/GaAs HBTs typically have an AlAs mole fraction of 20–30%. HBTs, in comparison with silicon, have higher electron mobility, resulting in higher cutoff frequencies; a larger bandgap, thus less thermal generation of charge carriers; a semi-insulating substrate for isolation; and the possibility of integration with optoelectronic components. Compared with GaAs-based MESFETs, HBTs have higher transconductance and output current. HBTs can have cutoff frequencies higher than 300 GHz.

The base of the Si/SiGe HBT is made of an alloy of silicon with about 10% germanium. Since germanium has a bandgap of 0.7 electronvolt (eV), as compared with 1.1 eV for silicon, the base will have a lower bandgap than the silicon emitter. A potential problem is that the lattice constant of germanium is larger than that of silicon, which could produce crystal defects at the interfaces. However, if the germanium content is relatively small and the base region is thin enough, the SiGe layer is strained to match the silicon crystal lattice and defects are not a serious problem. The bandgap difference is small and is in the valence band discontinuity. SiGe HBTs are compatible with standard silicon technology. See CRYSTAL DEFECTS; CRYSTAL STRUCTURE.

Microwave field-effect transistors. Field-effect transistors (FETs) operate by varying the conductivity of a semiconductor channel through changes in the electric field across the channel. The three basic forms of FETs are the junction FET (JFET), the metal semiconductor FET (MESFET), and the metal-oxide semiconductor FET (MOSFET). All FETs have a channel with a source and drain region at each end and a gate located along the channel, which modulates the channel conduction (Fig. 2). Microwave JFETs and MESFETs work by channel depletion. The channel is *n*-type and the gate is *p*-type for JFETs and metal for MESFETs. FET structures are well suited for microwave applications because all contacts are on the surface to keep parasitic capacitances small. The cutoff

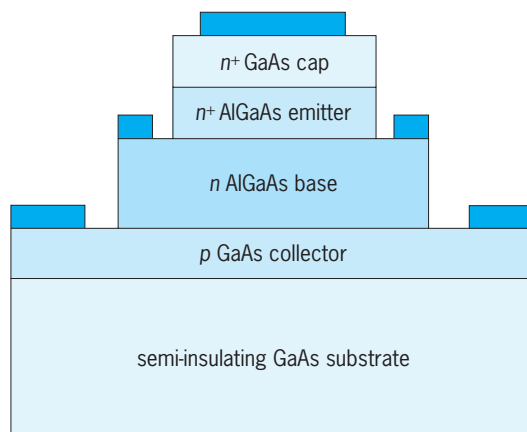


Fig. 1. Materials composition for a heterojunction bipolar transistor (HBT).

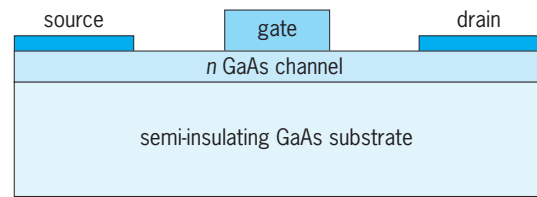


Fig. 2. Gallium arsenide metal semiconductor field-effect transistor (MESFET).

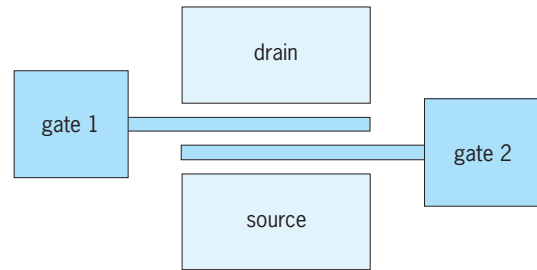


Fig. 3. Top view of a dual-gate MESFET structure.

frequency is mainly determined by the transit time of the electrons under the gate; thus short gate lengths (less than $1 \mu\text{m}$) are used.

Power devices consist of a number of MESFETs in parallel with air bridges connecting the sources. GaAs MESFET devices are used in low-noise amplifiers (LNAs), Class C amplifiers, oscillators, and monolithic microwave integrated circuits. The performance of a GaAs FET is determined primarily by the gate width and length. The planar structure of a MESFET makes it straightforward to add a second gate which can be used to control the amplification of the transistor (Fig. 3). Dual-gate MESFETs can be used as mixers (with conversion gain) and for control purposes. Applications include heterodyne mixers and amplitude modulation of oscillators. See AMPLIFIER; HETERODYNE PRINCIPLE; MIXER; OSCILLATOR.

The MOSFET has a highly insulating silicon dioxide (SiO_2) layer between the semiconductor and the gate; however, silicon MOSFETs are not really considered microwave transistors. Compared with the GaAs MESFET, MOSFETs have lower electron mobility (a decade smaller, resulting in much longer transit times), larger parasitic resistances, and higher noise levels. Also, since the silicon substrate cannot be made semi-insulating, larger parasitic capacitances result. MOSFETs therefore do not perform very well above 1 GHz. Below this frequency, MOSFETs find application mainly as radio-frequency (RF) power amplifiers where silicon MOS is competitive. Silicon lateral DMOS (diffusion metal-oxide semiconductor) is finding uses in the 10–200-watt power range for cellular base stations operating in the 0.8–1.8 GHz frequency range.

Heterojunction FETs. A disadvantage of the MESFET is that the electron mobility is degraded since electrons are scattered by the ionized impurities in the channel. By using a heterojunction consisting of *n*-type AlGaAs with undoped GaAs, electrons move from the AlGaAs to the GaAs and form a conducting channel

at the interface. The electrons are separated from the donors and have the mobility associated with undoped material. A heterojunction transistor made in this fashion has many different names: high electron mobility transistor (HEMT), two-dimensional electron gas FET (TEGFET), modulation-doped FET (MODFET), selectively doped heterojunction transistor (SDHT), and heterojunction FET (HFET). The HEMT has high power gain at frequencies of 100 GHz or higher with low noise levels. HEMTs are fabricated using ion implantation, molecular beam epitaxy (MBE), or metal organic chemical vapor deposition (MOCVD). GaAs/AlGaAs/InAlGaAs pseudomorphic HEMTs (PHEMTs) have higher mobility than standard MESFETs or HEMTs. In many applications, PHEMTs are replacing MESFETs. Compound semiconductor FETs have low noise levels. The main sources of noise are thermal noise of the parasitic resistances, thermal velocity fluctuations in the channel, fluctuations in the thickness of the channel, and scattering of electrons between valleys in the conduction band at high electric field strengths. See CRYSTAL GROWTH; ION IMPLANTATION.

Microwave integrated circuit amplifiers. A monolithic microwave integrated circuit (MMIC) can be made using silicon or GaAs technology with either BJTs or FETs. For high-frequency applications, GaAs FETs are the best choice. A MMIC has both the active and passive devices fabricated directly on the substrate. MMICs are typically used as low-noise amplifiers, as mixers, as modulators, in frequency conversion, in phase detection, and as gain block amplifiers. Silicon MMIC devices operate in the 100-MHz to 3-GHz frequency range. GaAs FET MMICs are typically used in applications above 1 GHz.

Active microwave diodes. Active microwave diodes differ from passive diodes in that they are used as signal sources to generate or amplify microwave frequencies. These include step-recovery, tunnel, Gunn, avalanche, and transit time diodes, such as impact avalanche and transit-time (IMPATT), trapped plasma avalanche triggered transit-time (TRAPATT), barrier injection transit-time (BARITT), and quantum well injection transit time (QWITT) diodes.

Step recovery diodes. A step recovery diode is a special PIN type in which charge storage is used to produce oscillations. When a diode is switched from forward to reverse bias, it remains conducting until the stored charge has been removed by recombination or by the electric field. A step recovery diode is designed to sweep out the carriers by an electric field before any appreciable recombination has taken place. Thus, the transition from the conducting to the nonconducting state is very fast, on the order of picoseconds. Because of the abrupt step, this current is rich in harmonics, so these diodes can be used in frequency multipliers. See FREQUENCY MULTIPLIER.

Negative resistance and transferred electron devices. For microwave power generation or amplification, a negative differential resistance (NDR) characteristic at microwave frequencies is necessary. NDR is a pheno-

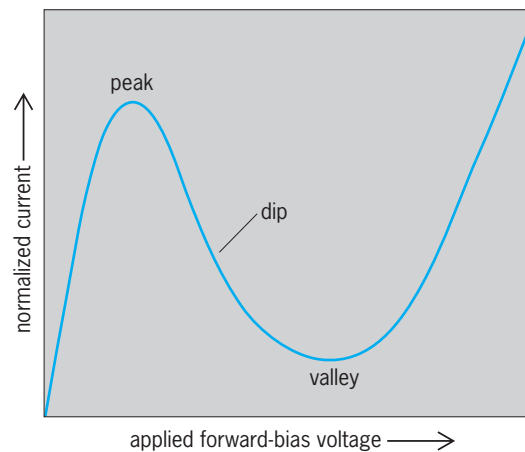


Fig. 4. Typical tunnel diode current-voltage (I - V) curve illustrating the negative differential resistance (NDR) phenomenon.

menon that occurs when the voltage (V) and current (I) are 180° out of phase. NDR is a dynamic property occurring only under actual circuit conditions; it is not static and cannot be measured with an ohmmeter. Transferred electron devices (TEDs), such as Gunn diodes, and avalanche transit-time devices use NDR for microwave oscillation and amplification. TEDs and avalanche transit-time devices today are among the most important classes of microwave solid-state devices. See NEGATIVE-RESISTANCE CIRCUITS.

Tunnel diodes. The tunnel diode uses a heavily doped abrupt pn junction resulting in an extremely narrow junction that allows electrons to tunnel through the potential barrier at near-zero applied voltage. This results in a dip in the current-voltage (I - V) characteristic, which produces NDR (Fig. 4). Because this is a majority-carrier effect, the tunnel diode is very fast, permitting response in the millimeter-wave region. Tunnel diodes produce relatively low power. The tunnel diode was the first semiconductor device type found to have NDR. See TUNNEL DIODE; TUNNELING IN SOLIDS.

A double-barrier resonant tunneling (DBRT) diode can be grown by MBE with thin layers of AlGaAs and GaAs having sharp interfaces. The layer widths are comparable to the Schrödinger wavelength of the electron, thus permitting resonant behavior. The I - V characteristic of a DBRT diode is quite similar to that of the ordinary tunnel diode, showing NDR. Since the structure is symmetric, the I - V characteristic is the same for both positive and negative bias. Asymmetry can be achieved using barriers of different widths or heights.

Avalanche diodes. Avalanche diodes are junction devices that produce a negative resistance by appropriately combining impact avalanche breakdown and charge-carrier transit time effects. Avalanche breakdown in semiconductors occurs if the electric field is high enough for the charge carriers to acquire sufficient energy from the field to create electron-hole pairs by impact ionization. For silicon and GaAs,

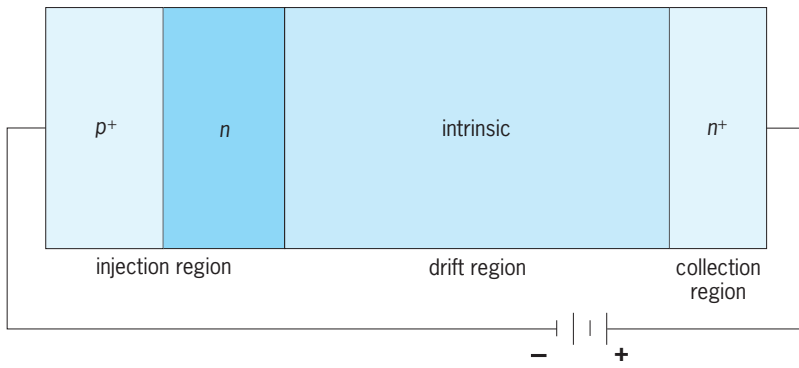


Fig. 5. Impact avalanche and transit-time (IMPATT) p^+nin^+ diode structure and DC bias (Read design).

the semiconductors commonly used to fabricate avalanche diodes, the threshold field for breakdown is on the order of a few hundred thousand volts per centimeter.

The avalanche diode is a pn -junction diode reverse-biased into the avalanche region. By setting the DC bias near the avalanche threshold, and superimposing on this an alternating voltage, the diode will swing into avalanche conditions during alternate half-cycles. The hole-electron pairs generated as a result of avalanche action make up the current, with the holes moving into the p region, and the electrons into the n region. The carriers have a relatively large distance to travel through the depletion region. At high frequencies, where the total time lag for the current is comparable with the period of the voltage, the current pulse will lag the voltage. By making the drift time of the electrons in the depletion region equal to one-half the period of the voltage, the current will be 180° out of phase. This shift in phase of the current with respect to the voltage produces NDR, so that the diode will undergo oscillations when placed in a resonant circuit.

Gunn diodes. A Gunn diode is typically an n -type compound semiconductor, such as GaAs or InP, which has a conduction band structure that supports negative differential mobility. Although this device is referred to as a Gunn diode, after its inventor, the device does not contain a pn junction and can be viewed as a resistor below the threshold electric field (E_{thres}). For applied voltages that produce electric fields below E_{thres} , the electron velocity increases as the electric field increases according to Ohm's law. For applied voltages that produce electric fields above E_{thres} , conduction band electrons transfer from a region of high mobility to low mobility, hence the general name "transferred electron device." Beyond E_{thres} , the velocity suddenly slows down due to the significant electron transfer to a lower mobility band producing NDR. For GaAs, E_{thres} is about 3 kV/cm. The Gunn effect can be used up to about 80 GHz for GaAs and 160 GHz for InP. Two modes of operation are common: nonresonant bulk (transit-time) and resonant limited space-charge accumulation (LSA). See ELECTRIC FIELD.

In the transit-time mode a charge dipole, consisting of an electron accumulation and a depletion layer

(Gunn domain), travels through the semiconductor at a frequency dependent on the length of the semiconductor layer and the drift velocity. For typical lengths, the transit time is about 0.1 nanosecond corresponding to a frequency of 10 GHz. The transit-time mode is very inefficient—power efficiencies are about 1–5%. In the LSA mode, the frequency of operation is set by a resonant circuit to be much higher than the transit-time frequency so that domains have insufficient time to form while the field is above threshold. As a result, most of the sample is maintained in the negative conductance state during a large fraction of the voltage cycle. This is a distinct advantage in achieving efficient conversion of power in the upper microwave range. LSA diodes can be operated over a very wide range of frequencies by choosing the proper doping concentration. The efficiency of the LSA mode can be as high as 20%, and LSA diodes can be operated at very high frequencies (100 GHz and above).

IMPATT diodes. Impact avalanche and transit-time diodes (IMPATTs) are NDR devices that operate by a combination of carrier injection and transit time effects. There are several versions of IMPATT diodes, including simple reverse-biased pn diodes, complicated reverse-biased multidoped pn layered diodes, and reverse-biased PIN diodes including those of the Read design (Fig. 5). The IMPATT must be connected to a resonant circuit. At bias turn-on, noise excites the tuned circuit into a natural oscillation frequency. This voltage adds algebraically across the diode's reverse-bias voltage. Near the peak positive half-cycle, the diode experiences impact avalanche breakdown. When the voltage falls below this peak value, avalanche breakdown ceases. A 90° shift occurs between the current pulse and the applied voltage in the avalanche process. A further 90° shift occurs during the transit time, for a total 180° shift which produces NDR. An IMPATT oscillator has higher output power than a Gunn equivalent. However, the Gunn oscillator is relatively noise-free, while the IMPATT is noisy due to avalanche breakdown.

TRAPATT diodes. A trapped plasma avalanche triggered transit-time (TRAPATT) diode is basically a modified IMPATT diode in which the holes and electrons created by impact avalanche ionization multiplication do not completely exit from the transit domain of the diode during the negative half-cycle of the microwave signal. These holes and electrons form a plasma which is trapped in the diode and participates in producing a large microwave current during the positive half-cycle.

BARRITT diodes. A barrier injection transit-time diode (BARRITT) is basically an IMPATT structure that employs a Schottky barrier formed by a metal semiconductor contact instead of a pn junction to create similar avalanche electron injection.

QWITT diodes. A major drawback of the IMPATT diode is the fact that the avalanche process, which depends on random impact ionization events, is inherently noisy. A variety of approaches have been investigated to find alternative methods for injecting

carriers into the drift region without relying on the avalanche mechanism. Quantum well injection transit-time diodes (QWITT) employ resonant tunneling through a quantum well to inject electrons into the drift region. The device structure consists of a single GaAs quantum well located between two AlGaAs barriers in series with a drift region of made of undoped GaAs. This structure is then placed between two n^+ -GaAs regions to form contacts.

Laurence P. Sadwick

Bibliography. K. Chang, *Microwave Solid-State Circuits and Applications*, Wiley-Interscience, 1994; S. Y. Liao, *Microwave Devices and Circuits*, 3d ed., Prentice Hall, 1996; T. G. Van De Roer, *Microwave Electronic Devices*, Chapman & Hall, 1994; J. S. Yuan, *SiGe, GaAs, and InP Heterojunction Bipolar Transistors*, Wiley-Interscience, 1999.

Microwave spectroscopy

The study of the interaction of matter and electromagnetic radiation in the microwave region of the spectrum. Microwaves are loosely defined as electromagnetic radiation with wavelengths between about 1 mm and 30 cm or frequencies between 1 and 300 GHz. The wavelengths are comparable to the dimensions of experimental apparatus. Experimental techniques make use of ideas from radio-frequency spectroscopy where wavelengths greatly exceed the dimensions of the apparatus, and also techniques from optics where wavelengths are much smaller than the size of the apparatus. See MICROWAVE; SPECTROSCOPY.

Apparatus. Microwave circuit elements (waveguides, resonant cavities, directional couplers) cannot be characterized by lumped capacitances (electric field regions) or lumped inductances (magnetic field regions), as in the case of radio-frequency circuits. Because of the short wavelengths of microwaves, both electric and magnetic fields are present in most circuit elements. When many wavelengths are present within a microwave circuit element, the momentum of the electromagnetic wave can become well defined, and momentum conservation or phase-matching conditions can have an important bearing on some spectroscopic techniques. An example is the mixing of microwaves and light waves, where the frequency of the mixed wave must equal the sum of the frequencies of the microwave and the light wave, while the momentum of the mixed wave (the inverse wavelength) must equal the vector sum of the momenta of the microwave and the light wave. See CAVITY RESONATOR; DIRECTIONAL COUPLER; WAVEGUIDE.

The transit times of electrons in ordinary electronic tubes are too long for efficient coupling to the rapidly oscillating fields of the microwave region, and special electronic tubes, klystrons, traveling-wave tubes, magnetrons, and so forth have been designed to overcome transit time limitations. Such tubes are often locked in frequency to a high harmonic of a stable quartz crystal oscillator. Solid-state

microwave devices are increasingly practical alternatives to electron tube devices for some spectroscopic applications. At the very highest microwave frequencies, alternate methods like time-domain spectroscopy with short-pulse sources and optical analogs like Fourier-transform spectroscopy are more practical than conventional microwave methods. Microwave receivers are extraordinarily sensitive, and 10^{-19} W of microwave power can be detected with a good heterodyne system in a 1-Hz bandwidth. See GYROTRON; INFRARED SPECTROSCOPY; KLYSTRON; MAGNETRON; MICROWAVE SOLID-STATE DEVICES; MICROWAVE TUBE; TRAVELING-WAVE TUBE.

Interaction of microwaves with matter. The interaction of microwaves with matter can be detected by observing the attenuation or phase shift of a microwave field as it passes through matter. These are determined by the imaginary or real parts of the microwave susceptibility (the index of refraction). The absorption of microwaves may also trigger a much more easily observed event like the emission of an optical photon in an optical double-resonance experiment or the deflection of a radioactive atom in an atomic beam. See MOLECULAR BEAMS.

Microwave energy at a frequency ν is absorbed or emitted according to the Bohr frequency condition, Eq. (1), where h is Planck's constant and E_f

$$h\nu = E_f - E_i \quad (1)$$

and E_i are the energies of the final and initial states of the absorbing system. The initial and final states may be discrete as in the case of rotational states of a molecule, or they may be continuous as in bremsstrahlung in a plasma.

Enhancement of population differences. The characteristic temperature θ of the energy splitting involved in a microwave transition is given by Eq. (2), where

$$\theta = \frac{h\nu}{k} = 0.05\text{--}15 \text{ K} \quad (2)$$

k is Boltzmann's constant. At room temperature T , the relative difference between the populations, N_i and N_j , of the states involved in the transition, $(N_i - N_j)/(N_i + N_j) = \tanh [\theta/(2T)] \sim \theta/(2T)$, is a few percent or less. The population difference can be close to 100% at liquid helium temperatures, and microwave spectroscopic experiments are often performed at low temperatures to enhance population differences and to eliminate certain line-broadening mechanisms.

The population differences between the states involved in a microwave transition can be enhanced by artificial means. For example, state selection of atoms or molecules in an atomic beam passing through inhomogeneous magnetic or electric fields can lead to very large population imbalances. Such large population inversions can be prepared so that stimulated emission cross sections for microwaves can exceed the absorption cross section. When the molecules or atoms with inverted populations are placed in an appropriate microwave cavity, microwave oscillations will build up spontaneously in the cavity

through maser (microwave amplification by stimulated emission of radiation) action. Optical pumping can also lead to large artificial population imbalances between the states involved in a microwave transition. *See* MASER; OPTICAL PUMPING.

Applications. The magnetic dipole and electric quadrupole interactions between the nuclei and electrons in atoms and molecules can lead to energy splittings in the microwave region of the spectrum. Thus, microwave spectroscopy has been used extensively for precision determinations of spins and moments of nuclei. The field-independent component of the 9.192-GHz hyperfine transition in the cesium atom is used as a time standard in many laboratories, and the second is defined as the time needed for a free cesium atom at zero magnetic field to make 9,192,631,770 oscillations. *See* ATOMIC CLOCK; HYPERFINE STRUCTURE; NUCLEAR MOMENTS; TIME.

Properties of molecules. The rotational frequencies of molecules often fall within the microwave range, and microwave spectroscopy has contributed a great deal of information about the moments of inertia, the spin coupling mechanisms, and other physical properties of molecules. The rotational frequencies of water and oxygen molecules are responsible for much of the attenuation of microwaves in the atmosphere. Many microwave transitions from simple atoms and molecules are seen from sources in interstellar space, and microwave astronomy has provided much information about the chemistry and molecular composition of various astronomical objects. Inversion frequencies in molecules, for example, the frequency of periodic motion of the nitrogen atom at the apex of the pyramidal NH₃ ammonia molecule through the three hydrogen atoms at the base, often lie in the microwave region. One of the inversion transitions in ammonia was used in the first maser. *See* INTERSTELLAR MATTER; MOLECULAR STRUCTURE AND SPECTRA.

Electron-spin resonance. The magnetic resonance frequencies of electrons in fields of a few thousand gauss (a few tenths of a tesla) lie in the microwave region. Thus, microwave spectroscopy is used in the study of electron-spin resonance or paramagnetic resonance. In the simplest cases, the spins may be well isolated in a dilute gas like oxygen (O₂) or cesium vapor. Then the microwave spectrum provides information about the internal spin couplings of a free atom or molecule. For denser gases or condensed phases, the paramagnetic resonance spectrum provides information about interactions between the spins and their environment. Both static interactions, which determine the resonance frequencies of the microwave transitions, and fluctuating random interactions, which determine resonance line widths, are of interest. Typical static interactions are the electrostatic interactions between the spin and the crystal field. Magnetic dipole interactions and spin exchange interactions between spins at neighboring sites can lead to broadening or narrowing of the resonance lines. Particularly strong interactions between neigh-

boring spins occur in ferromagnetic or antiferromagnetic materials, both of which can be investigated by microwave spectroscopy. *See* ELECTRON PARAMAGNETIC RESONANCE (EPR) SPECTROSCOPY; MAGNETIC RESONANCE.

Cyclotron resonance. The cyclotron resonance frequencies of electrons in solids at magnetic fields of a few thousand gauss (a few tenths of a tesla) lie within the microwave region of the spectrum. The effective mass and the cyclotron frequency of an electron in a solid are greatly modified from those of the free electron by the crystal interactions. Microwave spectroscopy has been used to map out the dependence of the effective mass on the electron momentum. *See* CYCLOTRON RESONANCE EXPERIMENTS.

Plasmas. Gaseous plasmas are strong emitters and absorbers of microwaves. The main coupling mechanisms of charged-particle motion to microwave radiation are bremsstrahlung and cyclotron radiation. Thomson scattering is also important in dense plasmas. Microwaves are completely reflected from plasmas where the electron density n is so high that the plasma resonance frequency f given by Eq. (3) ex-

$$f = 8980^{1/2} \text{ Hz} \cdot \text{cm}^{-3/2} \quad (3)$$

ceeds the microwave frequency. Electron densities and temperatures in plasmas can be determined by microwave spectroscopy. *See* PLASMA (PHYSICS).

Cosmic microwave radiation. Thermal microwaves at a temperature of about 3 K permeate all space and can be detected by sensitive microwave receivers. These microwaves are believed to be the cooled electromagnetic radiation left over from the big bang at the beginning of the universe. *See* COSMIC BACKGROUND RADIATION.

William Happer

Bibliography. G. W. Chantry, *Modern Aspects of Microwave Spectroscopy*, 1980; G. Grüner (ed.), *Millimeter and Submillimeter Wave Spectroscopy of Solids*, Berlin, 1998; C. Kittel, *Introduction to Solid State Physics*, 8th ed., 2005; E. A. Oks, *Plasma Spectroscopy: The Influence of Microwave and Laser Fields*, 1995; C. P. Poole and H. A. Farach, *Handbook of Electron Spin Resonance*, 2 vols., 1994, 1999; C. H. Townes and A. L. Schawlow, *Microwave Spectroscopy*, 1955, reprint 1975.

Microwave tube

A high-vacuum tube designed for operation in the frequency region from approximately 3000 to 300,000 megahertz. Two considerations distinguish a microwave tube from vacuum tubes used at lower frequencies: the dimensions of the tube structure in relation to the wavelength of the signal that it generates or amplifies, and the time during which the electrons interact with the microwave field.

Effect of tube geometry. In a vacuum tube the active region is where the electrons travel through the evacuated space from the cathode, through the grid, to the plate (**Fig. 1**). The circuit in which the tube

operates extends from this active region along the internal tube structure, through the enclosing vacuum-tight envelope, and onto the portions of the circuit external to the tube. In the microwave region wavelengths are in the order of centimeters; resonant circuits are in the forms of transmission lines that extend a quarter of a wavelength from the active region of the microwave tube. With such short circuit dimensions the internal tube structure constitutes an appreciable portion of the circuit.

For these reasons a microwave tube is made to form part of the resonant circuit. Leads from electrodes to external connections are short, and electrodes are parts of surfaces extending through the envelope directly to the external circuit that is often a coaxial transmission line or cavity. Design of the tube and of the circuit in which it is to operate thus become closely related. See CAVITY RESONATOR; TRANSMISSION LINES.

Effect of transit time. At frequencies well below the microwave region, the time during which any one electron travels through the active region within a tube is so short compared to the period of the signal as to be negligible. At microwaves the period of the signal is in the range of 0.001–1 nanosecond. If transit time is comparable to the period of the signal, an electron has an inappreciable net change in energy. Even if transit time is reduced to half the signal period, an electron that is in transit when the signal reverses polarity experiences little net change in energy (Fig. 2). Only if transit time is less than a quarter of the signal period do significant numbers of electrons exchange appreciable energy with the signal field.

Transit time is reduced in several ways in microwave tubes. Electrodes are closely spaced and made planar in configuration. Spacings of 0.025 mm (0.001 in.) are practical, but to be effective require that electrodes be closely parallel to each other. High interelectrode voltages also decrease transit time by their acceleration of electrons; however, the volt-

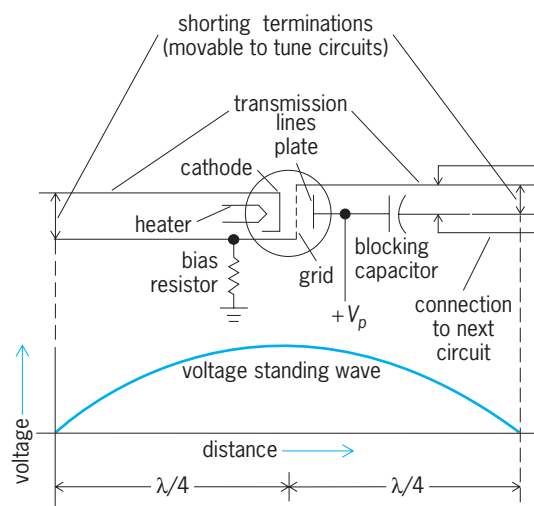


Fig. 1. Schematic of microwave triode in a circuit of two transmission lines. Usually the tube has circular symmetry and fits directly into coaxial lines.

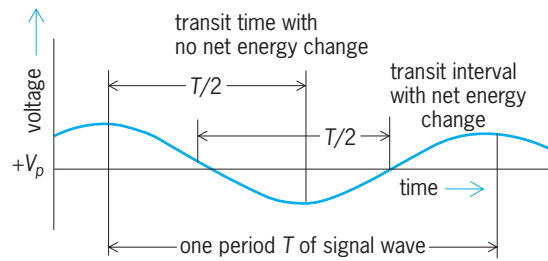


Fig. 2. Diagram of transit time. For efficient tube operation, the transit time during which an electron passes through the signal field needs to be short compared to the period of the signal.

age stresses that glass vacuum seals can withstand place a practical limit on the voltage. See VACUUM TUBE.

Alternative designs. Tubes designed by the foregoing principles are effective for wavelengths from a few meters to a few centimeters. At longer wavelengths lumped-constant circuits are effective and tubes can be designed for optimum internal characteristics. At shorter wavelengths different principles are necessary. To obtain greater exchange of energy between the electron beam and the electromagnetic field several alternative designs have proved practical.

Instead of collecting the electron beam at a plate formed by the opposite side of the resonant circuit, the beam is allowed to pass into a field-free region before reacting further with an external circuit. The electron could be deflected by a strong static magnetic field so as to revolve and thereby react several times with the signal field before reaching the plate. See KLYSTRON; MAGNETRON.

Instead of producing the field in one or several resonant circuits, the field can be supported by a distributed structure along which it moves at a velocity comparable to the velocity of electrons in the beam. The electron beam is then directed close to this structure so that beam and field interact over an extended interval of time. See TRAVELING-WAVE TUBE.

Such structures as these greatly extend the region over which useful gains, low signal-noise ratios, and significant powers can be produced, although the design of any one microwave tube can usually be optimized for only one of these characteristics at a time. Even so, dimensions and tolerances limit the wavelengths for which such tubes can be manufactured. For operation at shorter wavelengths entirely different techniques are necessary, such as those using quantum behavior within molecules or the intermodulation of signals within a nonlinear device. See MASER; MICROWAVE SOLID-STATE DEVICES; PARAMETRIC AMPLIFIER.

Frank H. Rockett

Bibliography. R. E. Collin, *Foundations for Microwave Engineering*, 2d ed., 2000; A. S. Gilmour, Jr., *Microwave Tubes*, 1986; C. A. Lee and G. C. Delman, *Microwave Devices, Circuits and Their Interaction*, 1994; S. Y. Liao, *Microwave Devices and Circuits*, 3d ed., 1990; D. M. Pozar, *Microwave Engineering*, 2d ed., 1997.

Mictacea

A proposed order of the Peracarida established for two small crustacean species, *Hirsutia bathyalis* and *Mictocaris halope* (see **illus.**). The two species share many features common to other peracaridans but differ sufficiently to justify their assignment to a distinct order with two monotypic families. Common peracaridan features include a brood pouch formed by basal lamellae of the pereopods (oöstegites) in the female; a small movable process (lacinia mobilis) on the mandible; free thoracic somites not fused to a carapace shield; a single maxilliped of typical peracarid form; and partially immobile pereopodal basal segments.

The body is slender and cylindrical, with the cephalon (head) fused to the first thoracic (trunk) somite. A carapace as such is not present, but small lateral carapace folds cover the bases of the post-mandibular appendages. Eyes appear to be completely lacking in *Hirsutia*. Eyestalks are present in *Mictocaris*, but again no trace of visual elements has been found. The first thoracic appendage is modified as a maxilliped, whereas pereopods 1–5 in *Mictocaris* or 2–6 in *Hirsutia* are biramous and provided with natatory exopods. Pereopod 7 of *Mictocaris* males is provided with a copulatory structure (males of *Hirsutia* have not been seen). Although both species possess a ventral brood pouch formed from oöstegites, those of *Mictocaris* lack the marginal setae typical of most peracaridans. The oöstegites of *Hirsutia*, while having marginal setae, differ from other peracaridans in that the oöstegites arise posteriorly from pereopod segments rather than medially. Small, one-segmented pleopods are present on the first five abdominal somites in *Mictocaris*, but the identity of similar structures on the abdomen of *Hirsutia* is questionable since no suture separating them from the body wall has been

identified. The sexes are distinct in *Mictocaris*, with eggs hatching at the typical peracaridan manca stage in which the seventh pereopod is not yet developed. *Hirsutia* is presently known from a single, presumably preparatory, female and one juvenile molt. Presumably it is also a sexually dimorphic species.

Mictocaris is a cave-dwelling species, whereas *Hirsutia* has been found only in soft muddy sediment of the deep sea. Nothing is known about the feeding habits of these crustaceans, but with its spined first pereopod and fossorial second, *Hirsutia* is thought to be carnivorous. In contrast, the feeding appendages in *Mictocaris* resemble those of thermosbaenaceans, which scrape food particles from the substrate.

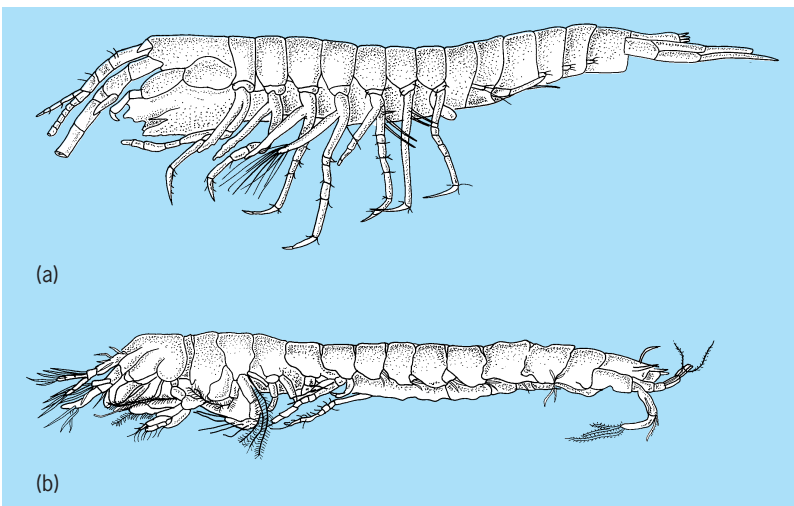
Mictacea appear to be most closely related to the Thermosbaenacea, Spelaogriphacea, and Mysidacea. See CRUSTACEA; MYSIDA; PERACARIDA; SPELAEOGRIPHACEA; THERMOSBAENACEA. Patsy A. McLaughlin

Bibliography. T. E. Bowman and T. M. Iliffe, *Mictocaris halope*, a new unusual peracaridan crustacean from marine caves on Bermuda, *J. Crust. Biol.*, 5:58–73, 1985; T. E. Bowman et al., Mictacea, a new order of Crustacea Peracarida, *J. Crust. Biol.*, 5:74–78, 1985; H. L. Sanders, R. R. Hessler, and S. P. Garner, *Hirsutia bathyalis*, a new unusual deep-sea benthic peracaridan crustacean from the tropical Atlantic, *J. Crust. Biol.*, 5:30–57, 1985.

Middle-atmosphere dynamics

The motion of that portion of the atmosphere that extends in altitude roughly from 10 to 100 km (6 to 60 mi). Interest in the middle atmosphere has been fueled by concerns over the Earth's environment, and in particular by the discovery of the ozone hole in the Antarctic lower stratosphere, together with predictions of greenhouse warming. Among the most difficult and important tasks in atmospheric science are unequivocal determinations of long-term changes in both amount of ozone and greenhouse-induced global temperature. Both are intimately related to the composition, structure, and dynamical motions in the middle atmosphere.

The Earth's climate is determined by a balance between incoming solar and outgoing Earth thermal radiative energy, both of which must necessarily pass through the middle atmosphere. The lower portion, the stratosphere, contains many greenhouse gases (ozone, water vapor, carbon dioxide, methane, nitrous oxide, chlorofluorocarbons, as well as others); and it is predicted to cool at the same time as the lower atmosphere is warmed by the greenhouse effect. The middle atmosphere is also a focus for effects of emissions from proposed commercial fleets of stratospheric aircraft. In addition to the chemistry involved, dynamical transport modeling and measurements are needed to predict the widespread transport of these important trace gases and emissions over the globe. See GREENHOUSE EFFECT; STRATOSPHERE; TERRESTRIAL RADIATION.



The two species of the order Mictacea. (a) *Hirsutia bathyalis* (after H. L. Sanders et al., *Hirsutia bathyalis*, a new unusual deep-sea benthic peracaridan crustacean from the tropical Atlantic, *J. Crust. Biol.*, 5:30–57, 1985). (b) *Mictocaris halope* (after T. E. Bowman and T. M. Iliffe, *Mictocaris halope*, a new unusual peracaridan crustacean from marine caves on Bermuda, *J. Crust. Biol.*, 5:58–73, 1985).

Middle-atmosphere structure. The stratosphere comprises the lower part of the middle atmosphere, from about 10 to 50 km (6 to 30 mi) altitude; from about 50 to 80 km (30 to 48 mi) or so lies the mesosphere. The location of the base of the stratosphere (called the tropopause) depends on meteorological conditions, varying on average from about 10 km (6 mi) in altitude at the poles to about 16 km (10 mi) at the Equator.

Actually, the middle atmosphere is not so far removed from everyday experience. The flat-topped anvils of thunderstorms are formed when rapidly ascending air flattens out like a pancake against the bottom of the stratosphere; further vertical penetration is inhibited by negative buoyancy effects due to the rising temperature with height in the stratosphere. A second example is experienced by jet aircraft travelers on middle-latitude or polar routes that often are flown near or just below the tropopause; a traveler looking up into the darker sky from the aircraft window is peering into the stratosphere. A third example involves eruptions of volcanoes. Often after a major volcanic explosion, a beautiful twilight purple glow can be seen a few minutes after sunset, caused by dust ejected into the stratosphere perhaps half a world away, and transported by strong winds in the stratosphere. Except occasionally in winter polar conditions, precipitation does not occur in the stratosphere, and the tiny dust particles many remain there for a year or more.

In the lowest atmospheric layer (called the troposphere) where rain and storms occur, temperature usually decreases with height. The stratosphere is generally a very stable layer in which temperature increases with height. In equatorial regions the tropopause is characterized by, and usually defined by, a narrow altitude range over which temperature stops decreasing with height and begins to increase with height. This definition is not as useful at high latitudes because of different temperature structure. A more general, dynamical definition of the interface between lower and middle atmosphere is in terms of a quantity called potential vorticity, which has large values in the stratosphere but small values in the lower atmosphere. In many cases of interest, the potential vorticity is also an approximately conserved quantity, making it useful for tracing air motions of air parcels.

Ozone. About 90% of the Earth's ozone shield lies in the stratosphere. Ozone is produced mainly at low latitudes (near the Equator) by solar ultraviolet radiation; however, larger concentrations of ozone are found at middle and polar latitudes (Fig. 1). The resolution of this apparent anomaly lies in transport from the equatorial source to the polar regions by dynamical processes in the stratosphere. For example, wave-motion-induced transport moves significant amounts of stratospheric ozone to the Antarctic polar region; furthermore, the wave-induced stratospheric conveyance transports the ozone-destroying (mainly chlorine) compounds from low latitudes (where they enter the stratosphere from below) to the polar regions. Without such transport,

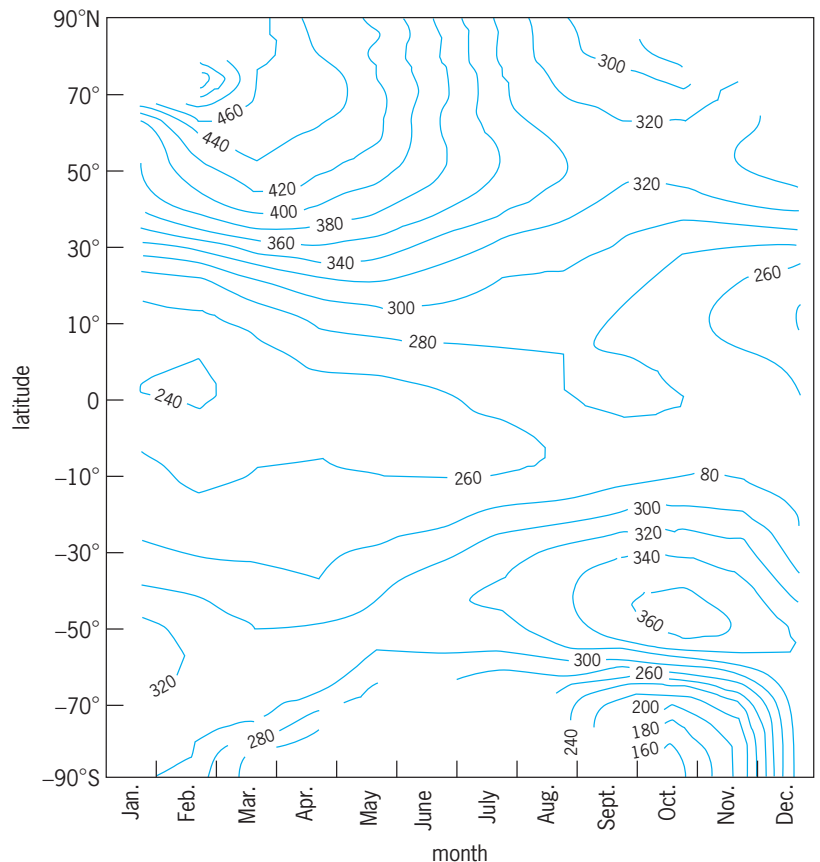


Fig. 1. Time-latitude section showing the seasonal variation of total ozone in Dobson units, based on the Total Ozone Mapping Spectrometer (TOMS) satellite instrument data for 1987. The maxima in middle to high latitudes of each hemisphere's respective spring are due primarily to dynamical transport from the equatorial source region. The very low values of ozone in October near 75–90°S represent the Antarctic ozone hole. (After World Meteorological Organization Global Ozone Research and Monitoring Project, Report of the International Ozone Trends Panel 1988, Rep. 18, vol. 1, 1988)

there would not be a significant Antarctic ozone hole.

Horizontal transport into or out of the polar stratosphere is greatly inhibited, however, when strong stratospheric jet winds develop into a winter vortex core centered over the pole. Shaped somewhat like a vertical cylinder, the winter vortex acts as a barrier to horizontal transport through its walls, dynamically isolating ozone and compounds relevant to ozone depletion. This isolation from outside influences during winter is a major ingredient in the formation of the Antarctic ozone hole, which forms at the end of the Antarctic polar night with the first sunlight in early spring. Although less continuous and less intense, a similar effect often occurs in the north polar winter stratosphere.

Understanding the dynamical and radiative physics underlying middle-atmosphere transport is thus critical for several aspects of predicting global climate change.

Middle-atmosphere waves. Atmospheric gravity waves (not to be confused with the gravity waves of general relativity) result from combined gravitational and pressure gradient forces. Typical characteristics are transverse polarization, vertical wavelengths of

0.1–10 km (0.06–6 mi), horizontal wavelengths of 1–100 km (0.6–60 mi) or more, and periods in the range of 5 min to several hours. These waves may be excited by airflow over orography (mountains) as standing lee waves, by growing clouds, and by large-scale storm complexes in the lower atmosphere; and then they propagate up into the middle atmosphere.

Planetary-scale Rossby waves are large and slowly moving waves affected by the Coriolis effect due to the Earth's rotation. They arise from the variation with latitude of the component of the Earth's rotation vector normal to the planetary surface, and they propagate against the planetary rotation (that is, westward) with periods of several days. They have vertical wavelengths of order 10 km (6 mi) and horizontal wavelengths from hundreds of kilometers up to as large as wave 1, where one wavelength fits around the Earth's circumference (wave 2 refers to two wavelengths fitting around the Earth's circumference, and so forth). Rossby waves are common at middle latitudes in winter, where they can propagate up into the middle atmosphere from excitation regions below. *See* CORIOLIS ACCELERATION.

Near the Equator, hybrid Rossby-gravity waves and also Kelvin waves (a special class of eastward-propagating internal gravity waves having no north-south velocity component) have been observed in the middle atmosphere.

A variety of global-scale normal-mode oscillations (somewhat like the vibrations of a kettledrum head) are also found in the middle atmosphere, prominent examples being wave 1, westward-moving waves with periods of about 5 and 16 days, and a wave 3, a westward moving feature with a period of about 2 days. Another observed oscillation in the middle atmosphere has been found with periods in the range

1–2 months (propagating up from the lower atmosphere).

Waves resulting from fluid dynamical instabilities are also observed: medium-scale (waves 4–7) eastward-moving waves, which are actually the tops of tropospheric storm systems, can dominate the circulation of the summer Southern Hemisphere lower stratosphere. The medium-scale waves have periods of 10–20 days. There are other possible instabilities as well. *See* DYNAMIC INSTABILITY.

Wave driving of the middle atmosphere. Perhaps the most subtle and complex middle-atmosphere phenomena are related to nonlinear effects of large-amplitude waves. An example is the mechanism responsible for the stratospheric-mesospheric jet stream. This striking feature of the middle atmosphere exhibits a nearly zonally symmetric (donut-shaped) vortex with strong winds that maximize in the upper stratosphere and lower mesosphere (Fig. 2). The stratospheric-mesospheric jet winds primarily blow from the west in the winter hemisphere and from the east in the summer hemisphere. The primary cause of the jet stream is a combination of the Equator-to-pole difference in solar heating and the Coriolis acceleration due to the Earth's daily rotation. Maximum observed winds (Fig. 2) are usually less than 100 m/s (around 200 mi/h), several times smaller than predicted by radiative transfer models without dynamics.

Radiative equilibrium model. If there were no wave dynamical effects, the structure of the Earth's middle atmosphere would be very different. A radiative equilibrium model using only incoming solar visible light and outgoing infrared radiation plus photochemistry was used to predict the atmospheric temperature field (Fig. 3). Wave dynamics were left out of this model on purpose to see what the radiative

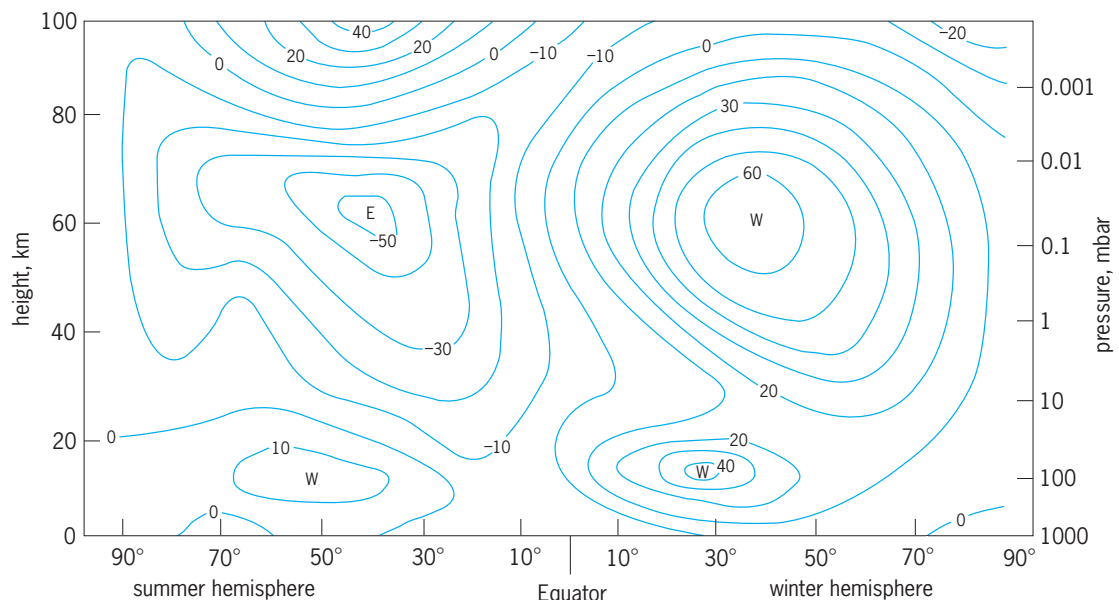


Fig. 2. Latitude-height section of typical observed zonal (east-west) mean wind speeds; contours in m/s ($1 \text{ m/s} \cong 2 \text{ mi/h}$). W and E denote centers of wind from the west and east. Negative values denote winds from the east. Prominent jets maximize in upper stratosphere and lower mesosphere, decreasing to weak values near 90–100 km (50–60 mi) altitude. (After D. G. Andrews, J. R. Holton, and C. B. Leovy, *Middle Atmosphere Dynamics*, Academic Press, 1987)

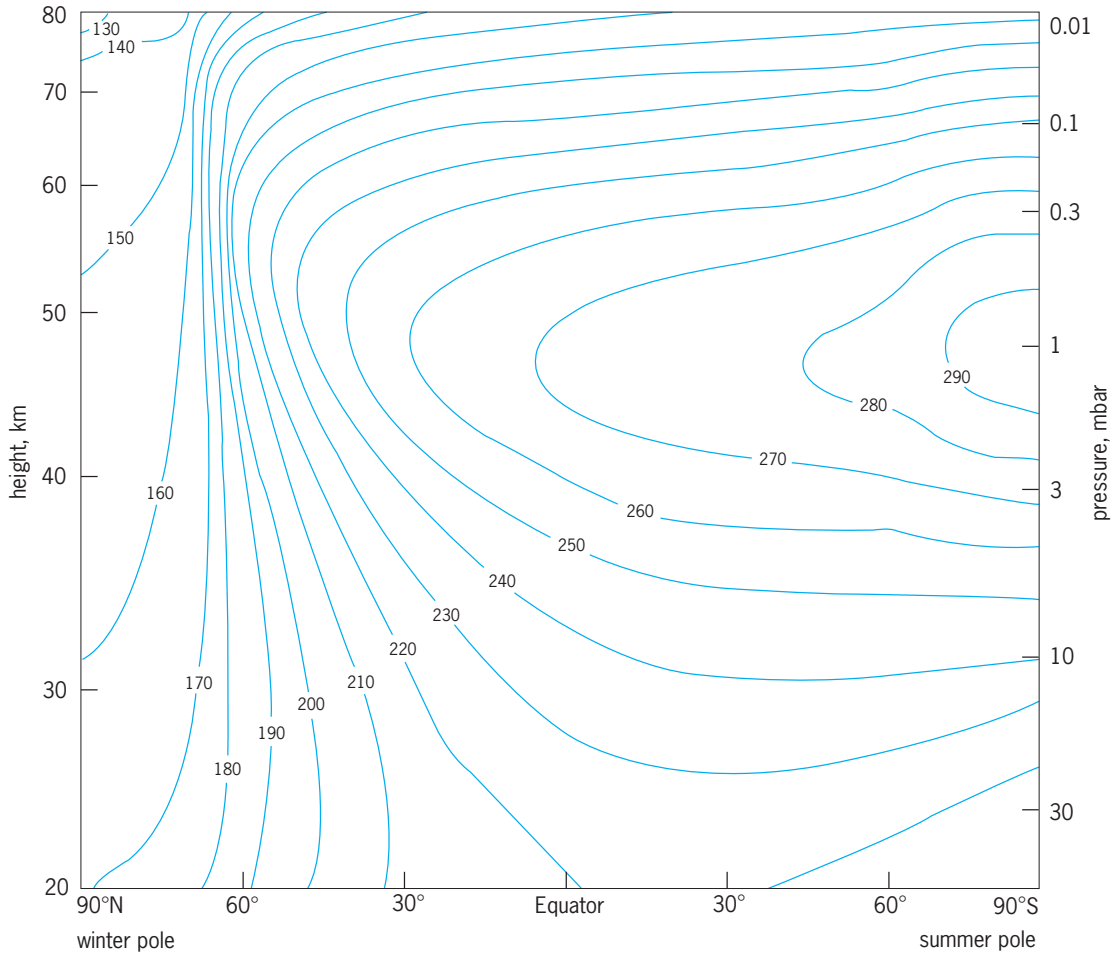


Fig. 3. Zonal mean middle-atmosphere temperatures (K) determined from a model with only radiative and photochemical effects. Vertical axis: height in kilometers (left), atmospheric pressure in millibars (right). $^{\circ}\text{F} = (\text{K} \times 1.8) - 459.67$. (After S. B. Fels, *Radiative-dynamical interactions in the middle atmosphere*, *Adv. Geophys.*, 28:277-300, 1985)

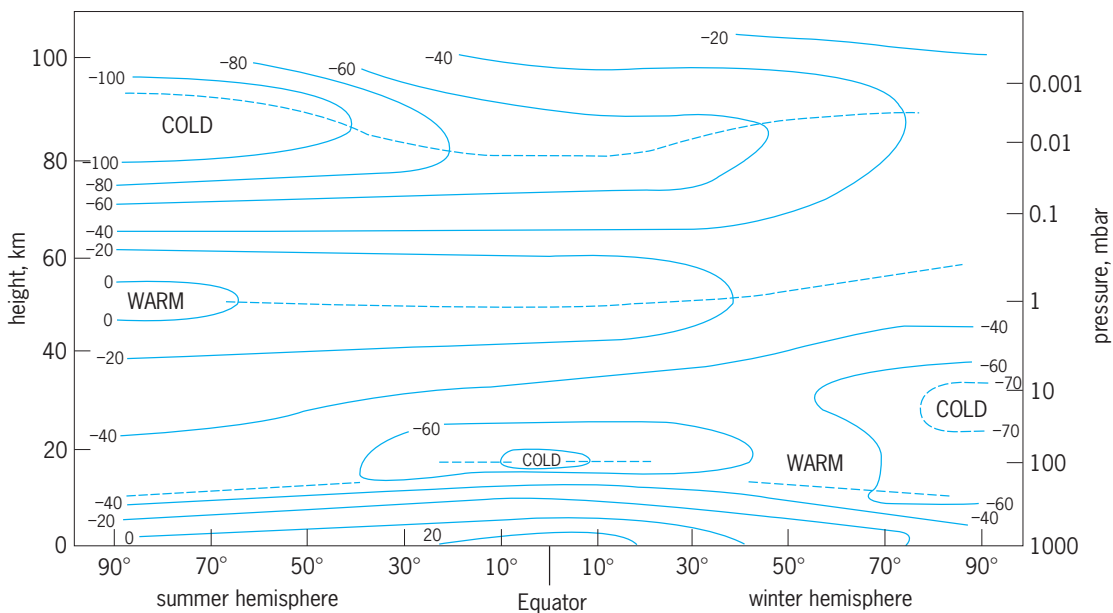


Fig. 4. Latitude versus altitude section of typical observed zonal mean temperatures for solstice conditions. Broken lines indicate, with increasing height, the tropopause, stratopause, and mesopause levels. The temperatures are in $^{\circ}\text{C}$ [$^{\circ}\text{F} = (^{\circ}\text{C} \times 1.8) + 32^{\circ}$], and winter and summer poles are reversed from those in Fig. 3. (After D. G. Andrews, J. R. Holton, and C. B. Leovy, *Middle Atmosphere Dynamics*, Academic Press, 1987)

equilibrium state would be. The radiative equilibrium state would have very cold temperatures in the sunless polar winter night, and a warm mesosphere (50–80 km or 30–50 mi altitude) in the continuous sunshine at the pole during its summer season. In contrast, observed mean atmospheric temperatures are nearly 100 K (180°F) warmer in the sunless winter mesosphere, and much colder in summer, than predicted by the model (Fig. 4).

Middle-atmosphere winds predicted with the radiative equilibrium model are also very different from observed winds (Fig. 5). The strong temperature gradients produce pressure gradients that couple with the Earth’s rotation (Coriolis torque) to produce strong jet winds. The purely radiative atmosphere would have very high jet-wind speeds of 200–300 m/s (400–600 mi/h) in the upper mesosphere in winter, several times larger than observed jet winds (Fig. 2).

Wave drag. The mechanism that “closes” the jet (preventing extremely large winds) is thought to result from vertical propagation of atmospheric gravity waves, which were not included in the radiative model. Variations in storms and the jet stream in the lower atmosphere excite the atmospheric gravity waves, which then propagate upward into the middle atmosphere. The waves grow in amplitude with increasing height because of decreasing atmospheric density. In the mesosphere the wave amplitudes become so great that they break, somewhat analogous to the breaking surf of ocean waves. When this happens, the waves’ momentum per mass (speed) is absorbed by the atmosphere at the breaking altitude. Because they are generated by slow-moving tropospheric storm systems or by airflow over fixed orography, the waves are moving much more slowly than the faster mesospheric jet. The absorption and mixing of the gravity waves causes the mesospheric

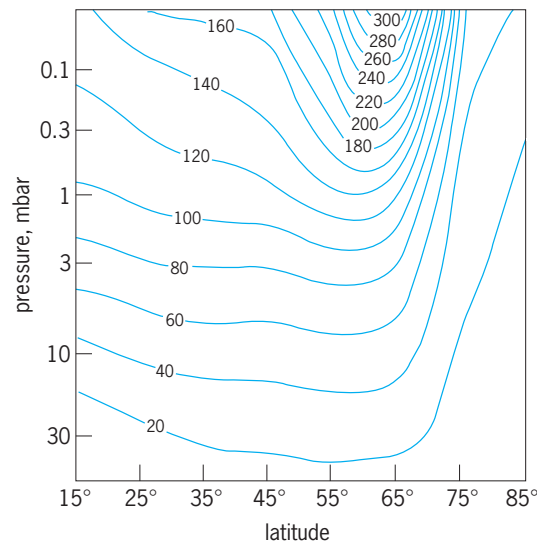


Fig. 5. Mean zonal (east-west) wind calculated from the radiative-equilibrium temperature field in Fig. 4. Contours in m/s (1 m/s \cong 2 mi/h). (After D. G. Andrews, J. R. Holton, and C. B. Leovy, *Middle Atmosphere Dynamics*, Academic Press, 1987)

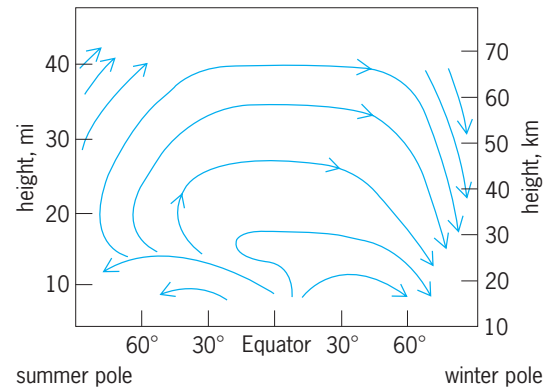


Fig. 6. Schematic flow in the mean meridional circulation of the middle atmosphere. (After T. Dunkerton, *On the mean meridional mass motions of the stratosphere and mesosphere*, *J. Atm. Sci.*, 35:2325–2333, 1978)

jet winds to be decelerated. This process, known as wave drag, acts as a dynamical brake. Without it the middle atmosphere jet would reach the very high speeds as predicted by the radiative-only model (Fig. 5).

In addition to wave drag, wave dynamics can alter the mean wind state of the atmosphere. There are a variety of wave types excited continually in the atmosphere. When these waves propagate vertically, they can encounter critical regions where the wave horizontal-phase speed matches the mean zonal background wind speed. Near such critical regions, waves are absorbed by the atmosphere, leading to dynamic forcing of the mean wind state. Studies have shown that the mean background jet wind can exhibit a seesaw pattern, increasing in strength when waves are absorbed, and decreasing when energy is drawn out of the jet to strengthen waves again. See WAVE MOTION.

Wave dynamics also induces a slow but important average meridional circulation of the atmosphere (Fig. 6). The basic structure of the meridional circulation has been confirmed, with meridional circulation at high altitudes from summer to winter hemisphere and downward transport into the upper stratosphere at the winter pole. After 1–2 months, air originally high in the summer mesosphere is transported across the Equator and descends deep into the winter polar stratosphere. Although the circulation resembles a simple hot-air-rises cold-air-sinks pattern, in reality the phenomenon requires a rather complex, nonintuitive dynamical wave-driving explanation. See JET STREAM; MESOSPHERE.

Sudden stratospheric warming. One of the most dramatic events in the stratosphere occurs about once every other year when the Northern Hemisphere winter polar stratosphere suddenly experiences dramatic warming in a few days. Such a sudden stratospheric warming may lead to a rapid reversal in direction of the stratospheric winter jet, changing it from its usual winter eastward motion (Fig. 2) to westward. Only once has a sudden stratospheric warming been observed in the Southern Hemisphere, the difference between hemispheres

apparently being related to stronger planetary-scale waves (waves 1 and 2) forced by the effects of more extensive mountain ranges and land-ocean temperature contrasts in the Northern Hemisphere.

Quasibiennial oscillation. An unusual, almost periodic, zonally symmetric, quasibiennial oscillation is found in the equatorial stratosphere, manifested as slowly descending winds at low latitudes, switching between westward and eastward directions with an irregular period averaging about 27 months.

The basic mechanism involves a subtle internal oscillation resulting from wave interactions with the mean stratospheric jet flow. Vertically propagating waves, such as gravity, Kelvin, and Rossby-gravity waves (and perhaps others), are radiatively or mechanically damped in the lower stratosphere, leading to the alternating wind regimes of the quasibiennial oscillation. Although the main idea and beauty of the fundamental principle is clear, aspects of the observations, especially why the two quasibiennial oscillation regimes are not symmetric, have yet to be elucidated.

Semiannual oscillation. Another middle-atmosphere phenomenon is the semiannual oscillation, with 6-month period. The equatorial semiannual oscillation, like the quasibiennial oscillation, is thought to involve vertical propagation and absorption of waves. Kelvin and gravity waves are thought to be important, with the latter playing the dominant role. Curiously, the semiannual oscillation winds are out of phase between the upper stratosphere and mesosphere. That is, when semiannual oscillation winds are from the west in the lower layer, they are from the east in the higher layer. More research is needed, but it is thought that this out-of-phase behavior is caused by wave absorption at critical regions in the stratosphere. The forcing waves are absorbed and thus prevented from propagating vertically, or they are allowed to do so, depending on whether the lower altitude winds are moving with or against the waves.

Waves due to instabilities. The dominant wave features in the Southern Hemisphere winter stratosphere are large scale (wave 2) and move eastward with periods of 2–3 weeks. These features are postulated to originate from instabilities that draw their energy from spatial variations in the background wind field. Instabilities have also been predicted to account for curious long-lived perturbations observed circling the pole in about 4 days in the winter middle atmosphere. *See* ATMOSPHERE; DYNAMIC METEOROLOGY; METEOROLOGICAL SATELLITES. J. L. Stanford

Bibliography. D. Andrews, A stratospheric transport system, *Phys. World*, 4(11):41–46, November, 1991; D. G. Andrews, J. R. Holton, and C. B. Leovy, *Middle Atmosphere Dynamics*, 1987; T. Dunkerton, On the mean meridional mass motions of the stratosphere and mesosphere, *J. Atm. Sci.*, 35:2325–2333, 1978; S. B. Fels, Radiative-dynamical interactions in the middle atmosphere, *Adv. Geophys.*, 28A:277–300, 1985; S. Solomon et al., Tracer transport by the diabatic circulation deduced from satellite observations, *J. Atm. Sci.*, 43:1603–1617, 1986.

Midnight sun

The phenomenon occurring when the Sun does not set, but only approaches the horizon at midnight. The effect occurs near the time of the summer solstice, on June 21, for latitudes north of the Arctic Circle. The same effect occurs near the time of the winter solstice, on December 21, for latitudes south of the Antarctic Circle. (Here “summer” and “winter” refer to the Northern Hemisphere; seasons are reversed in the Southern Hemisphere.)

Between the spring equinox and the summer solstice, in the Northern Hemisphere, the locations of sunrise and sunset move northward along the horizon. At the same time, the noontime sun climbs higher in the south each day. The highest altitude is reached at the summer solstice, after which the cycle reverses. On short summer nights, the Sun does not dip far below the northern horizon. A person who travels north along the Earth’s curved surface inevitably reaches a point where the Sun becomes visible over the pole above the northern horizon; the midnight sun is observed.

The Earth orbits the Sun on a plane called the ecliptic. The Earth’s Equator is inclined with the ecliptic by $23^{\circ}26'$. As a result, the North and South poles are in turn inclined toward the Sun for 6 months. Close to the summer solstice, on June 21, the Northern Hemisphere reaches its maximum inclination toward the Sun and the Sun illuminates all the polar area down to latitude $+66^{\circ}34'$. As seen from the polar area, the Sun does not set but only reaches its lowest altitude in the north at midnight. Latitude $+66^{\circ}34'$ defines the Arctic Circle, which is the southernmost latitude in the Northern Hemisphere where the midnight sun can be observed near the summer solstice. However, atmospheric refraction raises objects on the horizon by about $34'$, and the midnight sun can therefore be seen for a few days from locations 80 km (50 mi) south of the Arctic Circle. Observers at latitudes above the Arctic Circle see the midnight sun higher above the northern horizon, or correspondingly, see the midnight sun before the summer solstice. For example, at latitude $+70^{\circ}$, uninterrupted daylight starts 5 weeks before the summer solstice. At the North Pole, 6 month-long daylight starts around March 20, the moment of vernal equinox when the Sun travels above the Equator. *See* ECLIPTIC; METEOROLOGICAL OPTICS; REFRACTION OF WAVES.

Near the winter solstice, on December 21, the Earth reaches the opposite point in its orbit and the polar area in the Southern Hemisphere is now most inclined toward the Sun. The South Pole area is illuminated continuously, and the midnight sun phenomenon occurs south of about $-66^{\circ}34'$. *See* EARTH ROTATION AND ORBITAL MOTION; SEASONS.

Pekka Parviainen

Bibliography. G. O. Abell, D. Morrison, and S. C. Wolff, *Exploration of the Universe*, 6th ed., 1991; J. M. Pasachoff, *Astronomy: From the Earth to the Universe*, 5th ed., 1999.

Mid-Oceanic Ridge

A largely interconnected system of broad submarine rises totaling at least 60,000–80,000 km (37,000–50,000 mi) long, the precise length depending on what is included and how it is measured. Thus the Mid-Oceanic Ridge is the longest mountain range system on the planet. The origin of the Mid-Oceanic Ridge is intimately connected with plate tectonics. Wherever plates move apart sufficiently far and fast for oceanic crust to form in the void between them, a branch of the Mid-Oceanic Ridge will be created (Fig. 1). The plate boundary of the Mid-Oceanic Ridge comprises an alternation of spreading centers (or axes or accreting plate boundaries) interrupted or offset by a range of different discontinuities, the most prominent of which are transform faults. As the plates move apart, new oceanic crust is formed along the spreading axes, and the ideal transform fault zones are lines along which plates slip past each other and where oceanic crust is neither created nor destroyed. Bruce Heezen, who pioneered studies of the Mid-Oceanic Ridge, once referred to this feature as a “wound that never heals.”

The term Mid-Oceanic Ridge is somewhat a misnomer, having been applied before its tectonic significance (a system of spreading axes connected by transform faults) was appreciated. The ridge generally remains in the middle of an ocean basin only if that basin has formed between two continents

rifted apart, and the average spreading rates on each flank have been the same. Some spreading axes are located near the edges of ocean basins and behind island arcs. See PLATE TECTONICS; TRANSFORM FAULT.

Separation of plates causes the hot upper mantle to rise along the spreading axes of the Mid-Oceanic Ridge; partial melting of this rising mantle generates magmas of basaltic composition that segregate from the mantle and rise in a narrow zone at the axis of the Mid-Oceanic Ridge to form the oceanic crust. The partially molten mantle “freezes” to the sides and bottoms of the diverging plates to form the mantle lithosphere that, together with the overlying “rind” of oceanic crust, comprises the lithospheric plate. At the axis of the Mid-Oceanic Ridge the underlying column of crust and mantle is hot and thermally expanded; this thermal expansion explains why the Mid-Oceanic Ridge is a ridge. With time, a column of crust plus mantle lithosphere cools and shrinks as it moves away from the ridge axis as part of the plate. The gentle regional slopes of the Mid-Oceanic Ridge (typically from 3 to 50 parts per thousand near the axis, and decreasing smoothly toward the flanks) therefore represent the combined effects of sea-floor spreading (divergent plate motion) and thermal contraction. The height and shape of the average Mid-Oceanic Ridge profile in meters of depth are approximated by the formula $D = 2900 + 350\sqrt{T}$, where T is the age of the crust in millions of years.

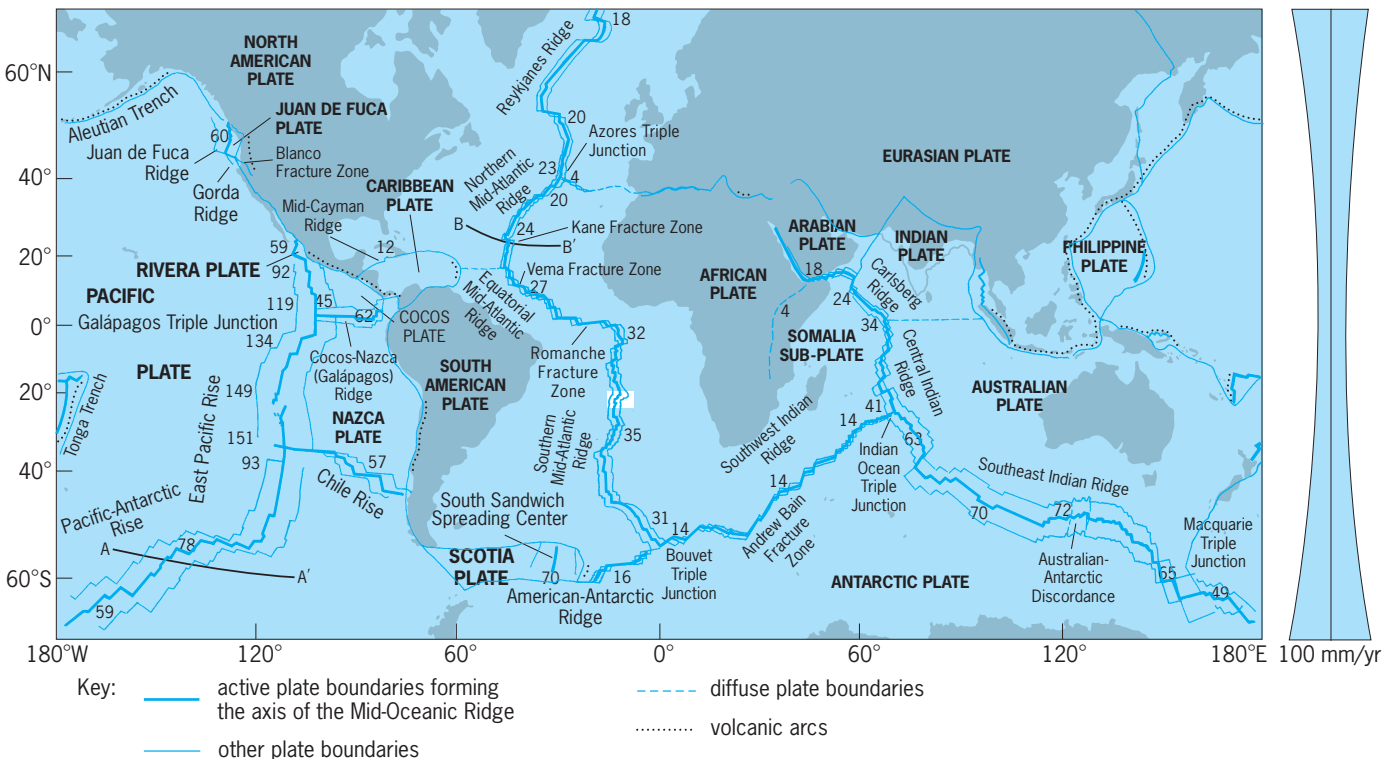


Fig. 1. Mid-Oceanic Ridge system. Paired lines on either side of the axis show the amount of crust generated in the last 10 million years at the current opening rates. Dotted lines show the volcanic arcs, which are lines of volcanoes and volcanic islands formed from magma rising from the subducted plate. The distortion caused by plotting these lines on a Mercator projection is indicated by the hourglass-shaped graph on the right, which gives the amount of crust generated at a 100 mm/yr opening rate (1 mm = 0.04 in.). Lines A-A' and B-B' show the locations of the profiles shown in Fig. 2. The white rectangle at 20°S, 12°W defines the location of the contour map of depths over a part of the axis of the ridge in Fig. 3.

For example, the depth of 4-million-year-old crust is $2900 + 350\sqrt{4}$, or 3600 m (12,000 ft). This formula predicts that the Mid-Oceanic Ridge will be bilaterally symmetric as long as spreading is symmetric (that is, crust is accreted at the same rate to both plates). See EARTH CRUST; LITHOSPHERE.

Width and related effects. The height and thermal contraction rate of the ridge crest are relatively independent of the rate of sea-floor spreading; thus, the width and regional slopes of the Mid-Oceanic Ridge depend primarily on the rate of plate separation (spreading rate). Where the plates are separating at 2 cm (0.8 in.) per year (for example, along the northern Mid-Atlantic Ridge; Fig. 1), the Mid-Oceanic Ridge has five times the regional slope but only one-fifth the width of a part of the ridge forming where the plates are separating at 10 cm (4 in.) per year (for example, the northern East Pacific Rise; Fig. 1). One consequence of the relation between the width and plate separation rate of the Mid-Oceanic Ridge is that more ocean water is displaced, thereby raising sea level, during times of globally faster plate motion. Observed long-term changes in global (eustatic) sea level probably reflect the combined effect of changing length of the total Mid-Oceanic Ridge and changing average rates of plate motion. About 80 million years ago, in the Late Cretaceous, sea levels were about 300 m (1000 ft) higher than today, primarily because of more rapid plate motion and a wider Mid-Oceanic Ridge.

The modern Mid-Oceanic Ridge is typically 1000–4000 km (600–2500 mi) wide, depending on the rate of plate separation and other factors. Actually, the ridge as a feature of thermal expansion has no sharp outer edge; the plate continues to cool and contract gradually and at ever-decreasing rates. However, the outer edge may be defined functionally as that line or zone beyond which the sea floor, deepening from the axis of the Mid-Oceanic Ridge, ceases to deepen further. Several processes can affect the location of the ridge's outer edge. Where postulated plumes of hot mantle material rise under the plates away from the axis of the Mid-Oceanic Ridge, the crust and mantle lithosphere become reelevated by as much as 1000–2000 m (3300–6600 ft). Whether formed by such a plume or not, an example of such a midplate swell is the Bermuda Rise.

Sedimentation is another process that helps give the Mid-Oceanic Ridge an outer edge. If hemipelagic sedimentation (sediment dropping down on the sea floor from the surface waters) were constant over a given part of the Mid-Oceanic Ridge, sediment thickness would increase linearly with crustal age away from the ridge axis. Although the sediment load depresses the lithospheric plate by a certain amount, the net effect of sedimentation is to make the ocean less deep. The outer edge of the Mid-Oceanic Ridge would be that line or zone where thermal subsidence of the lithospheric plate equals the shoaling effect of sedimentation. With greater distance from the Mid-Oceanic Ridge, the effect of constant sedimentation would exceed that of thermal contraction and the sea floor would rise. In some ocean basins

the more dramatic effects of turbidity (suspension) flows have overwhelmed hemipelagic sedimentation by depositing large numbers of turbidites (vertically and horizontally graded sheets of sand, silt, and clay) to form the abyssal plains between the continental margins and the Mid-Oceanic Ridge. As the abyssal plains were built up in this fashion, they simultaneously extended seaward, inundating the lower flanks of the Mid-Oceanic Ridge and displacing its outer edge. Turbidite deposition was greatly accelerated as a result of expanded Plio-Pleistocene glaciation and resultant low sea levels; and so the Mid-Oceanic Ridge, particularly in the North Atlantic, is somewhat narrower than it was prior to this glacial expansion. See BASIN; CONTINENTAL MARGIN; ISOSTASY; MARINE SEDIMENTS; TURBIDITE; TURBIDITY CURRENT.

Hot spots. Although the Mid-Oceanic Ridge exhibits little systematic depth variation along much of its length, there are several bulges (swells) of shallower sea floor. The shallow part of the Mid-Atlantic Ridge centered at Iceland is the most prominent example. Such bulges represent 500–3500-m (1600–11,000-ft) shallower-than-normal sea floor along sections of the Mid-Oceanic Ridge that are 500–3000 km (300–1900 mi) long. The features are attributed by some scientists to mantle hot spots, upwelling plumes with warmer-than-average mantle temperature. It may be that off-axis swells like the Bermuda Rise ultimately have the same origin as on-axis ones like the Iceland or Azores swells. For reasons not well understood, the sea-floor bulges are more prominent along parts of the Mid-Oceanic Ridge where the rate of plate separation (spreading rate) is slower; for example, along the northern Mid-Atlantic Ridge and the Southwest Indian Ridge. See HOT SPOTS (GEOLOGY).

Local topography. The average (or regional) depth and shape of the Mid-Oceanic Ridge is modulated by more local topographic relief, some of which equals or exceeds the regional ridge in amplitude (just as the local relief represented by Pike's Peak exceeds the regional topographic high of the Rocky Mountains) [Figs. 2 and 3]. Whereas the regional elevation of the Mid-Oceanic Ridge is determined by the thermal structure of the upper mantle and in part by average crustal thickness, the local relief is largely the result of tectonic deformation (principally normal faulting) and volcanism. The most prominent local topographic feature is the axis itself, which for opening rates less than 3.4 cm (1.3 in.) per year is a 20–40-km-wide (12–25 mi) rift valley 1000–2800 m (3300–9200 ft) deeper than the flanking rift mountains. This rift valley was first noticed on echosounding lines across the Mid-Atlantic Ridge by Marie Tharp (Lamont-Doherty Earth Observatory) in 1953. At opening rates above 9 cm (3.5 in.) per year, there is a topographic high at the axis, and at intermediate rates a rift valley tends to occur only at triple plate junctions and intersections with transform faults. Where a rift valley is present on the axis of the Mid-Oceanic Ridge, the flanks tend to exhibit rough (± 500 m or 1600 ft) topography, which is thought to originate by elongate blocks of young crust being dropped and rotated along normal faults

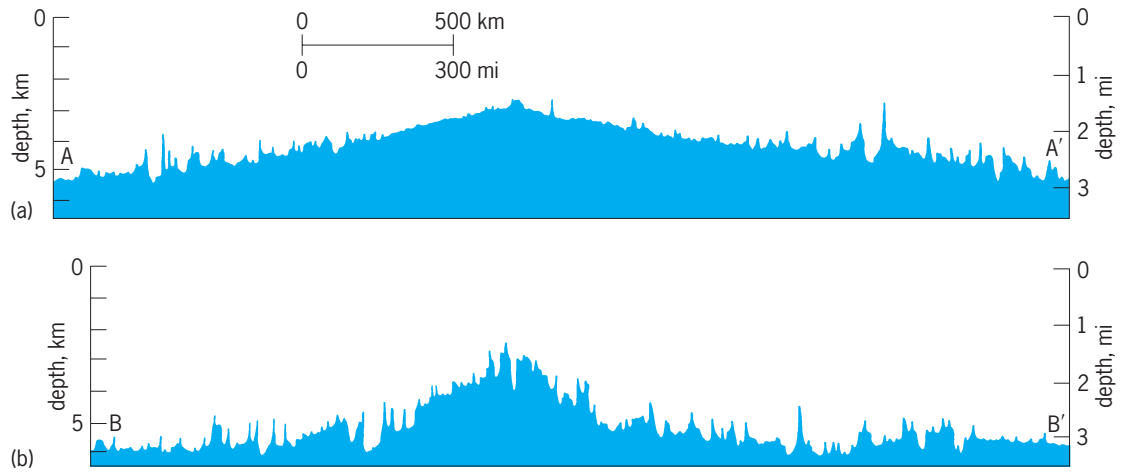


Fig. 2. Topographic profiles comparing (a) the broad, smooth, fast-spreading Pacific-Antarctic Ridge with (b) the narrow, rough, slow-spreading Mid-Atlantic Ridge. The more prominent ridges and valleys on the flanks are fracture zones (transform fault zones) that were crossed at an oblique angle. Vertical exaggeration 100:1.1 m = 3.3 ft. (After B. C. Heezen, *The deep-sea floor*, in S. K. Runcorn, ed., *Continental Drift*, Academic Press, 1962)

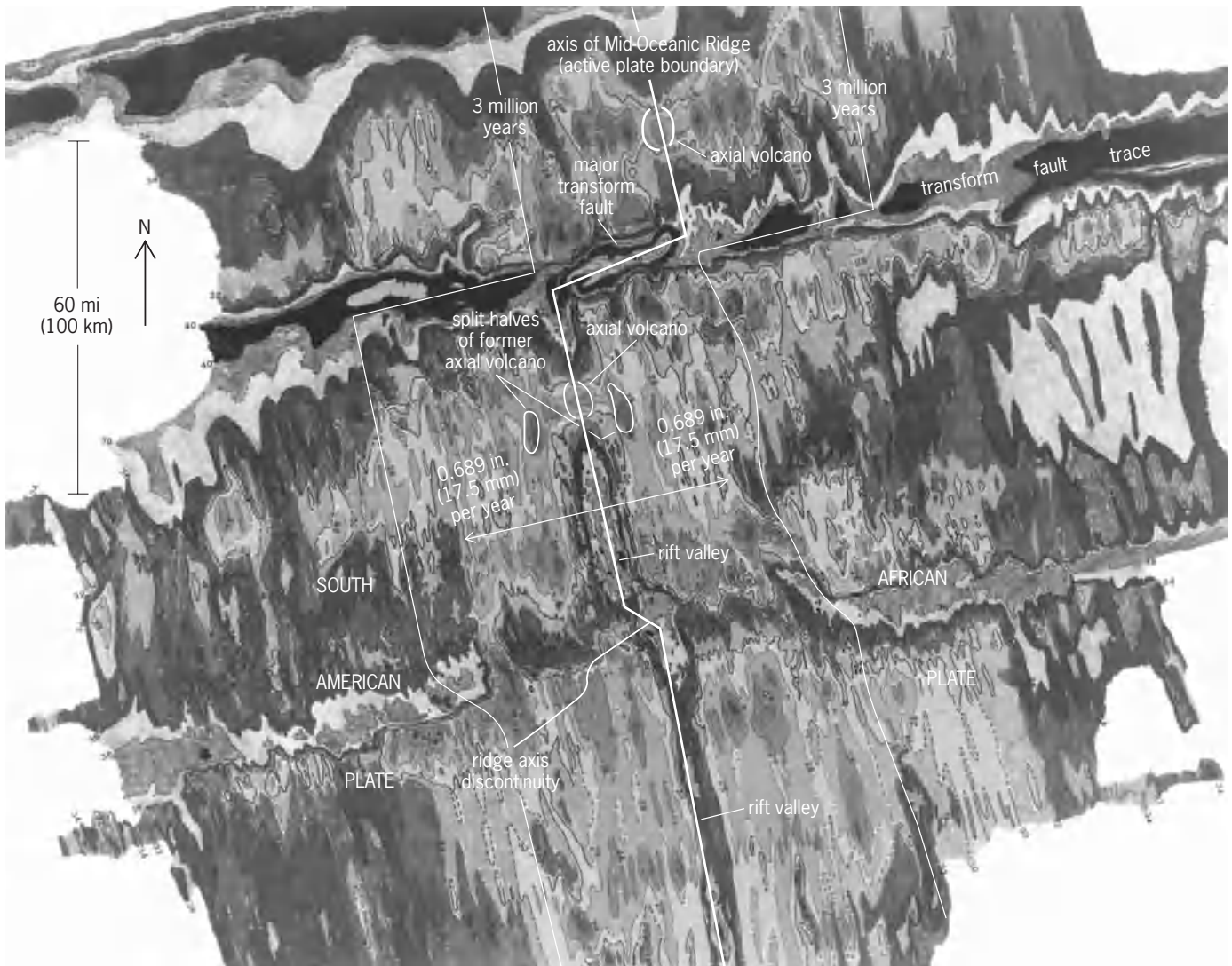


Fig. 3. Topographic (depth) contour over the crest of part of the southern Mid-Atlantic Ridge, illustrating several types of topographic features associated with the ridge. This map was the result of a joint expedition by the Naval Research Laboratory, the Lamont-Doherty Geological Observatory, and the University of Rhode Island. Depth contour intervals are 200 m with small numbers (on the perimeter) in units of 100 m (e.g., 36 = 3600 m).

dipping toward the rift axis. See FAULT AND FAULT STRUCTURES; RIFT VALLEY.

Aside from local failure of steep volcanic or tectonic escarpments—generally only on young crust near the axis of the Mid-Oceanic Ridge—and widely scattered off-axis (midplate) volcanism, the oceanic crust and its surface topography are passively carried away, slowly disappearing below the ever-thickening sediment cover. Except for the effects of sedimentation, the local topography on the flanks of the Mid-Oceanic Ridge is similar to that on the crest. Aside from the small area of oceanic crust exposed on Iceland, the topography of the Mid-Oceanic Ridge is essentially protected from significant erosion. Parts of the ridge may be exposed to deep currents up to about 1 knot (50 cm/s), but such currents suffice only to erode and redeposit sea-floor sediments, not the basaltic topography that may be exposed to the currents. See VOLCANOLOGY.

When local topographic effects are included, the maximum relief along the Mid-Oceanic Ridge axis is about 10,000 m (33,000 ft). The highest points are central volcanoes constructed on the hot-spot bulges (although generally some tens of kilometers to a hundred kilometers from the active plate boundary). Away from hot-spot bulges, the highest peaks of the Mid-Oceanic Ridge rarely breach the sea surface. More dramatic topographic relief forms near the intersections of transform faults and spreading axes (spreading centers). The rift valley deepens as it approaches its termination at a transform fault, forming a nodal basin. Transform faults are generally marked by prominent valleys, which serve as permanent traces of relative plate motion; these transverse valleys crosscutting the Mid-Oceanic Ridge are continuously generated at the site of the nodal basins. The greater the offset across a transform fault (offset being measured by the age of the old crust next to the nodal basin), the deeper the nodal basin and associated fracture zone valley.

Axial phenomena. The detailed shape of the Mid-Oceanic Ridge is not just a simple alternation of transform faults and spreading centers. Interspersed between transform faults are lesser irregularities ranging from ridge axis discontinuities (significant offset of the axis but without transverse fault development) down to minor departures from linearity. The smaller the irregularity, the more ephemeral the feature. When the offset is small, the tips of the offset spreading center may propagate past each other, forming so-called overlapping spreading centers. This is an unstable configuration and results in deformation of the crust between the overlappers. When one center continues to propagate at the expense of the offset spreading center, the result is a propagating rift. Such rifts tend to propagate away from the regional topographic bulges. Smaller-scale irregularities of the ridge axis are probably related to local volcanic centers developed several tens of kilometers to a hundred kilometers apart along the axis.

The axis of the Mid-Oceanic Ridge—that is, the active plate boundary between two separating plates—is a narrow zone only a few kilometers wide, char-

acterized by frequent earthquakes, intermittent volcanism, and scattered clusters of hydrothermal vents where seawater, percolating downward and heated by proximity to hot rock, is expelled back into the ocean at temperatures as high as 350°C (660°F). Surrounding such vents are deposits of hydrothermal minerals rich in metals, as well as exotic animal communities including, in some vent fields, giant tube-worms and clams. Sulfate-reducing bacteria, rather than photosynthesizing plants, are the base of the food chain at the vents. Life on Earth may have originated in hydrothermal vent environments similar to those presently found along the axis of the Mid-Oceanic Ridge. See HYDROTHERMAL VENT; VOLCANO.

Extinction. If a particular section of the Mid-Oceanic Ridge becomes inactive—through cessation of sea-floor spreading—it is referred to as an extinct (or abandoned) spreading axis or center. This may happen on the scale of an entire plate boundary, such as 40 million years ago, when spreading ceased along the Mid-Labrador Sea Ridge between Greenland and North America; as a result, the entire Greenland plate became part of the North American plate. More commonly, only short sections of accreting plate boundary are abandoned as a section of boundary jumps to a new location inside preexisting crust. Propagating rifts create a special category of extinct spreading centers known as failed rifts. As rifts propagate continuously, the extinct spreading center is abandoned continuously and therefore varies in age along its length.

After a section of Mid-Oceanic Ridge becomes extinct, its regional topographic high gradually decays as a result of cooling and thermal contraction. For example, the axis of a ridge that died 36 million years ago is no longer 2900 m (9500 ft) deep, but has subsided to a depth of $2900 + 350\sqrt{36}$, or 5000 m (16,400 ft), and thus is barely recognizable as a ridge. However, the shorter-wavelength topography does not decay, and the site of an extinct ridge will generally still be recognizable by its median rift valley, symmetric magnetic anomaly pattern, and other features.

Deep structure. The properties of the crust and upper mantle below the Mid-Oceanic Ridge have been inferred mostly from seismic waves generated by earthquakes and by artificial sources such as explosions and air guns. By contrast, the deepest oceanic borehole penetration (on the flanks of the Galápagos Ridge) is so far only about 1850 m (6100 ft); on Iceland the deepest penetration is 2820 m (9250 ft), 4300 m (14,000 ft) below the original uneroded land surface. None of these drill holes have penetrated to the Moho (seismic M-discontinuity), the boundary between the crust and upper mantle. However, it is believed that rock assemblages known as ophiolites, exposed in ancient mountain belts, represent remnants of former oceanic crust and uppermost mantle that are tectonically separated from the main down-going slab and thrust into the upper continental crust. At least in terms of physical properties such as seismic-wave velocity, the ophiolites

resemble the modern oceanic crust as determined by geophysical measurements. See MOHO (MOHOROVICIC DISCONTINUITY); OPHIOLITE.

Although magma (molten rock) erupts occasionally onto the sea floor at the axis of the Mid-Oceanic Ridge, seismic methods have failed to detect a continuous crustal magma chamber, at least below slow-spreading ridges. However, this hot mantle is partially molten at depths of between about 10 and 50 km (6 and 30 mi) below the axis of the Mid-Oceanic Ridge. Seismic tomography indicates that the anomalously warm, rising mantle below the Mid-Oceanic Ridge probably extends to depths of several hundred kilometers. See MAGMA; MARINE GEOLOGY.

Peter R. Vogt

Bibliography. J. Bird (ed.), *Plate Tectonics*, 2d ed., American Geophysical Union, 1980; J. R. Cann (ed.) et al., *Mid-Ocean Ridges: Dynamics of Processes Associated with the Creation of New Oceanic Crust*, Cambridge University Press, 1999; K. C. Condie, *Plate Tectonics*, 4th ed., Butterworth-Heinemann, 1997; A. Cox and R. B. Hart, *Plate Tectonics*, Blackwell Science, 1986; J. Kennett, *Marine Geology*, Prentice Hall, 1982; P. Vogt and B. Tucholke (eds.), *The Western North Atlantic Region*, Geological Society of America, 1986.

Migmatite

Rocks originally defined as of hybrid character due to intimate mixing of older rocks (schist and gneiss) with granitic magma. Now most plutonic rocks of mixed appearance, regardless of how the granitic phase formed, are called migmatites. Commonly they appear as veined gneisses. See GNEISS.

Several modes of origin have been proposed. (1) Granitic magma may be intercalated between thin layers of schist (lit-par-lit injection) to form a banded rock called injection gneiss. (2) The granitic magma may form in place by selective melting of the rock components. (3) The granitic layers may develop by metamorphic differentiation (redistribution of minerals in solid rock by recrystallization). (4) The granitic layers may represent selectively replaced or metasomatized portions of the rock.

Veined gneisses include two genetic types: arterites, in which the vein material was injected, and veinites, in which the vein material was secreted from the rock itself. There are many other types of migmatites. Some consist of blocklike masses of various shapes enclosed in granitic rock and resembling fragments of a breccia.

Migmatites are found in zones marginal to intrusive granite and in deep zones of ultrametamorphism. See METAMORPHISM; METASOMATISM.

Carleton A. Chapman

Migratory behavior

Regularly occurring, oriented seasonal movements of individuals of many animal species. The term migration is used to refer to a diversity of animal

movements, ranging from short-distance dispersal and one-way migration to round-trip migrations occurring on time scales from hours (the vertical movements of aquatic plankton) to years (the return of salmon to their natal streams following several years and thousands of miles of travel in the open sea). This article will concentrate on seasonal movements, usually between breeding and nonbreeding areas, that have clear directionality and are long-distance relative to the sizes and locomotory abilities of the species involved.

Evolution of migration. Organisms that live in highly seasonal environments have two basic options: they may become inactive or dormant for part of the year, or they may move elsewhere. Animals which had evolved means of rapid, long-distance locomotion were preadapted to take advantage of the latter strategy. There may often be significant advantages for an individual that can occupy environments that are available only part of the time. Food and other resources often occur in pulses of great abundance, and at high latitudes summer days are long, giving more time for feeding and caring for the young. In other species such as salmon (*Oncorhynchus*) and eels (*Anguilla*), migrations between fresh and salt water enable the animals to exploit vastly different environments at different stages in their life cycles. In its incipient stages, natural selection should favor the evolution of migratory behavior in a population if those individuals that move seasonally to a new area succeed in producing more offspring than their sedentary counterparts. Beyond this generalization, it has been difficult to reconstruct the details of the evolution of migratory behavior. It is obvious that Pleistocene glaciations had major influences on the patterns of long-distance bird migrations in northern latitudes, but there is no reason to think that migration did not have its origins long before that time. In fact, migratory behavior has likely evolved independently many times, even within the birds; indeed, it is continually appearing and disappearing in present-day populations.

Spectrum of migration. Migratory behavior is amazing in both the breadth of its occurrence in the animal kingdom and the magnitude of the journeys performed by individual species. The feather-light monarch butterfly (*Danaus plexippus*) departs from Canada and the northern United States in early autumn on a 1550-mi (2500-km) flight to Mexican wintering sites used year after year by individuals which have never been there before and which will never return. Some that survive the winter will initiate a return movement in the spring, often crossing the Gulf of Mexico. On the sea floor of the Bahama Bank in autumn, long queues of spiny lobsters (*Panulirus argus*) march southward to deeper water. Green sea turtle (*Chelonia*) hatchlings emerge from their beach nests on Ascension Island in the middle of the South Atlantic Ocean and head out to sea. For 2-3 years they remain at sea, some going as far as the coast of Brazil. Then they return across 1400 mi (2250 km) of seemingly trackless ocean to deposit their own eggs on their natal beaches. For the great whales entire ocean basins may constitute a home

range. Many species regularly move hundreds of kilometers between rich feeding areas and warmer calving sites. Great herds of ungulate mammals migrate over vast distances in many parts of the world. In the Far North, caribou (*Rangifer tarandus*) file along traditional routes across hundreds of kilometers of tundra between summering and wintering grounds. Even in tropical regions such as east Africa, large ungulates, such as the wildebeest (*Connochaetus taurinus*) and zebras (*Equus*), engage in mass movements synchronized with seasonal changes in their food supply.

Animals that have escaped the confines of terrestrial living through the evolution of flight have developed the greatest diversity of migratory strategies. Among mammals, bats are the longest-distance migrants, analogous in many ways to birds. A vast array of birds literally span the globe during their annual migrations. The short-tailed shearwater (*Puffinus tenuirostris*) of the Pacific and great shearwater (*P. griseus*) of the Atlantic breed on tiny islands in the Southern Hemisphere. During the austral winter they perform great clockwise circumnavigations of their respective oceans, returning to the nesting islands with remarkable precision. Arctic terns (*Sterna paradisea*) breed northward nearly to the Arctic Circle, and twice each year fly diagonally across the Atlantic Ocean to spend the winter off the southern coast of Africa, some even rounding the Cape of Good Hope into the Indian Ocean. The round-trip distance is about 25,000 mi (40,000 km). Even more remarkable is the autumn migration of the black-poll warbler (*Dendroica striata*), a 0.42-oz (12-g) songbird that nests in stunted conifers near the tree line in northern Canada. Following the nesting season, the birds move southeastward to the Atlantic coast of maritime Canada and New England. Then, usually following the passage of a cold front, they embark on a nonstop overwater flight to the mainland of South America, a journey that may take up to 4 days.

Control of migratory behavior. Many temperate zone species, including many migrants, are known to respond physiologically to changes in day length with season (photoperiodism). For example, many north temperate organisms are triggered to come into breeding condition by the interaction between the lengthening days in spring and their biological clocks (circadian rhythms). Similar processes, acting through the endocrine system, bring animals into migratory condition. Although these events of the annual cycle seem to be phased by photoperiod, the process is not simple. Many birds, for example, migrate southward in fall and cross the Equator into lengthening days. Yet they do not respond to the increasing photoperiod by coming into spring physiological condition; by some means their systems have become refractory to the potentially stimulating day length. Evidence that has been gathered on several species of vertebrate animals, including migratory birds, suggests that many major events of the annual cycle (hibernation, fattening, molt, migratory restlessness) are controlled by an endogenous timer (circannual rhythm) and will occur spon-

taneously under a constant photoperiod. See BIOLOGICAL CLOCKS; PHOTOPERIODISM.

Mechanics of bird migration. Evidence indicates that many birds imprint on or learn some feature of their birthplace, a prerequisite for them to be able to return to that area following migration. On its first migration, a young bird appears to fly in a given direction for a programmed distance. Upon settling in a wintering area, it will also imprint on that locale and will thereafter show a strong tendency to return to specific sites at both ends of the migratory route. Although some kinds of birds regularly migrate at very high altitudes (over 20,000 ft or 6000 m), and at least the bar-headed goose (*Anser indicus*) regularly crosses the highest Himalayas, most songbirds fly at night at altitudes of less than 3000 ft (900 m) above the ground. During migration, birds usually do not fly every night. The typical pattern for a songbird is to make a flight of 200 mi (300 km) or so one night, and then rest and replenish its fat stores for 2–3 days. Larger water birds often make much longer flights, but may remain on the ground for many days. The initiation of migration by most birds is highly dependent on the ambient weather. Large migratory movements are accompanied by favorable following winds, which can often substantially increase the ground speed of slow-flying songbirds. Many large water birds (ducks, geese, swans, sandpipers, plovers, cranes) migrate in flocks, and sometimes family groups. There is thus opportunity for young, inexperienced individuals to follow or learn from older birds. Most songbirds, however, appear to be randomly dispersed in the night sky, and in some species the sexes and age classes migrate at slightly different times of year.

Orientation mechanisms. To perform regular oriented migrations, animals need some mechanism for determining and maintaining compass bearings. Animals use many environmental cues as sources of directional information, and additional cues may remain to be discovered. Work with birds has shown that species use several compasses; there are backup systems. Much work is devoted to discovering the relationships among the several usable directional cues.

Sun compass. Many species of vertebrates and invertebrates possess a time-compensated Sun compass. With such a system, the animal can determine absolute compass directions at any time of day; that is, its internal biological clock automatically compensates for the changing position of the Sun as the Earth rotates during the day. This can be proved by resetting an animal's internal clock by placing it under an artificial light/dark cycle out of phase with Sun time. When the animal is tested under the Sun, its choice of bearings will be shifted (by about 15° for each hour of clock shift, corresponding to the average rate of change in the Sun's azimuth), because it takes its clock information to be correct and thus misinterprets the azimuth position of the Sun. Interestingly, animals using a Sun compass disregard the elevation of the Sun and use only its azimuth direction. Many arthropods, fish, salamanders, and birds can perceive the plane of polarization of sunlight.

Patterns of polarized skylight are a function of the position of the Sun and can thus be used as a compass. Polarized skylight and perhaps the position of the setting Sun seem to be important in the choice of migratory directions by some night-migrating birds.

Star compass. Only birds that migrate at night have been shown to have a star compass. Unlike the Sun compass, it appears not to be linked to the internal clock. Rather, directions are determined by reference to star patterns, which seem to be learned early in life by observing the axis of stellar rotation of the night sky. Birds in migratory condition will orient in the proper stellar direction even under a stationary planetarium sky.

Magnetic compass. Evidence indicates that several insects, fish, salamanders, birds, and mammals may derive directional information from the weak magnetic field of the Earth. The phenomenon has been best studied in homing pigeons and migratory birds. On cloudy days, the homing orientation of pigeons can be disrupted by attaching magnets to them. The orientation of the nocturnal migratory restlessness, or *Zugunruhe*, of some songbirds can be predictably changed in direction by shifting an Earth-strength magnetic field generated by coils placed around their orientation cages. Birds and salamanders, at least, do not seem to sense the polarity of the magnetic field. Rather, they possess an inclination compass. The horizontal component of the field provides a north-south axis, and in the Northern Hemisphere the axial direction that is characterized by the smaller angle between the resultant magnetic field vector and gravity is taken to be north. See MAGNETIC RECEPTION (BIOLOGY).

Wind direction. Wind provides directional information to birds, but because it changes with time, it cannot yield compass information directly. Nocturnal songbird migrants generally fly with a following wind; in part this is because they select a night with favorable winds to initiate migration. Birds sometimes fly in reversed or other inappropriate directions during migration. Usually these flights are downwind, and are especially likely to occur under overcast skies, which prevent the birds from seeing the Sun and stars. Under these conditions, birds often resort to wind direction as their main orientation cue.

Moon. Three beach-dwelling amphipod crustaceans (*Talitrus*, *Orchestoidea*, and *Talorchestia*) seem to possess a time-compensated lunar compass in addition to a Sun compass.

Homing and navigation. Many kinds of animals show the ability to return to specific sites following a displacement. The phenomenon can usually be explained by familiarity with landmarks near "home" or sensory contact with the goal. For example, salamanders and toads can return to capture sites following displacements of up to a few kilometers by following odor gradients. Salmon are well known for their ability to return to their natal streams after spending several years at sea. Little is known about their orientation at sea, but they recognize the home stream

by chemical cues (olfactory) in the water. The young salmon apparently imprint on the odor of the stream in which they were hatched. When they return, they move upstream, choosing the proper tributary at each branch point based on its odor. Young salmon can be imprinted to artificial odors introduced into the hatching stream and later lured into a different stream on the return migration. The origin of the natural odors used in stream recognition remains controversial.

Only in birds can an unequivocal case be made for the existence of true navigation, that is, the ability to return to a goal from an unfamiliar locality in the absence of direct sensory contact with the goal. This process requires both a compass and the analog of a map. Homing pigeons and several seabirds can return to their lofts or nests when experimentally displaced hundreds or thousands of miles. There are two main hypotheses to explain the map component that tells them which direction is toward home. Many experiments with homing pigeons during the 1980s support the olfactory map hypotheses. By perceiving odors at their home loft and associating those odors with the directions of the winds carrying them, the pigeons gradually learn the odor environment over an extensive area. Experiments suggest that such an odor map may be useful in determining the home direction even when the pigeons are displaced hundreds of miles to completely unfamiliar places. The other hypothesis is that pigeons use gradients in one or more parameter of the Earth's magnetic field as a map. The main supporting evidence is that pigeons released at magnetic anomalies are often completely disoriented. These hypotheses are not mutually exclusive.

Kenneth P. Able

Bibliography. R. R. Baker, *Bird Navigation*, 1984; P. Berthold (ed.), *Orientation in Birds*, 1991; M. A. Rankin (ed.), *Migration: Mechanisms and Adaptive Significance*, 1985; T. H. Waterman, *Animal Navigation*, 1988.

Military aircraft

Aircraft that are designed for highly specialized military applications. Fixed-wing aircraft, rotary-wing aircraft, free-flight balloons, and blimps have all been used in both crewed and crewless flight modes for military purposes. See DRONE.

Designations of military aircraft are standardized by the U.S. Department of Defense (Table 1). A letter indicates the purpose of the aircraft, and a following number refers to a particular model. An additional letter before the letter of purpose indicates a modification of function, while a letter following the model number shows a modification of the basic aircraft's equipment. For example, a reconnaissance version of the F-4 fighter with equipment modifications is designated as an RF-4C. Basic types of military aircraft include bombers, fighters, transports, patrol aircraft, trainers, and reconnaissance and observation aircraft.

Modern military aircraft are fitted with complex

TABLE 1. Designations for United States military aircraft

| Designation | Type |
|-------------|--------------------------------|
| A | Attack |
| B | Bomber |
| C, C* | Cargo and transport |
| D* | Director for drones |
| E* | Electronic, early warning |
| F, F* | Fighter |
| H | Helicopter |
| H* | Air rescue modification |
| K* | Refueling tanker |
| O | Observation |
| OV | STOL, observation |
| P | Patrol |
| Q | Target and drone |
| R* | Reconnaissance |
| SA | Rescue amphibian |
| SR | Strategic reconnaissance |
| T | Trainer |
| U, U* | Utility, observation |
| V* | Staff administrative transport |
| W* | Weatherreconnaissance |
| X* | Experimental |
| Y* | Service test |

*Used as prefix to basic type letter.



Fig. 1. Rockwell-B-1B strategic bomber. (North American Aircraft, Rockwell International Corp.)

equipment for successful accomplishment of their various missions. Many automatic devices assist the crews in obtaining results not possible with purely human abilities. Target-finding, tracking, and weapons-firing equipment to a great degree have replaced manual gunnery, missile launch, and bomb release. Electronic countermeasures equipment automatically counters enemy air defense systems. See AIR ARMAMENT; ARMY ARMAMENT; ELECTRONIC WARFARE; NAVAL ARMAMENT.

Operational aircraft are those in active inventory. They have been fully developed and tested in a production configuration (Table 2).

An aircraft proceeds through various stages of system engineering during development. The process starts with a definition of system requirements, which leads to concept definition, preliminary design, detailed design, limited-quantity prototyping, test and evaluation, and eventually production de-

sign. When an aircraft design is released for serial production, the development phase has been completed. See AIRCRAFT DESIGN.

Bombers. These are usually characterized by relatively long range, low maneuverability, and large weapon-carrying capability. Bombers are sometimes classified by their range capabilities, such as intercontinental, medium, or short. However, the use of aerial refueling by tanker aircraft gives most bombers a global range. Bombers may be equipped to deliver conventional or nuclear weapons in day, night, or adverse weather.

B-1B. The Reagan administration revived the B-1 bomber program in October 1981, and the first production B-1B (Fig. 1) flew in October 1984. All 100 of these aircraft were produced and delivered to the Strategic Air Command (now the Air Combat Command).

The B-1B was designed to operate at speeds up to Mach 0.9 (9/10 the speed of sound) at altitudes near 200 ft (60 m) in the penetration role. One of its defensive attributes is its ability to fly at low altitudes and at high speeds. Another defensive feature is the aircraft's low radar cross section. The B-1B has 1/100

TABLE 2. Table 2. Representative performance characteristics of military aircraft

| Designation | Country | Function | First flight | Maximum takeoff weight, lb (kg) | Maximum speed, Mach number (M) or knots* (km/h) | Unrefueled radius, nmi (km) |
|-------------|---------------|---------------------------|--------------|---------------------------------|---|-----------------------------|
| B-1B | United States | Bomber | Oct. 1984 | 477,000 (216,558) | M 1.25+ | 4047 (7500) |
| TU-160 | Russia | Bomber | June 1982 | 590,000 (267,860) | M 2.0 | 3940 (7300) |
| IL-76 | Russia | AEW & C [†] | Mid 1980 | 375,000 (170,250) | 434 (804) | 1241 (2300) |
| E-3A | United States | AEW & C | Oct. 1975 | 325,000 (147,392) | 462 (856) | 600 (1112) |
| F-117A | United States | Stealth fighter | June 1981 | 35,000 (15,890) | Subsonic | 400+ (740+) |
| F/A-18 | United States | Strike fighter | Nov. 1978 | 50,000 (23,000) | M 1.8+ | 500+ (927+) |
| F-16A | United States | Fighter | Dec. 1976 | 24,000 (10,896) | M 2.0 | 540 (1000) |
| F-15E | United States | Fighter/attack | Dec. 1986 | 81,000 (36,774) | M 2.5 | 500 (925) |
| TR-1A | United States | Reconnaissance | Aug. 1981 | 40,000 (18,160) | 375+ (695+) | 1303 (2415) |
| C-5B | United States | Heavy airlift | Sept. 1985 | 837,000 (379,998) | 495 (917) | 1490 (2761) |
| AH-64 | United States | Attack helicopter | Sept. 1975 | 17,400 (7,900) | 162 (300) | 130 (240) |
| B-2 | United States | Stealth bomber | July 1989 | 350,000 (158,900) | 435 (806) | 6000+ (11,119+) |
| C-17 | United States | STOL airlift [‡] | Mid 1990 | 570,000 (258,780) | 350 (649) | 1900 (3521) |

*Knot = 1 nautical mile per hour.

[†]AEW & C = airborne early warning and control.

[‡]STOL = short takeoff and landing.



Fig. 2. Tupolev TU-160 intercontinental bomber (NATO codename Blackjack). (TASS from SOVFOTO)



Fig. 3. Overhead view of Northrop B-2 stealth bomber. (Photograph by W. G. Hartenstein; from U.S. Air Force, Northrop unveils B-2 next-generation bomber, *Aviat. Week Space Technol.*, 129(22):20-23, November 28, 1988)

the cross section of a B-52 and 1/7 that of the FB-111. The aircraft has a flight crew of four and is powered by four afterburning turbofan engines. It is capable of carrying short-range attack missiles, nuclear and conventional gravity bombs, and air-launched cruise missiles. See TURBOFAN.

Tupolev TU-160. This Russian long-range strategic bomber (NATO codename Blackjack) is one of the world's largest and heaviest bombers. Designed to carry bombs and air-launched cruise missiles, it can cruise subsonically over long distances, perform high-altitude supersonic dash, and attack by utilizing low-altitude, high-subsonic penetration maneuvers.

As with the B-1, the TU-160 has a blended wing-body design with a variable swept wing and a single vertical stabilizer (Fig. 2). It has an unrefueled combat radius of about 3940 nautical miles (7300 km) and a maximum speed of Mach 2.0.

B-2. One of the most costly and secret United States aircraft development programs has been the B-2 advanced technology bomber or stealth bomber. The B-2 is designed to attack either fixed or mobile targets and fly at high altitudes over lightly defended areas or at low altitudes when warranted by heavy defenses. The flying wing design (Fig. 3) provides both long range and high payload capacity in a vehicle of low observability. All exterior features are designed to present the smallest radar, electrooptical, and infrared signatures possible when viewed from any aspect. By carefully integrating features of low observability, the aircraft does not have to rely on the sophisticated offensive and defensive avionics systems carried by the B-1 bomber.

The B-2 carries all weapons internally to minimize drag and radar cross section. The aircraft is powered by four nonafterburning turbofan engines buried in the aircraft body on each side of the bomb bay area. The two-crew-member cockpit is equipped with multifunctional displays and conventional control sticks. The fly-by-wire flight control system is quadruple-redundant. Aircraft range at high altitude is more than 6000 nmi (11,000 km) unrefueled and more than 10,000 nmi (18,500 km) with a single refueling. The bomber is capable of operating subsonically from very low altitudes up to the 50,000-ft (15,000-m) regime. See FLIGHT CONTROLS.

Airborne early warning and control. Aircraft that provide airborne early warning and control include a variant of the Russian Ilyushin IL-76 and the United States E-3A.

Ilyushin IL-76. Increasing numbers of the airborne early warning and control variant of the Russian Ilyushin IL-76 aircraft (NATO codename Mainstay) have been produced for early warning against low-altitude penetration and for air battle management (Fig. 4). This aircraft has a rotating saucer radome mounted on top of the aft fuselage, an identification-friend-or-foe (IFF) system, comprehensive electronic



Fig. 4. Ilyushin IL-76, airborne early warning and control variant (NATO codename Mainstay). (From U.S. Department of Defense, *Soviet Military Power: An Assessment of the Threat*, 1988)

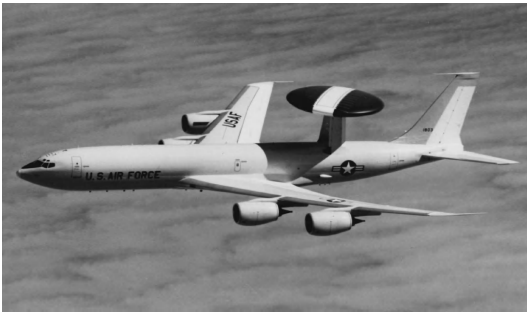


Fig. 5. Boeing Airborne Warning and Control System (AWACS) E-3A. (Boeing Aerospace Co.)

countermeasures, and a flight refueling probe. The combination of Mainstay and advanced Russian counterair fighters, such as the MIG-29, MIG-31, and Sukhoi SU-27, gives the capability to extend strategic air defense far beyond the borders of the country.

E-3A (AWACS). The U.S. Air Force Airborne Warning and Control System (AWACS), designated E-3A, is a versatile surveillance, command, and control center, designated to provide battle management in the conduct of air warfare (Fig. 5). AWACS provides full long-range surveillance over all air vehicles, crewed and crewless, at high and low altitudes, during all kinds of weather, and above all kinds of terrain. Its data-storage and processing capability provide an assessment of enemy action as well as the condition and location of available friendly resources. Its primary advantage is that it centralizes the coordination of complex, simultaneous air operations. It can command and control the total air effort: strike, air superiority, support, airlift, reconnaissance, and interdiction. In short, the airborne commander has available all the information needed to detect, assess, and counter any enemy threat.

The Air Force crew complement consists of 13 AWAC system specialists in addition to a flight crew of 4. The AWAC system is installed in a 707-320 aircraft carrying an externally mounted, 30-ft (9-m) rotodome and a powerful, concentrated payload of electronic detection and display equipment. On-board communication gear provides for reception, recording, display, transmission, and relay of a wide variety of signals, both digital and voice, and messages to and from ground and air stations.

The most important feature of AWACS is the amount of surveillance coverage supplied by the radar for command-and-control use. Target detection and other performance characteristics of AWACS are classified. However, it is known that a single AWACS can observe the entire airspace volume over several states. Even in an electronic countermeasures environment, AWACS is able to perform its mission acceptably. The AWACS radar subsystem includes some unique features such as the computerized control and signal-processing functions. Electronic systems are principally digital. The high energy levels required by the radar transmitter affect the design of aircraft power and cooling systems. Although the

radar contains approximately 100,000 components and parts, reliability is high. See AIRBORNE RADAR; RADAR.

Identification of a target aircraft as a friendly or potential threat vehicle is provided by the identification-friend-or-foe system, which uses an antenna mounted in the rotodome back-to-back with the surveillance radar antenna. This system is a highly directional, interrogate-receive, coded identification system that provides azimuth and range data on coded transponding targets.

External communication with supporting mission elements is provided by high-frequency (HF), very high-frequency (VHF), and ultrahigh-frequency (UHF) communication channels, with information transmitted as clear (unencoded) or secure voice and digital data. A high-powered UHF channel transmits commands to supporting aircraft by data link, and reply information is received on a separate UHF blade antenna. A second UHF high-powered channel is used for communication between AWACS airplanes in the same vicinity, employing voice or digital data. Medium-powered UHF channels also provide voice transmissions for terminals that are beyond the AWACS line of sight. VHF-AM (amplitude modulation) transmits in clear voice modes to friendly aircraft, and VHF-FM (frequency modulation) is used for coordination with ground forces. The AWACS flight crew has its own UHF transceiver and shares access to mission VHF-AM equipment.

Fighters. Unlike bombers and AWACS aircraft, fighters are relatively short-range, highly maneuverable, fast aircraft, designed to destroy enemy aircraft and to attack ground targets. They can carry machine guns, cannons, rockets, guided missiles, and bombs, depending upon the mission. They may be interceptor fighters, designed to shoot down enemy airplanes or missiles during day, night, or adverse weather conditions. Other fighters may be designated for close-in attack of mobile enemy ground forces to provide close support for friendly ground troops. Some fighters, called fighter-bombers, can carry conventional or nuclear weapons several hundreds of kilometers behind enemy lines to strike priority ground targets.

F-117A. This is a unique fighter aircraft in active service in the U.S. Air Force (Fig. 6). The F-117A, known as the stealth fighter, is a single-seat, twin-engine aircraft with low radar, infrared, optical, and acoustic



Fig. 6. Lockheed F-117A stealth fighter. (U.S. Air Force)



Fig. 7. Two McDonnell Douglas F/A-18 (Hornet) strike fighter aircraft, each carrying two heat-seeking Sidewinder missiles and an external fuel tank. (McDonnell Aircraft Co., McDonnell Douglas Corp.)

signatures. It first flew in 1981 and became operational in 1983. The aircraft is designed to evade enemy radar and infrared detection while conducting low-altitude strikes against high-priority targets. The F-117A is about the same size as an aircraft-carrier-based fighter and carries its weapons internally in two armament bays. The aircraft fuselage has a number of flat, vertical surfaces angled outward to deflect and diffuse radar returns. The F-117A also has radar-absorbent composite materials on its external surfaces. Engine exhaust nozzles are located on top of the wing root so that the exhaust flows over the top aft fuselage and is shielded from being viewed by infrared sensors below the aircraft. See COMPOSITE MATERIAL; RADAR-ABSORBING MATERIALS.

F/A-18. This aircraft, in service with the U.S. Navy and Marine Corps, is known as the Hornet (Fig. 7). It is called a strike fighter or swing aircraft because it can be rapidly reconfigured to fly either as a fighter or attack aircraft. Most of the conventional instrumentation in the cockpit has been replaced by three multifunction displays, an information control panel, and a head-up display in front of the pilot. Every critical switch required for air-to-air and air-to-surface engagements is either in the throttle (left hand) or on the control stick (right hand), so that the pilot is not distracted by having to move the hands to different controls. See AIRCRAFT INSTRUMENTATION.

During air-to-air combat the radar can track multiple targets, displaying up to eight target tracks while retaining up to ten in its memory. A raid-assessment mode enables the pilot to discriminate between closely spaced targets. The radar information is displayed on a clutter-free scope in either lookup or lookdown attitude; it also provides range-while-search capability, long-range search and track, and several modes for close-in combat.

A variety of survival enhancement features are designed into the F/A-18. The Hornet is powered by two low-bypass turbofan engines, which provide greater reliability than a single engine. Reticulated foam is used in the fuel tanks to suppress fires and explosions which might result from fragment penetration. The fuel tanks and fuel lines are also self-sealing. The aircraft has two completely separate hydraulic systems. It also has a quadruple, redundant, digital, fly-by-wire flight-control system with direct electrical backup to all control surfaces, and a direct mechanical backup to the stabilizers. See STABILIZER (AIRCRAFT).

F-16. The F-16 was developed to replace F-4s in the active force and to modernize the air reserve forces. Advanced techniques incorporated from the start in the single-seat F-16A and two-seat F-16B versions made them two of the most maneuverable fighters ever built. The advances include decreased structural weight through the use of composite materials, decreased drag resulting from a reduced static stability margin, fly-by-wire flight controls with a side-stick controller, a high-visibility cockpit with a 30° reclining seat and single-piece bubble canopy, blended wing-body aerodynamics with forebody strakes, and automatically variable leading-edge wing flaps (Fig. 8). The F-16 is powered by a single afterburning turbofan engine. Equipment includes a multimode radar with clutter-free look-down capability, an advanced radar warning receiver, a head-up display, internal chaff and flare dispensers, and a 500-round 20-mm internal gun. The aircraft also has provisions for electronic countermeasures equipment. The F-16 entered operational service in 1979. See STRAKE.

The F-16s has a multimode radar with increased range and advanced electronic countermeasures, and advanced cockpit displays including a wide-angle head-up display with forward-looking infrared video. Shrike antiradiation missiles and



Fig. 8. Lockheed F-16 fighter. (General Dynamics Corp.)



Fig. 9. McDonnell Douglas F-15E (Eagle) fighter, fitted with LANTIRN (Low-Altitude Navigation and Targeting Infrared for Night) pods. (McDonnell Aircraft Co., McDonnell Douglas Corp.)

multitarget Advanced Medium-Range Air-to-Air Missile (AMRAAM) compatibility have been added to the F-16Cs and Ds. System improvements were also introduced that include installation of an advanced navigation and attack system, digital flight controls, automatic terrain following, advanced identification-friend-or-foe capability, and advanced takeoff weight and maneuvering limits. An advanced radar warning receiver, improved defensive countermeasures, increased-performance engines, and full capability to launch the High-Speed Antiradiation Missile (HARM) and the Shrike missile are part of the current production aircraft.

F-16s are standard equipment with 15 units in the Air Combat Command, U.S. Air Forces Europe, and Pacific Air Forces, and are progressively replacing older aircraft in the Air Force reserves and the Air National Guard. F-16Cs also equip the U.S. Air Force's Thunderbird air demonstration squadron. Eleven hundred more have been delivered to or ordered for the air forces of 17 other countries and the U.S. Navy.

F-15E. The U.S. Air Force's primary all-weather fighter, the F-15, known as the Eagle, has been replacing the F-4 since the mid-1970s. The original single-seat F-15A and two-seat F-15B were followed in 1979 by the F-15C and the F-15D, respectively, with 2000 lb (900 kg) of additional internal fuel and provision for carrying conformal fuel tanks. Basic F-15 equipment includes a lightweight X-band pulse-Doppler radar for long-range detection and tracking of small, high-speed objects down to the tree-top level. *See DOPPLER RADAR.*

The F-15E (Fig. 9) is an advanced two-seat, dual-role, totally integrated fighter for all-weather, air-to-air deep interdiction missions. Production F-15Es also have changes in the front cockpit, which include redesigned controls, a head-up display with a wide field of view, and three multipurpose cathode-ray-tube displays. The F-15E is capable of carrying up to 24,500 lb (11,100 kg) of ordnance. The digital, triply redundant flight-control system permits coupled automatic terrain following. Navigational accuracy has been improved with the use of a ring laser gyro. For low-altitude, high-speed penetration and precision attack on tactical targets at night and in

adverse weather, the F-15E carries high-resolution APG-70 and LANTIRN (Low-Altitude Navigation and Targeting Infrared for Night) pods, with wide-field, forward-looking infrared (FLIR) imagers. *See GYROSCOPE; INFRARED IMAGING DEVICES.* Robert A. Strohl

Reconnaissance aircraft. The United States reconnaissance program provides capabilities to meet many peacetime and wartime information collection requirements. Reconnaissance resources include strategic, tactical standoff, and penetration aircraft systems that are flexible and responsive. Reconnaissance aircraft carry photographic, infrared, radar, and television sensors. These aircraft may be specially designed or may be modified from a basic fighter or bomber type. Some are equipped with special electronic gear for such purposes as submarine detection; others serve as picket planes for early warning of an enemy approach. Strategic airborne systems have included the U-2R, SR-71, RC-135, and EP-3E aircraft that carried a variety of long-range sensors used on standoff and overflight missions. Tactical aircraft standoff systems include the Air Force TR-1A (Fig. 10) and RC-135; the Army RC-12, RU-10, RU-21, EH-1, EH-60, RV-1D, and OV-1D; the Navy EP-3E, EA-3B, and EA-6B; and the Marine Corps EA-6B. The only Air Force tactical reconnaissance aircraft used in a penetrating role is the RF-4C with photo, infrared, and tactical electronic reconnaissance capabilities. The Navy uses the F-14 Tactical Air Reconnaissance Pod System (TARPS), and the Marines employ the RF-4B in a similar role.

The configuration of the TR-1A is basically that of a powered sailplane. Its unusual bicycle landing gear, combined with underwing balancer units, provides stability during takeoff, and both are then jettisoned. Range can be extended by shutting off the engine and gliding. Because of its configuration, the TR-1A requires unusually precise handling during takeoff and landing since there is an extremely small margin between approach speed and stalling speed. After touchdown the aircraft comes to rest on one of its down-turned wingtips, which serve as skids. The single-seat TR-1A is equipped with a variety of electronic sensors to provide continuously available, day or night, high-altitude, all-weather standoff surveillance. The TR-1A (Fig. 10) has the same



Fig. 10. Lockheed TR-1A tactical reconnaissance aircraft. (Lockheed Aeronautical Systems Co., Lockheed Corp.)



Fig. 11. Lockheed C-5B strategic transport aircraft. (Lockheed-Georgia Aircraft Co.)

basic airframe as the U-2R, but with the significant addition of an advanced synthetic aperture radar system (ASARS), in the form of a side-looking airborne radar (SLAR), and modern electronic countermeasures. The TR-1A is intended primarily for use in Europe, where its SLAR provides the capability to observe approximately 30 nmi (55 km) into hostile territory without the need to overfly an actual or potential battle area.

Transport aircraft. These provide dedicated logistic support to all types of military operations. Transport aircraft carry troops and war supplies. Many are adaptations of airplanes used by commercial airlines. Passenger seats or cargo space and tie-downs are provided as needed. Military transport aircraft, such as the C-5B (Fig. 11), are capable of moving large quantities of people and material rapidly to distant points. Cargo may be attached to pallets, which are easily loaded, secured for flight, and quickly unloaded for delivery. Cargo also may be discharged from flying aircraft on parachutes, eliminating the necessity for landing. The aerial tanker is a special-purpose transport aircraft. Fighters, bombers, and helicopters can refuel from tankers while in flight by means of special probe and drogue fittings. Any point on Earth can be brought within range of aerially refueled aircraft. Another special-purpose transport is a gunship. This aircraft is equipped with rapid-fire weapons for saturation attack on ground targets.

C-17. The United States Air Force has in operation a heavy-lift aircraft designated the C-17. The aircraft is the first airlifter designated from the outset to provide both inter- and intratheater airlift. A C-17, with its 172,200-lb (78,100-kg) maximum payload, can carry an M1 Abrams tank or three M2 Bradley fighting vehicles (Fig. 12). It is slightly larger in wingspan and length, and nearly 20 ft (6 m) taller, than a C-141. The C-17 is capable of airdropping outsized equipment as well as conducting parachute drops and low-altitude parachute extractions. On the ground the C-17 is able to fuel helicopters and other aircraft directly from the airlifter's wing tanks. The floor can be set up to carry either pallets or tie-down cargo. It can be completely converted in 50 min while in flight. The lower quarters of the fuselage are lined with a bulletproof,

composite material to protect troops sitting in the 54 permanently installed seats.

The C-17s are powered by four turbofan engines, each producing 40,700 lb (181 kilonewtons) of thrust. These engines are used to power many 757 civilian aircraft, and the power plants have been in commercial service since September 1987.

The airlifter's short-field landing capability is provided by an externally blown flap system. This technology was pioneered on the YC-15 prototype airlifter in the mid-1970s. The engine exhaust on the C-17 is blown through and down across huge flaps (each as big as the wing of a DC-9 jetliner), creating the effect of a much larger wing surface. This propulsive lift technology allows the C-17 to descend slowly and have a moderate landing speed of 115 knots (59 m/s).

Anotov An-225. The Soviets flight-tested a very heavy airlifter (Fig. 13) called the An-225 Mriya (Dream). The An-225 (NATO codename Cossack) has a maximum takeoff gross weight of 1,322,750 lb (600,000 kg), which is much larger than that of any other aircraft currently flying. It is powered by six turbofan engines and has a six-member flight crew. Each main landing gear has seven pairs of wheels in tandem, which gives it superior turning ability on narrow runways.

The aircraft was developed to carry cargo inside the fuselage, which has a large cross section, as well as accommodate oversized payloads externally on



Fig. 12. McDonnell Douglas C-17 transport aircraft. (Douglas Aircraft Co., McDonnell Douglas Corp.)



Fig. 13. Antonov An-225 Mriya transport aircraft, carrying the space shuttle orbiter Buran and accompanied by a fighter aircraft. (From D. North, J. Lenorovitz, and M. Dornheim, *Perestroika's changes grip Soviet aerospace industry*, *Aviat. Week Space Technol.*, 130(23):34-37, June 5, 1989)



Fig. 14. McDonnell Douglas AH-64 Apache helicopter. (Douglas Aircraft, McDonnell Douglas Corp.)

upper fuselage attach points. The Mriya played a major role in the Soviet space program by carrying loads such as the Buran space shuttle orbiter (Fig. 13) and sections of the giant Energylite rocket launch vehicle. The redesigned, twin-fin tail unit assures optimum directional control when such loads are in place. The An-225 is also designed to ferry heavy military equipment to remote and inhospitable regions.

Helicopters and VTOL aircraft. Helicopters deserve special mention as military aircraft. They are unexcelled for rescue work and for delivery of people and material to otherwise inaccessible areas. Some helicopters are armed and serve as attack aircraft, providing gun and rocket fire against ground targets. Other helicopters deliver assault troops to advanced combat areas and supply them with ammunition and other needs.

AH-64. A unique rotary-wing aircraft is the AH-64 Apache, the U.S. Army's primary attack helicopter (Fig. 14). The Apache is a quick-reacting airborne antitank weapon system. Terrain limitations and an unfavorable balance in armor dictate the need for a system that can fly quickly to the position of heaviest enemy penetration and destroy, disrupt, or delay the attack long enough for friendly armor and ground units to reach the scene. The Apache is designed to fly anywhere in the world and survive. It is equipped with a Target Acquisition Designation Sight and Pilot Night Vision Sensor (TADS/PNVS), which permits its two-member crew to navigate and attack in darkness as well as adverse weather conditions. Although the principal mission of the Apache is the destruction of enemy armor with the Hellfire missile, it is also equipped with a 30-mm chain gun and Hydra 70 rockets that are lethal against a wide variety of targets.

V-22. One of the most advanced United States vertical takeoff and landing (VTOL) aircraft in development is the V-22 Osprey tilt-rotor (Fig. 15). The aircraft is equipped with wingtip-mounted engines and three-bladed, 38-ft-diameter (11.6-m) rotors/propellers that allow it to take off and land vertically like a helicopter but to cruise at higher speeds as a fixed-wing turboprop aircraft.

The V-22 Osprey makes extensive use of corrosion-

resistant advanced composite materials, without which its capacity to perform like a helicopter and a conventional aircraft would not be possible. The aircraft also has triply redundant digital flight management and fly-by-wire flight-control systems.

The Osprey is being developed for all four armed services. The Marine Corps plans to use the aircraft for amphibious troop assault support missions. The Air Force will use its V-22s for long-range placement and pickup of special operations forces. The Navy will use the aircraft for combat search and rescue, special warfare, and fleet logistics support. The Army may order the aircraft at a later date for troop transport. See HELICOPTER; VERTICAL TAKEOFF AND LANDING (VTOL).

Research aircraft. These special-purpose aircraft are occasionally designed, assembled, and tested in order to experiment with advanced aerodynamic, structural, avionic, or propulsion concepts that must be validated before they can be applied to other aircraft designs. Research aircraft are usually well instrumented, with performance data telemetered on radio-frequency data links to ground stations located at the test ranges where they are flown. See AIRCRAFT TESTING.

X-31. An aircraft capable of maneuvering at extremely high angles of attack without loss of control would be highly advantageous in air combat. Under a joint technology program of the United States and Germany, the concept of enhanced fighter maneuverability has been realized in the form of the X-31 aircraft (Fig. 16). The aircraft's ability to perform extremely rapid three-dimensional maneuvers in the poststall regime is expected to give future fighter aircraft a decided tactical advantage in air combat. Head-on or near-head-on missile launches will dominate in the future air-to-air combat arena. For an aircraft to survive in such an environment, it will have to perform extremely agile and tight maneuvers. It must be able to point its nose swiftly in a new direction, especially in the up and down directions (as



Fig. 15. Bell Boeing V-22 Osprey tilt-rotor aircraft. (Bell Helicopter Textron)



Fig. 16. Rockwell-Messerschmidt X-31 enhanced maneuverability fighter. (North American Aircraft, Rockwell International Corp.)

opposed to side to side), to fire off a shot quickly. In many cases, this will mean moving abruptly into the poststall regime of the flight envelope. This regime is defined to include angles of attack higher than that at which the lift reaches its maximum value (about 35°).

Radius of turn is foreseen as the decisive parameter during close-in, frontal engagements. A smaller radius of turn is likely to lead to preemptive weapon launch as two aircraft are turning against each other. The smallest radius of turn is achieved at maximum lift conditions, which is very close to stall. With current design, the highest-performing contemporary fighter can turn in a radius of about 1600 ft



Fig. 17. Grumman X-29 forward-swept-wing aircraft. (From S. W. Kandebo, *Second X-29 will execute high-angle-of-attack flights*, *Aviat. Week Space Technol.*, 129(18):36-42, October 31, 1988)

(500 m). Further meaningful reductions in turn radius can probably be achieved only by penetrating the poststall regime at angles of attack much higher than those for maximum lift. Since the X-31 is intended to hold the advantage in a head-on combat engagement, it is being designed for maneuverability in this regime.

A particularly difficult challenge is posed by the design of the X-31's flight-control system for safe and reliable operation at poststall angles of attack. Beyond an aircraft's stall limits, conventional flight controls are insufficient to keep the aircraft from departing into uncontrollable flight, and some means of augmenting control power will be required. Therefore, a vectored thrust system has been integrated into the X-31's flight-control system.

X-29. This X-series research aircraft, built and tested during the 1980s (Fig. 17), was designed as a technology demonstrator for the Air Force and NASA to explore the performance potential of a forward-swept-wing design and associated technologies. The aircraft incorporates a composite wing swept forward at an angle of 29.3° . The wing is aeroelastically tailored and incorporates a thin, high lift-to-drag supercritical airfoil. The trailing edge of the wing features double-hinged control surfaces that provide variable camber. The X-29 has close-coupled canards that offer high levels of longitudinal control. The aircraft incorporates a triply redundant, digital, fly-by-wire control system with an analog backup. See AEROELASTICITY; SUPERCritical WING.

The first X-29 began flight tests in December 1984. The aircraft completed more flights than any previous X-series aircraft, providing extensive aerodynamic data on the basic forward-swept-wing configuration. It has flown up to 22.5° angle of attack, attained a maximum level speed of Mach 1.47, and reached a maximum altitude of 50,800 ft (15,500 m). The aircraft's dual design points, at which performance is optimized, are Mach 0.9 at 30,000 ft (9144 m) and Mach 1.2 at the same altitude, giving the X-29 a sustained transonic maneuvering capability. It appears that a forward-swept fighter aircraft can be constructed with 10-20% less drag and 5-25% less weight than a conventional aft-swept-wing aircraft.

Unlike the X-31A program, which will examine agility specifically in maneuvers performed at high angles of attack, the X-29 design goal is to provide high levels of maneuvering performance across a wide envelope of operation. Performance associated with the forward-swept wings is expected to permit a number of transonic maneuvering performance improvements, including a high, instantaneous rate of turn and roll control at high angles of attack. The X-29's unique three-surface control configuration (canards, wing control surfaces, and strake flaps) should also provide the aircraft with significant longitudinal control at high angles of attack.

Robert J. Strohl

Bibliography. *Jane's All the World's Aircraft*, revised periodically; U.S. Department of Defense, *United States Army Weapon Systems*, annually.

Military satellites

Artificial satellites used for a variety of military purposes. Of approximately 4000 satellites successfully launched between 1957 and 1999, about 50% have been either specifically military or usable for military purposes. Major functions of military satellites include communications, positioning and navigation, meteorology, reconnaissance and surveillance, early warning, remote sensing, geodesy, and research.

Communications. Although only certain satellites are used continuously for military purposes, all communication satellites, including commercial, may find use during conflict. All contain the necessary equipment to transmit a signal over great distances to assist in the command, control, administration, and logistic support of military forces. Military communication satellites differ from commercial satellites only in that they contain specialized components, certain capabilities, and multiple redundant systems designed to make them less vulnerable and more effective in a hostile environment.

United States satellites. Several constellations of United States military communication satellites are in operation. They may be divided into three general categories according to the frequencies at which they operate: ultrahigh frequency (UHF, 300 MHz–3 GHz), superhigh frequency (SHF, 3–30 GHz), and extremely high frequency (EHF, 30–300 GHz). See RADIO SPECTRUM ALLOCATION.

In the UHF arena are the Fleet Satellite Communication System and the Air Force Satellite Communication System. The Fleet System consists of five active satellites and several spares in geostationary orbit (22,300 mi or 35,800 km altitude), providing worldwide communications (**Fig. 1**). Each spacecraft has 23 channels operating at a variety of UHF frequencies. As these satellites wear out, they are replaced by a new constellation of UHF Follow-on (UFO) satellites with 42 channels providing narrow-band tactical communications and an antijam capability. The Air Force System provides worldwide communications for high-priority users. This system has no satellites of its own. Instead, its transponders ride as passengers on other satellites and provide a communication node independent of other systems on their host satellites.

The Defense Satellite Communication System (DSCS) operates in the SHF arena (**Fig. 2**). These satellites are designed to provide long-haul communications between both fixed and mobile stations for strategic and tactical users. Occupying geostationary positions over the Equator, the DSCS can operate on a great number of SHF channels and has independently operating transponders, a multibeam antenna (allowing the formation of a transmission beam of nearly any size and shape), limited antijam capability, and a 10-year life-span.

Milstar, the most advanced component of the United States' military satellite communication system, operates mostly in the EHF region and consists of extremely large and advanced satellites (**Fig. 3**). It was originally designed only for strategic emergency

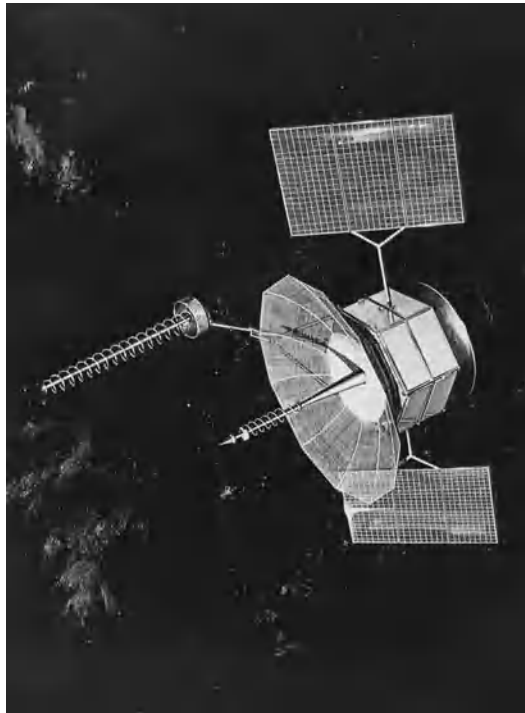


Fig. 1. Fleet Satellite Communication System satellite. (U.S. Air Force)

situations, but the emphasis now is on both strategic and tactical missions. It uses advanced techniques to achieve a high degree of reliability despite advances in electronic warfare. It is highly survivable and jam-resistant, and provides secure worldwide communications. Some advantages of EHF communication for the military user are that it is highly directional, makes better use of available bandwidth, and recovers quickly from the propagation degradation caused



Fig. 2. Defense Satellite Communications System, Phase III (DSCS III) satellite. (U.S. Air Force)

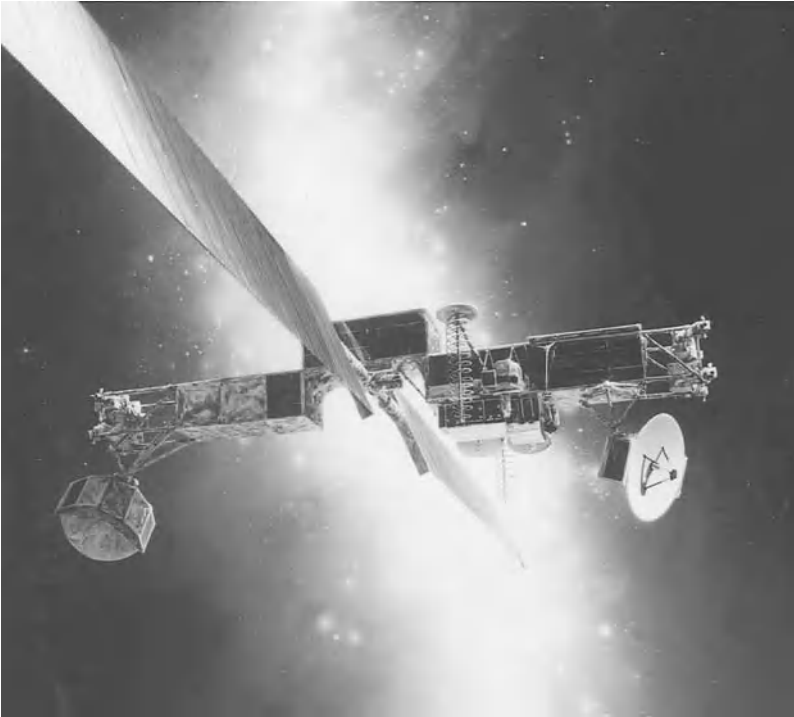


Fig. 3. Milstar satellite. (TRW)

by a high-altitude nuclear detonation. Additionally, each satellite can reconfigure its onboard equipment to make best use of multiple transponders and antennas, including spot-beam antennas. Milstar satellites can cross-link with other satellites so that reliance on intermediate ground stations is minimized. They are designed to be backward-compatible with the Fleet and Air Force Systems.

For training purposes, the U.S. Navy operates the Petite Amateur Naval Satellite (PANSAT) to provide operational communication experience for naval officers.

Russian satellites. Russia maintains a sophisticated three-tier network of satellites: small, low-altitude satellites; sophisticated *Molniya* satellites in semisynchronous (12-h) orbits; and satellites in geostationary orbit.

The *Molniya* spacecraft are notable for their innovative orbit, designed to maximize coverage over the high latitudes of the Asian landmass. They fly 12-h, highly elliptical orbits inclined at 63° with respect to the Equator so that any four can be spaced to give virtually continuous coverage over all of Asia, including areas north of 70° latitude that do not normally have satellite coverage. They repeat a ground trace each day over Russia and North America, lingering for 11 h over the Northern Hemisphere and then sweeping low and fast over the Southern Hemisphere.

In geostationary orbit are the *Raduga* spacecraft. At nearly 4500 lb (2000 kg), they provide telephone and telegraph relay and military communications.

Other countries and organizations. Skynet is a British military communication satellite system developed with the assistance of the United States. Several satellites have been launched to complete a constellation in

geostationary orbit. They are compatible with the United States' Fleet System and certain ground terminals.

China has a sophisticated satellite communication system of which some channels are devoted to military and government use. France uses a *Système de Radio Communications Utilisant un Satellite* (SYRACUSE) package on each of its TELECOM satellites to provide secure links for the French Ministry of Defense. Both Spain and Brazil (Brasilsat) have communication satellites in orbit with some channels devoted to government and military purposes. Chile has the *EASAT* satellite, which provides message relay for the Chilean Air Force. The North Atlantic Treaty Organization (NATO) has satellites positioned in geostationary orbit over the Atlantic Ocean that are used for command and control of NATO forces. *See COMMUNICATIONS SATELLITE.*

Positioning and navigation. Military forces must be able to quickly and precisely determine their position on the ground, in the air, or at sea.

GPS. The Navstar Global Positioning System (GPS) [Fig. 4] is the most recent addition to positioning and navigation satellite systems and is the most accurate and reliable satellite navigation system available. It includes 25 spacecraft in semisynchronous (12-h) orbits inclined at 55° to the Equator at 11,600 mi (18,700 km) altitude. Of these, 21 are operational and 4 are spares. The inclined orbits provide worldwide coverage, including the Poles. Control stations around the world keep GPS satellites precisely calibrated and their orbits aligned. Each GPS satellite contains an atomic clock and transmits a continuous time signal and other information to receivers on Earth. The receiver must acquire and track these signals, decode the data, and then make pseudorange and relative velocity calculations. The fully operational GPS system allows a user anywhere on Earth to receive the transmissions of at least four satellites at once. Triangulation with these satellites provides a very accurate reading of position and velocity in three dimensions. Advanced versions of GPS spacecraft offer increased redundancy, can be reprogrammed from the ground, and feature cross-link ranging to improve accuracy to less than 20 ft (6 m). GPS helps guide precision weapons such as cruise missiles and also aids search and rescue, air refueling, air combat missions, mapping, geodetic surveys, ground troop movements, and other missions.

Other systems. Russia also maintains global navigation satellite systems. Its Tsikada/Nadezhda low-Earth-orbit system functions similar to the United States' decommissioned Transit system. In addition, Russia operates the GLONASS navigation system. Similar to GPS, the system is less complex, but its satellites have proven less reliable than the United States' version. Receivers are available that will accept navigational data from either GPS or GLONASS.

The international COSPAS/SARSAT system was established in 1979 by a coalition of the United States, France, Canada, and the former Soviet Union. It is a satellite-based search and rescue system. With no satellites of its own, its sensors ride on several



Fig. 4. Global Positioning System (GPS) satellite. (U.S. Air Force)

satellites, including U.S. National Oceanographic and Atmospheric Administration (NOAA) weather satellites. See SATELLITE NAVIGATION SYSTEMS.

Meteorology. From orbit, it is possible to obtain a wide-field-of-view image of the Earth, its cloud formations, and their movements. This meteorological information is valuable for military planning and operations.

United States meteorological satellites. The *Television Infrared Observation Satellites (TIROS)* have, for decades, traveled in Sun-synchronous, low Earth orbits providing images of cloud cover, snow, ice, and the sea surface. The NOAA/TIROS system continues to operate.

The Defense Meteorological Satellite Program (DMSP) [Fig. 5] consists of several satellites in low-Earth, Sun-synchronous, polar orbits at an altitude of 517 mi (833 km). These satellites are spaced to provide complete coverage of the Earth at various times of the day and night. The Sun-synchronous orbit allows them to record images of the Earth at the same local time every day. They have some antijam and maneuver capability and view the Earth in infrared and microwave spectra providing information on clouds as well as water vapor, ozone, and temperature. Other sensors examine the near-Earth space environment. The satellites are designed to support military operations through all levels of conflict.

Broader images of an entire hemisphere for mil-

itary use may be obtained from United States' *Geostationary Operational Environmental Satellite (GOES)* spacecraft and other geostationary meteorological satellites.

European and Asian countries. Russia operates the Meteor system of satellites. As many as five satellites are in low Earth orbit similar to *TIROS* with an



Fig. 5. Defense Meteorological Satellite Program (DMSP) 5D-2 satellite. (U.S. Air Force)

orbital inclination of 81.2° . They image in the infrared spectrum.

Several other countries maintain a meteorological satellite capability that provides useful information for military operations, although not specifically designed for military use. Meteosat is a European system whose weather satellites are owned and operated by Eumetsat, a consortium of 17 nations. The satellites are positioned in geostationary orbit providing the best view of the European continent, and provide infrared, visible, and water-vapor imagery for weather forecasting, storm tracking, and data analysis.

China uses *Feng Yun*, a Sun-synchronous satellite that provides weather data for China and east Asia. The signal formats for this satellite are compatible with those of *TIROS*, so receivers designed for *TIROS* can also receive data from *Feng Yun*.

India operates *INSAT*, a multipurpose satellite in geostationary orbit. Among other capabilities, *INSAT* has weather sensors to monitor south Asia.

Japan's *Geostationary Meteorological Satellites (GMS)* operate in geostationary orbit over the Pacific Ocean. They have visible and infrared radiometers and space environment monitors. See METEOROLOGICAL SATELLITES.

Reconnaissance and surveillance. Military reconnaissance and surveillance satellites offer near-real-time unrestricted access over almost any area on Earth. Operating in many parts of the electromagnetic spectrum, they can be used to observe weapons development and deployment of forces, and to provide warning of attack by ground forces as well as targeting intelligence, technical intelligence on enemy capabilities, electronic intelligence, and bomb damage assessment. Space reconnaissance augments reconnaissance by ground and air systems, resulting in a clearer picture of the battlefield, a more complete intelligence preparation of the battlefield, and more informed strategic and tactical decision making.

Russia frequently launches photoreconnaissance and electronic intelligence satellites. France has the *HELIOS* reconnaissance satellite in low Earth, Sun-synchronous orbit, and China has the *Fanhui Shi Weixing* satellite and others in low Earth orbit for photoreconnaissance and remote sensing purposes.

Early warning. Early warning satellites provide information on missile launch and nuclear detonation that give governments time to make strategic military decisions.

The United States' missile early warning satellite system is the Defense Support Program (DSP) [Fig. 6]. Its satellites operate in geostationary orbit and survey the Earth with an infrared sensor that detects heat sources. During Operation Desert Storm in 1991, DSP satellites detected the launch of Iraqi Scud ballistic missiles.

The Nuclear Detonation Detection System (NDDS) sensors on DSP and GPS satellites are a means of detecting, locating, and reporting atmospheric nuclear detonations on a global, near-real-time basis. This system does not have any satellites of its own. Being part

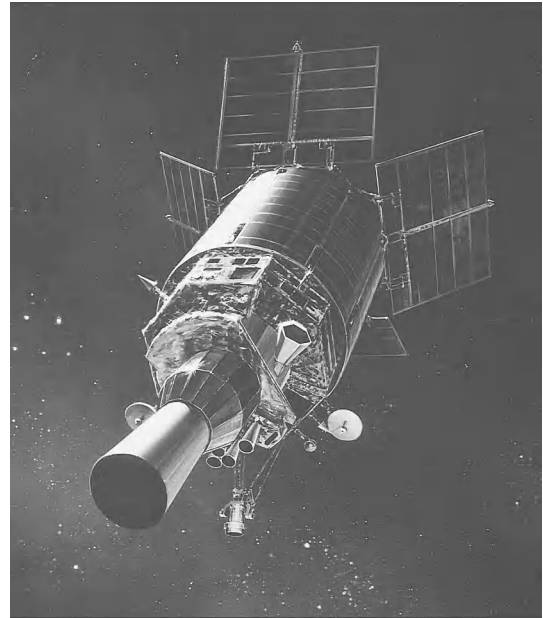


Fig. 6. Defense Support Program (DSP) satellite. (TRW)

of the GPS system allows more than one sensor to view an explosion and, by triangulation, enhances geolocation and quantification. The sensors detect the visible light, x-rays, and electromagnetic pulse given off by a nuclear explosion. See ELECTROMAGNETIC PULSE (EMP).

Russia maintains early warning satellites both in geostationary orbit and in orbital paths similar to that of the 12-h *Molniya* spacecraft.

Remote sensing. Meteorological and reconnaissance-surveillance satellites are not the only Earth-observing satellites used for military purposes. Remote-sensing satellites also afford a unique view of Earth, providing vital information to military forces. The images produced by these satellites are used to conduct route reconnaissance, analyze waterways, assist in exercise and strike planning, and provide up-to-date maps for forces deploying to unfamiliar areas.

Important information is gathered not only in the visible spectrum but in other bands of the electromagnetic spectrum. Multispectral imagery is a process of imaging the Earth at various wavelengths. These data yield information on types of minerals and vegetation as well as trafficability. Those areas that cannot be easily surveyed from the ground or by aircraft make ideal targets for remote-sensing satellites.

The United States maintains *Landsat* satellites in Sun-synchronous orbit. Each satellite carries a multispectral scanner and a thematic mapper. The scanner records images in four different spectral bands; the mapper, in seven. The best resolution is 100 ft (30 m).

Russia currently maintains the *RESURS* series of satellites in low Earth polar orbit for remote sensing photography in both visible and infrared wavelengths. Their best resolution is 7 ft (2 m). They are

designed to be maneuvered on orbit and typically have a 25-day mission life. After the photography period, a film canister separates, reenters the Earth's atmosphere, and is recovered for processing.

France owns and operates *Système Probatoire d'Observation de la Terre (SPOT)* in low Earth, Sun-synchronous orbit. Similar to *Landsat*, *SPOT* spacecraft record in fewer spectral bands but with a better spatial resolution of 33 ft (10 m). The Earth images they produce are mostly for use in agriculture, forestry, geology, and resource and environmental management. They orbit at 516 mi (832 km) and at an inclination slightly greater than that of *LANDSAT* at 98.7°.

The Japanese operate the *Marine Observation Satellite (MOS)* and *Japan Earth Resources Satellite (JERS)*. *MOS* is a multispectral imaging spacecraft and, although designed primarily for oceanic studies, its data are useful to topographers. *JERS 1* carries infrared optical sensors and a synthetic aperture radar, allowing it to see through clouds, which simplifies the remote-sensing data-gathering process.

India operates the *Indian Remote Sensing Satellites (IRS)* in a low Earth, Sun-synchronous orbit. Current versions of this spacecraft carry a panchromatic camera with 19-ft (5.8-m) resolution, a linear imaging camera for visible and infrared wavelengths at varying resolutions, and a wide-field sensor. See REMOTE SENSING.

Geodesy. Geodesy is the study of the Earth's size and shape. Geodetic data are important to the military in that the data affect position determination, navigation, map making, and a variety of other missions. Almost all satellites can be used for geodesy, provided their position in space can be accurately determined by optical or electronic means from the Earth.

GPS satellites are used for geodetic measurements as well as the *GEOS* satellites. This is a joint effort among the U.S. Department of Defense, NASA, the Department of Commerce, and other countries. In addition, the *GEOSAT* follow-on was launched for the U.S. Navy and has radar altimetry as its primary mission.

Russia uses the *Zeya* satellite in low Earth orbit for geodetic purposes. The satellite is a spheroid, slightly smaller than 3.3 ft (1 m) in diameter. It carries 20 laser reflectors for geodetic use and GLONASS and GPS receivers for precise orbit determination.

Research. In addition to the missions discussed above, there have been hundreds of military research and technology spacecraft, as well as thousands of experimental investigations for military purposes on space vehicles launched by all space-faring nations.

Examples of United States research satellites include a secondary use of the *DMSP* weather satellite to map the aurora borealis and australis with their low-light sensors and some military testing of the *Advanced Communication Technology Satellite (ACTS)*, which incorporates high-speed, high-volume satellite communications. The *Midcourse Space Experiment (MSX)* satellite tested the detection of missile signatures. The *Miniature Sensor*

Technology Initiative (MSTI) satellite tested sensors for missile defense. The *Space Technology Experiment Platform (STEP)* tested new technologies for the U.S. Air Force. *Mighty Sat* satellites, launched for the Air Force, test emerging technologies such as advanced solar cells and a composite bus. The 6000-lb (2700-kg) *Advanced Research and Global Observation Satellite (ARGOS)* provides the Air Force with a platform to collect data on the Earth's environment and to perform technology demonstrations for military space programs.

Israel uses the *OFEQ* satellite as a test bed for satellite reconnaissance and remote sensing.

Daniel F. Moorer, Jr.

Bibliography. *Interavia Spaceflight Directory*, annually; *Jane's Spaceflight Directory*, annually; TRW, *Space Log*, annually; U.S. Army Space Institute, *Space Reference Text*, 1993.

Milk

The U.S. Food and Drug Administration defines milk as the lacteal secretion, practically free from colostrum, obtained by the complete milking of one or more healthy cows and containing not less than 8.25% milk solids (not fat) and not less than 3.25% milk fat. Among mammals, humans utilize milk as a source of food. The dairy cow supplies the vast majority of milk for human consumption, particularly in the United States; however, milk from goats, water buffalo, and reindeer is also consumed in other countries. Without qualification, the general term milk refers to cow's milk.

Composition. Average composition of milk is 87.2% water, 3.7% fat, 3.5% protein, 4.9% lactose, and 0.7% ash. This average composition varies from cow to cow and breed to breed, as well as during the lactation period and the different seasons of the year, and is dependent upon the feed, nutritional level, age, and health of the animal and mammary gland. See DAIRY CATTLE PRODUCTION; LACTOSE; PROTEIN.

Nutritionists state that milk is the most nearly perfect food. Although this is true, there are some limitations, particularly in iron and vitamin C. Whole milk and skim milk are classified as excellent sources of calcium, phosphorus, and riboflavin because 10% of the daily nutritional requirement is supplied by not over 100 kilocalories (420 kilojoules). These two beverages are also classified as good sources of protein and thiamin; and whole milk is a good source of vitamin A. To be classified as good, the source must contribute 10% of a nutrient in not over 200 kcal (840 kJ). Milk is a good source of protein rich in the essential amino acids. See FOOD.

Standards for quality. The Food and Drug Administration publishes the *Grade A Milk Ordinance*, which is the basic standard for milk sanitation practice for most states and all interstate milk shippers. This ordinance defines milk and each milk product, establishes the requirements for the production of grade A raw milk, gives instructions for the inspection of dairy farms and processing facilities, and gives

sanitation requirements for the production of raw and pasteurized milk and milk products. The Food and Drug Administration certifies the sanitary quality of all raw milk shipped interstate. To accomplish this, the service inspects farms and collection depots to determine compliance to the code.

Grade A raw milk for pasteurization must be cooled immediately and maintained at 45°F (10°C) or less until processed, cannot exceed 100,000 bacteria/ml per producer, and must have no detectable antibiotic residues. Grade A pasteurized milk and milk products must be cooled immediately and maintained at 45°F (10°C) or less, contain no more than 20,000 bacteria/ml, contain no more than 10 coliforms/ml, and show a negative phosphatase test even at the pull date or expiration date. The laws of some states vary from those established by the code. Agriculture Handbook No. 51 (USDA, Agriculture Marketing Service) shows a compilation of all the federal and state standards for composition of milk products.

Fluid products such as whole milk; low-fat milk; skim milk; flavored milk; coffee, light, whipping, and heavy cream; and half-and-half are all defined by fat and solids contents. Filled (substitution of vegetable fat or milk fat) or imitation (made from nondairy ingredients) products are claiming increased sales. The lower cost of vegetable fat, and therefore lower cost to consumer, and the total lack of standards of identity are two major reasons for the increase. *See FAT AND OIL (FOOD)*.

Nearly all dairy processing equipment in use in the United States has milk contact surfaces made from stainless steel, and these surfaces legally must be sanitized before each use. The fact that stainless steel is used in construction also permits the use of strong cleaning solutions, both alkaline and acid, to properly remove residues on these surfaces. Negligible corrosion or leaching of constituents from stainless steel occurs when these detergents and cleaning compounds are used properly. After being cleaned, the entire milk contact surface is sanitized or essentially sterilized with a chemical usually containing either active chlorine or iodine to render the surface free of pathogenic and most other bacteria. *See DAIRY MACHINERY*.

Processing

Most raw milk collected at farms is pumped from calibrated and refrigerated stainless steel tanks into tank trucks for delivery to processing plants. The 10-gallon (38-liter) can that was common some years ago is found only on a few farms.

Collection and intake. The bulk truck drivers are required to check flavor, temperature, and volume of milk in the farm tank and to collect a sample of raw milk for analysis before pumping the milk into the truck. At the receiving area of the processing plant or receiving station, the milk in the farm truck is weighed and pumped into the plant through flexible plastic and stainless steel pipelines.

Separation and clarification. The actual processing of raw milk begins with either separation or clarifi-

cation. These machines are essentially similar except that in the clarifier the cream and skim milk fractions are not separated. These machines have large high-speed sealed bowls into which whole milk is introduced through ports at the bottom. In the separator milk passes upward through holes in the closely spaced conical discs. At this point the specific gravity is approximately 1.0. Milk fat (sp gr 0.93 at 68°F or 20°C) is forced to the center, but skim milk (sp gr 1.037 at 68°F or 20°C) is forced to the outside. Leukocytes, debris, some bacteria, and sediment carried with the skim milk fraction are deposited in the periphery of the bowl. This single function of a clarifier precludes sediment in homogenized milk. Whole milk or cream and skim milk travel upward to the top of the bowl. A separator contains an upper conical plate without holes to prevent admixing; the clarifier contains no such device. *See CENTRIFUGATION*.

Separators have two discharge pipes, one for cream and one for skim milk. Clarifiers have only one pipe for whole milk. Separators have a device called a cream screw by which the fat content in the cream is regulated. This screw allows more or less cream to pass out through the discharge pipe.

Many processors have units called standardizer-clarifiers which separate only a small fraction of the fat from the raw whole milk. Through manipulation of the cream screw, the amount of fat removed can be regulated. This facilitates the production of milk of standard fat content even though that in the raw product may vary. Recent modifications of separators, clarifiers, and standardizer-clarifiers permit the units to be cleaned without disassembly and the bowls to be self-cleaned during operation. *See MECHANICAL SEPARATION TECHNIQUES*.

Pasteurization. Milk is rendered free of pathogenic bacteria by pasteurization. This is accomplished in a manner so that every particle of milk is heated to a specified temperature and held at that temperature for a specified time. The U.S. Public Health Service stipulates at least 145°F (63°C) for 30 min when milk is pasteurized in a vat or at least 161°F (72°C) for 15 s when milk is pasteurized continuously. Cream and chocolate milk must be heated to either at least 150°F (66°C) for 30 min or 166°F (74°C) for 15 s. Frozen dessert mixes must be heated to at least 150°F (68°C) and held at that temperature for 30 min or 175°F (79°C) for 25 s. The greater heat treatments for the latter products are required because fat and sugar in greater concentrations than that found in whole milk provide heat resistance to bacteria normally found in milk. Most city and state codes follow these requirements.

Pasteurization requirements were originally established to provide for total destruction of *Mycobacterium tuberculosis* with a safety factor added; this bacterium will not survive 145°F (63°C) for 6 min or 155°F (68°C) for 30 s. Certain spore-forming pathogens will survive pasteurization; however, subsequent refrigeration precludes growth. These organisms, such as *Clostridium botulinum*, become hazardous only after production of toxin and growth.

The rickettsia causing Q fever in humans can be transmitted through milk, but is destroyed at 145°F (63°C) for 30 min.

Pasteurization on a batch operation requires a jacketed vat where steam or hot water can circulate and heat the milk. This temperature requires the longer times at lower temperatures to accomplish pasteurization (LTLT pasteurizer). If there is a tendency of the particular product to foam, then a space heater is utilized at the top of the vat to ascertain that every particle of product is adequately heat-treated. Modern methods of processing milk and milk products utilize the high-temperature short-time (HTST) pasteurizer. If milk passing through this pasteurizer is not at a high enough temperature, the flow diversion valve at the end of the holding tube is activated and the milk is diverted back to the surge or float tank to be reprocessed.

In Europe and to a limited amount in the United States, milk and milk products may be ultra-heat-treated (UHT). This process may use equipment similar to that used for HTST or tubular heating equipment. Aseptic equipment is first sterilized by circulating hot water at 295°F (146°C) for 30 to 45 min, then temperatures are adjusted to allow processing of milk, and packaging is done in sterile containers usually made of five laminated layers of paperboard, polyethylene, and aluminum foil. The UHT processing requires a minimum heat treatment of 280°F (138°C) for 2 s. UHT dairy foods have extended shelf life because all of the bacteria that would survive even HTST pasteurization have been destroyed, and thus the food needs no refrigeration. A typical packaging machine for UHT sterilized milk shows the need for absolute sterility.

Homogenization. Fat globules in fluid milk products are broken by homogenization into sizes that are 2 micrometers or less and thus are relatively unaffected by gravitational forces. Most fluid milk is homogenized. The flow of product into the homogenizer head is through suction valves, where it is forced by pistons through discharge valves and finally through valves that cause shearing of the fat globules. These latter two valves can be regulated at different pressures, depending on the product being homogenized. Total pressures as high as 3000 lb/in.² (21 megajoules) are used. This reflects 500–700 lb/in.² (3.4–4.8 mJ) on the second stage and 2300–2500 lb/in.² (16–17 mJ) on the first stage. In order to properly homogenize milk, the product must be heated to at least 140°F (60°C) to liquefy the fat globules and inactivate the enzyme lipase to prevent a rancid flavor. Homogenizers are commonly connected into the HTST pasteurizer where the hot milk comes from the regenerator or heater section.

Several theories explaining homogenization or globule fractionation include shearing, shattering, and cavitation. Shearing is the effect produced as whole milk is forced through a minute orifice at high speed. Shattering is the effect that occurs when whole milk under high velocity strikes a flat surface, such as the impact ring. Cavitation is the effect produced as whole milk changes abruptly from an area

of high pressure before the valve to an area of somewhat reduced pressure after the valve. All three are evident in a homogenizer.

Efficiency of homogenization can be ascertained by examining the size of the fat globules under 1000× magnification. The Food and Drug Administration specifies that the fat content of the upper 100 ml of a quart of homogenized milk that has been undisturbed for 48 h cannot differ by more than 10% from that of the remainder.

Vitamin fortification. Most milk is fortified with 400 international units (IU) of vitamin D per quart, and most low-fat and skim milk is fortified also with 2000 IU of vitamin A per quart. These vitamins as concentrates are added either by automatic dispensing with a peristaltic pump into a continuous flow of milk prior to pasteurization or as a single quantity in a batch operation. Vitamin concentrates have a potency range of 3000–200,000 IU/ml.

Nonfat dry milk can be enriched with vitamins A and D by adding the vitamin concentrates to the condensed milk just prior to drying or by blending in a dry beadlet form to the dry milk. *See* VITAMIN; VITAMIN A; VITAMIN D.

Defects

When milk and milk products are mishandled or good management practices with cows are not followed, some defects can be observed in the flavor and appearance of fluid dairy foods.

Flavor defects. There is an apparent delicate balance between the flavor constituents in milk. These are very subtle flavors, with some well below the threshold concentration where organoleptic detection is impossible. Some of the flavoring materials occurring naturally in milk serve only to potentiate other flavors. Humans vary considerably in their ability to detect flavors. Whereas milk normally has a slightly salty character as well as a sweet flavor, the degree to which these background flavors are observed organoleptically varies considerably. *See* TASTE.

Cooked or heated flavor. Prominent in sterilized milks and baby formulations is a heated flavor. To a much lesser degree this flavor can be detected in fluid milk, cream, and ice cream mixes. Conventional pasteurization generally will not cause a noticeable flavor. In fact, a slightly cooked flavor is now desired by most consumers. However, when temperatures and holding times greater than the minimums established for pasteurization are used, this defect is discerned.

Heat causes the whey proteins, β -lactoglobulin and euglobulin, to change from the normally helical or springlike structure to random structures. The energy supplied by the heat breaks the chemical bonds that hold these coils together. The result is that numerous sulfhydryl (SH) groups are exposed, which contribute to the flavor.

Prolonged exposure of milk to elevated temperatures will also cause some discoloration or browning. Most of this color change is caused by amino acid-sugar reactions or Maillard-type browning. *See* MAILLARD REACTION.

Feed flavors. Milk is a good absorbent for flavors whether the flavor itself is in the feed of the animal or in the barn air. Changes from one forage crop to another, feeding of silage within 2 h of milking, or allowing cows to feed on cabbage, onion, or other highly flavored crops or some weeds will cause objectionable flavors in raw milk. In the southern sections of the United States, deodorization of milk is mandatory to remove objectionable wild-onion flavors. This process involves flashing or injection of a thin stream of milk at pasteurization temperature into a vacuum chamber which can be included in an HTST pasteurizer between the heating chamber and the flowing diversion valve.

Oxidized and sunlight flavors. Oxidized flavor is probably the most important single flavor defect in milk and milk products. This flavor is also described by terms such as tallowy, metallic, and cardboard. Lipid material is the source, but the specified flavor compounds differ apparently among milk products.

After homogenization or some degree of churning, milk fat itself becomes exposed as the phospholipids no longer totally protect the fat globules; then the fat itself can be oxidized. Of the factors essential to the development of this flavor, atmospheric oxygen and copper are important. An adequate heat treatment is a commonly used process to retard the development of this flavor. The chemistry of this oxidation is a complex phenomenon.

When milk is exposed to light for any period of time, another defect can occur. This is sunlight or activated flavor and is attributed to a reaction with the amino acid, methionine, changing to an aldehyde, methional, in the presence of riboflavin (vitamin B₂). Milk in metal containers is resistant to this change because no light penetrates. Some years ago amber glass bottles were used to prevent this defect in homogenized milk. Glass of this type reduces the amount of energy transmitted but is not totally a preventative.

Rancid flavor. Modern automated methods of handling raw fluid milk promote the onset of this flavor. Each fat globule in milk has a surface coating composed of phospholipids (lecithin and cephalins) and proteins from the serum portion (euglobulin and some casein), among other constituents. Moreover, the enzyme lipase is also present in raw whole milk and presumably is associated with casein.

Raw whole milk is stable to the action of lipase until some physical force disrupts the membrane surrounding the fat globules. Forces such as excessive agitation through prolonged pumping in pipelines, mild heat treatment followed by cooling, or freezing can disrupt this membrane. Lipase acts on fat to cause hydrolysis of the triglycerides into glycerol and free fatty acids. Because milk contains a significant proportion of short-chain fatty acids which have strong aromas, the defect is easily detected by taste. Lipase, as with most enzymes in milk, is inactivated by proper pasteurization.

Physical defects. Fluid whole milk is a rather stable product; however, subjection to adverse conditions can cause some physical changes.

Destabilization. If milk is frozen and then thawed, the fat and protein will aggregate. This can be observed in the thawed fluid product by microscopic flakes of protein and macroscopic fat specks.

Feathering. A defect associated with coffee cream occurs principally because of a salt imbalance in the cream (calcium, magnesium, citrate, or phosphates), too high an acidity, or improper homogenization. Cream is said to feather when a grayish, flaky scum floats to the surface when the cream is added to hot coffee.

Cream plug. Cream and occasionally fluid milk will show some signs of a thick surface. The aggregation of fat that either floats on the surface of fluid milk or forms a thick mass on the surface of cream is caused by churning. Prevention of churning eliminates the defect. Therefore, warming and cooling, freezing, and excessive agitation should be avoided.

Testing for Quality

Laboratories associated with dairy processing operations, or private laboratories that analyze milk and milk products, perform many tests to determine compliance with local, state, and federal regulations; quality control; processing efficiency; and payment to farmers. The test performed in most laboratories for homogenization efficiency is described above in the discussion of homogenization. The subsequent discussion involves other commonly used tests.

Fat determinations. There are several methods used for the determination of fat in milk and its products. The Babcock test developed by S. M. Babcock in 1890 is classical. This test and the European counterpart, the Gerber test, are volumetric measurements of the quantity of fat in a sample. Sulfuric acid is used to digest the 18-g (0.63-oz) sample that was added to the special flask. The heat that develops liquefies the fat. Centrifugal force is used to bring the fat into the calibrated neck of the flask. The Babcock test is used as the basis for payment for raw milk received from farms.

Fat can also be determined by extraction from a sample using specific organic solvents and subsequent evaporation of these solvents to obtain a quantity of fatty residue which is weighed. This procedure is outlined in the Mojonnier and Roese-Gottlieb methods. The solvent extraction methods are considered by far the most accurate ever developed; however, they are time-consuming. One skilled operator can complete 30 samples per 8-h day.

An approved (Association of Official Analytical Chemists) method for fat determination in raw milk involves only the use of the Milko-Tester manufactured by Foss Electric in Denmark. This automated process consists of heating, homogenization to render fat globule size uniform, dilution, and photoelectric sensing of solution opacity. Each analysis requires less than 1 min. Frequent, daily calibration of this instrument is required. Other similar instruments use infrared adsorption to assay fat, protein, and lactose in milk. All automated instruments need calibration using pretested milk to assure accurate analysis.

Total solids or moisture. Whatever is not lost during analysis for moisture in a food product is considered as total solids. There are several different methods approved by the Association of Official Analytical Chemists as legal for the analysis of milk and milk products. Fundamentally, all involve a heat treatment applied to a sample and then vacuum desiccation or subjection of the sample to treatment in a vacuum oven for a stated period. Other less accurate ways involve the use of specialized hydrometers called lactometers which determine the specific gravity of a fluid sample.

Freezing point. The freezing point of milk is a constant at 31°F (−0.55°C) and consequently can be used for the detection of the addition of water. For this determination a cryoscope is employed which contains a bath to lower the temperature of the sample and a precise thermometer or electrical readout system to determine the temperature at which stable freezing occurs. The sample will be slightly supercooled, and when crystallization occurs, the temperature will rise and plateau. This is the moment of measurement.

Pasteurization efficiency. To determine whether milk or cream has been pasteurized properly, a test to determine the amount of remaining phosphatase enzyme is employed. The test involves the mixing and incubation of a small sample of milk, buffered solution, and disodiumphenylphosphate. An indicator is added to show colorimetrically the amount of disodiumphenol liberated. Properly pasteurized milk normally gives a negative test because only 0.1% of the phosphatase remains. However, if this milk were separated and a phosphatase test performed on a sample of the cream, then a positive test would result. This is because this phosphatase enzyme is associated with the fat phase. The phosphatase test is so sensitive that it can determine underpasteurization by 1°F (1.8°C) or adulteration with 0.1% raw milk.

Inhibitory and other foreign substances. On occasion raw milk becomes contaminated with antibiotics used in the control of mastitis and other bovine diseases. Milk containing such foreign substances may not be legally sold.

Detection of antibiotics is mandatory, particularly in cheesemaking and other manufacturing procedures that require bacterial fermentation because approximately 0.2 unit of antibiotic will inactivate most of the fermenting bacteria. Agar which has been inoculated with *Bacillus stercorophilus* spores is poured into a petri dish and hardened, and filter disks dipped in raw milk are placed on the surface. After incubation, the plates are examined for clear zones around the small disc and a comparison is made with a standard to determine the approximate concentration of antibiotic, if any.

Bacteriological determinations. The microbial content of milk and milk products is an important determination because it relates directly to the sanitary quality and the conditions under which the raw and finished products were handled. A direct microscope technique or an agar plate method is used.

Milk is an excellent growth medium for microorganisms which, if permitted to grow, will produce changes that render the milk unfit for human use. The U.S. Public Health Service Code, which most states and municipalities have adopted, stipulates that pasteurized milk and milk products can contain no more than 20,000 bacteria/ml and no more than 10 coliform bacteria/ml at pull or expiration date. To ascertain compliance with the law and general product quality, processing plants and regulatory agencies analyze samples of milk and its products for bacterial content. A representative sample of the milk is diluted in sterile buffered water and a quantity of the solution (usually 1 or 0.1 ml) is transferred to a petri dish. Bacteriological growth media (plate count agar) is poured into the dish, swirled to distribute the sample, and allowed to gel. The dishes are inverted and incubated at 90°F (32°C) for 48 h. Colonies of bacteria show as white or pigmented growth on the surface or subsurface. Each is counted as one colony times the dilution factor to give total count.

Coliform counts are performed in a similar manner except that a direct sample without dilution is used. Violet-red bile agar is added, swirled, allowed to gel, and overlaid with more agar to assure subsurface growth. After the agar has gelled, the plates are inverted to prevent the collection of moisture on the surface of the growth media and the spreading of colonies, and are then incubated 24 h at 90°F (32°C). Coliform bacteria are recognized by their red pigmentation.

One frequent test performed by laboratories associated with processing facilities is to determine the storage stability of the finished product. A container of fluid milk, for example, is stored at 45 or 50°F (7.2 or 10°C) for 1 week. At the end of this period, a standard plate count would determine the quality of the product. If bacterial analyses show the product to be inferior, then an active program is initiated to correct the problem and protect quality.

The significance of bacteria in milk and milk products depends upon the organism and the product. Coliform organisms usually are totally destroyed by pasteurization; therefore, the presence of these organisms in the finished product indicates contamination after pasteurization. With present-day facilities for refrigeration, the psychrophilic bacteria, those that multiply at refrigerator temperatures, present problems. These are normally destroyed by pasteurization and are therefore postpasteurization contaminants. Meticulous cleaning and sanitizing in a processing plant is necessary to keep these microorganisms at a minimum in finished products.

Pasteurization practically precludes the presence of pathogens, and those that survive will not survive refrigerated storage. Milk and its products also contain thermoduric and thermophilic bacteria that do not grow at pasteurization temperatures but can grow at 131°F (55°C), respectively. Thermoduric organisms are transmitted to milk from poorly cleaned equipment, while thermophilic bacteria are usually contaminants from the farm.

Milk Microbiology

Aseptically drawn milk from healthy cows is not sterile. The interior of the udder is open to invasion by bacteria when the opening of the teat comes in contact with the air and bedding. The bacteria present in the udder are distributed internally by their own growth as well as by physical movement. However, only small numbers, averaging about 1000/ml of milk, of a few types are normally found in aseptically drawn milk, although much lower and much higher counts are often reported. During the milking procedure, bacteria are in largest numbers at the beginning and gradually decrease. See FOOD MICROBIOLOGY.

Contamination from external sources. There are many sources from which the milk can be contaminated by microorganisms during milking as well as during the subsequent handling of the milk. The most important are given below.

Stable air. This may contain dust in considerable quantities, especially in a dirty stable or when hay is distributed before the milking.

Flies and other insects. A fly may carry as many as 1,000,000 bacteria. Such a fly, falling into a liter of milk, will increase the bacterial count of the milk by 1000/ml, even without bacterial reproduction.

Coat of the animal. Soil, feed, and manure adhere to the cow's coat. During the milking process this material may fall from the coat and a portion of it may get into the milk. Dry manure is a source of heavy contamination. Proper care and preparation of the cow prior to milking precludes most contamination from this external source.

Feed. Hay and silage often contain a great number of spores. Milk may be easily contaminated with these when airborne.

Milk equipment. Equipment, such as pails, cans, coolers, pipelines, bulk tanks, and milking machines, is the most serious source of bacterial contamination. It is very important that utensils be made without seams and sharp corners to facilitate cleaning.

Milking personnel. If the milking personnel are not in good health or have infections on their hands, pathogenic bacteria may be added to the milk. Milk may serve as a carrier of human pathogens from one person to another or animal to human, hence the need for pasteurization.

Kinds of microorganisms. The saprophytic and pathogenic microorganisms in milk are discussed in this section.

Saprophytes. Saprophytic microorganisms live on dead or decaying organic matter. The important ones found in milk and dairy products are as follows:

1. Certain species of bacteria found in milk convert milk sugar into lactic acid and other by-products. These bacteria are cocci as well as rods, and belong to such genera as *Streptococcus*, *Leuconostoc*, and *Lactobacillus*. Important species are *S. lactis*, causing spontaneous souring of milk and widely used for the making of cheese; *S. thermophilus*, found in fermented milks like yogurt; *Leuconostoc cremoris*, responsible for the butter-

like aroma; and *Lactobacillus casei*, present in hard cheese.

2. Certain gram-negative asporogenous (non-sporeforming) rod-shaped bacteria commonly occur in the large intestine of animals. Best known are the species *Escherichia coli* and *Enterobacter aerogenes*. Their presence in pasteurized milk serves as a sensitive index of contamination after pasteurization.

3. The genus *Pseudomonas* is well known for causing spoilage, frequently with pronounced biochemical activity, especially on proteins and fats.

4. The genus *Bacillus* contains aerobic bacilli, like *B. subtilis* and *B. cereus*. Because their spores can survive pasteurizing and sometimes even the sterilizing treatment of milk, such bacteria can be important in causing spoilage of pasteurized and sterilized milk. Some species of the anaerobic genus *Clostridium* attack proteins (*C. sporogenes*) and some produce gas (*C. butyricum*). They are well known for causing defects in cheese.

5. Yeasts sometimes ferment carbohydrates and produce gas, and sometimes are lipolytic (hydrolyze fats). They occur as contamination in sour milk products like yogurt and butter and on cheese rinds.

6. Molds which actively live on carbohydrates, fats, and commonly proteins often spoil milk and dairy products, but some are useful. For example, *Penicillium roqueforti* is used in making blue-veined cheese (Blue, Bleu, Roquefort, Stilton, and Gorgonzola cheese).

Pathogens. The milk of diseased animals may contain living germs or pathogenic microbes, and the consumption of such milk (without heat treatment) by other animals and humans may then cause the disease to be transmitted. Tuberculosis, brucellosis (undulant fever in humans), Q fever (caused by *Coxiella burnetii*), foot-and-mouth disease, and causative agents of mastitis may be so propagated. See BRUCELLOSIS; FOOT-AND-MOUTH DISEASE; TUBERCULOSIS.

In addition, milk of healthy animals may become contaminated with pathogens of other origin. Milk has been known to transmit in this manner typhoid and paratyphoid fevers, septic sore throat, diphtheria, and scarlet fever.

Some pathogens do not thrive well in milk but remain active a fairly long time. These are easily destroyed by pasteurization.

Cultured Products

Many fermented or cultured products are produced from milk. These fermentations require the use of bacteria that ferment lactose or milk sugar. These bacteria are of two general categories: homofermentative, those that produce only lactic acid from lactose; and heterofermentative, those that produce acetic acid, ethyl alcohol, and carbon dioxide in addition to lactic acid from lactose and a flavor precursor, acetoin, from citric acid. See BACTERIAL PHYSIOLOGY AND METABOLISM; FERMENTATION.

Mother or starter cultures. Bacteria such as *Streptococcus lactis*, *S. cremoris*, *S. diacetylactis*, *S. thermophilus*, *Leuconostoc cremoris*, *Lactobacillus*

acidophilis, and *Lactobacillus bulgaricus* are among the common organisms used in cultures. Cultures of these organisms purchased from a commercial supply house are transferred several times to produce an active growth. Selection of the proper culture to use depends upon the talent and judgment of a trained technologist. Larger quantities of milk are inoculated and ripened to develop the appropriate acidity which is directly related to the populations in the culture, and then large vats of milk are inoculated to produce different fermented products. See BACTERIOLOGY.

Among the problems associated with fermentations, contamination of a culture with bacteriophage is probably the most devastating. Bacteriophages are viral agents that attack and destroy specific bacteria. To preclude this, manufacturers use special media for control of phage growth. Furthermore, rotation of the strains of the particular organisms used in a culture program is advantageous. Obviously, antibiotics and improper manipulations of cultures can be of serious consequence. See BACTERIOPHAGE.

Cultured buttermilk. Skim milk or low-fat milk is pasteurized at 180°F (82°C) for 30 min, cooled to 72°F (22°C), and inoculated with an active starter culture containing *S. lactis* and *Leuconostoc cremoris*. This mixture is incubated at 70°F (21°C) and cooled when acidity is developed to approximately 0.8%. This viscous product is then agitated, packaged, and cooled. The desired flavor is created by volatile acids and diacetyl; the latter is produced by *L. citrovorum*.

So that the modern product will resemble the classical buttermilk, resulting from the churning of cream to butter, manufacturers often spray or otherwise introduce tiny droplets of milk fat into the product during agitation.

Cultured sour cream. Cream containing 18–20% milk fat is processed and inoculated similar to the milk used in the manufacture of buttermilk. One additional procedure involves homogenization of the hot, pasteurized cream prior to cooling to inoculation temperature to control the consistency. Because of its high viscosity sour cream is often ripened in the package to an acidity of 0.6%.

Yogurt. One of the oldest fermented milks known is yogurt. Historically the people of the Middle East relied on yogurt as an important food item. Later, consumption increased rapidly in Europe because of the suspected correlation with longevity.

Yogurt is prepared using whole or low-fat milk with added nonfat milk solids. The milk is heated to approximately 180°F (82°C) for 30 min, homogenized, cooled to 115°F (46°C), inoculated with an active culture, and packaged. Yogurt cultures are mixtures of *S. thermophilus* and *L. bulgaricus* in a 1:1 ratio. Balance of these organisms in the culture is important for production of a quality product. The product after inoculation is incubated until approximately 0.9% acidity has developed and then cooled.

In the United States, greater sales are realized in yogurts that contain added fruit than in the unflavored product. Three types are marketed: fruit mixed throughout, fruit on top, and fruit on bottom. Man-

ufacturing procedures and costs demand an automated process where the product is continuously handled. After the proper acidity is developed during quiescent incubation, the product is agitated, fruit is added, the product is cooled to prevent further acid development, and finally it is pumped to a packaging machine to be filled mostly into small cups.

Yogurt is custardlike in consistency and is generally eaten with a spoon.

Other fermented milks. A few other types of fermented milks produced in other parts of the world are relatively obscure to people in the United States. Bulgarian buttermilk is made by the inoculation of whole milk with a culture of *L. bulgaricus*. It is incubated until approximately 0.9% acidity develops, and then packaged. This product is similar to yogurt except that it is fluid.

Kefir is a fermented milk native to the Caucasus Mountain area of southeastern Europe. Little is produced in the United States. During the fermentation process, lactic acid is produced by *Streptococcus* and *Lactobacillus* organisms, and alcohol (approximately 1%) is produced by lactose-fermenting yeasts. The latter collect in grains that are about the size of wheat and float on the surface.

Kumiss is another milk beverage in which both acid and alcohol are developed. Commonly, milk from mares is used, and the finished product resembles kefir except that no visible grains float on the surface. Kumiss is a mildly acid milk product with relatively little alcohol (0.5–1.5%) and is native to southern Russia.

Concentrated and Dried Milk Products

To reduce costs of transportation and handling, either part or all of the water is removed from milk. Moreover, the partly dehydrated milk can either be sterilized or dried to permit unrefrigerated storage for prolonged periods. Many different milk products are produced for these specific reasons. The composition of some of these is controlled by standards of identity and some by request of the commercial buyer.

In the United States nonfat dry milk is defined by Public Law 244, March 2, 1944, as amended by Public Law 646, July 2, 1956. Such milk is the product resulting from the removal of fat and water. In the dry form, nonfat dry milk can contain no more than 5% water and not over 1.5% fat.

Dry whole milk is the result of removing water from whole milk until the dry product contains no more than 4% water and not less than 26% fat.

Dry buttermilk results from the removal of moisture from fluid buttermilk drained from the churn during the manufacture of butter. It contains no more than 5% water and not less than 4.5% fat.

Evaporated milk is the product that results from removing moisture from whole milk so that the final product contains no less than 7.9% fat or 25.9% total milk solids. Emulsifying salts, to reduce the effect of heat on the protein during sterilization, and vitamin D are optional additives and cannot exceed 0.1% and 400 IU (reconstituted), respectively.

Baby formulations are produced to give a product, when reconstituted, that has the same composition as mother's milk. These formulations contain nonfat milk solids, salts, sugars, and vegetable fat. Baby formulations are marketed in either dry or condensed sterilized liquid form.

Sweetened condensed whole or skim milks are prepared to contain not less than 8.5% fat and 28% total milk solids or 24% total milk solids, respectively. Enough sugar is added to prevent spoilage, usually 42–45%.

Condensed milks and buttermilk are the result of removing some water from the fresh fluid product. The concentration of milk solids is generally requested by the food manufacturer; however, if the concentration of skim milk exceeds 36% and this product is stored, lactose will crystallize.

Condensing or evaporating. Water is removed from the exposed milk product by condensing or evaporating. Initially the fluid product is given a heat treatment to destroy bacteria and enzymes and to impart particular properties to the finished product. For example, ice cream manufacturers want a low-heat product, whereas bakers need a high-heat product to give loaf volume and crust color to bread. Heat treatments vary from 150 to 210°F (66 to 99°C), for a few seconds to as much as 60 min. The preheated product is drawn into the vacuum pan from the hot well or surge tank and is boiled at reduced pressure. To increase the efficiency of this operation, two, three, or four units or effects are connected together. The vapor energy from one pan is used to cause boiling in the next. The vacuum applied to each effect is increased as the final effect is reached.

Sterilization. Evaporated milk is standardized to meet federal specifications, canned, sealed, and sterilized. Temperatures and times of 238–245°F (114–118°C) for 15–17 min are used. To prevent coagulation of the protein during sterilization, the salt balance of the milk must be examined. If the salt balance or concentration of calcium and magnesium is not in proportion to the phosphates and citrates content, then the defect occurs. However, coagulation can be prevented by the addition of phosphate or citrate salts not to exceed 0.1% in the finished product.

Spray drying. Most condensed milks, baby formulations, whole milk, and buttermilk are dried. Condensed fluid product is pumped directly while still hot from the vacuum pan into a surge tank and then to the spray dryer. This piece of equipment can either be in the form of a horizontal stainless steel box or a vertical cylinder that is cone-shaped on the bottom. The fluid product is atomized to facilitate drying. Atomization is accomplished either by centrifugal force where the product is discharged from a rapidly turning horizontal disk, or by forcing the product through a nozzle with a minute orifice. Drying occurs almost instantaneously in the heated atmosphere in the dryer.

The dried product is collected from the drier by several means. One method involves the use of a cyclone, where the dried particles are removed by centrifugal force from a swirling air stream. Other

methods used in horizontal dryers involve mechanical sweeps to push the dried powder across the bottom of the dryer to an auger for discharge and a series of cloth filter bags in a separate housing.

The Department of Agriculture has assigned grades to spray-dried nonfat dry milk on the basis of results from flavor, physical, and laboratory examinations. These are U.S. Extra, U.S. Standard, and U.S. Grade Not Assignable. A product in the latter classification is used as animal feed. Further grading is assigned to denote heat treatments that the product receives during preparation. These are U.S. Low Heat, U.S. High Heat, and U.S. Medium Heat, and these reflect the amount of whey protein change or denaturation.

Roller drying. Little milk for human use is roller-dried. The process involves feeding the concentrated milk between two heated and narrowly spaced stainless steel rollers. A thin film of milk adheres and is dried as these rollers turn. A scraper or doctor blade removes the dried product. The chief disadvantage to this process is the excess exposure to heat incurred by the product. The finished product has a characteristic flavor and limited reconstituting characteristics.

Instantizing. To render spray-dried nonfat dry milk more easily reconstituted, the product is exposed to another heat treatment. Unfortunately, the process decreases the solubility of milk proteins. Instantizing creates an agglomerated powder; therefore, the particles are heavier and sink and wet faster than a noninstantized powder. Nonfat dry milk directly from the spray dryer is fed into a hopper and metered into a steam current to cause the product to lump together or agglomerate. This wet (10–15% water) product is redried, and the agglomerates are sieved to the required size. In some commercial operations, sizing rollers are used. Any fine particles are recirculated to the inlet hopper. Most of the nonfat dry milk for fluid consumption is sold in this form. See FOOD MANUFACTURING; ICE CREAM. Robert L. Bradley, Jr.

Bibliography. A. W. Farrall, *Engineering for Dairy and Food Products*, 2d ed., 1980; J. L. Henderson, *The Fluid Milk Industry*, 3d ed., 1971; Y. H. Hui (ed.), *Dairy Science and Technology*, 3 vols., 1993; G. H. Richardson (ed.), *Standard Methods for the Examination of Dairy Products*, 16th ed., 1992; N. P. Wong, *Fundamentals of Dairy Chemistry*, 3d ed., 1988; U.S. Department of Agriculture, *Federal and State Standards for the Composition of Milk Products*, Handb. 51, revised 1977.

Milky Way Galaxy

The large disk-shaped aggregation of stars, gas, and dust in which the solar system is located. The term Milky Way is used to refer to the diffuse band of light visible in the night sky emanating from the Milky Way Galaxy. Although the two terms are frequently used interchangeably, Milky Way Galaxy, or simply the Galaxy, refers to the physical object rather than its appearance in the night sky, while Milky Way is used to refer to either.

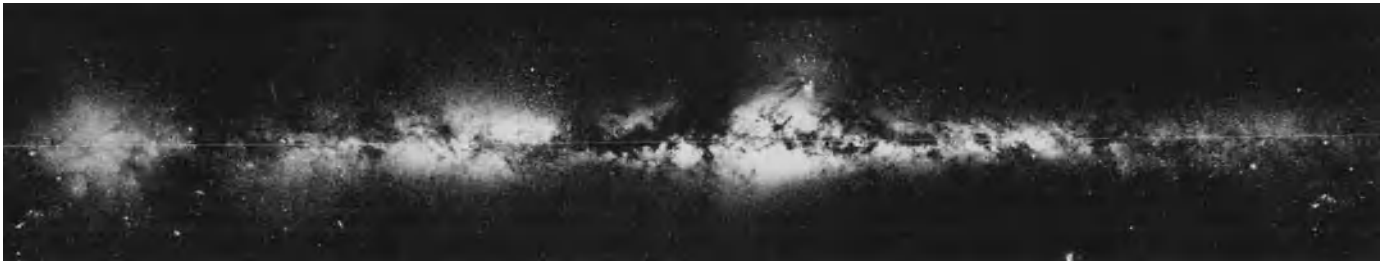


Fig. 1. Panoramic photograph of the entire Milky Way. The dark patches are nearby clouds containing interstellar dust that block out the background starlight. The photograph conveys the overall impression of the flattened distribution of stars that make up the disk. The direction of the galactic center is in the center but it is hidden by foreground dust. The galactic bulge can be seen as the thickening of the disk in the center. (*European Southern Observatory*)

Appearance. The Milky Way is visible in the night sky to the unaided eye as a broad diffuse band of light stretching from horizon to horizon when viewed from locations away from bright city lights (**Fig. 1**). Nearly all of the visible light is due to individual stars, which in many directions are too numerous to be resolved without a telescope. The patchy appearance of the Milky Way is due to collections of microscopic dust particles which block the light of more distant stars. All of the stars seen by the unaided eye are part of the Milky Way Galaxy and lie relatively close to the Sun. The overall appearance is due to the Sun's location near the midplane of the galactic disk; the diffuse band of light is seen toward directions close to the midplane where there are many more stars along the line of sight than in directions away from the plane.

Structure and contents. The Milky Way Galaxy contains about 2×10^{11} solar masses of visible matter. Roughly 96% is in the form of stars, and about 4% is in the form of interstellar gas. The gas both inside the stars and in the interstellar medium is primarily hydrogen (roughly 87–90% by number of atoms) and helium (about 10%) with a small admixture of all of the heavier atoms (0–3% depending on the location within the Galaxy). The mass of dust is about 1% of the interstellar gas mass and is an insignificant fraction of the total mass of the Galaxy. Its presence, however, limits the view from the Earth in the plane of the Galaxy to a small fraction of the Galaxy's diameter in most directions. *See* INTERSTELLAR MATTER.

The Milky Way Galaxy contains four major structural subdivisions: the nucleus, bulge, disk, and halo. The Sun is located in the disk about halfway between the center and the indistinct outer edge of the disk of stars. The currently accepted value of the distance of the Sun from the galactic center is 8.5 kiloparsecs, although some measurements suggest that the distance may be as small as 7 kpc. (A parsec is equal to 2.06×10^5 times the average distance between the Sun and the Earth, and is approximately equal to the average distance between stars in the solar neighborhood; it is also equal to 3.26 light-years.) **Figure 2** is an infrared image of the Milky Way made from the Sun's position in the disk, showing the bulge and disk of the Galaxy. The infrared image minimizes the obscuring effect of dust. **Figure 3** shows a spiral galaxy

that is structurally similar to the Milky Way, giving an approximate idea of how the Milky Way would appear if viewed obliquely. *See* INFRARED ASTRONOMY; PARSEC.

Nucleus. The nucleus of the Milky Way is a region within a few tens of parsecs of the geometric center and is totally obscured at visible wavelengths. The nucleus is the source of very energetic activity detected by means of radio waves and infrared radiation.

At the galactic center, there is a very dense cluster of hot stars observed by means of its infrared radiation. In 1997, astronomers confirmed the existence of a black hole with a mass of about 2.5 million times the mass of the Sun at the position of an unresolved source of radio emission known as Sgr A* in the middle of the central star cluster. The black hole appears to be the dynamical center of the Milky Way, and evidence for its existence appears to be unequivocal. Many spiral galaxies show evidence of black holes at their centers. At a distance of about 1.5 pc from Sgr A* is a ring of gas consisting primarily of molecular hydrogen that surrounds the central star cluster. Within this ring is a three-armed "minispiral" of ionized gas that appears to be falling into the central cluster and possibly onto the black hole



Fig. 2. Infrared image of the Milky Way as seen from the location of the Sun, made from the 2MASS infrared sky survey. The image shows both the bulge and the disk of the Milky Way. Absorption by dust is minimized in the infrared, which accounts for the difference between this image and Fig. 1. However, the dust remains visible as the dark band across the disk and other dark patches. The Large and Small Magellanic Clouds are visible in the lower right. The faint vertical wisp coming down from the left part of the bulge is the Sagittarius Dwarf. (2MASS)



Fig. 3. NGC 4603, a galaxy with a gross morphology similar to the Milky Way, seen obliquely. The spiral arms and the bulge are visible in the center. The bulge appears to be elongated because the galaxy also contains a bar, an extended feature containing primarily old stars that rotate collectively about the center as if they were a solid body. (NASA)

(Fig. 4). Molecular hydrogen gas, which is characterized by low temperatures and relatively high densities, is found in great profusion in the inner few hundred parsecs of the Milky Way. Great arcs of gas

resulting from the interaction of cosmic rays and magnetic fields have been mapped in the central tens of parsecs of the Milky Way, attesting to the energetic activity taking place there. See BLACK HOLE; COSMIC RAYS; RADIO ASTRONOMY.

Bulge. The bulge is a spheroidal distribution of stars centered on the nucleus which extends to a distance of about 3 kpc from the center. It contains a relatively old population of stars, very nearly as old as the Milky Way itself. The bulge can be seen by the unaided eye as a thickening of the diffuse band of light that constitutes the Milky Way in the direction of the galactic center, toward the constellation Sagittarius. There is little gas and dust throughout most of the volume of the bulge. Direct imaging with infrared satellites has demonstrated that the bulge is actually an elongated barlike structure with a length about two to three times its width. The Milky Way is thus classified as a barred spiral galaxy, a classification that includes about half of all disk-shaped galaxies. See STELLAR POPULATION.

Disk. The disk is a thin distribution of stars and gas orbiting the nucleus of the Galaxy. The disk of stars begins near the end of the bar and can be identified to about 16 kpc from the center of the Galaxy; the disk of gas can be identified to about twice this distance, about 35 kpc from the center. The thicknesses of the disks are characterized by a scale height: the distance from the midplane at which the density of gas and stars falls by a factor of e (2.718...). In the solar vicinity, the scale height is different for each of the components of the disk, varying from about 75 pc for the molecular gas to about 350 pc for the lowest-mass stars. The faint, low-mass stars make up most of the mass of the disk. There is also a thick disk of stars and gas with a scale height of about 1.5 kpc for the stars and about 1 kpc for the atomic gas. The origin of these thick disks is not known. The thin disk of stars contains most of the mass and has a



Fig. 4. Radio image of the ionized gas within about 1.5 parsecs of the center of the Galaxy. The image, made with the Very Large Array (VLA), shows what appears to be a three-armed spiral of gas. Other observations indicate that this gas may be falling into the central star cluster (not seen in this image). The position of Sgr A* is indicated by the elongated bright dot seen above the horizontal bar of emission. The dot is thought to be radio emission from hot gas falling into the black hole. (NRAO, courtesy of N. Killeen and K.-Y. Lo)

thickness relative to its diameter similar to that of a commercial compact disk. At this scale, the bulge would have the size of a sausage about 2.5 cm (1 in.) in length superimposed on the nucleus.

The star and gas disks are both warped; the gas disk deviates from a true plane by about 3 kpc (about 10%) in its outermost parts. The disk also becomes thicker with increasing distance from the galactic center. The thickening occurs because there is less gravity to confine the gas and stars to the midplane. The reason for the warping is not yet understood, but warping is a common feature of spiral galaxies.

The disk is the location of the spiral arms that are characteristic of most disk-shaped galaxies, as well as most of the present-day star formation. Attempts to map the location of the spiral arms are hampered by the Sun's location in the disk, but in the parts of the Galaxy beyond the Sun's distance from the center, several long coherent spiral arms have been identified. The inner regions appear to be more chaotic, and no agreed-upon spiral arm structure has been identified. The spiral arms are where giant clouds of molecules are primarily found. These clouds are the most massive objects in the Milky Way, ranging up to about 5×10^6 solar masses; they are the sites of all present-day star formation. The Sun presumably once formed in a giant molecular cloud. *See* MOLECULAR CLOUD.

Halo. The halo is a rarefied spheroidal distribution of stars nearly devoid of the interstellar gas and dust that surrounds the disk. The stars found in the halo are the oldest stars in the Galaxy. The stars are found individually as "field" stars as well as in globular clusters: spherical clusters of up to about a million stars with very low abundances of elements heavier than helium. The extent of the halo is not well determined, but globular clusters with distances of about 40 kpc from the center have been identified. Dynamical evidence suggests that the halo contains nonluminous matter in some unknown form, commonly referred to as dark matter (see below). The dark matter contains most of the mass of the Galaxy, dominating even that in the form of stars. *See* STAR CLUSTERS.

Companion galaxies. The Milky Way has 11 companion galaxies clustered within about 250 kpc of its center. The largest and brightest of these, the Large and Small Magellanic Clouds, are seen as diffuse patches of light roughly 7° and 5° in diameter and separated from the plane of the Milky Way by about 30° and 40° , respectively. They are visible mainly from the Earth's southernmost latitudes. The distance to the Magellanic Clouds is about 55 kpc, not quite twice the diameter of the stars in the disk of the Galaxy. A third small galaxy, known as the Sagittarius Dwarf, was discovered in 1994 and is the nearest galaxy to the Milky Way. It went undiscovered for such a long time because it lies close to the obscuring dust in the midplane. The Sagittarius Dwarf is seen in the general direction of the center of the Galaxy, and its distance from the Sun is about 24 kpc, about three times the distance of the Sun from the center of the Galaxy.

All three galaxies are interacting with the Milky Way. The Magellanic Stream, a narrow band of gas drawn out of the Magellanic Clouds, which stretches over much of the sky, is produced either by the tidal forces of the Milky Way or by the ram pressure of the clouds going through a tenuous halo of galactic gas. The Sagittarius Dwarf shows evidence of being stretched apart by the strong tidal forces of the Milky Way, and is about to collide with the disk, a process that will take millions of years. *See* MAGELLANIC CLOUDS.

Dynamics and kinematics. The stars and gas in the disk of the Milky Way rotate around the galactic center in nearly circular orbits in the plane of the disk. The deviations from circularity are generally small, and give rise to oscillations of stellar motions perpendicular to the disk, epicyclic motions in the plane of the disk, and other, more complex motions. The overall rotation of the disk is differential rather than like a solid body in that the rotation period of the stars and the gas increases with increasing distance from the center. Large-scale deviations from circularity are associated with the spiral arms of the disk.

The barlike bulge produces large noncircular motions in the disk of gas contained within it, providing independent confirmation that the Milky Way contains a "bar" in its inner parts. A bar is a large collection of stars in the inner part of a galaxy in which the stars collectively rotate about the center as a solid body would, even though the individual stellar orbits are quite complex.

The orbits of the stars and clusters in the bulge and halo are not confined to the galactic plane and are, in general, fully three-dimensional. Although the motions of most stars are too small for the orbits to be determined from their changes in position alone, orbits can be inferred from the measured radial velocities of the stars (the velocities measured along the line of sight) and from their positions in the Galaxy. The orbits of stars close to the Sun can be observed directly from their apparent motions in the sky measured over many years. Motions of small numbers of stars as distant as the center of the Galaxy can be measured directly by special techniques.

In the outermost parts of the Milky Way, one expects that the orbital velocities of stars and gas around the galactic center should decrease in a well-determined way, as they do, for example, in the solar system. Measurements show, however, that in the outermost regions of the disk the velocities of the gas and stars do not decrease with increasing distance from the galactic center (**Fig. 5**). These measurements suggest that unseen matter is pulling on the visible matter in the disk and producing the anomalously large velocities. The constancy of the rotation velocity of the Milky Way is generally thought to be the most compelling evidence for large quantities of material in the Milky Way in a form other than stars, gas, and dust. This dark matter is common to most galaxies for which rotation measurements are possible.

Dark matter. The dark matter in the Milky Way constitutes by various estimates $2\frac{1}{2}$ to 10 times the total

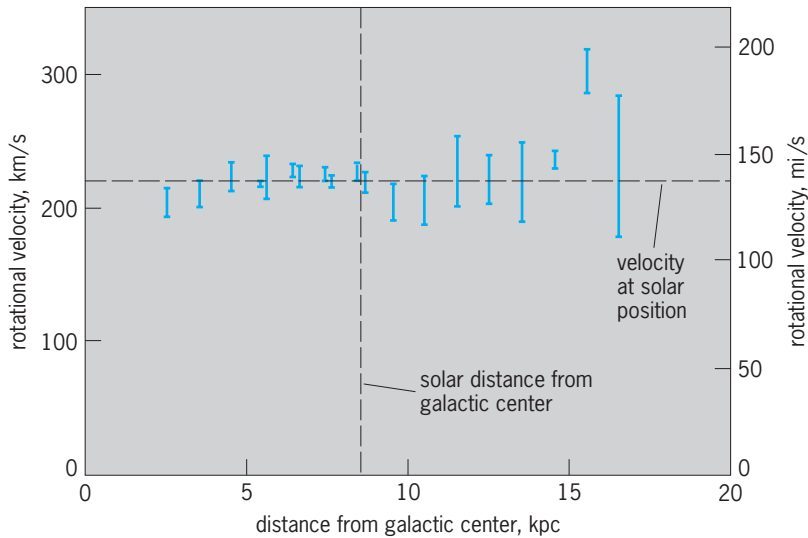


Fig. 5. Plot of the measurement of the rotational velocity of gas and stars about the center of the Milky Way Galaxy as a function of distance from the galactic center, showing that the rotational velocities are essentially constant. Beyond the solar distance, one would expect that the rotational velocity would decrease markedly to a value about two-thirds the value measured at the Sun's distance if only visible matter contributed to the gravity of the Milky Way. (Courtesy of M. Fich, L. Blitz, and A. A. Stark)

amount of known matter in the Milky Way, and is consequently the dominant component of mass in the Galaxy. Because the kinematics of most galaxies indicate that they too contain dark matter as their largest mass constituent, the dark matter appears to be the dominant form of matter in the universe. The composition of the dark matter is currently unknown and is one of the major unsolved problems in astronomy.

In the Milky Way, various forms for the dark matter have been ruled out. It cannot be in the form of ordinary stars or in remnants such as white dwarfs, neutron stars, or black holes that are the end products of ordinary stellar evolution. It cannot be in the form of gas in any form, nor can it be in the form of small, solid dust particles. Although a small fraction of the dark matter may reside in the disk, most of it appears to reside in the halo. See NEUTRON STAR; WHITE DWARF STAR.

Two possibilities that have not yet been ruled out are that the dark matter consists of small planet-sized bodies that are insufficiently luminous to be detected with present instruments, and that it consists of primordial black holes in the halo. Still other possibilities include weakly interacting massive particles (WIMPs) that are predicted by various elementary particle theories. Observations of the supernova 1987A have apparently ruled out the possibility that neutrinos (massless or very low mass neutral particles) constitute most of the dark matter in the universe. An alternative theory to newtonian gravitation has also been proposed to explain the rotation curve of the Milky Way and other galaxies, but this is generally considered to be a particularly radical explanation. See COSMOLOGY; GRAVITATION; NEUTRINO; SUPERNOVA; WEAKLY INTERACTING MASSIVE PARTICLE (WIMP).

Location. The Milky Way is part of a small group of galaxies known as the Local Group. The Local Group also contains two large spiral galaxies, the great nebula in Andromeda (M31) and M33, and 40 small irregular, spheroidal and elliptical galaxies. Some of these small galaxies have been discovered only in the last few years, and it is likely that the census of galaxies in the Local Group is not yet complete. About 95% of the mass in the Local Group is associated with either the Milky Way or M31. The Local Group is itself part of a large supercluster of galaxies known as the Virgo supercluster containing about 1000 known galaxies. The Local Group is an outlying collection of galaxies in the Virgo supercluster and is about 15 Mpc from its center. The Virgo supercluster is one of many such clusters in the universe and occupies no special location within it. See ANDROMEDA GALAXY; GALAXY, EXTERNAL; LOCAL GROUP; UNIVERSE.

Formation and evolution. Inferences about the formation and evolution of the Milky Way can be drawn from a large variety of sources, including the theory and observation of the kinematics and dynamics of the gas and stars in the Milky Way, the chemical abundances of the stars, their locations in the Galaxy, and stellar evolution theory. Current evidence suggests that the ages of the oldest stars in the Milky Way are within about 10% of the age of the universe as a whole; thus parts of the Milky Way must have formed early in the history of the universe, about 12 billion years ago. See STELLAR EVOLUTION.

There is increasing evidence that the Milky Way formed as a result of the coalescence of small galaxies and protogalaxies, objects with the masses of small dwarf galaxies that are thought to have been among the first objects to form in the universe. The coalescence would have proceeded rapidly at first and then more irregularly as relatively large pieces merged to form the Milky Way. According to this picture, the formation of the Milky Way is not yet complete, and both small galaxies (such as the Magellanic Clouds and the Sagittarius Dwarf) and starless clouds of gas are continuing to rain in on the Milky Way. Direct evidence for this picture comes from the orbit of the Sagittarius Dwarf, which shows it to be colliding with the Milky Way, and from the presence of clouds of hydrogen gas with high velocities and low abundances of heavy elements, which appear to be accreting from intergalactic space.

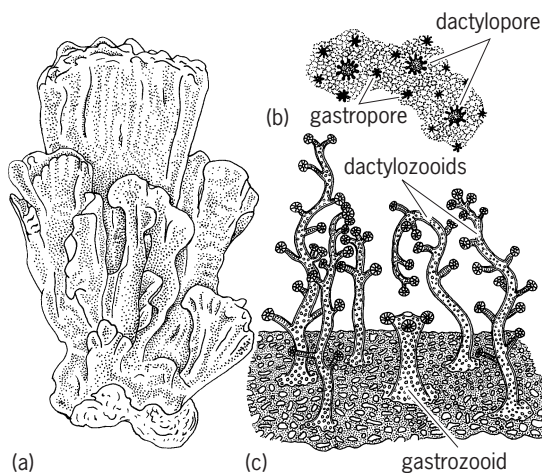
When the Galaxy was about two-thirds its present age, 4.6 billion years ago, the Sun and the planets formed from the interstellar medium in the disk, the Sun being a rather unremarkable star in an unremarkable location. The part of the disk that has not yet been used up in the process of star formation continues to form stars. See SOLAR SYSTEM; SUN.

The Milky Way and M31 are approaching each other at a velocity of 125 km/s (78 mi/s). At that velocity, the two galaxies will collide in about 5 billion years if the transverse velocity of M31 is not too large. The collision will not be violent; the distance between the stars is too vast for more than a few stars to actually collide. The likely outcome is a merged system that resembles an elliptical galaxy. Leo Blitz

Bibliography. J. Binney and M. Merrifield, *Galactic Astronomy*, Princeton University Press, 1998; L. Blitz et al., The centre of the Milky Way, *Nature*, 361:417–424, 1993; L. Blitz and P. Teuben (eds.), *Unsolved Problems of the Milky Way*, Kluwer Academic Publishers, 1996; B. J. Bok and P. F. Bok, *The Milky Way*, 5th ed., Harvard University Press, 1981; A. Eckart and R. Genzel, Stellar proper motions in the central 0.1 pc of the Galaxy, *Mon. Not. Roy. Astron. Soc.*, 284:576–598, 1997; B. K. Gibson, T. S. Axelrod, and M. E. Putman (eds.), *The Galactic Halo*, Astronomical Society of the Pacific, 1999; I. King, G. Gilmore, and P. C. van der Kruit, *The Milky Way as a Galaxy*, University Science Books, 1994.

Milleporina

An order of the class Hydrozoa of the phylum Cnidaria. These are the “stinging corals” of the shallow tropical seas. Their structure is similar to that of hydroids except for the addition of a calcareous exoskeleton (illus. a). Because of this skeleton, they resemble true corals, which belong, however, to a different class of coelenterates, the Anthozoa. The skeleton is covered by a thin layer of tissue, is penetrated by interconnecting tubes, and is perforated by tiny holes through which the bodies of the “coral animals,” or polyps, are extended (illus. b).



Milleporina. (a) Piece of dry *Milleporina*, showing typical flabellate shape. (b) Same, magnified, showing pores. (c) Polyps of *Milleporina*. (After L. H. Hyman, *The Invertebrates*, vol. 1, McGraw-Hill, 1940)

The polyps are of two types (illus. c): nutritive gastrozooids with tentacles and mouth; and protective polyps, which are long and armed with stinging cells but have no mouth. Their sting is very painful to humans, unlike the relatively harmless sting of most true corals. Millepores produce medusae, or jellyfish, in which sex cells develop.

Some authorities combine the Milleporina with the Stylasterina in a single order, the Hydrocorallina. See HYDROZOA. Sears Crowell

Millerite

A mineral having composition NiS and crystallizing in the hexagonal system. Millerite usually occurs in hairlike tufts and radiating groups of slender to capillary crystals (see **illus.**). There is rhombohedral cleavage,



Millerite crystals on calcite from Keokuk, Iowa. (Specimen from Department of Geology, Bryn Mawr College)

but it is difficult to observe on the hairlike crystals. The hardness is 3–3.5 (Mohs scale) and the specific gravity is 5.5. The luster is metallic and the color pale brass yellow. Millerite forms at low temperatures, often in cavities and as an alteration of other nickel minerals. It is found in many localities in Europe, notably in Germany and the Czech Republic. In the United States it is found with pyrrhotite at the Gap Mine, Lancaster County, Pennsylvania; with hematite at Antwerp, New York; and in geodes in limestone at Keokuk, Iowa. In Canada large cleavable masses are mined as a nickel ore in Lamotte Township, Quebec. See NICKEL; PYRRHOTITE. Cornelius S. Hurlbut, Jr.

Millet

A common name applied to at least five related members of the grass family grown for their edible seeds: foxtail millet (*Setaria italica*), proso millet (*Panicum miliaceum*), pearl or cat-tail millet (*Pennisetum typhoideum*; formerly *P. glaucum*), Japanese barnyard millet (*Echinochloa frumentacea*), raggee or finger millet (*Eleusine coracana*), and koda millet (*Paspalum scrobiculatum*). Millets have been used since prehistoric times as food crops, primarily in regions where the warm growing season is short (60 to 120 days), or in dry regions where rainfall periodicity provides a short period when soil moisture permits growth and ripening of a short-season crop. Under these climatic conditions, one or more millets are grown in such diverse geographic regions as Russia, China, India, Africa, and Latin America.

Use. As a crop for human food, pearl millet is grown widely in the tropics and subtropics in regions of limited rainfall where there is a growing season of 90 to 120 days. In Africa and Asia some 1.8×10^7 tons (2×10^7 metric tons) of grain are produced yearly in 40 different countries. This millet is grown

where the limited rainfall or length of growing season is inadequate for grain sorghum or maize. Pearl millet grain yields average 520–620 lb/acre (600–700 kg/hectare), but much higher yields are possible with favorable rainfall and soil fertility. The naked seeds are yellowish to whitish in color and about the size of wheat grain. The dried grain is usually pulverized to make a meal or flour and then cooked in soups, in porridge, or as cakes.

In some Asian countries, finger millet is grown as a short-season food grain crop and is used in similar manner as pearl millet. The other millets also are grown as food grains in dry regions, mostly in subtropical zones or where there is a very short growing season of 45 to 90 days.

Millet grains are high (55–65%) in starchy components and thus serve as energy foods. The protein content and quality vary greatly among the various types, but they are generally low in lysine amino acid and must be used along with animal products or fish to balance the protein diet. Where widely grown, the millets are major food crops.

In North America millets are used to some extent for forage (foxtail millet in northern regions and pearl millet in warm regions). Proso millet is grown for grain, mostly for livestock feed, in the northern Plains region. The feeding value is about equivalent to corn. There is also an increasing demand for proso and foxtail millet for use as birdseed. See GRASS CROPS.

Howard B. Sprague

Diseases. In the United States four general types of disease occur on the millets. These are foliage diseases, head mold, smut, and seedling and root diseases.

Foliage diseases such as *Helminthosporium* and *Cercospora* leaf spot diseases of pearl millet usually occur late in the season and, as a rule, damage is slight. Lesions caused by *Helminthosporium* vary in size from brown flecks to large oval or rectangular spots (Fig. 1a), in contrast to the small dark brown, gray, or tan-centered spots caused by *Cercospora*

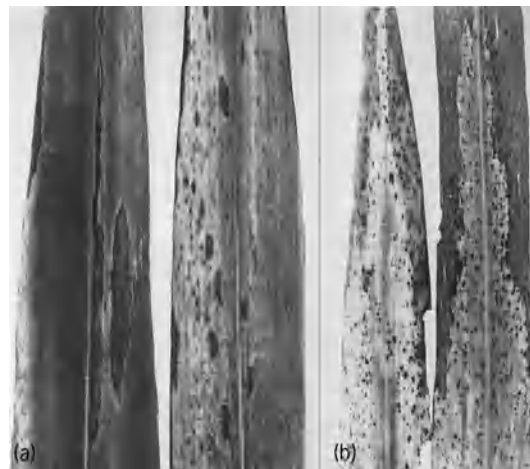


Fig. 1. Leaf diseases of millet. (a) *Helminthosporium* leaf spot of pearl millet (*Helminthosporium stenospilum* or *H. sacchari*). (b) *Cercospora* leaf spot of pearl millet (*Cercospora penniseti*). (From E. S. Luttrell et al., *Diseases of pearl millet in Georgia, Plant Dis. Rep.*, 38(7):508, 1954)

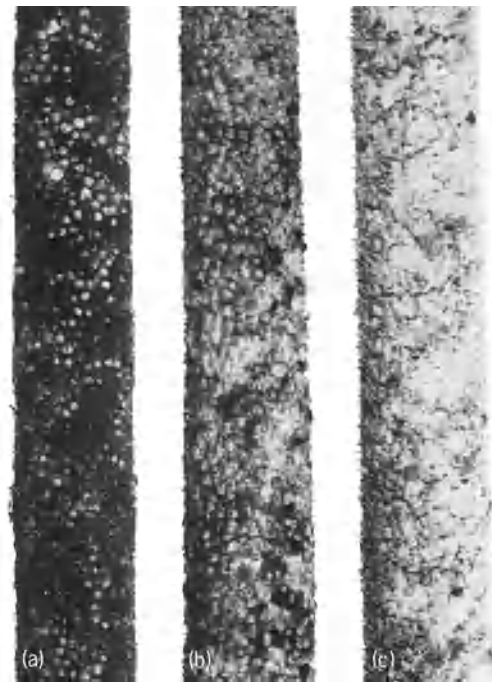


Fig. 2. Head molds of pearl millet. (a) Black mold, which is caused by *Helminthosporium* and *Curvularia*; (b) orange mold, which is caused by *Fusarium*; and (c) white mold, which is caused by *Oidium*.

(Fig. 1b). Head molds caused by several fungi may seriously affect seed production, particularly in moist weather (Fig. 2). Individual seeds or entire heads may be covered with woolly mycelial mats which are black, orange, or white, depending on the specific fungi present. Smut diseases, common in other countries, also affect the seed in the heads of foxtail and proso millet in the United States. These smuts may be controlled by seed treatment. Seedling blights and root rots, caused by a number of soil-inhabiting fungi, may reduce stands under unusually moist conditions. See PLANT PATHOLOGY. Herman A. Rodenhiser

Mineral

A naturally occurring solid phase with a restricted chemical composition and a definite atomic arrangement. This is really an idealized definition insofar as some minerals (especially those that form at low temperature) have poorly defined atomic structures. A few solid geological materials (such as coal and obsidian) are not minerals because they have no ordered crystal structure and a completely variable composition. Rocks are usually made up of one or more minerals. In a thermodynamic sense, a mineral is a separate phase that is made up of one or more components. For example, the mineral olivine [$(\text{Fe}, \text{Mg})_2\text{SiO}_4$] can be viewed as being made up of fayalite (Fe_2SiO_4) and forsterite (Mg_2SiO_4) in a solid solution.

Classification and systematics. New minerals are still being discovered. At present, the International Mineralogical Association recognizes over 4000

| Examples of important minerals | |
|--|--|
| Silicates | |
| Orthosilicates | |
| Olivine $(\text{Fe,Mg})_2\text{SiO}_4$ | Mafic to ultramafic igneous rocks; a major phase in the upper mantle |
| Garnet $(\text{Ca,Mg,Fe})_3(\text{Al,Fe})_2(\text{SiO}_4)_3$ | Metamorphic rocks, ultramafic rocks |
| Kyanite, andalusite, sillimanite $\text{Al}_2\text{O}(\text{SiO}_4)$ | Metamorphosed shales: schists and gneiss |
| Single-chain silicates: pyroxenes | |
| Orthopyroxenes $(\text{Mg,Fe})_2\text{SiO}_3$ | Mafic to ultramafic igneous rocks |
| Clinopyroxenes $(\text{Ca,Fe,Mg})(\text{Mg,Fe})\text{SiO}_3$ | Mafic to ultramafic igneous rocks |
| Double-chain silicates: amphiboles | |
| Hornblende $(\text{Ca,Na})_{2-3}(\text{Mg,Fe,Al})_5\text{Si}_6(\text{Si,Al})_2\text{O}_{22}(\text{OH})_2$ | Metamorphosed mafic rocks, granodiorites |
| Sheet silicates | |
| Muscovite $\text{K}_2\text{Al}_4(\text{Al}_2\text{Si}_6)\text{O}_{20}(\text{OH})_4$ | Metamorphic rocks, some granites |
| Biotite $\text{K}_2(\text{Mg,Fe})_6(\text{Al}_2\text{Si}_6)\text{O}_{20}(\text{OH})_4$ | Granites, metamorphic rocks |
| Montmorillonite $(\text{Ca,Mg,Na})_{0.5-0.7}(\text{Mg,Al,Fe})_4(\text{Al}_2\text{Si})\text{O}_{20}(\text{OH})_4$ | Clay mineral: altered volcanic ash, mudstone, soil |
| Kaolinite $\text{Al}_2\text{Si}_2\text{O}_5(\text{OH})_4$ | Clay mineral: altered granites, soil |
| Framework silicates | |
| Quartz SiO_2 | Major phase in most felsic rocks, sandstone, soil |
| K-feldspar KAlSi_3O_8 | Major mineral in the Earth's crust: occurs in granites and metamorphosed sediments |
| Plagioclase feldspar $(\text{Na,Ca})(\text{Al,Si})\text{AlSi}_2\text{O}_8$ | Most common mineral in the crust, nearly all igneous and metamorphic rocks |
| Nonsilicates | |
| Oxides and hydroxides | |
| Magnetite Fe_3O_4 | Minor phase in most igneous and metamorphic rocks |
| Hematite Fe_2O_3 | Soils and sediments, banded-iron formations |
| Goethite FeOOH | Soils and sediments |
| Sulfides | |
| Pyrite FeS_2 | Hydrothermal deposits, anoxic sediments |
| Chalcopyrite CuFeS_2 | Hydrothermal deposits |
| Sphalerite ZnS | Hydrothermal deposits |
| Galena PbS | Hydrothermal deposits |
| Carbonates | |
| Calcite CaCO_3 | Sedimentary rocks, limestone |
| Aragonite CaCO_3 | Biogenic shells |
| Dolomite $\text{CaMg}(\text{CO}_3)_2$ | Sedimentary rocks, hydrothermal deposits |
| Sulfates | |
| Gypsum $\text{CaSO}_4 \cdot 2\text{H}_2\text{O}$ | Evaporite deposits |
| Barite BaSO_4 | Evaporites, hydrothermals deposits |
| Phosphates | |
| Apatite $\text{Ca}_5(\text{PO}_4)_3(\text{F,Cl,OH})$ | Minor mineral in igneous, sedimentary, and metamorphic rocks: biogenic (bone and teeth). |
| Halides | |
| Halite NaCl | Evaporite deposits |
| Fluorite CaF_2 | Hydrothermal veins |
| Tungstates | |
| Scheelite CaWO_4 | Granite pegmatites, high-temperature hydrothermal deposits |
| Native elements | |
| Diamond C | Kimberlites |
| Gold Au | Hydrothermal and placer deposits |

known minerals. Most of these are very rare, and only a few minerals are common enough to be geologically significant. As developed from the scheme of J. D. Dana (1848), minerals are viewed as nominally ionic compounds (salts) and are classified by the nature of the anion (see **table**). The most abundant elements in the Earth's crust are oxygen (O), silicon (Si), and aluminum (Al). These elements combine to form a variety of anionic units based on SiO_4 and AlO_4 tetrahedra. Such aluminosilicate minerals are, by far, the most volumetrically important in the Earth's crust. The silicate minerals are classified by the degree of polymerization of the $(\text{Si,Al})\text{O}_4$ tetrahedra. The SiO_4^{4-} tetrahedra can polymerize by sharing oxygen atoms to form $[\text{Si}_2\text{O}_6^{6-}]$ dimers (sorosilicates), $[(\text{Si}_2\text{O}_6)^{4-}]_n$ single chains, $[(\text{Si}_8\text{O}_{22}(\text{OH})_2^{4-})_n]$ double chains, $[(\text{Al,Si})_4\text{O}_{10}(\text{OH})_2^{6-}]_n$ sheets, and three-dimensional $[\text{Si}_{2-n}\text{Al}_n\text{O}_4^{n-}]$ frameworks. See SILICATE MINERALS.

There are many nonsilicate minerals based on anions such as sulfide, oxide, hydroxide, sulfate, and carbonate, but by volume they are not as significant as silicates in the Earth's crust. Nearly all important ore minerals are nonsilicates. Moreover, the second most abundant mineral in the Earth's lower mantle is the oxide ferroperricline $[(\text{Mg,Fe})\text{O}]$.

Thermodynamics and phase transformations. The variety of minerals found in the Earth reflects not just the chemical heterogeneity of crustal rocks but also the variations in pressure and temperature. Minerals formed near the Earth's surface (for example, in sedimentary rocks) will transform to other phases when subjected to the high pressures and temperatures accompanying the collision of crustal terrains by plate tectonics. With increasing pressure, a mineral will tend to transform to a denser phase that is characterized by an atomic structure with higher coordination numbers and increased edge sharing

of coordination polyhedra. With increasing temperature, a mineral will transform to another phase with a structure that has greater entropy. This can result from increased disorder of cations over the crystallographic sites or by adopting a structure with higher symmetry. Consequently, many minerals occur in several different polymorphs that form at different pressures and temperatures. For example, kyanite, andalusite, and sillimanite are all polymorphs of Al_2SiO_5 . Phase transformations of minerals can be sluggish. Consequently, the high pressure–high temperature phases that formed deep in the crust or mantle are usually preserved when rocks are brought to the surface during the uplift or erosion of continental crust. The phase transformations that minerals in the Earth's crust undergo are invaluable for the geologist since they provide clues about the temperature and pressure history of a rock. The structural layers of the Earth's deep interior (upper mantle, transition zone, and lower mantle) appear to be defined, at least in part, by pressure-induced phase transformations of the major minerals. The mineral olivine $[(\text{Mg},\text{Fe})_2\text{SiO}_4]$ transforms to the polymorphs wadsleyite and ringwoodite at high pressure (13.4–23.8 GPa). Because olivine is the dominant mineral in the upper mantle, these phase transformations result in seismic discontinuities at about 400 and 600 km (250 and 370 mi) depth and separate the upper mantle from the transition zone. A further transformation of $(\text{Mg},\text{Fe})_2\text{SiO}_4$ into ferropervicite $[(\text{Mg},\text{Fe})\text{O}]$ and silicate perovskite $[(\text{Mg},\text{Fe})\text{SiO}_3]$ occurs at about 660 km (410 mi) depth and defines the onset of the lower mantle. *See* ANDALUSITE; EARTH INTERIOR; KYANITE; OLIVINE; PHASE TRANSITIONS; POLYMORPHISM (CRYSTALLOGRAPHY); SILLIMANITE.

Minerals also react with each other, and at chemical equilibrium the number of minerals that can coexist in a rock is limited by the phase rule. The phase rule states that the number of degrees of freedom (such as temperature and pressure) with which an assemblage of minerals can coexist is given by $f = c - p + 2$, where c is the number of chemical components and p is the number of phases (each mineral being a separate phase). If a mineral assemblage is to be stable over a range of temperature and pressure, $f = 2$, so that $c = p$. Hence, the number of minerals in a rock is limited by the number of independent chemical components. Very rarely, rocks will be found where $p > c$ (or $f < 2$). If the mineral assemblages in such rocks are really at chemical equilibrium, the rock must have formed along the pressure-temperature boundary that separates one phase assemblage from another. For a metamorphic geologist, a map of where such rocks occur can be used to reconstruct pressure-temperature regimes during mountain-building episodes.

Occurrence and formation. Igneous rocks are those that form from a molten state. Partial melting of upper mantle rocks (peridotite) will yield a melt with a basaltic composition. As this melt cools, minerals such as olivine $[(\text{Mg},\text{Fe})_2\text{SiO}_4]$ and plagioclase feldspar $[(\text{Ca},\text{Na})(\text{AlSi})\text{AlSi}_2\text{O}_8]$ form. If the crys-

tals that form can be separated from the melt, the melt composition will evolve to become more silica rich, and from it minerals such as amphibole, mica, and potassium feldspar may form. The compositional evolution of igneous melts was noted early on in Bowen's reaction series. An interesting system often develops as felsic (granitic) magmas evolve. As the magma solidifies, a fluid-rich phase will remain in which many exotic elements will be concentrated. The high fluid content favors the growth of large crystals. Gem minerals such as tourmaline can be found in pegmatites. *See* AMPHIBOLE; BASALT; FELDSPAR; GEM; IGNEOUS ROCKS; MAGMA; MICA; PEGMATITE; PERIDOTITE; PETROLOGY; TOURMALINE.

When rocks are exposed to the Earth's surface, minerals are broken down by reactions with water and carbon dioxide. These weathering reactions give rise to clay minerals such as montmorillonite and kaolinite. These minerals always occur as colloids and cannot form large crystals. Because of their small particle sizes, clay minerals have very high surface areas. This allows them to play an important role in controlling the concentrations of dissolved ions in seawater and soil solutions. *See* CLAY MINERALS; COLLOID; KAOLINITE; MONTMORILLONITE; SEDIMENTARY ROCKS; WEATHERING PROCESSES.

As crustal plates collide, rocks are slowly subjected to high pressures and temperatures. The original mineral assemblage in a rock will reequilibrate to form new minerals that are stable at the pressures and temperatures of the metamorphic environment. *See* METAMORPHIC ROCKS; METAMORPHISM; PLATE TECTONICS.

Some minerals, especially sulfides, are deposited by hydrothermal solutions. These are usually H_2S - NaCl -rich fluids that are able to dissolve metals such as lead (Pb), zinc (Zn), and copper (Cu) by the formation of metal-chloride complexes. Hydrothermal fluids may have originated by exsolution from magma, or they may result from seawater and ground water that has circulated in a geothermal regime. As these fluids react with minerals such as feldspar (KAlSi_3O_8) or calcite (CaCO_3) the pH rises and metal sulfides will precipitate from solution either in vein deposits or as disseminated sulfide crystals in the host rock. Subsequent alteration of primary sulfide minerals by oxidized acidic ground water will yield a complex variety of sulfate and oxide minerals. Percolating ground water may also dissolve disseminated sulfides in granitic host rocks and reprecipitate new sulfides at the water table boundary by a process known as supergene enrichment. *See* HYDROTHERMAL ORE DEPOSITS; ORE AND MINERAL DEPOSITS; SULFIDE AND ARSENIDE MINERALS.

Relatively soluble minerals such as halite (NaCl) and gypsum ($\text{CaSO}_4 \cdot 2\text{H}_2\text{O}$) can form in evaporite deposits. Marine evaporates result when seawater is trapped in isolated basins. Smaller evaporite deposits can form on continents when drainage basins feed into extremely arid environments. Evaporation of such waters yields unusual minerals such as nitrates and borates. *See* BORATE MINERALS; GYPSUM; HALITE; NITRATE MINERALS; SALINE EVAPORITES.

Some minerals are formed by biological processes. The most important is aragonite (CaCO_3), which is formed by mollusks and planktonic organisms such as foraminifera and coccolithophorids. Aragonite dissolves and reprecipitates as the more stable polymorph calcite (CaCO_3) which makes up the vast deposits of limestone, a major rock in the Earth's crust. Iron(III) and manganese(IV) oxides are often formed by Iron- and manganese-oxidizing bacteria. Other bacteria use these oxides, along with sulfate minerals, as electron acceptors during respiration, leading to the formation of various iron(II) oxide and sulfide minerals such as pyrite (FeS_2). The mineral apatite [$\text{Ca}_5(\text{PO}_4)_3(\text{F},\text{Cl},\text{OH})$] is the inorganic phase in bone and teeth. See APATITE; ARAGONITE; CALCITE; LIMESTONE.

Properties. The physical and chemical properties of minerals can be used for their identification. A number of classical properties can be used by amateur collectors and geologists in the field, including color, streak, luster, habit, hardness, and specific gravity (density). The external crystallographic properties of a mineral reflect the symmetry of the internal structural arrangement of atoms. In the laboratory, the most diagnostic means of identifying minerals is by x-ray diffraction and chemical analysis. The optical properties of minerals are very useful for identifying silicate minerals. These properties can be observed when rocks are cut in thin section (35 micrometers) and viewed through transmitted polarized light in a petrographic microscope. See PETROGRAPHY; X-RAY DIFFRACTION.

The physical properties of minerals are of fundamental interest insofar as they determine the physical properties of the Earth. The elastic properties of minerals determine the velocities of seismic waves generated by earthquakes. Inversely, seismic velocities are used to infer the elastic properties and mineralogy of the Earth's interior. Most silicates are electrical insulators; however, most sulfides and transition-metal oxides are semiconductors. Their presence in geological bodies can be detected by measuring the conductivity of rocks in the field. This is an important tool for prospecting for sulfide ore bodies. A few minerals are ferrimagnetic. These minerals can record the direction of the Earth's magnetic field that was present when the minerals formed. Such paleomagnetic signatures were an essential line of evidence for continental drift and plate tectonics. See PALEOMAGNETISM; ROCK, ELECTRICAL PROPERTIES OF; ROCK MAGNETISM; SEISMOLOGY.

Economic aspects. A few minerals are used as gemstones, and this was the probable first use of minerals (as opposed to mineraloids such as flint) in the Paleolithic. Since the Bronze Age, minerals have provided the raw, natural material from which we extract useful metals. Most economically important metals (such as copper, lead, and zinc) occur as sulfide minerals. Iron ore minerals are oxides such as hematite (Fe_2O_3) and magnetite (Fe_3O_4). The primary ores of aluminum are oxides and oxide hydroxide minerals [for example, diaspore (AlOOH)]. See MAGNETITE; OXIDE AND HYDROXIDE MINERALS.

Industrial minerals are those that are economically useful in their natural state. Clay minerals such as kaolinite [$\text{Al}_2\text{Si}_2\text{O}_5(\text{OH})_4$] are used to make china, pottery, bricks, and tile. Gypsum [$\text{CaSO}_4 \cdot 2\text{H}_2\text{O}$] is the main component of plaster and plasterboard. Corundum (Al_2O_3) is a common abrasive. Because of its fibrous habit, chrysotile [$\text{Mg}_3\text{Si}_2\text{O}_5(\text{OH})_4$] is the main mineral of asbestos. The mineral talc [$\text{Mg}_3\text{Si}_4\text{O}_{10}(\text{OH})_2$] is a lubricant owing to its layer structure held together by weak van der Waals bonds. See ASBESTOS; MINERALOGY; TALC.

David M. Sherman
Bibliography. H. Blatt et al., *Petrology: Igneous, Sedimentary, and Metamorphic*, W. H. Freeman, 3d ed., 2005; C. Klein, *Manual of Mineral Science*, 22d ed., Wiley, 2001; R. V. Gaines et al., *Dana's New Mineralogy: The System of Mineralogy of James Dwight Dana and Edward Salisbury Dana*, Wiley-Interscience, 8th ed., 1997; H. Paquet et al. (eds.), *Soils and Sediments: Mineralogy and Geochemistry*, Springer, 1998.

Mineral resources

Mineral deposits, including ore bodies and potential ore, that are presently recoverable and may be so in the future. Much of what is consumed in modern society has to be obtained from mineral extraction. Countless items, whether coffee cups, camera film, electronic equipment, glasses, motor vehicles, or jewelry, contain minerals that must be extracted from the earth.

Types. Mineral resources generally are categorized into energy and nonfuel resources. Energy resources can be in solid (coal, lignite, uranium) or gaseous and liquid (natural gas, petroleum) form (see **table**). The many uses of these resources have evolved over time. Some resources such as gold have been used from the time of the earliest societies, while other minerals used in products from automobiles to missiles, although also available, have been needed only in recent times. See COAL MINING; ENERGY SOURCES.

The demand for minerals remains dependent on technological change and the perceived need for the final product. For example, a decline in the need for missiles would decrease the demand for certain minerals such as aluminum, as would the increased recycling of aluminum cans. Eliminating missiles or replacing aluminum cans with glass containers would have an even more drastic impact on the need to produce aluminum products. For mineral resources, the end-product demand rather than supply determines whether or not they are extracted from the earth.

Sources. Minerals may or may not be easy to find. Some parts of the world have been well explored for these resources, while others have not. Mineral resources can be located in politically stable or unstable countries. Minerals deemed vital for military production and defense purposes have been called strategic minerals. When the minerals are located largely in politically unstable or unfriendly countries, the United States has responded by setting up

| Classification of selected metals and minerals and some some of their major uses | |
|--|---|
| Materials | Uses |
| <i>Fuels</i> | |
| Bituminous and anthracite coal | Direct fuel, electricity, gas, chemicals |
| Lignite | Electricity, gas, chemicals |
| Petroleum | Gasoline, heating, chemicals plastics |
| Natural gas | Fuel, chemicals |
| Uranium | Nuclear power, explosives |
| <i>Metals</i> | |
| <i>Ferrous metals</i> | |
| Chromium | Alloys, stainless steels, refractories, chemicals |
| Cobalt | Alloys, permanent magnets, carbides |
| Columbium (niobium) | Alloys, stainless steels |
| Iron | Steels, cast iron |
| Manganese | Scavenger in steelmaking, batteries |
| Molybdenum | Alloys |
| Nickel | Alloys, stainless steels, coinage |
| Tungsten | Alloys |
| Vanadium | Alloys |
| <i>Nonferrous metals</i> | |
| Copper | Electrical conductors, coinage |
| Lead | Batteries, gasoline, construction |
| Tin | Tinplate, solder |
| Zinc | Galvanizing, die casting, chemicals |
| <i>Light metals</i> | |
| Aluminum | Transportation, rockets, building materials |
| Beryllium | Copper alloys, atomic energy field |
| Magnesium | Building materials, refractories |
| Titanium | Pigments, construction, acid-resistant plumbing |
| Zirconium | Alloys, chemicals, refractories |
| <i>Precious metals</i> | |
| Gold | Monetary, jewelry, dental, electronics |
| Platinum metals | Chemistry, catalysts, automotive |
| Silver | Photography, electronics, jewelry |
| <i>Industrial</i> | |
| Asbestos | Insulation, textiles |
| Boron | Glass, ceramics, propellants |
| Clays | Ceramics, filters, absorbents |
| Corundum | Abrasives |
| Feldspar | Ceramics, fluxes |
| Fluorspar | Fluxes, refrigerants, acid |
| Phosphates | Fertilizers, chemicals |
| Potassium salts | Fertilizers, chemicals |
| Salt | Chemicals, foods, glass, metallurgy |
| Sulfur | Fertilizers, acid, metallurgy, paper, foods, textiles |

strategic stockpiles of minerals in case the supply becomes disrupted.

In the United States the use of stockpiles has ranged from storing petroleum in underground salt domes along the Louisiana and Texas coast to maintaining hoards of strategic alloys or even silver in government warehouses. The need for such stockpiles diminishes as the availability of the supply increases or, in the case of strategic minerals, as the threat decreases. For example, the collapse of the Soviet Union (and its diminished threat to the United States) decreases the need for such minerals and their stockpiles. The increase in the number of nations with more democratic governments from South Africa to Chile also provides more stability to the provision of minerals in a worldwide trading system.

Minerals are found by various exploration techniques. Traditionally these techniques have been based on geological theories, on field surveys, and on drilling in test sites. However, remote sensing and analysis of satellite imagery have become more important in prospecting for minerals. *See* GEOPHYSICAL EXPLORATION; PROSPECTING; REMOTE SENSING; SCIENTIFIC AND APPLICATIONS SATELLITES.

Mineral reserves. There have been predictions since before World War II that the United States will not have access to necessary mineral resources during major wars. During the 1960s, another major debate raged on the limits to growth and the depletion of world resources. During the 1970s, prices on a wide range of minerals rose rapidly.

The fear of insufficient resources keeps recurring, because there are not good estimates of how much of any mineral is in reserve throughout the world, or how and when the demand will increase. The uncertainty is over how much there is, and how much can be recovered given current costs and technology.

Estimates of resources include those proven to exist, those that can be recovered, and probable, possible, and potential reserves. The boundaries between these categories can be quite fluid. New discoveries or changes in world politics can dramatically change the availability of resources. New technologies can make feasible the extraction of minerals previously considered too costly to develop.

Development and governmental considerations. The minerals industry has an image and history of disturbing the landscape and creating environmental problems. The image of strip-mining large tracts of land to expose coal seams is not accurate, since most hard-rock mining is done below the surface and disturbs only a small portion of the landscape. Environmental problems associated with such mining have been contamination of ground water and waterways, the disposal of mining wastes, and air pollution from smelting operations. *See* AIR POLLUTION; WATER POLLUTION.

The burden for controlling the negative impacts from mining and mineral processing falls on local, state, and federal governments to regulate the environmental costs resulting from mineral development. In the United States this has been done at the federal level by the passage of environmental and reclamation laws restricting the amount and type of pollution and requiring the restoration of lands to near-original conditions. Certain public lands such as federal wilderness areas have been put off limits to future mineral exploration and development. *See* LAND RECLAMATION.

The danger is that countries with less developed governmental controls will permit environmental degradation in order to foster economic development. The potential for environmental destruction is heightened by the economic restructuring in the minerals industry in the more developed countries such as the United States.

Economic restructuring conservation. The minerals industry in the United States has been in decline and

a period of consolidation because of a decrease in the extraction of resources and in their processing. Much of this activity has moved to countries that are less developed, with lower labor costs, and less stringent or laxly enforced environmental laws and regulation. Instead of importing extracted mineral resources, processing them, and then exporting the products, the United States increasingly imports the processed minerals.

The demand for minerals will increase with world population growth and as other countries become more consumer oriented. The increasing pressures on the world environment will magnify in intensity. This will require more personal efforts and government policies to foster conservation of resources and environmentally sound extraction and processing of minerals. Without such efforts, mineral resource development ultimately will contribute to the decline of quality of life. *See* CONSERVATION OF RESOURCES.

Gundars Rudzitis

Bibliography. S. L. Cutter, H. L. Renwick, and W. H. Renwick, *Exploitation, Conservation, Preservation: A Geographic Perspective on Natural Resource Use*, 3d ed., 1998; H. E. Johansen, O. P. Matthews, and G. Rudzitis (eds.), *Mineral Resource Development: Geopolitics, Economics, and Policy*, 1987; B. O'hUallachain and R. A. Matthews, Economic restructuring in primary industries: Transaction costs and corporate vertical in the Arizona copper industry, 1980–1991, *Ann. Ass. Amer. Geog.*, 84(3):399–417, September 1994.

Mineralogy

The study of the crystalline phases (chemical compounds or pure elements) that make up the Earth and other rocky bodies in the solar system—that is, the terrestrial planets (Mercury, Venus, and Mars), meteorites, asteroids, and planetary satellites. In universities, nearly all mineralogists are found in earth science or geology departments.

History. Before the twentieth century, the science of mineralogy was concerned with identifying new minerals and determining their chemical compositions and physical properties. Investigations of the chemical compositions of minerals played a central role in the early development of chemistry and atomic theory. One of the most striking aspects of minerals is that they often occur as well-formed crystals. This motivated the hypothesis that the external symmetry reflected an internal symmetrical arrangement of atoms. The development of x-ray diffraction in 1915 enabled mineralogists to determine the internal crystal structures of minerals (that is, how the atoms are arranged). Once the chemical compositions and crystal structures of minerals were known, it became possible to systematically determine the thermodynamic properties of minerals and understand their stability as a function of pressure, temperature, and composition. This enabled geologists to understand the stabilities of mineral assemblages and the origins of the different kinds of igneous and

metamorphic rocks. *See* CRYSTAL STRUCTURE; GEOLOGY; MINERAL; PETROLOGY; X-RAY DIFFRACTION.

By the 1960s, the crystal structures of most of the rock-forming minerals had been determined by x-ray diffraction. New techniques, however, were used in pursuit of a deeper understanding of the structural chemistry of minerals. Transmission electron microscopy, developed in 1961, has revealed the nanoscale (10–1000 Å) structures of minerals that result from twinning, dislocations, and phase transformations. Starting in the 1960s, a variety of spectroscopic methods (such as Mössbauer, nuclear magnetic resonance, infrared, and Raman) were used to understand how atoms are ordered among the crystallographic sites in minerals and to identify the nature of defects and impurities. *See* ELECTRON MICROSCOPE; SPECTROSCOPY.

With the advent of increasingly powerful computers, mineralogists began investigating the structures and properties of minerals using methods in computational chemistry. In the 1980s and 1990s, these calculations were largely based on classical models for interatomic interactions. In recent years, calculations of mineral stabilities and properties have been done using first-principles calculations based on either the Hartree-Fock approximation or density functional theory. *See* COMPUTATIONAL CHEMISTRY; QUANTUM CHEMISTRY.

Current research areas. Today, mineralogy is concerned with understanding the physics and chemistry of minerals insofar as they control processes in geochemistry and geophysics. Indeed, the boundaries between mineralogy, petrology, and geochemistry have largely disappeared. *See* GEOCHEMISTRY; GEOPHYSICS.

Phase equilibria and thermodynamic properties of minerals. The important minerals that occur deep in the Earth's crust have been discovered; however, their thermodynamic properties are often unknown. Laboratory experiments using high-pressure apparatus (the multianvil press and the diamond-anvil cell) can be used to determine the stable phases at the conditions of the Earth's deep interior. Laboratory experiments show that, in the extreme pressures of the Earth's lower mantle, the dominant mineral phases appear to be (Mg,Fe)SiO₃ and CaSiO₃ with a perovskite structure, along with (Mg,Fe)O with the rock-salt structure. There are still many other minerals, composed of minor chemical components, which may also be stable in the Earth's mantle. Although these phases may not contribute much to the seismic profile of the Earth, they may be geochemically important as reservoirs of trace components used to understand the chemical evolution of the Earth. The crystal structures of minerals allow a variety of solid solutions and order–disorder transformations. As we understand the thermodynamics and kinetics of such processes, we can develop tools to infer the temperatures and pressures at which a rock formed. Order–disorder transformations can give clues on the rate at which rocks cooled after their formation. These geothermometers, geobarometers, and geospeedometers are invaluable tools for geologists

seeking to unravel the processes of mountain building and metamorphism. See GEOLOGIC THERMOMETRY; HIGH-PRESSURE MINERAL SYNTHESIS; METAMORPHISM; OROGENY.

Physical properties of minerals at high pressure and temperature: understanding the composition of the Earth from seismology. A major goal in the earth sciences is to determine the chemical composition of the Earth. An important constraint is provided by seismic data on the density and seismic velocities of the Earth as a function of depth. To derive compositional information from such data, we need to know what mineral phases are stable at high pressures and temperatures and their densities and elastic properties. The multianvil press and the diamond-anvil cells also allow in-situ measurements using x-ray diffraction of minerals confined at high pressure and heated to high temperature. From such experiments, we can determine structural changes and densities of minerals under the conditions of the deep Earth. Elastic properties can be measured using Brillouin scattering; because such experiments are difficult, theoretical calculations based on first principles (quantum mechanics) are also being used. See SEISMOLOGY.

Water in nominally anhydrous minerals at high pressure. Some high-pressure minerals such as wadsleyite, $\beta\text{-Mg}_2\text{SiO}_4$, have defect structures that allow some oxygen atoms to be replaced by OH. Many investigators are proposing that a significant amount of water is chemically stored in the Earth's mantle by this mechanism.

Colloid and interface chemistry of minerals: controls on aqueous and environmental geochemistry. As rocks weather, the minerals that formed at high temperature will dissolve and reprecipitate to form a variety of clay minerals along with several oxide and hydroxide minerals. These secondary minerals are important phases in soils and sediments. Because they occur as colloidal-sized (nanocrystalline) particles, they have a very high surface area. The electrostatic charge of the mineral surfaces promotes the sorption of dissolved ions from aqueous solutions. Consequently, clay minerals and authigenic oxide and hydroxide minerals exert major controls on the chemistry of natural waters. In particular, chemical processes at the mineral-water interface control the fate of toxic metals in soil and ground water and the availability of important micronutrients in the oceans. Understanding these processes at a molecular level is an active area of research. The interfacial properties of sulfide minerals are investigated for the development of flotation techniques to separate ore minerals from their host rocks. Reactions at the surfaces of sulfide minerals also are responsible for the formation of acid mine drainage. A variety of analytical techniques are being used to investigate the mineral-water interface. The availability of synchrotron radiation sources has enabled mineralogists to characterize the coordination chemistry at the mineral-water interface using a variety of spectroscopic techniques. Scanning tunneling microscopy and atomic force microscopy are unraveling the atomic structures of mineral surfaces. See AUTHIGENIC MINERALS; CLAY MIN-

ERALS; OXIDE AND HYDROXIDE MINERALS; SCANNING TUNNELING MICROSCOPE; SULFIDE AND ARSENID MINERALS.

Biom mineralization. Current research is investigating the mechanisms by which organisms form inorganic minerals. Biom mineralization processes may provide important clues to Earth's history: variations in trace metals and isotopic compositions are found in the shells of marine planktonic organisms, such as foraminifera. It is believed that these proxies may be giving a record of past seawater temperatures.

David M. Sherman

Bibliography. C. Klein, *Manual of Mineral Science*, 22d ed., Wiley, 2001; S. Mann, *Biom mineralization*, Oxford University Press, 2002; W. D. Nesse, *Introduction to Mineralogy*, Oxford University Press, 1999.

Minimal principles

In the treatment of physical phenomena, it can sometimes be shown that, of all the processes or conditions which might occur, the ones actually occurring are those for which some characteristic physical quantity assumes a minimum value. These processes or conditions are known as minimal principles. The application of minimal principles provides a powerful method of attacking certain problems that would otherwise prove formidable if approached directly from first principles.

One simple minimal principle asserts that the state of stable equilibrium of any mechanical system is the state for which the potential energy is a minimum. Other general theorems of classical dynamics that are related to minimal principles are Hamilton's principle and the principle of least action. See HAMILTON'S PRINCIPLE; LEAST-ACTION PRINCIPLE.

Minimal principles are important in branches of physics other than mechanics. Fermat's principle in optics, for example, states that, of all possible paths of light transmission between two points, the actual path is the path for which the transmission time is a minimum. Minimal principles also find wide application in thermodynamics. See CHARGED PARTICLE OPTICS.

Dudley Williams

Bibliography. H. Goldstein, C. E. Poole, and J. L. Safko, *Classical Mechanics*, 3d ed., 2002; B. A. Kuperschmidt, *The Variational Principles of Dynamics*, 1992; C. Lanczos, *The Variational Principles of Mechanics*, 1970, reprint 1986.

Minimal surfaces

A branch of mathematics belonging to the calculus of variations, differential geometry, and geometric measure theory. A surface, interface, or membrane is called minimal when it has assumed a geometric configuration of least area among those configurations into which it can readily deform. Soap films spanning wire frames or compound soap bubbles

enclosing volumes of trapped air are common examples. See CALCULUS OF VARIATIONS; DIFFERENTIAL GEOMETRY; MEASURE THEORY.

Minimal surface equation. If a surface S in xyz space is represented as the graph of a function $z = f(x, y)$ over a region A in the xy plane, the area of S can be computed by formula (1), and the mean curvature of S at a point can be computed by formula (2).

$$\int \int_A \sqrt{1 + f_x^2 + f_y^2} \, dx \, dy \quad (1)$$

$$\frac{(1 + f_y^2)f_{xx} - 2f_x f_y f_{xy} + (1 + f_x^2)f_{yy}}{2(1 + f_x^2 + f_y^2)^{3/2}} \quad (2)$$

Geometrically, the mean curvature of a surface at a point is the difference between the maximum upward curvature there and the maximum downward curvature; in particular, a surface of zero mean curvature has such principal curvatures equal and opposite and hence typically appears “saddle-shaped.” It turns out that S is a minimal surface; that is, it cannot be perturbed to less area leaving its boundary fixed, provided the mean curvature is zero at each of its points; such a surface could occur, for example, as a soap film spanning a wire frame. The corresponding minimal surface equation is partial differential equation (3). If, alternatively, S were part of a

$$0 = (1 + f_y^2)f_{xx} - 2f_x f_y f_{xy} + (1 + f_x^2)f_{yy} \quad (3)$$

soap bubble enclosing trapped air, the mean curvature of S would be proportional to the difference in air pressure between the two sides of S . In the calculus of variations, typically area-minimizing properties of minimal surfaces are emphasized. In differential geometry, minimal surfaces are defined as surfaces of zero mean curvature; surfaces of constant mean curvature are also extensively studied.

Historical development. The minimal surface equation was first written down by J. L. Lagrange in 1762. A list of problems about minimal surfaces posed by J. D. Gergonne in 1816 stimulated mathematical interest in the subject. By the second half of the nineteenth century, many noted mathematicians were studying minimal surfaces; new minimal surfaces were found, and it was discovered that any piece of minimal surface could be parametrized by utilizing a suitable pair of complex functions. In 1873 the physicist J. Plateau published a treatise on the geometry of the interfaces of liquids in equilibrium; in his honor a collection of mathematical problems about the existence of area-minimizing surfaces bear the name Plateau’s problem.

The first “solution” to Plateau’s problem was given in 1931 by J. Douglas, who showed that every closed curve in space is the boundary of some minimal surface (formally, a harmonic mapping from a disk in the xy plane into xyz space). A decade of renewed mathematical activity followed.

Alternative formulations of Plateau’s problem were initiated in 1960, based on geometric measure

theory. Several different kinds of mathematical minimal surfaces were invented, including ones resembling actual soap films and compound soap bubbles; these surfaces carry names such as integral current, flat chain, and varifold.

Soap bubble geometry. The possible configurations of soap films and soap bubbles are governed by elementary rules. Any such configuration consists of flat or smoothly curved pieces of surface (each of constant mean curvature) smoothly joined together. Furthermore, these surfaces meet in only two ways; either three surfaces meet along a smooth curve, or six surfaces (together with four curves) meet at a vertex. Finally, when surfaces meet along curves or when curves and surfaces meet at points, they do so at equal angles; in particular, when three surfaces meet along a curve, they do so at angles of 120° , and when four curves meet at a point, they do so at angles of close to 109° . These rules, first formulated by Plateau from his observations, are now known to be a necessary mathematical consequence of area minimization alone.

Minimal surface forms in nature. The two-dimensional surface is dominant in determining shape whenever the energy of a system is changed significantly by a displacement or a change in area of the surface. Such surfaces include the interfaces between crystals in a typical rock or metal, the film of soapy water between the air cells in a soap froth, the membrane separating the cells in living tissue, and the cracks separating basalt columns. Minimization of surface area plays a role in determining the shape of many living organisms. See FOAM.

Minimal surfaces in art and design. Soap bubbles have appeared in the works of a number of classical painters, and minimal surfaces have influenced modern artists and sculptors. The architect F. Otto has used soap film surfaces as models for a number of his dramatic roof structures, and the designer P. Pearce has constructed elaborate playground labyrinths inspired by periodic minimal surfaces.

Directions of research. Minimal surfaces are being studied extensively. They provide a tool for understanding the geometry of curved spaces and of curved space-time. The study of higher-dimensional minimal surfaces in higher-dimensional spaces is leading to a wealth of new geometric discoveries and constructions. Methods are also being developed and improved for the computation of minimal surfaces. There is increasing understanding of the relationships between minimal surfaces and their boundary curves; with probability one, for example, a randomly chosen closed curve bounds exactly one surface of least area. Finally, there is a continuing interest in understanding relationships between minimal surfaces and their generalizations and forms of the natural world. Frederick J. Almgren, Jr.

Bibliography. U. Dierkes, *Minimal Surfaces I: Boundary Value Problems*, 1992; A. T. Fomenko and A. Tuzhilin, *Elements of the Geometry and Topology of Minimal Surfaces in Three-Dimensional Space*, 1991; J. C. Nitsche, *Introduction to Minimal Surfaces*, vol. 1, 1989.

Mining

The taking of minerals from the earth, including production from surface waters and from wells. Usually the oil and gas industries are regarded as separate from the mining industry. The term mining industry commonly includes such functions as exploration, mineral separation, hydrometallurgy, electrolytic reduction, and smelting and refining, even though these are not actually mining operations. *See* HYDROMETALLURGY; METALLURGY; ORE DRESSING.

Methods. Mining is broadly divided into three basic methods: opencast, underground, and fluid mining. Opencast mining is done either from pits or gouged-out slopes or by surface mining, which involves extraction from a series of successive parallel trenches. Dredging is a type of surface mining, with digging done from barges. Hydraulic mining uses jets of water to excavate material.

Underground mining involves extraction from beneath the surface, from depths as great as 10,000 ft (3 km), by any of several methods.

Fluid mining is extraction from natural brines, lakes, oceans, or underground waters; from solutions made by dissolving underground materials and pumping to the surface; from underground oil or gas pools; by melting underground material with hot water and pumping to the surface; or by driving material from well to well by gas drive, water drive, or combustion. Most fluid mining is done by wells. In one experimental type of well mining, insoluble material is washed loose by underground jets and the slurry is pumped to the surface. *See* COAL MINING; OPEN-PIT MINING; PETROLEUM ENGINEERING; PLACER MINING; SOLUTION MINING; SURFACE MINING; UNDERGROUND MINING.

The activities of the mining industry begin with exploration, which, since accidental discoveries or superficially exposed deposits are no longer sufficient, has become a complicated, expensive, and highly technical task. After suitable deposits have been found and their worth proved, development, or preparation for mining, is necessary. For opencast mining, this involves stripping off overburden; and for underground mining, the sinking of shafts, driving of adits and various other underground openings, and providing for drainage and ventilation. For mining by wells, drilling must be done. For all these cases, equipment must be provided for such purposes as blasthole drilling, blasting, loading, transporting, hoisting, power transmission, pumping, ventilation, storage, or casing and connecting wells. Mines may ship their crude products directly to reduction plants, refiners, or consumers, but commonly, concentrating mills are provided to separate useful from useless (gangue) minerals. *See* PROSPECTING.

Depletable resources. A unique feature of mining is the circumstance that mineral deposits undergoing extraction are "wasting assets," meaning that they are not renewable as are other natural resources. This depletable nature of mineral deposits requires that mining companies must periodically find new deposits and constantly improve their technology in order to stay

in business. Depletion means that the supplies of any particular mineral, except those derived from oceanic brine, must be drawn from ever-lower-grade sources. Evan Just

Underground Excavation

In all types of mining, there are at least three major operations: the ore or waste rock is broken from place, the broken material is transported to its final destination, and the opening caused by the excavation is supported, either naturally or artificially. In underground mining, where it is necessary to go below the surface of the earth to mine the ore body, three other operations must be considered: removing the material from the mine; controlling and pumping underground water; and providing a suitable and healthful working environment for the miners. Generally, all of these operations are involved in underground excavation. Depending on circumstances, some of these operations may be much more important than others. For example, in a dry mine with very strong ore and waste, support and pumping would be of minor importance.

There are two general methods used to mine the ore body below the surface of the earth. If the terrain is amenable, it is possible to excavate horizontal openings called adits into the mountain (**Fig. 1a**). However, if the mine is located in relatively flat country or the ore lies deeper than the lowest adit level, it becomes necessary to begin by sinking an opening called a shaft from the surface or adit level into the earth (**Fig. 1b**). This shaft may be vertical or inclined even to some small angle (such as 10°) to the horizontal. Low-angle shafts are frequently called declines.

From the shaft at some predetermined spacing, horizontal openings called levels are excavated toward the ore body to prepare the ore for mining. In most cases, the ore is mined so that gravity will cause the broken ore to fall or at least to be directed toward the level below. It can then be transported to the surface through the adits or shafts. Even though a mine may be started through adits, it is usually necessary to sink a shaft inside the mine if the mine is a success. This is called a *winze* (**Fig. 1a**).

To develop the ore for mining, it is usually necessary to drive openings known as raises from a lower level to a higher level. (**Fig. 1a, b**). The raises provide access for personnel and equipment into the stope (the zone where the ore is actually mined) or to pass ore from the mining area to the level below. The other openings except the stope, shafts, levels, and raises are known as development headings.

There are many different types of stoping methods. In practically all methods, the ore must be broken from place and transported to the level below. In some types of stoping, the working area must be made safe for the personnel. In other types, such as caving methods, people must work in the vicinity of the stope in a safe manner so that the ore can be removed from it. Two common methods of stope preparation for ore removal involve work at the bottom of the stope (**Fig. 1c and d**). Some ore bodies are nearly horizontal and similar to coal deposits. In

this situation, the ore is mined in a similar way to the room-and-pillar coal mining method.

A brief discussion of underground excavation requires some understanding of shafts, levels, raises, and stopes, because the excavation procedures vary in each of these types of openings. See COAL MINING.

Breaking. Ore and waste are broken from place by mechanical means, blasting, or natural forces.

Mechanical means. Mechanical means consist of drilling (augering, percussion, rotary, and abrasion such as with diamonds), dozing or scraping, ripping, chipping, and cutting. Drilling methods are usually employed to place explosives where blasting is necessary. In underground mining, machines designed to auger, scrape, chip, rip, or cut or combinations of these are used to mine softer ores and rocks. These machines are classified as continuous mining machines or long-wall cutting machines. They are used to a great extent in mining coal and some non-metallic minerals such as trona and salt, or other soft minerals.

Blasting. In most development headings such as shafts, levels, and raises, and in many stopes, the solid material is broken by the force of explosives. It is necessary to place the explosive charge in the rock mass where it can break the solid material efficiently. Hence, holes are drilled in the solid material with some predetermined pattern called drill rounds (Fig. 2). The explosives break the material toward a free face.

There is a hole pattern (Fig. 2a) that is commonly used in stopes where the explosive force can break toward a free face, which is either horizontal or vertical depending on the type of stoping operation. The holes are then drilled parallel to the free face. This procedure allows the explosive charge to be placed parallel to the face for maximum efficiency.

In a development heading, there is one free face where the holes are drilled; thus, the holes are positioned so they will either break toward this free face or create another free face. The holes in this pattern are known as the cut (Fig. 2b-e). The cut and hole pattern in a round varies greatly from mine to mine depending on the nature of the ground. As a rule, the maximum depth that a round will break is the shortest dimension of the opening (Fig. 2c-e).

The burn cut (Fig. 2b) is designed to break deeper than the other types of rounds. In the cut, one or more holes are drilled and left unloaded, without an explosive charge. Some holes are left uncharged with explosives. Thus, the unloaded hole is the free face that the loaded hole can break toward. Successive holes then break to this opening. The unloaded hole or holes are often drilled at a larger diameter than the regular loaded holes so that there is a bigger volume in the free face. The depth of burn-cut rounds are generally not limited to the smaller dimensions of the opening.

In all rounds, the order or sequence of detonating the explosive charges, called firing order, is extremely important. The holes nearest the free face must explode first to create and expand the opening (Fig. 2a and e). See EXPLOSIVE.

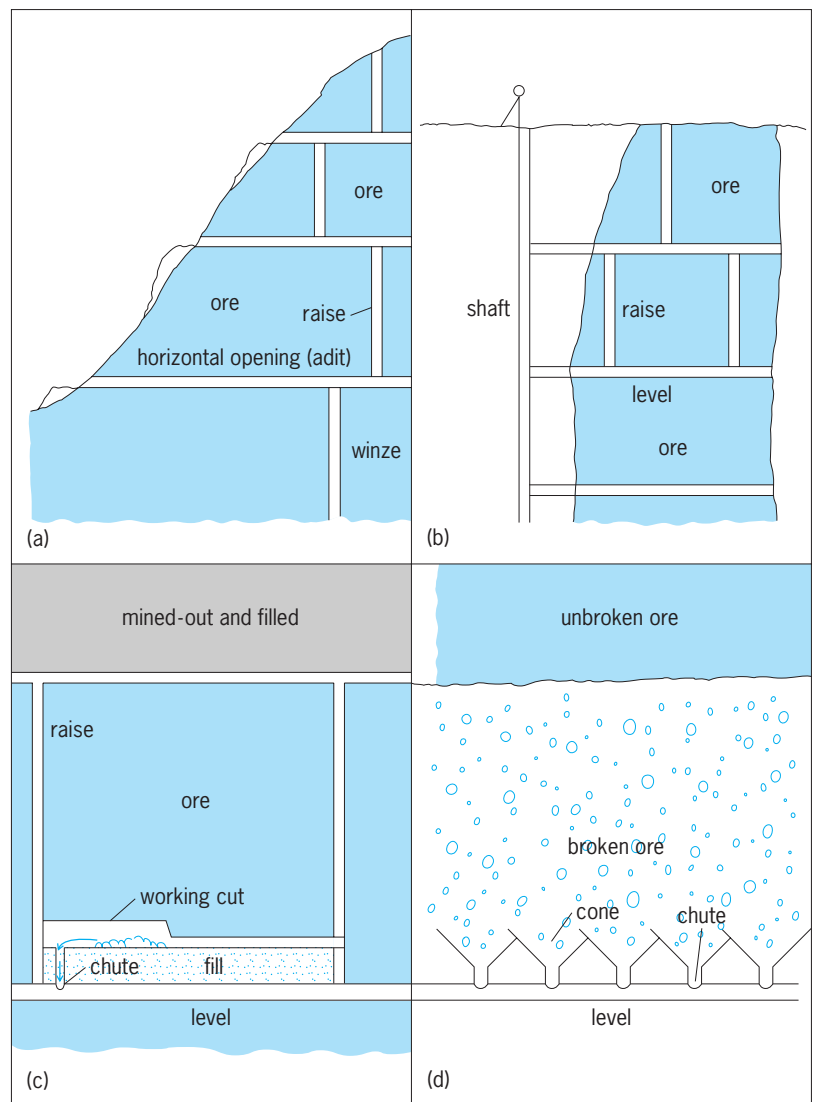


Fig. 1. Long sections of underground mines. (a) Ore body developed by adits. (b) Ore body developed by a shaft. (c) Bottom of stope prepared for horizontal ore movement. (d) Bottom of stope prepared by cone-shaped openings for ore movement.

Natural means. Breaking also occurs by natural means if the material surrounding the opening is not adequately supported. Caving methods take advantage of this phenomenon. The general plan is to remove underlying support from a block of ore, which allows the ore to cave into the opening left by the removed supports. The broken ore is directed into a series of cones at the bottom (Fig. 1d). As the broken ore is removed, an opening below the unbroken ore develops and the ore caves into this opening. As a rule, broken rock expands to about 140% of its original volume; therefore, as the broken ore develops, it fills in the opening, which stops the caving action. As the broken ore is pulled from the bottom, a new opening in the stope occurs and caving starts again. Where applicable, this system is very cost-effective and produces high tonnage rates of ore mined per day.

Loading and transporting. When the ore or waste is broken, it must be moved or transported to some

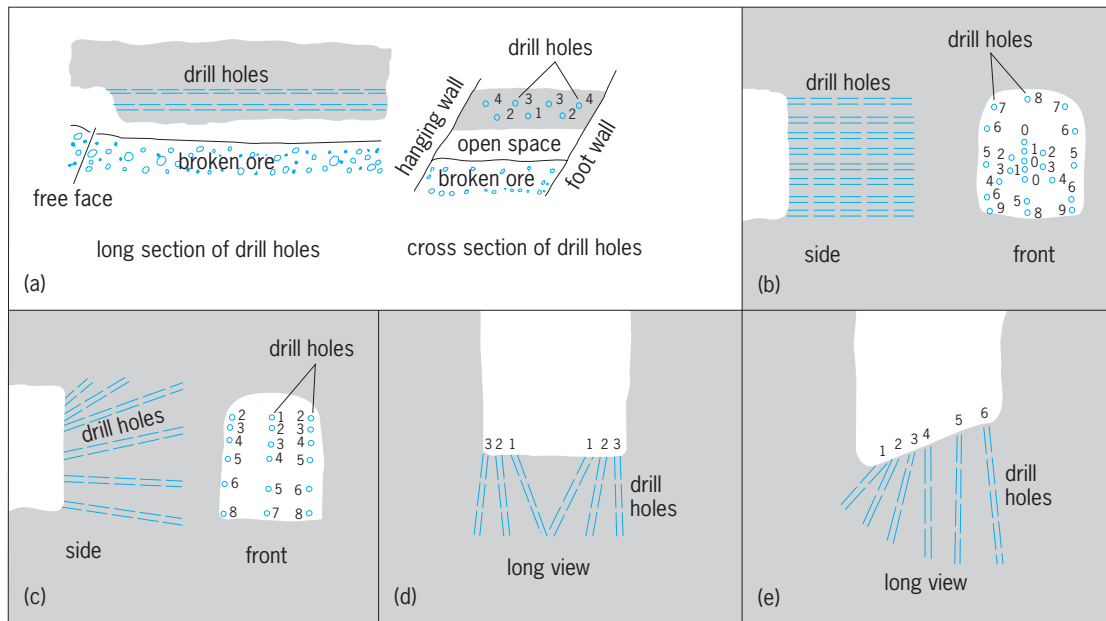


Fig. 2. Long and cross sections of drill rounds. (a) Typical stope round. (b) Burn-cut round (drift), the 0 holes are left uncharged. (c) Drift round with top draw cut. (d) V-cut shaft round. (e) Sump or bench shaft round. The order of firing follows the number sequence.

other place, either in the mine or to the surface of the mine. Usually this operation consists of loading some type of conveying device such as a truck or railroad track type of car. The mining industry has developed combined loading and transporting equipment called load-haul-dump units (Fig. 3a).

The typical hand shovel, power shovel, and wheel

or crawler loader and some underground loaders have a specific type of action (Fig. 3b). The bucket or shovel is pushed or crowded into the loose material as it is also lifted through the material, filling the bucket. The bucket is then directed to and dumped into the hauling conveyance. The hauling unit takes the broken material to its next destination point.

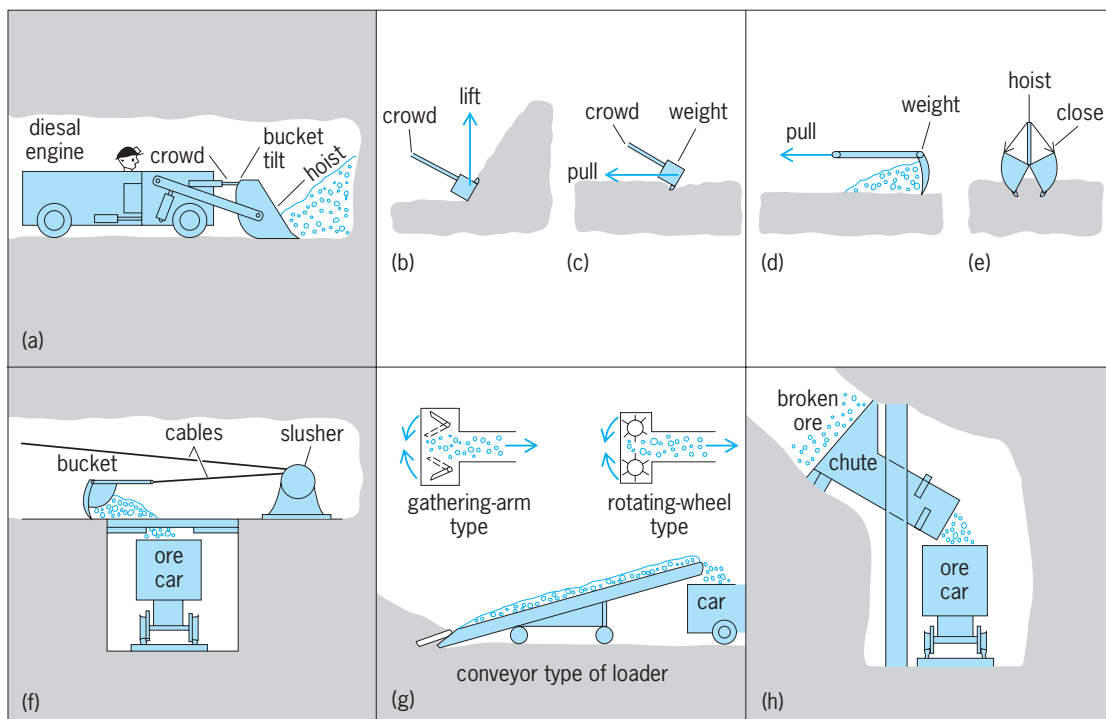


Fig. 3. Loading principles. (a) Load-haul-dump unit. (b–e) Types of loading action: (b) filling bucket by shovel action; (c) filling backhoe bucket; (d) slusher bucket transporting; (e) filling clamshell bucket. (f) Scraper and gravity loading. (g) Gathering and conveyor loading. (h) Gravity through chute loading.

In another action (Fig. 3c and d), the weight of the bucket or the combined weight of the bucket and machine causes the bucket to crowd into the loose material from the top while the pulling action fills the bucket. The loaded bucket can then either be hoisted, as in a pull-shovel or dragline (Fig. 3e), or be dragged along a hard surface (Fig. 3f) and dumped into a designated conveyance or area. Underground slushers and scrapers use this action and are popular in underground use.

Some machines use a combined crowding, raking, and lifting action (Fig. 3g). Machines using this action are high-capacity loaders, more commonly found in coal and nonmetallic soft ore types of mines. Chutes take advantage of gravity to load underground conveyances (Fig. 3b).

Loading in larger underground mines is more frequently done by diesel-powered wheel or tractor types of loaders that load into diesel-powered large trucks specially equipped for underground use. However, not all underground mines lend themselves to efficient use of wheel loaders and trucks, and some must rely on the standard types of loaders and conveyances.

Support. In underground excavation, one of the problems frequently encountered is supporting the excavation either during or after the excavation process. This process disturbs the natural forces that occur in the ground mass, and in the readjustment process forces act on the sides of the excavation (Fig. 4a). If the surrounding material is strong enough, no support is required. In other situations, the opening may stay open initially, but force readjustment in the ground mass or surface deterioration may cause the opening to fail at some time after it is excavated.

Rock bolts are probably the most frequently used means of development heading support, either temporary or permanent. Drill holes are placed into the rock mass (Fig. 4b), and the holes are filled with rock bolts that are anchored into the rock mass by various means. A head board on the surface of the opening is attached to each bolt and tightened firmly against the ground mass. This action keeps the groundmass firmly in place. Rock bolts are commonly made of steel rods or tubes; occasionally steel cable and fiberglass are used, being anchored in the holes by wedging, expansion shields, concrete, and various types of epoxy glues, and by expansion of split hollow tubes.

Rock bolts are sometimes used as temporary support in stopes before a more permanent support can be installed. However, in some development headings, such as shafts and semipermanent levels, a more lasting support is required. Wood (Fig. 4c) and concrete, with or without rock bolts, are frequently used. Timber provides good support for a considerable time and has the advantage of being easily repaired if a failure occurs. In mining situations, ground forces are continually changing; consequently, failures to support systems do occur. Concrete is a more permanent support, but it is very expensive. It is often used for shaft support, but less frequently for level support except in block caving mines where nothing else is satisfactory or where the level is to stay open for a long time.

The type of support required often dictates the mining method used for stoping out the ore. In some cases, the ground will support itself over large spans, and this is called open stoping. In other cases, minimal support is required, and wooden props (stulls) may be enough support until the zone is mined out;

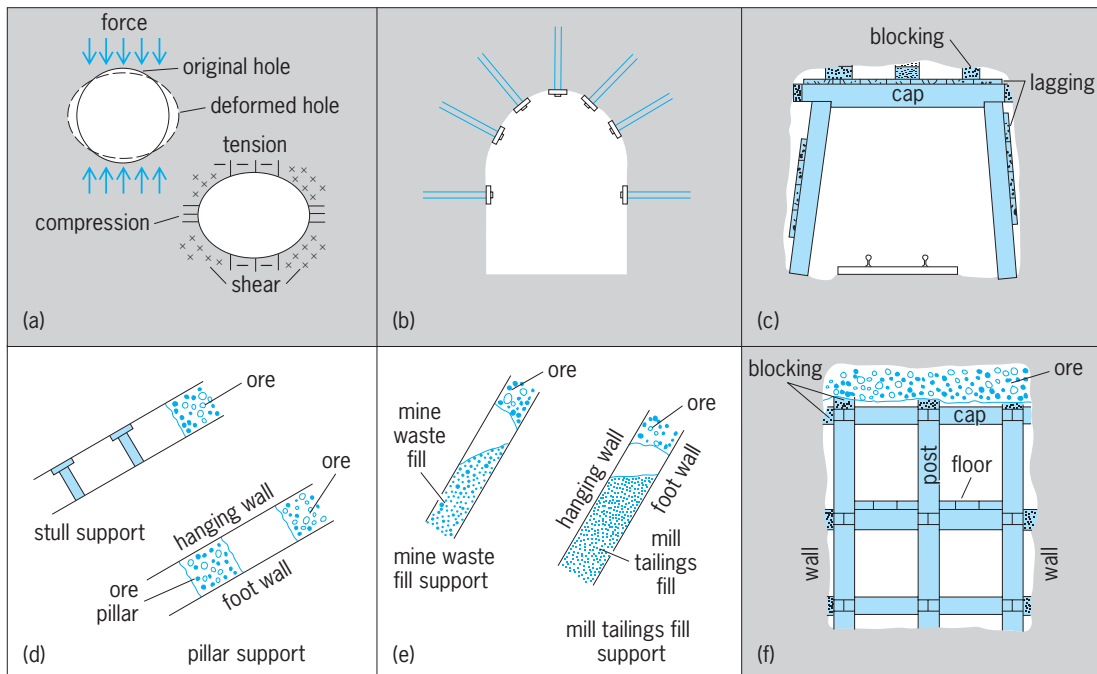


Fig. 4. Sections showing various types of support. (a) Deformation caused by forces. (b) Bolted cross section of level. (c) Timbered cross section of level. (d) Support in stopes. (e) Mine-waste support. (f) Timber support.

pillars usually give more support because they are bigger, but they are composed of ore and not all of the ore can be removed in pillar mining (Fig. 4*d*). Broken waste, ground, or fine mill tailings are often used for stope support (Fig. 4*e*); this is classified as a cut-and-fill system. Timbered systems are used for high-grade ore and in very weak or heavy ground (Fig. 4*f*); the timbered stopes are also filled with mill tailings for additional support. Although, technically speaking, caving systems use no support, the development headings required to extract the caved ore often require very heavy support. Koehler Stout

Machinery

Machinery is used in removing and transporting valuable solid minerals from their place of natural origin to a more accessible location for further processing or transportation. Many of the machines are identical to, or minor adaptations of, those used for excavating in the construction industry. In a wider sense, mining machinery could also include all equipment used in finding (exploring and prospecting), removing (mining: developing and exploiting), and improving (processing: ore dressing, milling, concentrating or beneficiating, and refining) valuable minerals; it could even include metallurgical (smelting) and chemical processing equipment used in extracting or purifying the final product for industry. The term mining machinery is also applied to special equipment for recovery of minerals from beneath the sea. In its usual context, the term does not include apparatus used principally in the petroleum industry. Perhaps those machines most often considered as uniquely mining machinery are drills, mechanical miners, and specially adapted materials-handling equipment for use in mining underground or on the surface (where a large proportion of mines are located). In addition, some unique auxiliary equipment and processing equipment are used in the mining industry.

Design and construction. All components of mining machinery—including primary mechanism, controls, means of powering, and frame—require the following features to a much greater extent than do other machines (with the possible exception of some military, construction, oil well, and marine units).

1. *Ruggedness.* Equipment is handled roughly, frequently receiving severe and sudden shocks from dropping, striking, and blasting vibration; overloading is common, and long life is demanded by the economics of mining.

2. *Weather resistance.* Operations extend over a wide range of climate and altitude.

3. *Abrasion resistance.* Minerals include some of the hardest substances known. Dust and fine particles are always present.

4. *Water and corrosion resistance.* Moisture and water, often acidic, are common in mining operations.

5. *Infrequent and simple maintenance.* Equipment is often widely scattered and in locations with restricted access. Trained mechanics and repair parts are generally limited in availability because of the remoteness of operations.

6. *Easy disassembly and reassembly.* Access to the machinery at the site of operation is frequently limited. Also, the working space near it and mechanical aids to moving or lifting it may be limited or nonexistent.

7. *Safety.* Because mining environments can expose workers to physical risk, most governments have testing bureaus, inspection agencies, and enforcement laws for the approval of mining equipment.

Simplicity of operation, low initial cost, and low operating costs are also desirable features of mining machinery.

Exploration machinery (vehicles, drills, and accessory equipment) is subjected to the same operating conditions as other mining equipment. Mineral-beneficiating equipment must have, above all, abrasion resistance. Smelting equipment has the requirement of heat resistance. Chemical refining process equipment must be highly resistant to corrosion. Reliability of all processing equipment is critical because slurries are commonly handled, and they can cause considerable difficulty in restarting after shutdowns.

Underground requirements. Machinery operated underground must meet special design requirements:

1. *Low-ventilation demand.* Quantity and geometry of passageways for air are rather rigidly fixed, so that high air consumption is a problem and noxious gases cannot be readily dispersed. Heat removal is a problem in deep mines.

2. *Compactness.* Space is at a premium, especially height, particularly in bituminous coal mines.

3. *Easy visibility.* Most operating areas are lighted only by individual cap-mounted or hand-held lamps.

4. *Hand portability.* Units or components must frequently be hand-carried into an operating area.

5. *Absence of spark and flame.* Equipment is often used in or near explosives, timber supports, and combustible gas or dust. In the presence of hydrocarbons, as well as of certain metal ores such as some sulfides, complete absence of open sparks or flames is a major requirement.

Power source. Mining machinery is very commonly powered by compressed air, but electricity is also widely used and is often the basic source. Compressed air has the advantages of simplicity of transmission and safety under wet conditions. It is especially advantageous underground as an aid to ventilation. Machines powered by compressed air can be easily designed to accommodate overloads or jamming, which is desirable on the surface as well as underground. Larger central compressors and extensive pipeline distribution systems are common, especially at underground mines.

Electric power, purchased from public sources or locally generated at a large central station, is common in open-pit and surface mines and dredging operations. Underground coal and saline-mineral mines often use electric-powered production machinery, but in other underground mines electricity is normally used only for pumps and transportation

systems in relatively dry or permanent locations. Direct-current devices have been dominant because of simplicity of speed and power control, but alternating-current apparatus has become common. Mobile equipment is often either battery or cable-reel (having a spring-loaded reel of extension power cable mounted on the machine) type. Processing machinery units are almost exclusively powered by individual electric motors.

Diesel engines are popular for generating small quantities of electric power in remote areas and for transportation units. Underground, abundant ventilation is essential, as well as wet scrubbers, chemical oxidizers, and other accessories to aid the removal of noxious and irritant exhaust gases. Hydraulic (oil) control and driving mechanisms are widely used. Transfer of power by wire rope is common, especially for main vertical transportation.

Drills. Drills make openings, of relatively small cross section and long length, which are used to obtain samples of minerals during exploration, to emplace blasting explosives, and to extract natural or artificial solutions or melts of minerals. Exploration holes are generally vertical or inclined steeply downward, less than 6 in. (15 cm) in diameter and up to 10,000 ft (3 km) long. Blastholes range from 0.75 to 12 in. (2 to 30 cm) in diameter and usually are under 50 ft (15 m) long in any direction, with the larger usually downward. Solution wells normally are vertical and 6 to 12 in. (15 to 30 cm) in diameter. They sometimes reach depths of several thousand feet, and are equipped with several concentric strings of pipe.

Rock drills. Percussion, rotary, or a combination action of a steel rod or pipe, tipped with a harder metal chisel or rolling gearlike bit, chips out holes up to 12 in. (30 cm) or more in diameter by 125 ft (38 m) or more long from the surface, and 1–3 in. (2.5–8 cm) by 5–200 ft (1.5–60 m) from underground. Crawler or wheeled carriers are used, and the smaller drills are often attached to hydraulically maneuvered booms. Air or liquids flush out the chips.

Diamond drills. Rotation of a pipe tipped with a diamond-studded bit is used in exploration to penetrate the hardest rocks. Large units make holes up to 3 in. (8 cm) in diameter and more than 5000 ft (1.5 km) deep; at the other extreme are units so small that they can be pack-carried. A cylindrical core is usually recovered.

Water-jet drills. For exploration and blasting in loose or weakly bonded materials, a water jet washes out a hole as a wall-supporting pipe is inserted.

Jet flame drills. For economical surface blastholes in hard abrasive quartzitic rock, a high-velocity flame is used to spall out a hole.

Mechanical miners. There are many machines designed to excavate the valuable mineral or the access openings by relatively continuous dislodgement of material without resorting to the more common practice of intermittent blasting in drill holes. These units also frequently transport the mineral a short distance, and when designed for weakly bonded minerals, they often become primarily materials-handling equipment.

Continuous miners. For horizontal openings in coal and saline deposits, toothlike lugs on moving chains, or rotating drums or disks rip material from the face of the opening as the assembly crawls ahead.

Longwall mining systems. A longwall mining system consists of a drum shearer (also called a plow), shield supports (movable roof supports) and controls, face conveyors, and crushers. The shearer is used to mill coal from the seam as it is driven along the wall, which may be many hundreds of meters long.

Augers. Coal and soft sediments are mechanically mined by augers up to 5 ft (1.5 m) in diameter and 100 ft (30 m) long, usually used horizontally.

Shaft and raise drills and borers. For vertical and inclined openings up to 8 ft (2.4 m) or even larger in diameter, various rotary coring and fullface boring equipment is used, both in an upward direction (raising) for several hundred feet or downward (sinking) for several thousand feet. These units usually use many rigid teeth or rolling gearlike bits to chip out the mineral.

Tunneling machines. There are rotary boring units for horizontal, or nearly so, openings of any length and up to 35 ft (11 m) in diameter (in soft rock).

Rock saws. To remove large blocks of material, narrow slots or channels are cut by the action of a moving steel band or blade and a slurry of abrasive particles (sometimes diamonds) rather than teeth. Small flame jets are also used.

Hydraulic monitors. Water jets of medium to high pressure (some to 5000 lb/in.² or 34 megapascals) are used to excavate weakly cemented surface material and brittle hydrocarbons on both the surface and underground. In some instances, the jets are pulsed.

Special materials-handling equipment. Loose material (muck) is picked up (mucked or loaded) and transported (hailed or hoisted) by a wide variety of equipment.

LHD (load-haul-dump). LHDs are loaders designed specifically to work in confined underground spaces. They resemble the front-end loaders often found at construction sites, but are not nearly as tall. They are used to transfer muck (broken ore) from the face (or draw point) and deposit it at the ore pass (dump site), often by dumping it through a screen to exclude rocks over a certain size. Because of the confined environment, LHDs are often driven in reverse, and the operator sits sideways to the travel direction to see ahead and behind the vehicle.

Excavator loaders. For confined places underground there are various unique grab-bucket shaft muckers and overcasting shovel tunnel muckers, and gathering-head loading-conveyor units having eccentric arms, lugged chains, and screws or oscillating pans for handling muck in horizontal openings.

Dragline scrapers. Scrapers (slushers) with a flat plowlike blade or partially open bucket pulled by a wire rope are commonly used to move muck up to a couple of hundred feet, especially in underground mines.

Dipper shovels and dragline cranes. In surface mining, single-bucket loads can handle up to 100 yd³ (75 m³) of material. Many of the intermediate size (20–40 yd³ or 15–30 m³) units move on unique walking shoes.

Bucketline and bucketwheel excavators. These are for surface use and can dig up to 5000 yd³/h (3800 m³/h), using a series of buckets on a moving chain or a rotating wheel supported on crawlers, railcars, or floating hulls (dredges).

Suction dredges and pipelines. On the surface, up to 3700 yd³/h (2800 m³/h) of moderately loose mineral up to several inches in diameter can be picked up and moved as a slurry (mix of water and solid material) by pumps mounted on floating hulls.

Trucks. Diesel and electric shuttle cars (short-haul trucks) of unusual design, often having very low profile and conveyor bottoms, are used underground. At surface mines, there are diesel and other dump trucks with 330 metric tons capacity.

Railroads. Underground locomotives with electric (storage battery, cable-reel, or trolley), diesel, and sometimes compressed-air power units are used to transport ore cut of the mine. Cars are usually of special design for automatic dumping.

Conveyors. Unique movable, self-propelled, sectional, and extensible conveyors are used in underground mining, especially in longwall and continuous mining operations. In surface mines, conveyors are often used to carry ore over land.

Wire rope hoists. Hoists or winders are used in shafts for vertical or steeply inclined transportation in single lifts of as much as 6000 ft (1.8 km). Of various particular designs, there are two basic types: drum, simply a powered reel of rope; and friction, in which the rope is draped over a powered wheel and a counterweight is attached to one end and a conveyance to the other.

Auxiliary equipment. Drainage pumps handling hundreds of gallons of water per minute at heads of 1000 ft (300 m) or more are used underground. Ventilation fans are capable of moving several hundred thousand cubic feet of air per minute. Crushers can handle pieces of hard rock several feet across in two dimensions. In processing, minerals are sorted by size or density, or both, by a variety of screens, classifiers, and special concentrators, using vibration, fluid flow, centrifugal force, and other principles. Froth flotation and magnetic and electrostatic equipment take advantage of other special properties of minerals. *See* BULK-HANDLING MACHINES; GRINDING MILL.

John P. H. Steele; Lloyd E. Antonides

Mine Automation

Mining automation encompasses all aspects of the mining process from exploration, development, and extraction, to milling and product refinement. Automation of mining processes includes surface and underground operations as well as their supporting technologies, such as autonomous vehicles, communications, remote control, teleoperation, and robotics. Computers and advanced sensing devices are changing the way that materials are mined and processed. In many respects, because mobile mining machines manipulate the earth directly, the new autonomous mining systems being developed represent the most advanced robots produced so far. *See* AUTOMATION.

Underground. Load-haul-dump (LHD) vehicles, which are related to the front end loaders used in construction, are the main tool for transporting ore and waste rock in metal and nonmetal underground mines. Presently, LHDs are operated by miners who are physically located on the machine and drive it much like an automobile, except that the path is a drift and the operator sits sideways to see both ahead and behind the vehicle. In locations where there is significant danger of rock falling from the roof (roof fall), for example in open stopes (ore bodies), the LHD may be operated by remote control using either a tether (cable) or radio signal. The operator stands some distance away but in view of the vehicle and controls the vehicle by manipulating joysticks and switches. While remote operation has successfully protected miners from roof falls, there have been a number of accidents in which operators have been struck by the vehicle due to the tight operating quarters. *See* REMOTE-CONTROL SYSTEM.

New robotic LHDs are being developed and tested in Finland, Sweden, Canada, the United States, and Australia. In Kiruna, Sweden, one mine has several driverless LHDs working in production that perform automated hauling and dumping, and remote control loading. Automated dumping is relatively easy. After arriving at the dump site (typically an ore pass, which is a vertical shaft that funnels the ore to a lower level where it is conveyed out of the mine), the vehicle raises and tips its bucket, dumping the ore into the opening. Various sensors let the vehicle know when it has arrived at the ore pass. Navigating the drift is more challenging because the LHD must traverse the tunnel without hitting the walls, which in many mines are only tens of centimeters away. The most favored sensors for detecting drift walls are scanning lasers in which the time of flight of the light is used to determine the distance from the sensor to the wall. This technique allows the LHD to build up a two-dimensional model of the drift as the vehicle travels through the tunnel. Relatively simple control algorithms are then used to steer the vehicle, and its speed is adjusted as the passage becomes wider or narrower. Automated vehicles can travel through the drift at speeds comparable to those of experienced operators (60 km/h or a little less than 40 mi/h). Automated vehicles are less likely to hit the walls, once the control system is properly tuned, because of the closed-loop control between the laser sensors and the vehicle steering. *See* ALGORITHM; ROBOTICS.

The loading operation is more challenging and so far only teleoperated, because the operator must plan the loading operation and manipulate the bucket and boom while driving into the rock pile, called a muck pile (broken ore or waste rock underground; LHDs are sometimes referred to as muckers). In LHD automation, the operator is stationed in a control shack somewhere other than where the mucker is working (in some cases, on the surface, operating the machine underground). Remote control from a remote location is called teleoperation since the information about the operation must be transmitted, typically as an electronic signal to the operator, who

sends commands back to control the machine. *See* REMOTE MANIPULATORS.

There has been work on automating the loading operation as well. For automated loading, the load in the bucket must be sensed by measuring the pressure in the hydraulic cylinders that lift the boom and bucket. By measuring the forces acting on the bucket and understanding how the load is elevated, a control program (controller) can determine the next action of the boom and bucket. The successful approaches to this problem will likely use artificial intelligence techniques (fuzzy logic). To complete the automation of the loading operation, the robot (LHD) must be able to sense the muck pile and to plan its approach to the pile in order to load it, which requires the ability to sense the muck pile in three dimensions. Researchers are developing stereo vision methods for this sensing. *See* ARTIFICIAL INTELLIGENCE; FUZZY SETS AND SYSTEMS.

Automated loading is dependent on communication with a robot as it works in another location. The ability to send a signal through the earth is limited, so all underground mining operations rely on communications that follow underground passageways (drifts, raises, passes). A leaky feeder is a coaxial cable that allows signals to exit cables all along its length. Such a cable is strung throughout the mine, and two-way communications is possible. However, the technology limits the bandwidth of the signal, and techniques such as spread spectrum communication, like that used for cell phones, are being used to transmit video from the mine. *See* COAXIAL CABLE; MOBILE COMMUNICATIONS.

Once the loading operation is fully automated, there will be true robots working underground. The interaction between the operator and the LHD underground will be at the level of scheduling—directing the LHD to work in certain locations at certain times.

Another underground mining operation that has been automated is drilling. At the Kiruna mine in Sweden, production raise drills (steep-angle drilling machines) are operated via remote control from the surface. The drill rigs are driven into position, and the locations for the drill holes are sited and sent to the machine. Control of the operation is turned over to the machine, and the drilling operation proceeds with the remote operator acting as a supervisor while the drilling of the complete pattern of holes proceeds (these holes will be used to place the blast charges used to fragment the rock). This allows the operator to oversee the operation of several drills at the same time.

Surface mining. Surface mining involves the removal of covering soil (called overburden), the fragmentation of rock or coal by blasting, the transportation of the ore from the mine site to another location for processing (often to a mill), and the processing of the ore to enhance the ore content or to size-reduce the material for shipment. Automation of surface mines has taken several forms, such as automation of the conveyors that are sometimes used to move the ore over long distances, or the automa-

tion of the milling operation to reduce the ore to the proper size for the next processing step. The automation that has captured the imagination of many engineers is the development of driverless haul trucks, used to move ore from the active mining site to the next stage in processing. The routes for this transfer are well defined and can be planned ahead of time. The key to making the trucks driverless is the ability to measure a truck's location with great accuracy. This has become possible with the advent of GPS (Global Positioning System) and its refinement, Differential GPS. Depending on the number of satellites within the view of the GPS receiver and their distribution in the sky, GPS can be used to locate a truck within approximately 30 m (100 ft), a solution that is adequate for human navigation, but not good enough to steer a haul truck in a mine. The accuracy of the position measurement can be improved by including the error in the measure as determined by another GPS receiver at a known location. Because the longitude, latitude, and elevation are known at this location, the computed location, based on the signals received from the satellite, includes the measurement error. This error can be passed along to a truck's GPS receiver so that it can correct its computed location. This process reduces the positional error to less than a meter (more accuracy measurements are possible, but will not be discussed). This level of accuracy is adequate for large haul trucks to navigate mine roads without a driver. This technology is currently under development in a mine in Arizona, as well as other locations. *See* SATELLITE NAVIGATION SYSTEMS.

GPS technology is also being applied to dozers. By having a Differential GPS receiver on the dozer and knowing the relationship of the dozer blade to the location of the GPS receiver, and having a map that includes elevation information of the mine area, the dozer operator can be informed about the elevation of the material to be moved relative to the desired final level. In fact, the information from the computer-based map is better than that available from more traditional surveying techniques.

A third activity that has been improved by the development of GPS is the positioning of drill rigs. GPS receivers are placed on the top of the drilling towers so that the location of the hole being drilled (in preparation for loading it with a blast charge) can be more accurately located and the direction of the drilling can be monitored. This results in better control of the blasting operation and more uniform size of the blasted material. Such improvements save energy in the downstream processes and lower production costs.

Another area where automation is having an important impact on mining, both on the surface and underground, is machine health monitoring, sometimes called machine condition monitoring. By having sensors on the machine as it works, a much better understanding of the machine's condition can be developed. Trends or changes in the data can be tracked, and failures anticipated. By knowing when the condition of the machine is changing, catastrophic failures can be prevented and unplanned

downtime can be reduced. This leads to a more efficient and safer mining operation. This technology is making its way into a number of mining systems.

The implementation of automation in mining operations will lead to safer and healthier conditions for miners and to more efficient operations. The nature of the mining enterprise with its constantly changing environment and its lack of structure provides a great challenge to automation developers. Mining continues to challenge engineers to come up with new and cost-effective methods. Automation is one of the principal methods that will be used to continue this tradition of innovation. John P. H. Steele

Mining Power

Mining power refers to the means by which mining equipment is powered from some energy source. The process of extracting minerals from modern underground and surface mines is highly mechanized, requiring mobile mining machines that have variable power demands and adhere to stringent safety requirements. In addition, the surface activities of any mine, which may include shops, changing rooms, offices, ventilation fans, hoisting equipment, mills, preparation plants, and so forth, can have large power requirements. These facilities may receive power from the mine distribution system or at times from a separate substation.

The two predominant power sources for today's mining operations are diesel engines and electricity.

Diesel power. Diesel-powered equipment is used extensively in surface mining for front-end loaders, drills, trucks, locomotives, small shovels, and small draglines, with application essentially the same as in other industries. There is also wide use of diesel-powered equipment in many underground metal and nonmetal mines for front-end loaders, drills, trucks, shuttle cars, and locomotives. The use in underground coal mines is limited and, in these few mines, usually restricted to shuttle cars and locomotives. Because of the confined underground environment, the U.S. Mine Safety and Health Administration enforces strict standards on toxic components in engine exhaust gases, mine-ventilation dilution of exhaust gases, and the handling and storage of diesel fuel. In underground coal and other gassy mines, the diesel equipment used in mining-face areas must also adhere to additional stringent (permissibility) standards so that the machines cannot cause an explosion in the mine atmosphere. See DIESEL ENGINE.

Electric power. By far the most popular source of power in underground coal mines is electricity, and one reason is the high horsepower require-

ments of continuous miners and longwall machinery. Diesel engines cannot supply the mechanical power needed within the available space restrictions on these machines. Electricity is also favored over diesel power for large drills, shovels, and draglines in surface mines because of space restrictions. Large surface excavators can have 12,000 or more connected horsepower (9000 kW), and the largest is over 30,000 hp (22,500 kW).

Three-phase alternating-current (ac) power distribution and motors are used to power the various mining machines that were introduced into underground mines. Alternating current can be distributed at a higher voltage, thereby reducing current, voltage drops, and transmission losses, and providing adequate power delivery to the mining equipment. The use of silicon-diode rectifiers with large current capabilities has made possible the application of direct current (dc) for traction and of ac distribution and utilization on high power loads. For example, while continuous miners normally use ac, part of the ac supply at the power center is rectified to dc, primarily for powering the shuttle cars.

Simplified mine electrical distribution. Most underground-mine power systems have two voltage levels: distribution and utilization (Fig. 5).

In remote locations, a mining operation may generate its own power, but it is common practice for a mine to purchase all or most electric power from utility companies whenever it is available. Because utility voltages can range from 24 to 138 kV, a substation transforms the utility voltage down to distribution levels, most often at a voltage greater than 1000 V. In addition to the transformer, substations contain a complex of switches, protection apparatus, and grounding devices, which function as safety equipment. Power at the distribution voltage is distributed from the substation through conductors to a power center; hence, it is termed the distribution system. The power center, actually a portable substation, transforms the voltage to the correct level for equipment use, typically at less than 1000 V. It is at this level that power is normally delivered through conductors to the underground mining equipment. See ELECTRIC POWER SUBSTATION.

Originally, ac distribution was made at 2300 or 4160 V. In most mines, these levels were later increased to 7200 V, but a few operations have increased the distribution voltage to 12,470 or 13,200 V for both longwall and continuous mining operations. At first 440 V was the most popular voltage for ac equipment. When the continuous miner proved so successful, its horsepower was progressively increased; this resulted in an increase in the size of its trailing cable, until the weight was almost more than personnel could handle. To compensate, the most usual solution was to raise the rated motor voltages to 550 V. Manufacturers then produced machines with 950-V motors to further overcome the trailing-cable problems.

Underground mine distribution. In an underground mine, there is little freedom for varying system arrangements and the distribution system must almost

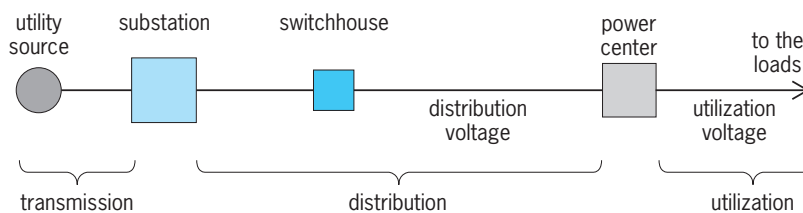


Fig. 5. Simplified electrical distribution system for mining.

always be radial; that is, a single power source and substation supply all equipment, and the system forms a treelike structure spreading out from the substation (Fig. 6). Power and mine grounding are fed underground in insulated cables, through a shaft, a borehole, or an entry. From a central disconnect switch, which allows total removal of underground power in an emergency, the power is distributed through cables to power centers that are located as close to the machinery as practical. The centers can power several face machines through couplers and trailing cables.

If belt haulage is used, small power centers are located close to all major conveyor-belt drives. With rail haulage, dedicated dc power centers (rectifiers) are placed along the rail to convert ac from the distribution system to dc. The distribution system can also serve large single motors directly through switchgear. All the power equipment used underground must be rugged, portable, self-contained, and specifically designed for installation and operation in limited spaces. As with diesel power, electrical equipment used in the face areas of gassy mines must be legally approved as permissible.

The primary purpose of any distribution system is to provide a flexible, easily moved or modified power source for the highly mobile mining equipment. The power system must be designed to function as an integral part of the total mine operation. The distribution system in any surface or underground mine that serves portable mining machinery is subject to damage from being run over by mining machinery, and as a result the system must be designed with optimum flexibility and consideration for personnel safety. See DIRECT-CURRENT TRANSMISSION; ELECTRIC DISTRIBUTION SYSTEMS; ELECTRIC POWER SYSTEMS.

Surface-mine power requirements. A combination of motor-generator sets driven by synchronous or induction ac motors, the Ward-Leonard speed control system, and dc motors is used on most mining excavators, especially the larger varieties. The standard 13,800-V distribution and machine voltages are for excavators larger than 100 yd³ (76 m³), while 23,000-V systems have been designed for machines greater than 200 yd³ (153 m³). Production shovels up to 18 yd³ (14 m³) commonly utilize 4160 or 7200 V, while in general 4160 V has become standard for machinery with 1500 hp (1125 kW) or less. As a result, more than one voltage level can be required at a mine when excavators of different sizes are employed. See ALTERNATING-CURRENT GENERATOR; ALTERNATING-CURRENT MOTOR.

Surface-mine distribution. Simple surface-mine power distribution (Fig. 5) is a very common arrangement in small surface operations where the distribution voltage is usually 4160 V. In the smallest mines, power is purchased at low-voltage utilization (often 480 V ac) and fed to a distribution box to which motors and equipment are connected. At times, simple distribution is also employed in large surface mines where only one machine must be served or where an extensive distribution network

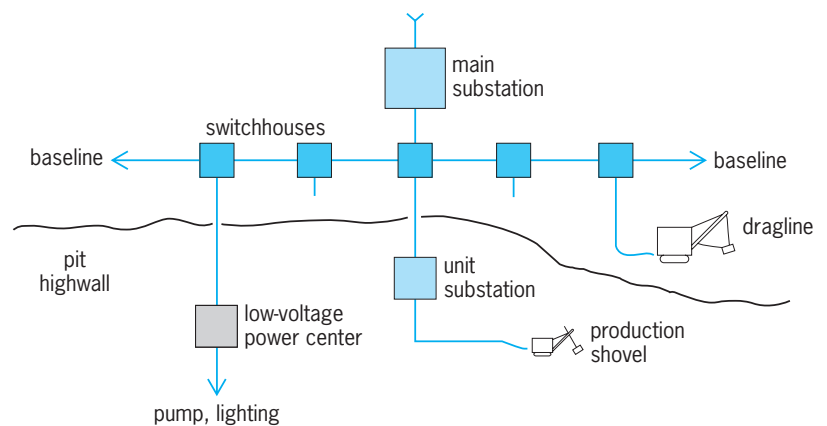


Fig. 6. Radial electrical distribution system for an underground mine.

cannot be established, as in some contour mining operations.

The majority of surface mines employ radial distribution. In some area-type surface (strip) mines, the distribution system forms a base line which is usually located on the highwall, paralleling the mine pit (Fig. 7). Cables connected to the base line deliver power and grounding to skid-mounted switchhouses located on the highwall or in the pit. Specially designed trailing cables then complete the circuit to the machines. As the machines move along the pit, the base-line connections are successively changed to convenient locations. Open-pit power systems are similar with one main exception: the distribution system typically establishes a ring bus or main power line that partially or completely encloses the pit (Fig. 8). Radial ties to the bus complete

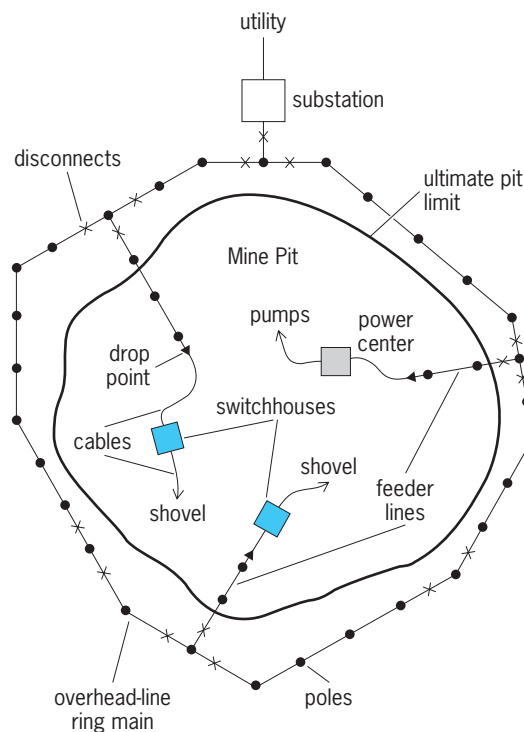


Fig. 7. Electrical distribution system for a surface mine.

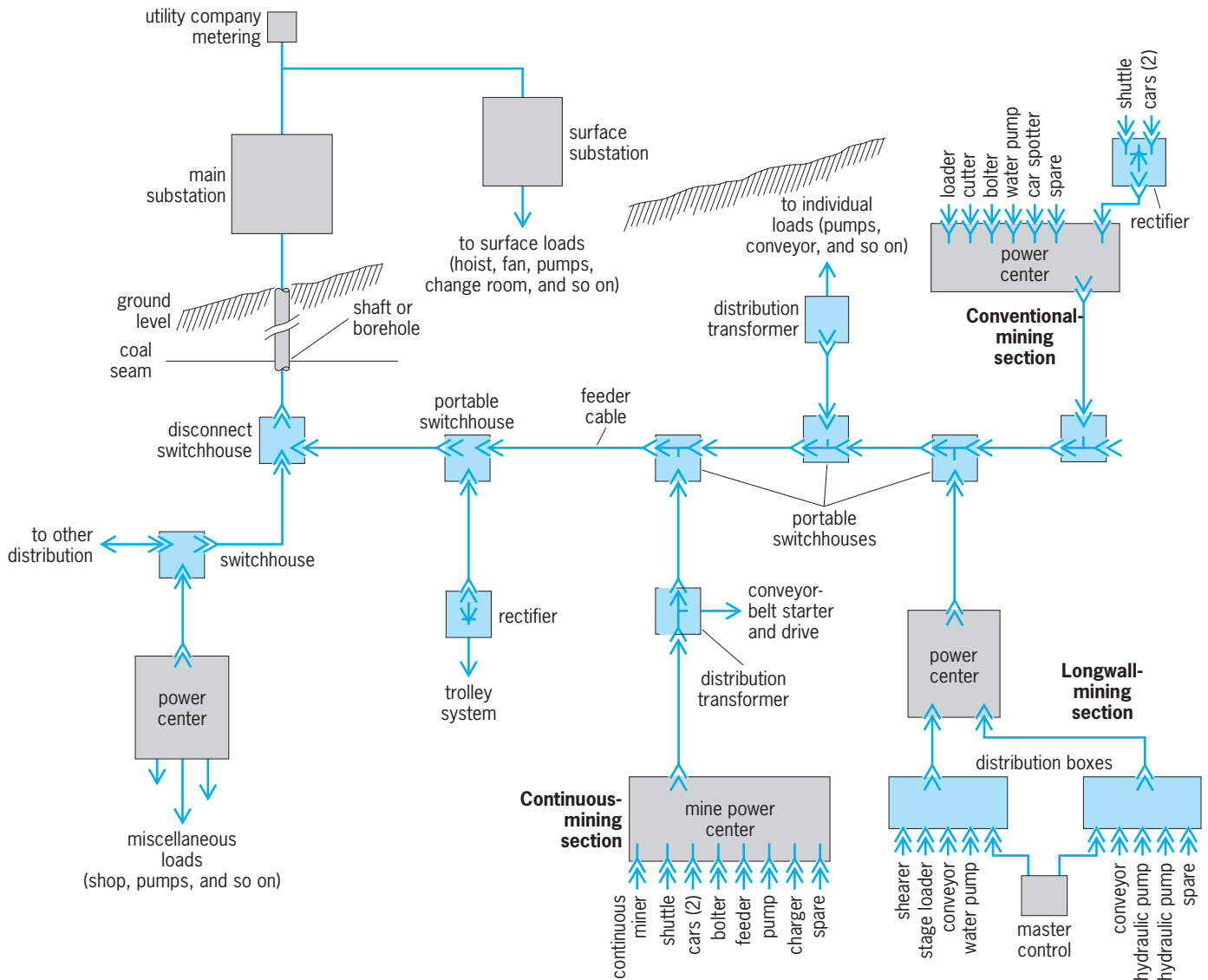


Fig. 8. Open-pit mine power system.

the circuit to switchhouses located in the pit, and portable equipment again uses trailing cables.

Lloyd A. Morley

Mine Ventilation

Mine ventilation consists of the tools, techniques, and methods used to create, maintain, and enhance the underground mine atmosphere for human health and safety. The objective of mine ventilation is to provide adequate quantities of fresh air to workers everywhere in the mine, as well as to render harmless and carry away toxic, noxious, and explosive contaminants, and to maintain heat and humidity at desirable levels. With increasing depth of mining, both ground pressure and virgin rock temperature increase, and are recognized as the limiting factors for operating mines at great depths. Among the areas of concern for worker health and safety underground, none is more critical to human survival than provision of a life-sustaining atmosphere in the hostile subsurface environment. See VENTILATION.

The principal atmospheric contaminants in underground mines are particulates (liquids or solids) and nonparticulates (gases, vapors, heat, and radiation). Liquid particulates include mists and fogs, and solid particulates include dust, fumes, smoke, pollen, and bacteria. The most common type of contaminants, however, are gases and dusts, and in deep mines, heat and humidity. Mine air contains several gases and particulates not encountered in normal air. Elevated concentrations of these contaminants are hazardous to worker health and safety. See GAS AND ATMOSPHERE ANALYSIS.

Minimum ventilation requirements for underground coal, and metallic and nonmetallic ore mines are prescribed in the Code of Federal Regulations (CFR), Title 30. The velocity and quantity of air required at designated sites (or spaces) are specified, as well as the quality of the air (such as the minimum or maximum concentrations of various gases and dusts). At selected intervals and locations, monitoring of the quality and quantity of a mine's air is required.

Several controls, such as legal, medical, engineering, and administrative, are applied to ensure compliance with standards. Achieving total mine air conditioning requires designing and controlling the quality, quantity, and temperature-humidity of the flowing air. Engineering principles fundamental to the quality and temperature-humidity control are, in order of preference, prevention (or avoidance), removal (or elimination), suppression (or absorption), containment (or isolation), and dilution (or reduction) of the contaminants. Often two or more of these principles are used together to achieve total air conditioning. Mine ventilation, primarily a dilution measure, is the solution to many mine air contaminant control problems. Severe air quality problems can be controlled by application of the other principles as well.

For quantity control, the principles of fluid dynamics are applied to analyze and design airflows in a mine. Airflow is induced from the atmosphere through the mine workings and back to the atmosphere by creating a pressure difference between the mine's openings. The mine ventilation system consists of fans (low pressure pumps creating the pressure differences), shafts, adits, and slopes (openings from the surface to the deposit being mined), airways (interconnections in the deposit between the openings to the surface and the mine working areas), and flow-regulating devices in the airways that control the direction and magnitude of airflow rates in the airways, shafts, and slopes.

Air-quality control. Air contains, by volume, about 78% nitrogen, 21% oxygen, 0.03% carbon dioxide, and less than 1% of argon and other rare gases. Mine air must contain at least 19.5% oxygen and no more than 0.5% carbon dioxide. It may also contain several other gases that are explosive, toxic, or both, and a host of particulates that are also toxic, explosive, or otherwise harmful to health. The limits to the airborne concentrations of these contaminants and durations of exposure are called the threshold limit values (TLVs). The TLVs are recommended by the American Conference of Governmental Industrial Hygienists (ACGIH), based on industrial experience, and animal and human studies. *See* INDUSTRIAL HEALTH AND SAFETY.

The formula to calculate the dilution air quantity required to bring the concentration of a pollutant to below its TLV at any location is Eq. (1). Here Q =

$$Q = \frac{Q_g}{(TLV - B)} \quad (1)$$

the minimum amount of dilution air quantity (m^3/s), Q_g = the rate of generation of the pollutant (m^3/s , mg/s), TLV = threshold limit value of the pollutant ($\%$, mg/m^3), and B = the concentration of the pollutant in the dilution air ($\%$, mg/m^3).

When the amount of air quantity required for dilution is prohibitively high, measures such as prevention, removal, suppression, and containment are used to prevent the contaminant from entering the air stream to which miners are exposed. *See* AIR CONDITIONING.

Mine gases. The gases encountered in mining include methane, carbon monoxide, hydrogen sulfide, oxides of nitrogen, hydrogen, and radon. These gases emanate from several sources, including strata, blasting, internal combustion engines, broken ore, oxidation, incomplete combustion, and fires.

Methane is one of the most dangerous strata gases found in coal and noncoal mines. It does not support life and is explosive in the range 5–15%. The limits for methane are set at 1% for working areas. In some deep coal mines, 5 to 20 metric tons of air are circulated for every metric ton of coal mined to achieve methane control. To decrease this volume of air, and to make mines safer, several deep mines are practicing methane drainage, a process for removing methane through boreholes and pipelines, thereby decreasing the amount of methane entering the ventilating air stream. Radon, hydrogen sulfide, and carbon dioxide are other strata gases, less common than methane.

Carbon monoxide, sulfur dioxide, and oxides of nitrogen are highly toxic gases that arise from blasting and internal combustion engines. The first two can also come from slow oxidation, incomplete combustion, and fires that can occur in the deposit, broken ore, and the mine. Proper selection and use of blasting agents, prevention of oxidation and fires, and proper operation of diesel engines are essential to ensure that gas problems do not evolve into major health and safety hazards.

Mine dusts. The airborne dust in a mine atmosphere is likely to contain the minerals being mined. These particulates are generated during the cutting, drilling, blasting, loading, hauling, and other operations associated with the extraction of the deposit and the subsequent transport of the mined product to the surface.

On the basis of their health effects, mine dusts have been classified as fibrogenic, carcinogenic, toxic, radioactive, explosive, and nuisance. Coal dust is both fibrogenic and explosive. Silica dust is fibrogenic and a suspected carcinogen. Uranium, radium, and thorium dusts are radioactive and toxic. Particulates in diesel engine exhaust contribute to the respirable dust load in the mine air, and are suspected carcinogens.

Coal worker's pneumoconiosis is an occupational disease associated with long-term exposure to high concentrations of coal dust, leading to the formation of fibrous tissues in the lungs. Silicosis is a lung disease associated with inhalation of particulates containing sufficient quantities of silica. For coal mines in the United States, the current permissible exposure limit (PEL) for airborne respirable coal dust is $2 \text{ mg}/m^3$. It is an 8-hour, time-weighted average concentration. This standard is reduced to (10 ÷ percent silica) mg/m^3 when the respirable quartz content exceeds 5% in the airborne dust sample. The ACGIH permissible exposure limit for respirable quartz is $100 \mu g/m^3$.

Respirable dust, once airborne, is difficult to control. Therefore, all control principles are applied to decrease the generation, reduce the entrainment, dilute the concentration, and control exposure. In

modern coal mining, the use of sharp cutting tools with appropriate power input is combined with the use of adequate water sprays to reduce the generation and entrainment of dust. Further, the cutting machines are equipped with dust extractors, and directional sprays to capture the dust, and if necessary, to steer the dust away from the miners. Operating the machines from a safe distance upwind, using remote control devices, and wearing personal protective equipment, such as respirators and air stream helmets, decrease the concentration of the dust. Despite these advances, respirable dust remains a hazard.

Temperature-humidity control. In very hot and humid working conditions, the body may be unable to cope with the heat stress. Miners' health may be threatened by heat illnesses, such as dehydration, heat rash, heat cramps, heat exhaustion, and heat stroke. Mine air conditioning for heat and humidity control becomes essential when airflow alone is inad-

equate to control the temperature. See AIR COOLING; HUMIDITY CONTROL.

Mine air cooling. The major sources of heat in mines, in order of their contributions, are the heat from the strata, the autocompression of the air as it flows down the intake shafts, and the powered equipment used in the mining process. The rate of change of the virgin rock temperature (VRT) with depth (called the geothermal gradient) is quite variable in different mining districts. The worldwide range is 0.7–6°C for every 100 m depth (Fig. 9). The effect due to autocompression is to increase the temperature at approximately 9.66°C for every 1000 m depth. Nearly all the energy input to machinery ends up as heat in the mine air.

The heat removal problem is acute in very deep mines due to the high VRTs, and an increased autocompression effect. As shown in Fig. 9, at an approximate depth of 1980 m, called the critical depth,

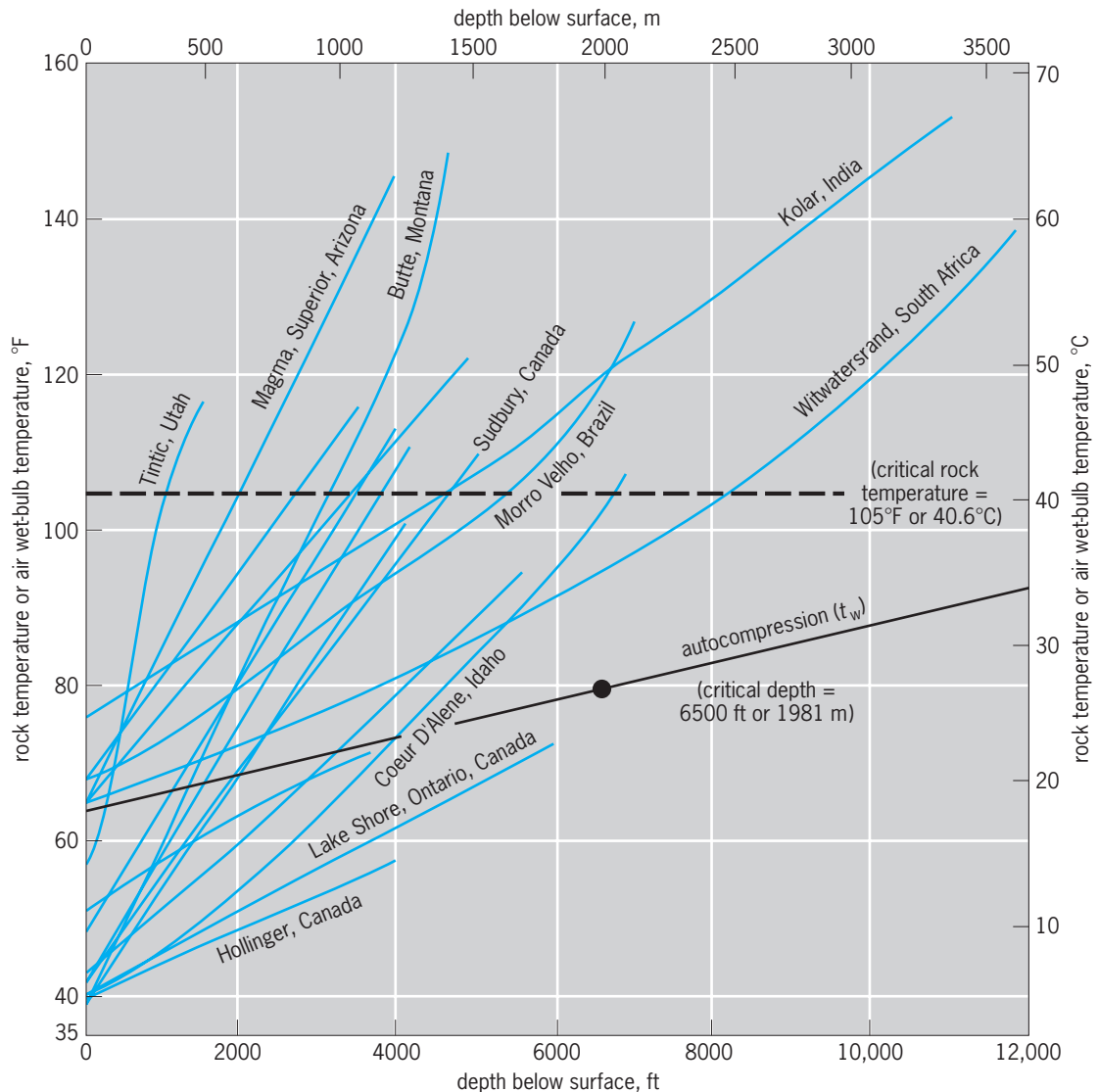


Fig. 9. Average geothermal gradients of various worldwide mining districts and effects of autocompression on air wet-bulb temperature (assumed surface $t_w = 64^\circ\text{F}$ or 18°C , and $\delta t_w = 2.4^\circ\text{F}/1000\text{ ft}$ or $4.4^\circ\text{C}/1000\text{ m}$), both as a function of depth below surface. Critical rock temperature and critical depth for t_w are also shown. (After H. L. Hartman et al., *Mine Ventilation and Air Conditioning*, Wiley, New York, 1997)

autocompression alone governs that air conditioning be introduced for worker health and safety. Similarly, a VRT of 40.6°C (called the critical temperature) dictates the installation of air-conditioning systems. The depth associated with this critical temperature, however, varies as a function of the local geothermal gradient.

Mine air heating. Mines in very cold climates or in high altitudes may have very low temperatures. In rare cases, the combination of extremely low temperature and high-velocity air intake can create high wind chill factors, making it necessary to heat the air for the comfort of the miners. Such heating is not a common practice as miners can wear warm clothes. Where required, heating the air to 4–7°C is generally adequate.

Heating the intake air is more often required to protect the equipment and facilities in the intake shafts and slopes from damages due to freezing of water mains, ice buildup on the walls, and cyclical freezing and thawing of concrete linings. Equipment safety and production can be severely affected due to increased operational and maintenance problems in subzero temperatures. Heating is required only when subfreezing temperatures occur, and raising the intake air temperature to 1°C is sufficient.

As opposed to heating in winter, in summer months, mines in the arctic region may require refrigeration techniques to maintain the intake air temperatures to below the melting point of the permafrost.

Quantity control. Quantity control in the mine ventilation system means achieving optimally in the mine airways and working sections the desired quantities of airflow in the desired directions. The term “optimally” as used here is global and not restricted to ventilation only. The ventilation system must complement the mining method, and offer flexibility in meeting emergencies and future needs.

The mine airflow distribution is completely defined by (1) the physical parameters of the airways, including shape, area, length, and surface characteristics; (2) the layout of the mine openings; (3) the pressure sources (for example, fans) in the system, their location and characteristics; and (4) the interconnections between the airways, mine openings, and pressure sources using control devices, such as seals, stoppings, air crossings, regulators, and doors. Most air quality control activities take place at the mine’s working faces, where the emanation of contaminants and the need for dilution are generally the greatest. The calculated dilution air must be conducted from the atmosphere to the working sections through the intake openings of the mine. After this air has served its purpose, it must be conducted back to the atmosphere through the return openings of the mine. In the working section, several methods of face ventilation are employed to direct the incoming air to the areas where workers are present and production operations are under way. The face ventilation in a room-and-pillar mine in coal is shown in Fig. 10.

The pressure necessary to create and maintain the airflow is provided by mechanical or natural sources.

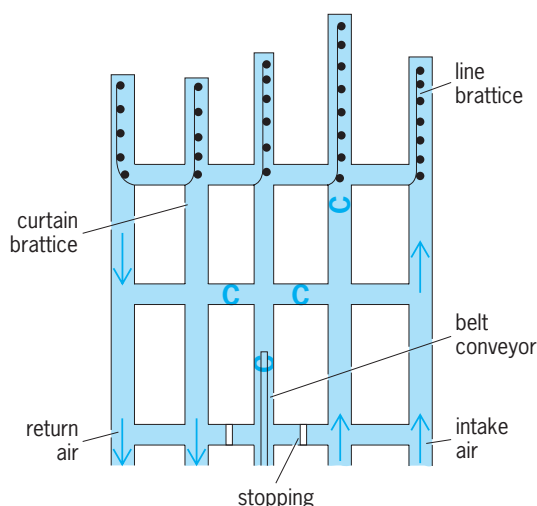


Fig. 10. Horizontal ventilation plan for room-and-pillar mine in coal. (After H. L. Hartman, *Mine Ventilation and Air Conditioning*, 2d ed., p. 433, Wiley, 1982)

The differences in the elevations and air temperatures in the mine openings create a natural draft, called natural ventilation pressure, which varies in both magnitude and direction. Therefore, all underground mines in the United States must be equipped with mechanical ventilators, which in most cases are fans. Compared to the atmospheric pressure, fan pressure is rather small, in most cases of the order of 2–3%. Therefore, it is generally adequate to treat airflow in mines as an incompressible phenomenon.

Some of the air flowing in the higher-pressure intake airways will leak directly into the lower-pressure return airways through the control devices. The quantity of air that is needed for dilution must be adjusted upward to allow for this leakage, and additional air for such needs as the ventilation of idled workings, shops, and battery charging stations must be provided. The volumetric efficiency (defined as the ratio of the volume of air flowing through the working areas to that flowing through the fans) is usually 50–60%. Estimation of the correct fan quantity is critical because the fan pressure and the fan power are proportional to the square and cube of the quantity, respectively. See FAN.

Air velocities are critical from two points of view. From the safety viewpoint, inadequate velocities in the airways can cause undesirable accumulations and layering of gases, and very high velocities can raise dust clouds. From the economic viewpoint, the head loss in a mine airway is proportional to the square of the velocity. Definition of the velocity and quantity requirements will automatically determine the number of airways needed to conduct the air once the shape and size of an airway are determined by mining-machine and ground-control conditions.

The frictional resistance of an airway to airflow is given by Eq. (2). Here R = resistance of the airway

$$R = \frac{KLO}{A^3} \quad (2)$$

($N\text{-s}^2/\text{m}^8$), K = friction factor of the airway ($N\text{-s}^2/\text{m}^4$),

L = length of the airway (m), O = perimeter of the airway (m), A = area of the airway (m^2).

The frictional pressure loss for a quantity flow is given by Eq. (3). Here H_f = frictional pressure loss (Pa), V = velocity (m/s), Q = quantity (m^3/s).

$$H_f = \frac{KLOV^2}{A} = \frac{KLOQ^2}{A^3} = RQ^2 \quad (3)$$

In addition to the friction loss, there are shock losses in an airway due to changes in airway cross section and/or airflow direction, as well as to obstructions in the airway. Shock losses can be as high as 30% of the total pressure loss in a mine ventilation system. The parabolic relationship between the quantity and the pressure or head loss for a given mine resistance is called the mine characteristic.

A desirable feature of any quantity distribution plan is to provide each working area with its own supply of fresh air. This is accomplished by splitting the intake air. Airways can be combined in series or in parallel or in combination to course the air to the various mine locations (Fig. 11). The combined resistance of n airways in series (R_s) is given by Eq. (4),

$$R_s = R_1 + R_2 + \dots + R_n \quad (4)$$

and that of n airways in parallel (R_p) is given by

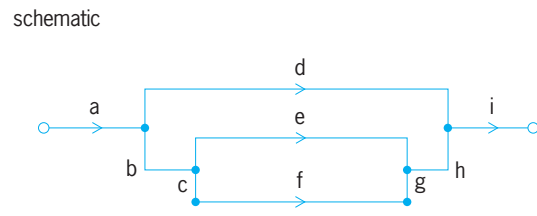
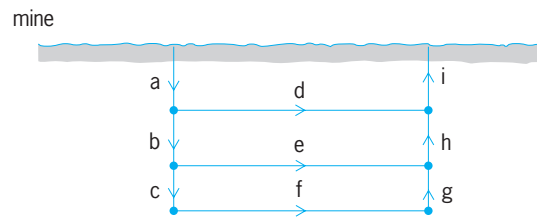
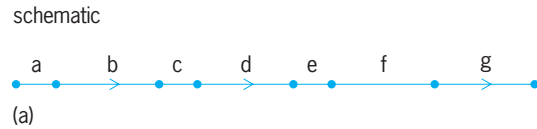
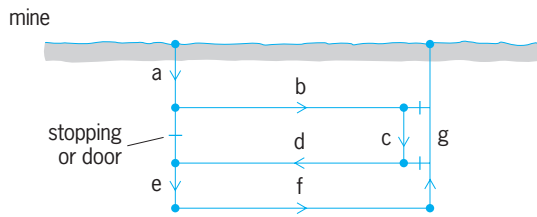


Fig. 11. Ventilation circuits. Letters designate airways. (a) Series circuit (after H. L. Hartman, *Mine Ventilation and Air Conditioning*, 2d ed., Wiley, p. 174, 1982). (b) Parallel circuit (After H. L. Hartman, *Mine Ventilation and Air Conditioning*, p. 128, Ronald, 1961).

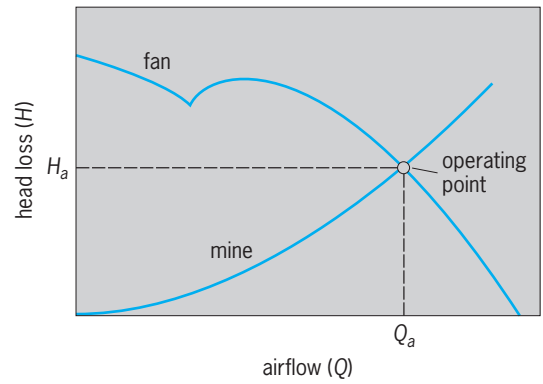


Fig. 12. Fan and mine operating point.

Eq. (5). Parallel flow requires multiple airways be-

$$\frac{1}{\sqrt{R_p}} = \frac{1}{\sqrt{R_1}} + \frac{1}{\sqrt{R_2}} + \dots + \frac{1}{\sqrt{R_n}} \quad (5)$$

tween locations but, as shown above, reduces the resistance to flow. It also enhances the opportunity to escape in the event of an emergency in one of the airways. In practice, combinations of series and parallel flows are used to split the intake air.

The pressure losses in individual splits are calculated once the resistance of the airways and the volumes of air circulating in them are known. The head loss in the system is calculated by accumulating the pressure losses in the individual splits using series and parallel flow laws as applicable. A fan of sufficient capacity that will operate at high efficiencies is selected, based on the quantity of the pressure head required for a mine. The pressure generated by a fan for various quantities of flow is called the fan characteristic. The operating point of a fan in a mine occurs at the intersection of the fan and mine characteristics (Fig. 12).

Mine ventilation systems. The mine ventilation principles apply equally to coal, metal, and non-metal mining. In the United States, underground coal mining is confined to relatively flat-lying seams at shallow-to-modest depths. Similar conditions are encountered in some limestone, salt, trona, and lead-zinc ore mines, and their ventilation systems are similar to those in coal mines. However, most metallic ore deposits also have a relatively large vertical dimension, resulting in mining and ventilation practices that are not similar to those used in coal mining.

In coal mines, the primary pollutants are methane gas and coal dust. In metal mines, a wider range of pollutants is often encountered, such as diesel exhaust, radioactivity, and heat, as well as toxic-gas emissions from mining and blasting. In United States coal mines, the two air streams—the intake and the return—must not be allowed to mix, whereas in metallic and nonmetallic mines, under certain conditions, recirculation of some of the return air to the workings is allowed.

Although underground main fans are prohibited in United States coal mines, they are used in coal and metal mines abroad and in metal mines in the

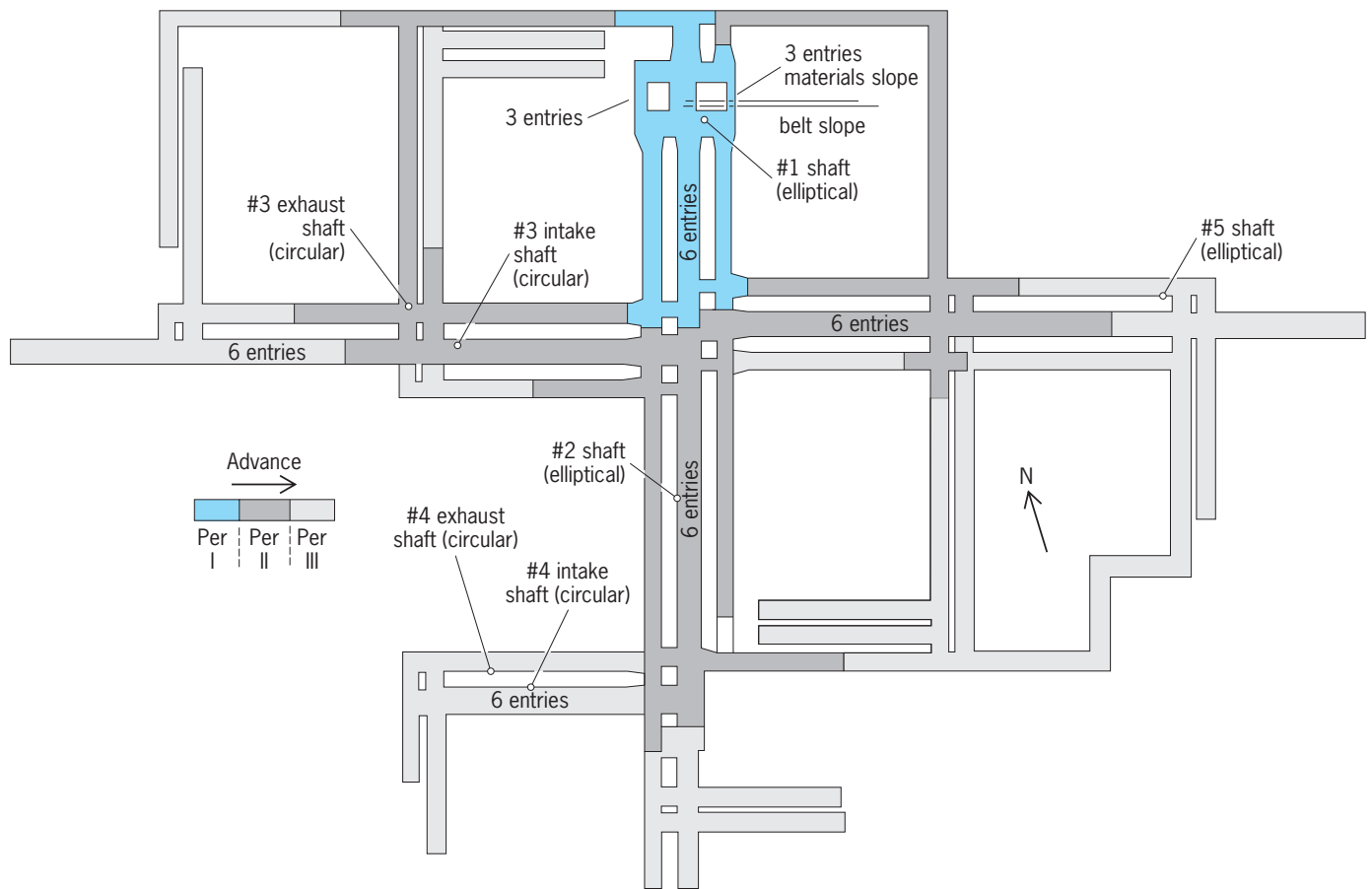


Fig. 13. Ventilation system in a southwestern Pennsylvania coal mine, showing the location of the shafts and slopes, and mine development.

United States. The location of the main fan at the surface is preferable. Still, the complexity of multilevel metal mine ventilation can inhibit the exclusive use of surface fans. Irrespective of coal or noncoal operations, the metric tons of air circulated per metric ton of mined product can vary from 3 to 20 depending on the severity of the air quality and temperature-humidity problems encountered.

Coal mine ventilation system. The ventilation system of a coal mine in the Pittsburgh seam in southwestern Pennsylvania is shown in Fig. 13. The seam is about 2 m thick, is fairly level, lies at a depth of 250 m, and is very gassy. The mine was designed for an annual production of 3 million metric tons from 13 continuous miner sections. The evolution of the mine to full production and the associated ventilation system is shown in three periods. In period I, there are five continuous miner sections that are ventilated using two slopes, and one dual compartment shaft that is divided by a partition running down the middle into an intake compartment and a return compartment. In period II, all the continuous miner units are engaged in driving the mains and developing the production panels. At this stage, the mine is ventilated by the two slopes, and four shafts, two of which have dual compartments. In period III, some of the continuous miner units are engaged in pillar extraction. Ventilation is provided by seven shafts and the

two slopes. Three of the shafts have dual compartments.

During the mains development, three sections are driven (two outlying and one central), each having in its own belt, intakes, and returns. Entries are 2 m high, 4.5 m wide, and 20 m apart. A solid pillar of coal, about 25 m thick, is left between the sections. At the end of development, the central section, which has six entries, serves as the main intakes, with an isolated belt in one of the entries. The outlying sections, which have five entries each, serve as the main returns.

At peak operation, the mine circulated $1180 \text{ m}^3/\text{s}$ (air) with a volumetric efficiency of 60% and connected fan power of 4500 kW. The intake mass air-flow rate per metric ton of coal produced is over 10 tons. While the mine was initially planned for room-and-pillar mining, the longwall method was later introduced as the main production method. The annual mine production increased to about 6 million metric tons. However, the major aspects of the mine ventilation system did not change significantly.

Metal mine ventilation system. The Homestake gold mine located in the Black Hills of South Dakota is more than 120 years old and is one of the deepest and most extensively developed mine in North America. The ore body consists of six major ledges.

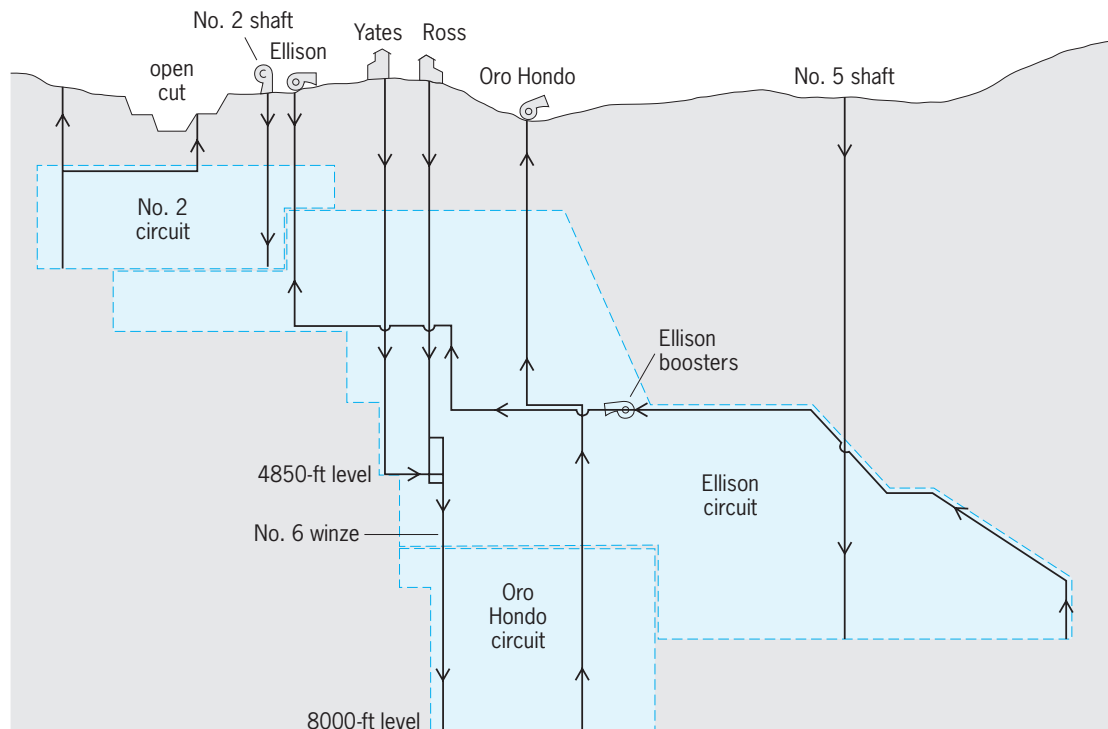


Fig. 14. Homestake Gold Mine, showing the three primary ventilation circuits and major airways.

The mine extends 2500 m below ground, annually producing 10 metric tons of gold from 1.5 million metric tons of ore. The predominant mining methods are mechanized cut-and-fill and vertical crater retreat. The geothermal gradient at Homestake is 2.19°C per 100 m in depth. The mean rock breaking depth is 1750 m, and the virgin rock temperature at this depth is 41°C. The designated reject temperature for the work areas is 29.4°C.

Homestake uses several cooling techniques—chilled water service, portable spot coolers, portable direct-contact spray coolers, refrigeration with water cooling coils, and refrigeration machines in a recirculation circuit. In the deepest mining section, the Virgin rock temperature is between 48 and 56°C. A controlled recirculation circuit, incorporating a refrigeration plant at 6950 ft (about 2000 m below surface).

The mine is ventilated by three relatively independent circuits (Fig. 14). For the upper part, No. 2 shaft is the intake. The fan is installed at the 1700 ft level, and exhaust is partly via raises that break into an open-cut mine, and partly through the Ellison shaft. The middepth part of the mine intakes via No. 5, Yates, and Ross shafts with exhaust up the Ellison shaft. Two underground booster fans at the 3500 ft level and two surface fans at the collar of the Ellison provide the motive force. For the deepest part of the mine (the Oro Hondo circuit), intake is via the Yates and Ross shafts, and through No. 6 winze. The exhaust is up the Oro Hondo shaft through an exhaust fan at the shaft collar.

The total intake air quantity is 391 m³/s, and the total exhaust 445 m³/s. The intake air mass per metric ton of ore production is 10.92 metric tons. In the

recirculation circuit, the fresh intake air volume is 66 m³/s, whereas the total air volume is 220 m³/s. The name-plate diesel power used in the mine totals 6200 kW. The intake air volume per diesel kilowatt is 0.063 m³/s. The total fan power is nearly 5500 kW, and the total air conditioning and refrigeration plant power is also 5500 kW. The mine has an extensive fire detection system, with carbon monoxide sensors at selected locations. The ventilation cost as a fraction of the total underground cost is estimated at 6.2%, and in terms of cost per underground ounce of gold it is \$21.20.

Raja V. Ramani

Mine Illumination

Electric lighting of mine working areas may be developed by permanently mounted light sources, machine-mounted lights, and portable light sources such as the miner's battery-powered cap lamp. Different tasks are performed throughout the mine, and different levels of light are needed for the safe performance of the tasks. Typical locations for permanent light sources are repair and maintenance garages, pump rooms, fuel storage areas, crusher stations, and high back stopes. Underground mining machinery may be equipped with headlights and spotlights similar to those on surface vehicles. Portable light sources are used in a wide variety of locations.

The United States has the most stringent lighting standards of any country. Minimum standards for the illumination of the working places in underground mines have been established in the United States and are described in *Code of Federal Regulations*, Title 30 (CFR 30). Illumination system design and implementation must satisfy the requirements of the

miner, operator, machine manufacturer, illumination equipment manufacturer, and the Mine Safety and Health Administration (MSHA) in order to assure a system's acceptance. Each has similar but differing criteria for acceptance. The goal to be obtained is a system that is practical and maintainable and that will enhance the safe and efficient performance of the miner's tasks.

Standards. The standards specify that illumination shall be provided wherever personnel and mobile mining equipment are operating. This illumination is in addition to that provided by personal battery-powered cap lamps and must provide all surfaces in the miner's normal field of vision with a luminous intensity of at least 0.06 footlambert (0.2 candela/m²). The area to be illuminated varies with the type of mining machinery. For continuous miners and loading machines, the face and ribs, roof, floor, and exposed surfaces of the mining equipment from the face to the outby end (away from the working face) of the machine's bumper is the minimum area to be illuminated. For self-loading haulage equipment such as a scoop, the face and ribs, roof, floor, and exposed surfaces of mining equipment that are between the face and a point 5 ft (1.5 m) outby of the machine is illuminated (Fig. 15). Cutting machines and drills have similar requirements. The area to be illuminated for longwall mining equipment consists of the face, the entire length of the self-advancing roof support system, the control station and head and tailpiece of the face conveyor, and the roof and floor for a distance of 5 ft (1.5 m) horizontally from the control station and head and tailpiece (Fig. 16). Other pieces of mining machinery have differing illumination area requirements.

All lighting fixtures must be permissible (an MSHA designation), either machine-mounted or stationary. A machine-mounted fixture is powered by the machine. Because of rough service conditions, these fixtures are required to remain effectively grounded when loosened or separated from the machines. Stationary lighting fixtures are attached to the mine roof, ribs, or roof-support equipment, such as chocks or shields on longwalls. Energized stationary lighting fixtures can be contacted by miners; thus, the entire system from power source to light must be designed to avoid hazards.

To increase the visibility for miners, the standards require that paint used on exterior surfaces of mining machines have a minimum reflectance of 30% and that red reflectors or tape be installed on each end of the mining machine. Each person going underground is required to wear an approved cap lamp or equivalent portable light and have a minimum of 6 in.² (39 cm²) of reflecting material on each side and back of the hard hat.

Lamps. Four basic types of lamps have been tested and are in use in underground mines: common incandescent, fluorescent, mercury-vapor, and high-pressure sodium-vapor. Each lamp has a system of ballasts and cables, and the entire system must gain permissibility before it can be used in an underground mine. The common incandescent system has

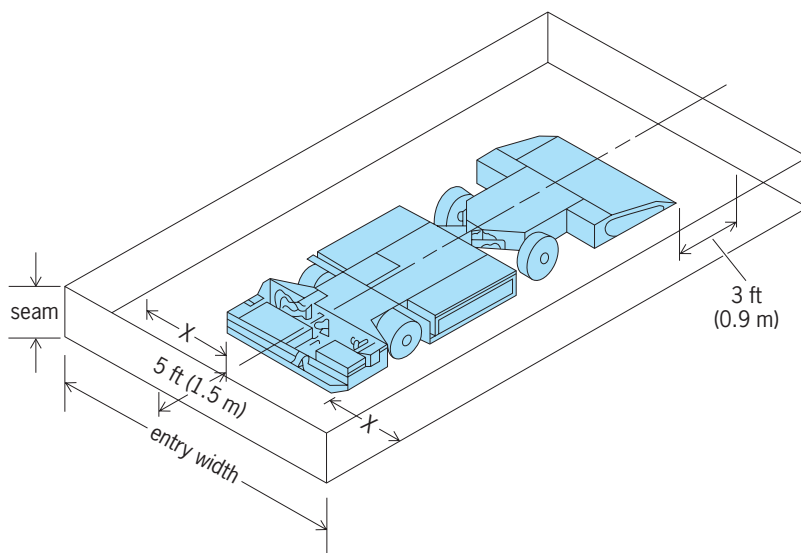


Fig. 15. Area required to be illuminated around a scoop used as a loading machine. Distance X is the distance from the side of the machine to rib or sidewall of the mine opening and is variable. (After Proceedings of the 4th West Virginia University Conference on Coal Mine Electrotechnology, August 2-4, 1978)

been thoroughly tested in practice and is widely used; however, other systems are more efficient and can better withstand shocks. The fluorescent system is ruggedly constructed and instantly started without delicate filaments. Fluorescent lamps give off a soft white light with little glare and have three to nine times the life of incandescent lamps. Fluorescent systems accept common machine voltages and provide more lumens per watt than incandescent systems. Mercury-vapor systems are rugged, long-lived, and efficient, but require 3-5 min warmup time and 5-7 min hot restrike delay time. The mercury-vapor lamp produces a blue-white light and provides nearly three times the lumens per watt compared with an incandescent lamp. The high-pressure sodium-vapor lamp is the most efficient of all, providing five times the lumens per watt compared with an incandescent lamp. The sodium-vapor system goes to full

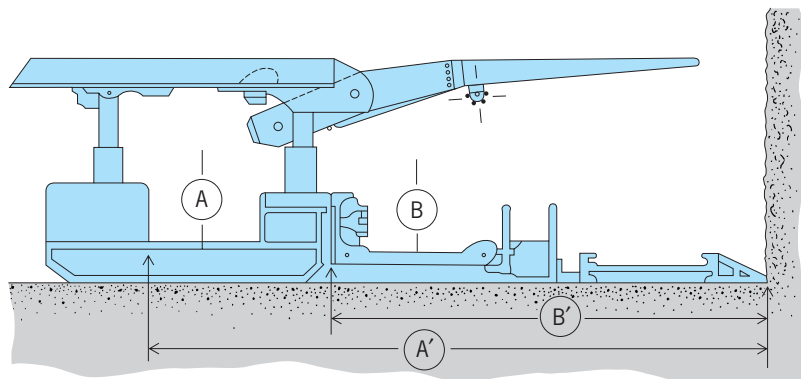


Fig. 16. Longwall illumination system showing area to be illuminated is from the face to the gob side of the travelway, which defines the path by which the miners operating the equipment move from one end of the longwall face to the other. Travelway A is between the hydraulic supports. For the travelway A, the area to be illuminated is A'. Travelway B is under the cantilever roof beam adjacent to the conveyor. For the travelway B, the area to be illuminated is B'. (After Proceedings of the 4th West Virginia University Conference on Coal Mine Electrotechnology, August 2-4, 1978)

brightness in 3 min and requires only 30 s for a hot restart. *See* FLUORESCENT LAMP; INCANDESCENT LAMP; MERCURY-VAPOR LAMP; SODIUM-VAPOR LAMP.

Field evaluation of illumination systems indicates that underground lights often dazzle the workers with glare. Louvers and other light diffusers are added to the system to correct this condition. *See* ILLUMINATION. Malcolm T. Wane

Mine Drainage

Water that is encountered in underground and surface mines has an important effect on the cost and engineering design of many excavations. Its severity varies with local conditions, which relate to geological and hydrological factors, and can be aggravated by topography and climatic conditions. In extreme circumstances, water inflow into mine workings may necessitate the temporary suspension of operations during certain periods of the year. Many mines, even at considerable depth, experience minimum difficulties with water inflow and can be called dry mines.

The existence of water may present hazards in mining and may dictate the methodology of the mining process. Water sources can be planned, such as water delivered to sprays for dust suppression on mining machinery, or can be unplanned, such as water flow from regional rainfall and snowfall, underground aquifers, and abandoned mines.

The dramatic rescue at the Quecreek mine in Somerset County, Pennsylvania, in July 2002 underscores the hazards posed by flooded underground mines. Nine miners accidentally breached a barrier between Quecreek and an abandoned mine thought to contain over 50 million gallons (190,000 m³) of water. After 77 hours of rescue efforts, including deployment of over a dozen pumps whose combined peak pumping rate exceeded 32,000 gallons/min

(120 m³/min), the trapped miners were brought to safety.

Losses from flooding can be devastating to production, life, and property. Therefore, an assessment of the mine drainage problem must be made early in the planning stage. Geologists and hydrologists are needed to identify the sources of water, the probable volume that will be handled, and the corrosive characteristics of the water. Mining engineers take that information and design the mine workings, and install the necessary pumps, discharge lines, and collection points to minimize the impact on the overall mine operations. *See* GROUND-WATER HYDROLOGY.

Pumping principles. Centrifugal pumps are the most commonly used pumps in mine drainage (Fig. 17). When a centrifugal pump is primed, or completely filled with water, its rotating impeller creates a partial vacuum. Atmospheric pressure forces more water into the suction side of the pump. Because centrifugal pumps merely create the vacuum that allows atmospheric pressure to push the water into the casing, they must be installed in proximity to the location from which the water will be drawn. Theoretically, if a perfect vacuum could be achieved, the maximum suction lift has a limiting value of about 34 ft (10 m) vertically at sea level, and depends directly on the atmospheric pressure. Since a perfect vacuum cannot be achieved, the height to which atmospheric pressure can push the water in suction is limited to about 22 ft (7 m) at sea level. Pump deficiencies reduce the theoretical height. If the pump is located below the water level, the intake is under positive pressure, and these limitations do not apply.

Centrifugal pumps may be directly connected to an electric or diesel motor and operate at 1750–3500 revolutions per minute. Single-stage pumps can pump about 5000 gallons/min (19 m³/min) against a 500-ft (150-m) head. Multistage centrifugals, in which pressure is built up by feeding output of the pump successively into several single-stage units in series, can pump to heights over 5000 ft (1500 m). Because the high speeds of rotation invite erosion when gritty water is handled, provisions must be made either to clarify the water or to equip the pump with an erosion-resistant lining and impellers. *See* CENTRIFUGAL PUMP.

Submersible pumps may also be used, and they find their greatest application as well pumps in water supply and dewatering. Special submersible small-diameter motors drive them.

Underground mine drainage. The problems related to water seepage and handling in underground mines are usually more complex than for surface mines. Because operations are conducted at greater depth, the geology and hydrology associated with underground excavations are more varied, and the openings are often located below the regional water table. The water table, in its simplest form, is a nearly horizontal plane when viewed over short distances. However, where a succession of rock layers is present and their permeabilities differ greatly, the geologic cross section may reveal multiple water tables. Thus, it is possible for a deep mine to be completely dry,

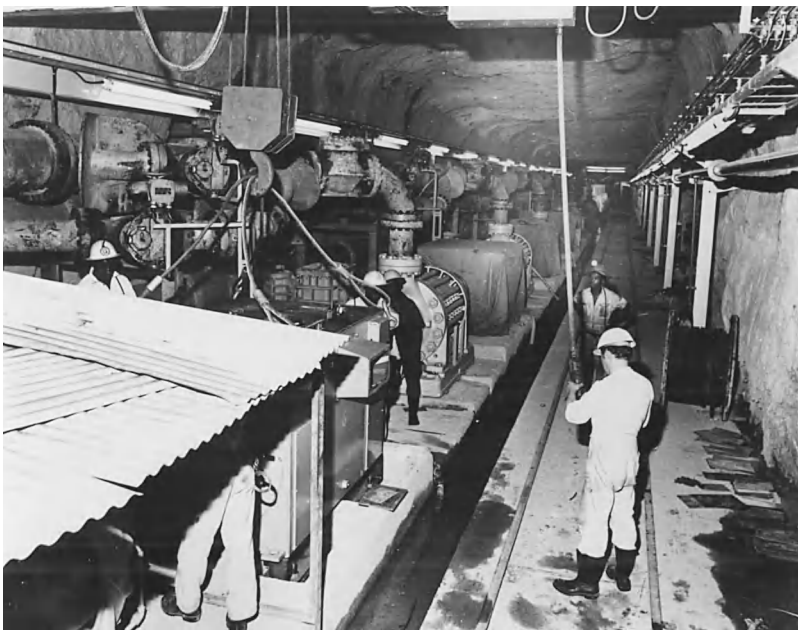


Fig. 17. Pumping installation at a copper mine in Zambia. (Zambia Consolidated Copper Mines Ltd.)

operating in impervious rock, while under several water horizons, aquifers, lakes, and even the sea. In general, some water seepage into openings may be expected in most underground mines. Where water flows are large and persistent, an evaluation must be made to strike a balance between the costs of incrementally increased drainage against the benefit of mining in a drier environment. In addition, wet ore often plugs transfer openings and sticks in mine cars, hoist containers, storage bins, trucks, and stockpiles. The general objective is to ensure an adequate drainage scheme for efficient and safe mining, while keeping the operating cost within tolerable limits.

Underground mine drainage strategies consist of two parts: (1) drainage of the mine entrance (portal), which could be a horizontal adit, inclined slope, or vertical shaft, and (2) drainage of the mine itself.

Adits. It is usually cheaper to handle surface precipitation at ground level than to pump the water out of the mine after it has seeped underground. The expected surface inflow can be caught in adequately placed and designed ditches that divert it from the areas where the water enters the mine. A study of surface topography, surface rock lithology, and permeability, and a definition of the drainage basin contributing to the possible inflow are necessary to establish an optimal surface drainage scheme.

Shafts. Shafts, which provide access to the mineral to be extracted, sunk through water-bearing strata and aquifers, present special drainage problems. A number of methods may be applied during the shaft-sinking operation, depending on the thickness of the aquifer, its permeability characteristics and volume of water, the depth of the shaft, and the geologic rock column (Fig. 18). The shaft site may be drained through a local depression of the water table by the installation of a system of wells. Pumping is continued until the shaft is completed and made watertight.

An alternative solution, where practicable, is to grout the critical sections of the shaft either before or during shaft sinking. There are many types of grouts, such as chemical grouts or cement slurries. Chemical grouts are low-viscosity resins that are able to pene-

trate fine cracks and then cure. The grout is pumped through boreholes, specially sunk for this purpose, which form a ring around the shaft. The expectation is that the grout will find its way through cracks and fissures in the rock and effectively seal them off to water. In many cases, grouting is difficult to control, and grouting pressure must be kept within limits to preclude the creation of more fractures than are being sealed. *See* GROUT.

Under some circumstances, a decision may be made to freeze the area through which the shaft is to be sunk. Time-consuming and costly, this procedure involves drilling holes outside the perimeter of the proposed shaft and installing special pipes through which a refrigerant is pumped until the strata and rock are frozen to the depth desired. The shaft is then sunk through the frozen ground and sealed to water inflow. Thawing of the ground, which takes considerable time, follows.

Mine workings. Water that enters the mine workings is allowed to drain downgrade through ditches toward a collection and pumping point excavated in the rock called a sump. The sump must have sufficient volume to contain the water inflow over a period during which power is off or the pumping facilities are down for repair. Sumps must have provisions for separating sediments from the water. Multiple-level mines can have several sumps. Engineering considerations dictate that the haulage system in an underground mine be located toward the shaft. This is the main reason why sumps are located in proximity to the shaft, a location with easy access for pump inspection and maintenance.

Surface mining. Water inflow in open pits does not present as many difficulties as in underground mines, although inflow fluctuations tend to be greater. The effect of water on the stability of the pit slopes is an important factor. The locations in which equipment must operate and the adequacy of the haulage roads are greatly affected by the nature of the inflow and the fluctuations in its volume.

Surface mines that operate above the water table and in rock of low permeability contend with surface runoff during rainfall and snowmelt only. The pit layout allows the water to be gathered near the pit bottom, where pumps dispose of it through pipelines with comparatively low pumping heads. This scheme, which is simple and low in cost, places the pumping installation in proximity to the mining operation, making it vulnerable to damage from blasting. The water collection site must be moved as mining progresses.

For open pits that are subjected to large inflows, especially those located below the regional water table, and to constant percolation of water into the pits, consideration must be given to lowering the water table around the mine excavation by means of wells, thereby reducing or eliminating the need to use pumping equipment in the pit. The wells may be located in the pit and/or outside the perimeter of the excavation. If wells are located outside the pit, water can be pumped without interruption during blasting. If the pit is cleared of pumping equipment,



Fig. 18. Dewatering through a mine shaft at a copper mine in Zambia. (Zambia Consolidated Copper Mines Ltd.)

pipes, and power lines, ease of maintenance is assured. In addition, the rate of pumping tends to be more uniform, the water is likely to be clearer (because of filtration through the strata), and the walls of the pit tend to be more stable.

Pit stability and the slope at which mining is done affect the overall cost and, in most cases, are directly related to the water residing in the rock being mined. The presence of water also has a deleterious effect on the haulage roads, resulting in increased operating costs. However, pumping through wells is expensive, and good engineering evaluation and planning must precede the decision to resort to such a dewatering scheme. Christopher J. Bise; Stefan Boshkov

Mine Safety

In the United States, mining operations include underground and surface mines, quarries, sand and gravel pits, and preparation plants. Although accidents in such facilities have been greatly reduced over the years, mining remains a dangerous occupation. During the period 1988–1997, more than 500 lives were lost in the United States coal mining industry and another 500 in metal and nonmetal mines. It may be assumed that 2000 worker-hours represents full-time employment of one miner for 1 year. On this assumption, the average coal mine employee working for one year during the period 1988–1997 had about 1 chance in 10 of experiencing a reportable injury; in the metal and nonmetal industry the corresponding odds were about 1 in 16.

Though they still occur, major mine disasters are no longer common. Most fatal mine accidents involve only one victim. The three most frequent types include falls of ground, powered haulage accidents, and machinery accidents. Less common but significant causes of mine fatalities include electricity; hoisting; falling, rolling or sliding material; and slips or falls of persons. The keys to preventing many fatal accidents are following safe work procedures, regular maintenance, thorough cleanup, and use of personal protective equipment—for example, safety belts and lines where there is danger of falling.

In nonfatal accidents, the parts of the body most frequently injured are the back, fingers, and knee, in that order. The most frequent causes are handling materials (for example, lifting) and slips and falls of persons. The majority of nonfatal accidents can be prevented by such measures as planning the job to fit the human factors, following safe work procedures, personal protective equipment (for example, eye shields), machine guarding, and cleanup.

Safety programs. Safety records of individual mines vary. The safest mines in the United States operate for hundreds of thousands of hours without a serious injury. Several common factors seem to characterize those mine operations that are most successful in preventing accidents. One is commitment to safety by top-level managers, who hold supervisors accountable for safety as well as production. At most mines, management has a formal safety policy and requires that safety rules be followed. Many mines employ a safety director who is likely to be a

qualified professional. Many companies have safety departments with technicians and industrial hygienists or other health and safety specialists. The safety function may be linked with loss control or training. Mines with fewer resources may receive assistance with safety and training programs from state or federal programs. Another common factor in safe mines is a cooperative relationship between management and labor. A critical element is thorough, effective, and continued training for all mine employees.

Effective mine safety programs have been found compatible with high productivity. A study published by the National Academy of Sciences (1982) reported that in a survey of underground coal mines the more productive mines also tended to be safer mines. In the 1990s, safety records in United States coal mines continued to improve as coal production reached the highest figures in history.

Under the law, safety in United States mines is primarily the mine operator's responsibility, with the assistance of the miners. Labor also takes a strong interest in safety, often cooperating with management in such areas as training and safety inspections. Safety provisions are included in some union contracts. Many states with extensive mining have state mine agencies whose activities may include training, consultation, rescue, inspections, and investigations. In the federal government, mine safety regulations are administered by the Labor Department's Mine Safety and Health Administration. The National Institute for Occupational Safety and Health conducts research into mine safety and health topics. Others in the mining community who influence safety include equipment manufacturers, educators, safety organizations, industry associations, contractors, and consultants. *See* INDUSTRIAL HEALTH AND SAFETY.

Major safety hazards. These include explosion, fire, toxic or asphyxiating atmospheres, falls of ground, unguarded moving machine parts, operating heavy equipment, and electricity.

Explosion. Methane gas seeps into the air in all underground coal mines and in some metal and nonmetal mines as well, presenting the hazard of a disastrous explosion. Unexpected outbursts may release large volumes of methane. Some underground coal mines liberate over 10^6 ft³ (28,000 m³) of methane in 24 h. Methane is explosive at a 5–15% concentration in air; any spark, electric arc, or open flame may set off an explosion. Coal dust suspended in the air increases the explosiveness of a methane-air mixture. It can also be explosive in the absence of methane and may propagate an explosion for hundreds of feet. Winter is the most frequent period for coal mine explosions, due to dry air and air pressure drops preceding storms.

Prevention measures include sound mine design, ventilation, frequent tests with a methanometer, prohibition of smoking and open flames, use of explosion-proof electrical or diesel equipment, cleanup of coal dust accumulations, and blanketing the mine floor and walls with an incombustible rock dust such as powdered limestone.

Fire. Extensive underground fires may be difficult to extinguish, especially in mines where the mineral itself fuels combustion, such as coal operations. In such cases, the mine may have to be sealed to exclude oxygen until the fire can burn itself out. Water or chemical fire extinguishers may be injected through boreholes. On the surface, fires in buildings or on large mining equipment may also endanger life. *See* FIRE EXTINGUISHER.

Causes of fire include faulty electrical equipment, belt conveyor friction, smoking, welding, explosives, and spontaneous combustion. Cleanup of coal dust and refuse is important in preventing fires and discouraging their spread. Flammable materials should be stored in closed containers, and fire-resistant materials should be used for ventilation curtains and conveyor belting. Fire extinguishers are required in buildings, in shops, and on mobile mining equipment. Automatic systems to detect, and in some cases automatically extinguish, fires are used in many underground coal mines, especially along belt lines, and on some large surface equipment. *See* AUTOMATIC SPRINKLER SYSTEM; FIRE DETECTOR.

Toxic or asphyxiating atmospheres. In fires and explosions, many miners have succumbed to smoke inhalation, carbon monoxide, or lack of oxygen. Such atmospheres also may develop in abandoned mines or unventilated areas of active mines. Other sources of this hazard include blasting and the use of internal combustion engines with inadequate ventilation. Carbon monoxide and oxygen-deficient atmospheres are the most frequent of these hazards, but nitric oxide, nitrogen dioxide, sulfur dioxide, and hydrogen sulfide are formed in some mines. Portable and stationary detectors are available to monitor for the presence of these gases.

Adequate ventilation is the key to preventing accumulations of dangerous gases. Miners should not enter unventilated areas and should allow enough time after blasting for gases to disperse. Unexpectedly breaking into a nearby, abandoned mine has caused deaths due to asphyxiation. These accidents—as well as underground mine flooding—can be prevented by accurate mapping and by long-hole or other exploratory drilling in advance of mining.

In emergencies, protection from toxic or suffocating atmospheres is provided to underground coal miners by self-contained self-rescuers, breathing devices that provide an hour's oxygen for escape. Underground metal and nonmetal miners are required to carry self-rescuers that filter carbon monoxide from the air. Mine rescue teams, which may have to enter areas with hazardous gases to locate missing miners, must use an approved self-contained breathing apparatus.

Falls of ground. Historically, these have been the single most frequent cause of death in mining. In underground mines, they are falls of the mine roof, face, or ribs (walls). In surface mines, they are falls of the highwall.

Ensuring roof stability can be a difficult task, especially in coal mines where thin, fragile shale strata

often overlie the coal seam. Geologic faults, fossil inclusions, and sandstone channels add to roof instability. Steel roof bolts are generally used to pin roof strata to more solid rock above. Bolts, which may be over 10 ft (3 m) long, are inserted by machine into drill holes in the mine roof, and held in place by an expansion shell, a resin grout, or a combination of methods. They may be supplemented with timbers, beams, steel arches, and lagging (planking laid over timbers or arches to prevent small rocks from dropping between them), depending on the roof conditions. Roof control plans should be carefully formulated to match the geological and mining conditions. Loose rock should be taken down before the roof is supported, and the roof's condition should be monitored.

Many roof fall deaths have occurred when miners go under unsupported roofs. Hydraulic roof supports, attached to roof bolting machines, or along the mining face in longwall operations, are used in many underground coal mines. Canopies on mining machines and shuttle cars help protect the operators of mobile equipment from falling rock. Remote-controlled mining machines can allow miners to stay under a supported roof while the machine advances.

In some underground metal and nonmetal mines, the rock requires little or no support; in others the ground may be very loose and extensive ground support is needed. Roof bolts of various types, posts, beams, arches, lagging, and wire mesh are among the support methods employed. In surface mines, rock fall accidents can be prevented by maintaining the highwall at a safe height and angle, stripping off loose material, and minimizing the need for miners to work at the foot.

Unguarded moving machine parts and heavy equipment operation. These are the most frequent causes of death at surface mines and preparation plants. Victims may be struck, crushed, or pulled into machinery. An effective maintenance program is important in guarding against these types of accidents, which may be caused by equipment malfunction or human error. The operation should be planned to minimize these risks; guards, railings, or barriers should protect workers from contact with machinery in motion; good communications should be maintained; and repairs should never be done on equipment in motion.

Electricity. Modern mines have complex electrical systems. Accidents may occur through malfunction or human error. Many electrical accidents occur when personnel knowingly or unknowingly work on energized electrical equipment. Others occur, for example, if high-profile equipment such as a crane contacts an overhead power line. Only qualified persons are allowed to install, maintain, and repair electrical equipment.

Major health hazards. Both mining and mineral processing present health hazards, of which lung diseases are the most important.

Coal mine dust can lead to coal workers' pneumoconiosis (black lung disease), which has killed or disabled thousands of miners. Another lung disease, silicosis, is caused by exposure to crystalline silica

(quartz) dust, which can be present in coal mine strata as well as in metal and nonmetal mines. Cases of silicosis have been identified at silica flour plants, especially in bagging operations, and at coal mines. In underground uranium mines and some nonuranium mines, radon daughters are found. These radioactive decay products of radon gas may cause lung cancer. Lung disease may also be caused by asbestos in the rock or in industrial use, and by large concentrations of other nuisance dusts. *See* ASBESTOS; RADON; RESPIRATORY SYSTEM DISORDERS.

The federal government has set standards limiting exposure to these respiratory hazards. Standards have also been proposed to control diesel particulate, another potential respiratory hazard. The basic method of controlling all these hazards is ventilation. Water sprays on drills, mining machines, and other equipment help to reduce dust. Cleanup of dust accumulations is important, and vacuum-type dust collectors are used in some operations. However, dust control remains a problem in some operations—notably in longwall mining and in the silica flour plants. When miners cannot be adequately protected by engineering controls, respirators are required. Research into better dust control methods is continuing. *See* RESPIRATOR.

A problem in many mines and plants is noise, which can cause hearing loss. Federal regulations limit noise exposure to a maximum of 90 dBA (decibels measured on the A scale of a sound-level meter) over an 8-h shift. Major sources of excessive noise include pneumatic drills, channel burners in dimension stone operations, continuous mining machines, muckers, load-haul-dump machines, bulldozers, draglines, front-end loaders, scrapers, graders, and trucks. At preparation plants, shakeouts used to load railroad cars are the worst sources, generating up to 118 dBA; other noisy equipment includes crushers, sorting screens, chutes, rod-and-ball mills, and kilns. Achieving adequate noise control has not been easy, and a machine-by-machine approach has sometimes been necessary. Compliance also may be achieved by isolating either machinery or operator in a noise-reducing enclosure or by limiting the worker's time in the noisy area. Hearing protection is used when engineering techniques are not adequate. *See* ACOUSTIC NOISE; EAR PROTECTORS.

Another health hazard is heat stress, a danger in some open-pit mines during the summer; prevention measures include acclimatization, drinking liquids, and rest periods. An additional hazard exists in the form of toxic chemicals, such as cyanide and mercury, which are used in some mining and milling processes or contained in the ore.

Training. Inexperienced miners have accounted for a large proportion of fatal accidents. Training and retraining in specific safety and health topics have been required for new and experienced miners by law since 1978. Federal law provides that new underground miners must receive 40 h of training, and new surface miners 24 h. Eight hours' annual refresher training is required for all experienced miners, along with new-task training when changing jobs. Some states have additional training regulations, and some

operators' training programs go beyond the legal requirements.

In addition, mine personnel must be qualified or certified to perform critical tasks such as electrical work, hoisting, gas tests, respirable coal mine dust sampling, and work on permissible diesel equipment in coal mines. Most coal mining states have certification requirements for coal mine forepersons and other supervisors, which usually include several years' experience, specialized training, and lengthy examinations.

Specialized training also is required for mine rescue teams. Each underground mine is required to have two such teams on call in case of emergency; federal regulations specify their qualifications, which include 20 h of initial instruction and 40 more hours of training and practice annually. To sharpen the teams' skills, mine rescue contests are held. Local and regional contests are usually sponsored by industry groups, assisted by federal and state agencies; government-sponsored national contests are held for coal and metal-nonmetal teams in alternate years.

Various types of mine safety training are supplied by a number of sources: mining companies, private firms, state agencies, universities, community colleges, vocational-technical schools, and the National Mine Health and Safety Academy.

Legislation. In the United States, legal responsibility for mine safety and health rests primarily with mine operators. The first federal mine safety legislation, in 1910, founded the Bureau of Mines, whose original role was strictly advisory. During World War II, the federal government took over operation of the coal mines during a strike and issued its first code of safety standards. Enforcement powers came in 1952, when Congress provided for closure of underground coal mines employing 15 or more in cases of imminent danger or continued noncompliance with standards. Successive laws expanded these powers; the Federal Mine Safety and Health Act of 1977 provides for enforcement of safety and health standards at virtually all coal, metal and nonmetal mines, quarries, sand and gravel pits, and mineral preparation plants.

Most important mining states also have safety laws, some of which predate federal involvement. Some states, such as Pennsylvania, Kentucky, Arizona, Nevada, and West Virginia, maintain their own inspection program. Others, such as North Carolina, concentrate on training and consultation, leaving most inspections to the federal government.

Federal mine safety and health rules, issued and enforced by MSHA, are published in Title 30 of the *Code of Federal Regulations*. In most cases, separate rules are provided for different types of operations. In issuing new or revised rules, MSHA encourages industry, labor, and others to provide input through written comments, public conferences, and hearings.

The law requires four complete federal inspections of each underground mine per year and two inspections of each surface mine. Additional inspections are to check for correction of violations found on a regular inspection, make spot checks at mines with excessive methane or other special hazards, or focus on specific areas such as electricity or dust

control. The law provides for a representative of the miners to accompany federal inspectors on each regular inspection.

A federal inspector who finds a violation of safety or health standards must issue a citation, which entails a fine that may range from \$55 to \$55,000, depending upon seriousness and other factors. If violations remain uncorrected or there is imminent danger, the mine or section may be closed. Criminal penalties are possible for knowing and willful violations.

MSHA also investigates fatal accidents, publishes current safety and health information, and supplements enforcement with a wide variety of programs to provide penalty-free help on engineering problems and training and safety programs.

J. Davitt McAteer; Katharine Snyder

Transportation and Storage

Transportation and storage in any mine, surface or underground, involves the safe, efficient movement of personnel and material at the least possible cost. Total cost includes the cost of lost-time accidents, the expenditure of time and energy, and the effect on the environment. Material includes ore, waste, mining equipment, powder, steel, track, timber, cement, maintenance and plumbing supplies, and other necessities for an operating mine.

Transportation. Loading, hauling, and unloading of ore and waste may involve transfer in stopes, chutes, or ore passes to intermediate levels; haulage to shaft pockets; skip loading or caging of ore cars; and hoisting, haulage, and feeder transfer to storage or active mill operations.

Gravity is employed wherever possible to move the mined rock: from the working face to haulage levels; out of skip pockets, ore bins, or stockpiles; and from one conveyor to another.

Handling of ore and waste at the shaft station and surface is governed by several factors: the underground transport of the ore; the types and sizes of the cars; the inclination of the shaft; the method of hoisting; the tonnage and physical characteristics of the ore; the number of ore classes that must be kept separate during transport; and the surface topography of the site.

The type and size of equipment employed depend upon the different mining conditions and the required results. An equally important factor is the operational cost, as dictated by the character of the material to be moved, the daily tonnage, the methods of loading and dumping, and the distances involved.

In smaller mines, haulage may be accomplished by hand-tramming or animals. Mechanical haulage can be by rail (Fig. 19), diesel vehicle (Fig. 20), scraper or slusher, or rope conveyor. Compressed-air machines have an obvious advantage in mines that need ventilation. Haulage by rail is the primary method used in large-scale mines, both surface and underground (Table 1).

Conveyors. If there is sufficient dump room or storage capacity near the mine, a system of belt conveyors can handle material at high rates and relatively



Fig. 19. Hopper cars being loaded by a belt conveyor. (Burlington Northern, Inc.)

low costs, but only if proper feed control of a sized material allows a continuous, even flow that matches the system design. Other factors that determine the practicality and size of a conveyor system are the rate at which the material must be handled, the material's density and stickiness, the dusting or degradation on transfer, and the need for the system to handle more than one product.

Slurry transportation. Pipelines have been successfully used to transport many different solid materials: coal, phosphates, limestone, fly ash, copper concentrates, gilsonite, mill tailings, slags, and sand fills. The dry material is first pulped to form a slurry and then pumped to the destination, where it is dewatered. Additional examples of hydraulic movement of material are the loading and unloading of finely divided ore into ore carriers such as ships or sealed cars, the dredging and placing of fill, and the emplacement of backfill in mine stopes.

One notable slurry pipeline completed in 1983 is a 237-mi (382-km) pipeline in northern Mexico, which carries iron ore concentrates from two separate mines to the pelletizing plant. See PIPELINE.

Miscellaneous. Rope or cable haulage, particularly aerial trams, are employed to move ore carriers over very rough terrain or along grades that are too steep for locomotives and trucks. The cost of transporting



Fig. 20. A 16-ton (14.5-metric-ton) diesel locomotive with exhaust water scrubber for use in coal mines. (Brookville Locomotive Division of Pennbro)

TABLE 1. Summary of haulage systems*

| Conditions | | Equipment† | | | | | | | | | | | | |
|-------------------------------------|--|------------------|-----------------------|------------------------------------|------------------------------------|-------------------------|---|------------------|-----------------------|-------------------------|------------------|----------------------------|---------------------------------|---------------------------------|
| | | Bulldozers | Scrapers | | | | | Truck | | | Train | Conveyor | Skip | Pipeline |
| | | | Tractor-drawn scraper | Underpowered, rubber-tired scraper | Full-powered single-engine scraper | All-wheel-drive scraper | Rubber-tired tractor with trailer scraper | Rear dump | Semitrailer rear dump | Semitrailer bottom dump | | | | |
| Material | Rough, blocky Max. 36 in. (91 cm) Max. 24 in. (61 cm) Fines | 1 1 1 1 | 1 1 2 1 | 2 2 1 1 | 2 1 1 1 | 3 2 1 1 | 3 3 1 1 | 3 3 1 1 | 1 1 1 1 | 3 2 2 2 | 1 1 1 1 | 4 3 1 1 | 1 1 1 1 | 4 |
| Length of haul | 0–300 ft (0–90 m) 300–500 ft (90–150 m) 500–1000 ft (150–300 m) 1000–1500 ft (300–450 m) 1500–5000 ft (450–1500 m) 5,000–10,000 ft (1500–3000 m) 10,000–15,000 ft (3000–4500 m) Over 15,000 ft (4500 m) | 1 2 | 1 1 2 3 | 2 1 1 2 | 3 2 1 1 | 3 2 1 1 | 3 3 1 1 | 3 3 1 1 | 3 2 1 1 | 3 2 1 1 | 2 2 1 1 | 4 4 4 4 3 4 | 4 4 4 4 4 3 3 | 4 4 4 4 4 3 2 |
| Ground conditions | Good Wet, soft | 1 1 | 1 1 | 1 1 | 1 3 | 1 1 | 1 1 | 1 1 | 1 1 | 1 1 | 1 1 | 1 3 | 1 1 | 4 2 |
| Maximum adverse grade | + 3% + 5% + 10% + 15% + 20% + 20% | 1 1 1 1 | 1 1 1 1 | 1 3 3 3 | 1 2 3 3 | 1 1 2 3 | 1 1 1 1 | 1 2 3 1 | 1 2 3 1 | 1 2 3 1 | 1 2 3 1 | 1 4 3 2 1 4 | 4 4 4 4 4 4 | |
| Flexibility under varied conditions | Good Fair Poor | 1 1 1 | 1 1 1 | 1 1 1 | 1 1 1 | 1 1 1 | 1 1 1 | 1 1 1 | 1 1 1 | 1 1 1 | 2 3 1 | 3 1 1 | 1 1 1 | 4 4 4 |
| Daily production rate | Low Medium High | 1 | 1 3 | 1 3 | 1 2 2 | 1 1 1 | 1 1 1 | 1 1 1 | 1 1 1 | 1 1 1 | 2 1 1 | 2 2 2 | 3 2 2 | 4 4 4 |
| Total tonnage | Small Medium Large | 1 | 1 3 | 1 3 | 1 1 1 | 1 1 1 | 1 1 1 | 1 1 1 | 1 1 1 | 1 1 1 | 3 2 1 | 2 3 1 | 3 1 1 | 4 4 4 |

*After Society of Mining Engineers of AIME, *SME Mining Engineering Handbook*, vol. 2, pp.17–23, 1973.
 †1. Should be considered. 2. May be considered. 3. May be considered under certain conditions. 4. May be considered, special situation.

material over very rough terrain is essentially confined to the initial cost of the installation.

A scraper is ordinarily employed as a mucking and loading machine, but it can also be used as a conveyor to transport ore and waste hundreds of feet when large quantities of ore must be moved comparatively short distances. Scrapers include bulldozers and scraper-loaders. When dragging is done at higher speeds, scrapers require a smooth base or bottom.

Economics. In any earth-moving operation, more profit may be earned or lost by equipment scheduling than by any other single facet of the project. Ideally, every item of the equipment fleet must be continuously working.

If mechanical haulage can be utilized full-time, the savings in operating costs can be substantial. Idle machines pay no dividends. Open-pit mines in which haulage is by rail should have large ore reserves. The construction of truck and handling facilities for effective

operation and maintenance of trains is expensive, however, as is the acquisition of locomotives and cars.

A U.S. Bureau of Mines report has estimated that handling and transportation of ore and waste has accounted for one-fourth of total mining costs.

Proper handling of waste rock can also be economically important. Waste rock is usually disposed of as cheaply and conveniently as possible. This rock, however, could be a submarginal, low-grade ore. Consideration should therefore be given to possible future processing, when improved economics or technical advances allow profitable handling. Many old mine dumps and mill-tailing piles, originally regarded as worthless, have produced valuable products in later years.

Storage. Stockpiles or intermediate storage bins are used to accumulate the product when carrier schedules cannot be diverted to useful development



Fig. 21. A 72-in. (22-m) conveyor belt at the Caridad Mine, Nacozari, Sonora, Mexico.

work or until a means for further transfer of the material becomes available.

It is customary to store large tonnages of crushed ore in piles on the ground because storage in elevated bins is costly. Depending upon the weather and type of ore (for example, dusty or soluble), storage areas may be covered. Each installation is a study in itself and can affect the entire mine operation.

Ore can be reclaimed from a stock-pile by draglines, scrapers or dozers, bucket-wheel excavators combined with stacker and tripper conveyors or mechanical feeders; or from controlled draw points onto conveyors in a tunnel below the stock-pile. Again, each installation varies with the type of mine.

A traveling gantry stacker is frequently employed to spread the ore in the manner that is most advantageous for future use. For example, the ore might be carefully bedded for blending, leaching, or expediting reclamation or reloading.

Technological advances. Electronic identification of haulage equipment and amounts of material carried has been under development. In a computerized truck-dispatching system, unit and product data are collected and transmitted without reliance on human operators.

Paved haulage roads are safer and more economical than unpaved roads. Trolley-assisted power systems have come into use in the steep graded areas.

Modern technology has developed multimile conveyors that follow terrain undulations and allow major savings in the civil engineering and earth moving required in construction of mining and milling operations. Regenerative conveyors or declined con-

veyors can produce, rather than consume, power; one example is the primary ore conveyor belt at La Caridad, Nacozari, Sonora, Mexico (Fig. 21), which carries 79,000 tons (72,000 metric tons) per day a distance of 1070 ft (960 m). Under normal conditions, it can generate 1.2 MW of electrical energy per day because of the downhill character of the installation.

Trucks have continued to grow in capacity, efficiency, and operational adaptability.

The planning, packaging, storing, and timing required for the safe, efficient transportation of personnel, equipment, materials, and supplies will always be as essential to a successful mining operation as the movement of ore and waste. David D. Rabb

Field Sampling and Ore Estimation

Sampling is a critical activity in the evaluation, development, and mining of a mineral deposit. The results of sampling are used in defining ore bodies, estimating ore reserves, mine planning, and controlling the quality of ore delivered from the mine to treatment plants. The information provided by sampling is organized into computer databases for geographic information systems (GIS) and ore estimation. See GEOGRAPHIC INFORMATION SYSTEMS.

Sampling. A sample is a small portion of material from a mineral deposit, and an accurate and unbiased collection of samples is intended to represent the entire deposit or a specified portion of the deposit.

The quality most commonly determined by sampling and by the laboratory testing or assaying of samples is the content of one or more particular elements or compounds. Amounts of impurities such as phosphorus in iron ore and ratios between certain metals such as chromium and iron in chrome

ore are also determined. Material with a content or grade sufficient to be economically minable is designated as ore. The grade of nonferrous metal ore is generally stated in percent of one or more contained metals such as copper, lead, or zinc. The grade of certain materials such as titanium ore, molybdenum ore, and potash is stated in percent of an index compound: titanium dioxide (TiO_2), molybdenum disulfide (MoS_2), or potassium oxide (K_2O). The grade of precious metal ore and of the precious metal content in other ores is stated in troy ounces per short ton or grams per metric ton of material. The grade in placer deposits is expressed in grams of the recoverable metal (gold, tin, titanium) per cubic meter of material.

Additional qualities determined by sampling a mineral deposit and by laboratory analysis deal with mineralogy and with metallurgical characteristics. An ore may consist of several minerals that contain the desired element but have different recovery characteristics, and an ore may have a composition or texture that affects its amenability to processing. Samples of industrial mineral deposits such as clay, limestone, and sand are tested for desired physical properties as well as for mineralogical components.

The most common methods of taking samples in a mineral deposit involve channel sampling, grab sampling, and bulk sampling in exposed ore and drilling holes into hidden ore.

Channel sampling. In vein and stratiform deposits, the basic procedure is to cut or chisel a linear slot or channel at a uniform width and depth across a fresh and cleaned exposure of the mineralization (Fig. 22). The channel is most commonly cut in the face or across the back of a mine working, with the broken material collected in sacks or on a plastic sheet. Where the mineralization consists of distinctive zones, a separate sample is taken from each zone so that the assay results can be composited in various ways to permit selective mining. Where the wall rock on each side of the recognized mineralization may also contain some values and may have to be incorporated as dilution in mining, additional samples are taken. Chip-channel sampling is similar to a more precise channel sampling but is less time-consuming. Chip channels consist of fragments broken in uniform amounts from a broad band across the exposed face or back. Channels are repeated at appropriate intervals along the ore zone.

Grab sampling. This procedure is used more commonly in massive and disseminated deposits where the face, back, or walls of a mine working are entirely in ore mineralization. The sample may consist of fragments chipped at uniform intervals from an exposure (chip samples), material scaled from an exposure (panel samples), a randomly chosen collection of fragments from a freshly blasted mine face, or amounts of material taken in succession from a trench. Samples taken by hand at random from cars and from conveyor belts are also referred to as grab samples.

Bulk sampling. In this procedure, large amounts of material are taken, primarily for the evaluation of dispersed mineralization and for metallurgical test-

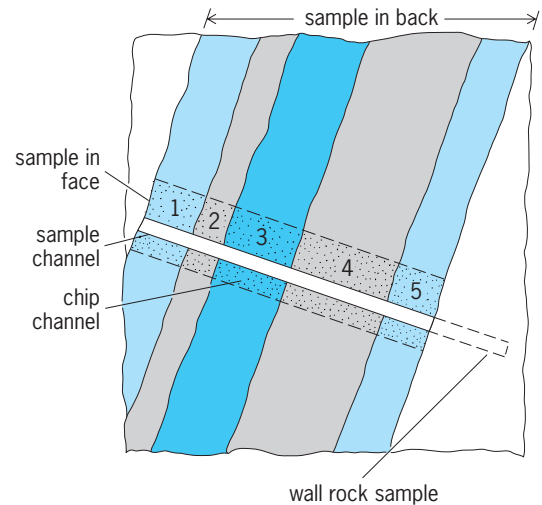


Fig. 22. Sampling channel and chip channel in a zoned vein exposed at a mine face, with a sample from each zone.

ing. In metallurgical pilot plant testing, the entire ore content of an underground mine working or of a surface trench, amounting to hundreds or thousands of tons, may be collected.

Drill-hole sampling. Drill-core and rotary or percussion drill-hole cuttings furnish most of the samples used in delineating ore bodies, evaluating prospects, and extending ore reserves beyond existing mine workings. In addition, cuttings from blastholes are often sampled and studied or quickly assayed for grade control during mining. After geologic examination, sections of drill core are commonly split lengthwise, with half of the core sent to the laboratory for assay and half kept for record. Drill-hole cuttings—chips of material loosened by the drill-hole bit—are examined in sequence at the drilling site and taken in their entirety or in part as samples. See DRILLING, GEOTECHNICAL.

Sampling patterns. In defining ore bodies and calculating reserves, the choice of a sampling method and a sampling pattern are dictated by the accessibility, nature, and size of the mineral deposit. The outcropping portion of a deposit and the portion exposed in underground workings will permit channel, grab, and bulk sampling. Hidden deposits and ore projections from exposed deposits permit sampling by drill holes, either from surface sites or from stations in existing underground workings. An exposed vein-type or tabular stratiform deposit with well-defined walls is suited to channel sampling, with preliminary channels taken at wide intervals and subsequent channels taken for detail at intervals of a few meters. The same kind of deposit may also be sampled in successive “fences” of drill holes across the trend; the holes may be drilled from a few sites in fanlike groups designed to cut the ore body at appropriate intervals (Fig. 23). Large massive and disseminated deposits are commonly sampled in a systematic grid of drill holes that pass through the ore at intervals of 100–400 ft (30–120 m) and by bulk sampling in underground exploration workings. Placer deposits are bulk-sampled in trenches across the trend, in pits or

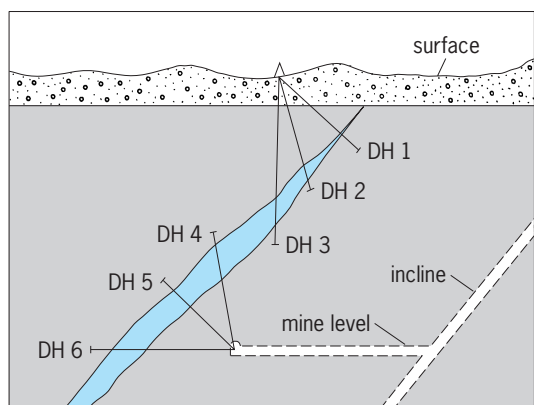


Fig. 23. Cross section of vein sampled by drill holes (DH) from surface and underground sites.

shafts at regular intervals, or by large-diameter non-core drill holes.

Grade control. Sampling for grade control, selectivity, and teleming automation during mining makes use of grab samples, blasthole samples, and samples of broken ore in transit. Broken ore in mine cars or on conveyor belts is commonly sampled by automated and robotic mechanical devices. The grade of uranium and potash mineralization in broken material and in blastholes is commonly estimated on site by the use of gamma-ray spectrometer probes. The grade of silver, lead, zinc, and copper mineralization in broken ore, mine faces, and drill holes can be estimated by the use of x-ray fluorescence scanners or probes. Similar instruments have been used in estimating the in-place grade of gold, nickel, and other metallic ores. See X-RAY FLUORESCENCE ANALYSIS.

Sample treatment. Samples must be large enough to be representative of their location in a mineral deposit. In a relatively uniform and fine-grained ore, an appropriate channel sample may weigh 5 lb (2.3 kg); in an irregularly mineralized and coarse-grained ore, a sample on the order of 10–20 lb (4.5–9 kg) may be needed. The sample must subsequently be reduced in amount for laboratory analysis; this is done in stages during which the ore fragments are crushed to successively finer size and thoroughly mixed. In the laboratory, a small portion of finely ground and homogenized sample is taken for instrumental, chemical, or fire assay.

Assaying. An assay (quantitative analysis) of the constituents in a mineral material is made through various combinations of laboratory procedures. The procedures incorporate classic methods of gravimetric and volumetric chemical analysis, while a variety of instrumental methods are used with the requisite accuracy and precision. See ANALYTICAL CHEMISTRY.

Fire assaying, the historic and current standard as a gravimetric method, is unique to mining. This method is used for determining the content of gold, silver, and platinum metals in ores. The steps in making a fire assay involve a furnace fusion of the sample with a lead flux, a recovery of the precious metals into a metallic bead by further heating (cupellation), and a weighing of the bead. The precious metals are then identified and quantified by further chemical

separation or by instrumental analytical techniques.

Sample and ore calculations. In assigning a composite assay value to the samples taken within a designated channel or interval of drill core, the length of each sample is accounted for in calculating a weighted average (Table 2).

In determining the average grade of a deposit or of a zone within a deposit, values at point, channel, and bulk sample locations are projected to reference levels and attitudes and are assigned specific volumes of material. With the use of computer hardware for data collection and graphically oriented mining software, the volumes of material are taken with measurements of density or specific gravity and are calculated into tonnage. In conventional geometric techniques, values at the sample locations may be connected to form a network of triangular prisms, or the value at each sample location may be given a surrounding area and volume of influence extending half-way to the next adjacent sample location. In the latter approach, the deposit is divided into a network of polygonal blocks. In combined geometric-statistical techniques, a search area representing a block of material is moved across the map of sample locations, and the value at the center of each successive search position is influenced by all of the included sample values. In the weighted moving average technique, each sample value is weighted by an inverse power of its distance from the center of the search area so that the nearer samples receive the greater influence.

Statistical techniques for calculating grades within specific volumes of ore make use of grids and polynomial trend surfaces fitted to the overall pattern of sample data, or make use of weighting factors determined from directional trends within the deposit. The latter is a widely used geostatistical approach in which a graph or variogram of values determined for increased sampling intervals in each of several directions is fitted to an idealized curve to identify the range of influence between samples. The estimation of values at individual locations employs the variogram function and kriging, a moving weighted average technique. See STATISTICS.

Coal sampling. Coal deposits are sampled to determine their properties as sources of fuel and coke. Characteristics are determined by proximate analysis for volatile matter, fixed carbon, moisture, ash, and heating value, and by ultimate analysis for chemical

TABLE 2. Composite assay values*

| Sample | Length, m | Silver, g/metric ton | Length × assay product |
|--------|-----------|----------------------|------------------------|
| 1 | 0.5 | 220 | 110 |
| 2 | 0.3 | 270 | 81 |
| 3 | 0.7 | 550 | 385 |
| 4 | 1.0 | 350 | 350 |
| 5 | 0.5 | 200 | 100 |
| Total | 3.0 | | 1026 |

*Based on sampling channel shown in Fig. 22. Average grade (1026/3) is 342 g/metric ton.

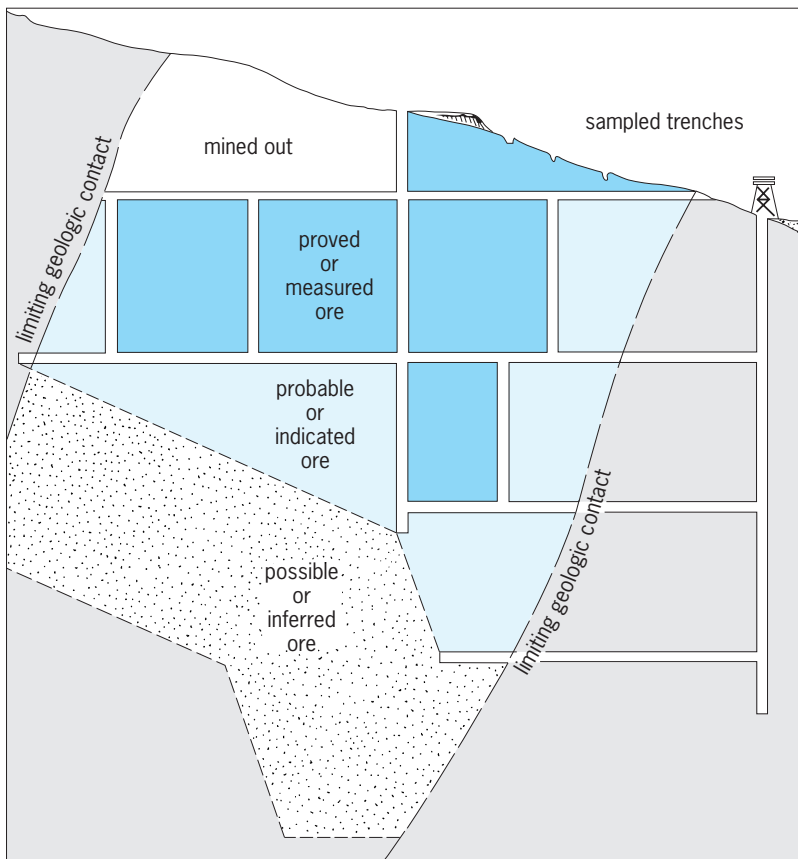


Fig. 24. Longitudinal section of a vein deposit, with ore reserve classifications.

composition. Engineering properties of coal seams are determined from samples by direct physical tests and indirect laboratory tests. Sampling is most commonly done by drill hole and channel methods.

Drill-hole sampling of coal involves core and rotary (chip) drilling, generally from surface locations. Sample intervals are selected according to recognizable coal units within the seam, and the samples of core or cuttings are placed in sealed containers to preserve natural moisture and minimize oxidation. Bulk samples are taken from drill holes by large-diameter core and by the washing of cuttings from induced cavities.

In trenches and underground workings, channel samples across a coal seam are taken as composites of shorter individual samples. Larger bulk samples are taken by blasting of the full seam.

Ore estimation. The quantity of ore in a block of mineralization and in an entire deposit is estimated by measuring areas and thicknesses and by considering the calculated volumes, tonnages, and grades.

A tonnage and grade estimate of the resource or geological reserves in a mineral deposit amounts to an inventory of the in-place mineralization. With graphical software for modeling and mining simulation, an estimate of ore reserves takes the minability of the deposit into account, with appropriate provisions for the method of mining, the percentage of ore that will actually be removed, the dilution of the ore in mining, and the grade and tonnage of material above the

minimum economic value or cutoff grade. The rated economic value of ore mineralization will also take into account the level of its metallurgical recovery. For the part of a mineral deposit so thoroughly sampled that its outline, tonnage, and grade are assured, the estimated reserves are called proved or measured. In other parts of the deposit where sampling has been less thorough but where geologic information is sufficient to make reasonably secure projections, reserves are probable or indicated (Fig. 24). On the fringes of a deposit where sampling is sparse but where there is geologic evidence of continuity, the reserves are possible or inferred. William C. Peters

Mine Evaluation

Mines represent the culmination of successful investment in discovering and developing economic mineral deposits. To understand mine evaluation, it is necessary to consider the various stages of the mineral supply process, whereby initially unknown mineral deposits are discovered, developed, produced, and reclaimed (Fig. 25). In the broad sense, mine evaluation considers assigning value to mineral deposits at all stages of the mineral supply process. Because every mineral deposit is unique, the task of assigning value is complicated. Furthermore, the mineral industry is subject to the offsetting forces of depletion and technology. The finite size of mineral deposits means that mines eventually run out of ore and are closed. Over time, this depletion effect should increase the value of remaining deposits. However, technological advances at all stages of the mineral supply process work to offset the impact of depletion by allowing us to find new deposits and lower the costs of production at existing mines.

Exploration. As shown in Fig. 25, exploration proceeds through several possible stages from early reconnaissance work, where targets are identified in preparation for mine development. At the primary exploration stage, the value of a mineral deposit is highly speculative. This reflects the fact that very few mineral deposits have the requisite geological and technical characteristics to become economic mineral deposits (mines). Several hundred mineral properties may have to be examined for every one that becomes a mine. Overcoming these high odds is both time-consuming and expensive.

At each stage of the exploration process, money is spent to acquire information. This information is then used to make a go/no-go decision. Does the project merit the cost of further exploration? If so, the expectation must be that the probability of success times the expected return is higher than the cost of continuing the exploration program. As projects proceed through the various stages of exploration, the probability of becoming an economic deposit increases, as some of the discovery risk has been resolved. The value assigned to any given property should increase proportionately with the reduction in uncertainty and the increase in information.

Feasibility. As projects move to the delineation phase, the economic factors become more prominent. Deposits that are deemed to have potential

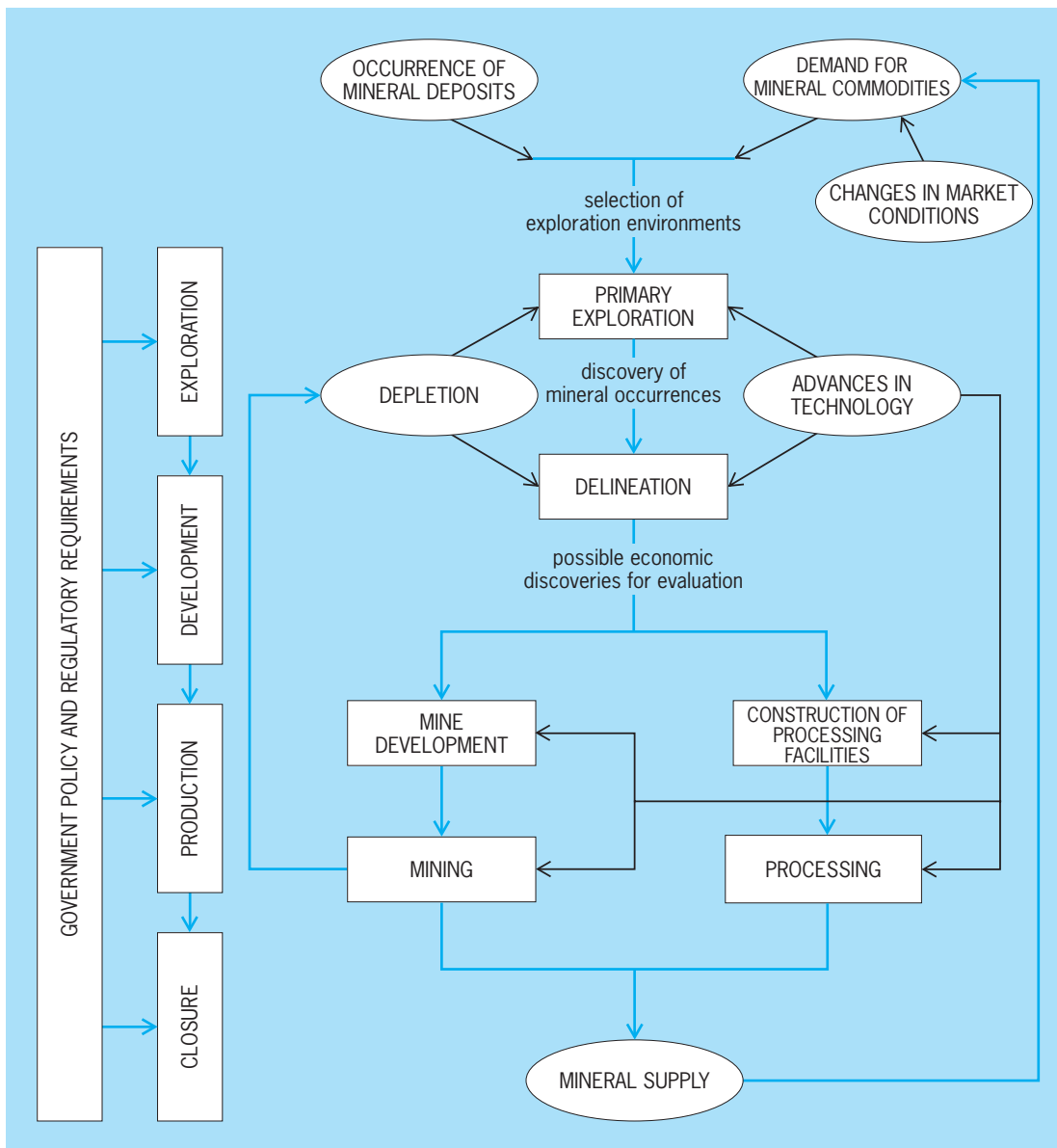


Fig. 25. Mineral supply process.

to become mines are assessed with scoping studies. These early technical and economic assessments are another stage in the go/no-go process. If go is the result, a feasibility study is done to determine if a project should be developed. Feasibility studies usually involve detailed technical evaluations of ore reserves, mining methods, processing recoveries, and waste disposal. Extra information may be collected in the form of core sampling, bulk sampling, or pilot tests. Feasibility studies may take many months or even years to complete and cost many millions of dollars. In the economic portion of a feasibility study, detailed projections are made for the revenues and costs associated with development and production. With respect to revenues, forecasts must be made of metal prices, inflation rates, exchange rates, and annual production of the various payable metals. On the cost side, estimates of capital expenditures are made based on the size of the

ore reserve and the production capacity to be installed. These costs can be broadly categorized as mine development, mine equipment, processing facilities, and infrastructure. Production or operating costs are usually subdivided into mining, processing, and overhead/administration, but include detailed breakdowns on individual components such as supplies, labor, power, water, transportation, and refining. Estimates also must be made of the impact of taxes and royalties on project cash flows.

Discounted cash flow measures. The feasibility study is the last step in the go/no-go process. A positive feasibility results in the recommendation that a mine be established on the deposit, with the expectation that the returns from the mine will outweigh the cost of building and financing its construction, as well as closing the mine and reclaiming the site. The feasibility go/no-go decision is traditionally based on a set of discounted cash flow indicators, including

net present value and rate of return on investment. A net present value of greater than zero indicates a positive investment opportunity, meaning that the benefits (positive cash flows) outweigh the costs (negative cash flows), including an allowance for the cost of the capital used in the investment. The expected net present value determined in the feasibility study represents the inherent economic value of the mine. In addition to a positive net present value, most companies impose a profitability requirement in terms of return on investment. For many mining companies, an after-tax rate of return of 15% is considered the minimum level of profitability.

Once a project is in production, the value of the mining operation at any given time is the difference between benefits and costs over the remaining life of the mine. The net present value at any given time, looking forward over the remaining mine life, should represent the value of the mine to the owners. Further investment decisions made after a mine has been put into production are considered in the same manner. The cost of further investment, such as an expansion of mine capacity, is made on its own merits. If the anticipated increase in benefits outweighs the cost of the expansion, it should proceed.

Sensitivity and risk analysis. Uncertainties are a reality of the exploration and mining business. Feasibility study estimates are made on the basis of incomplete and uncertain information with respect to the physical, technical, and financial parameters of mine development. Expected value results need to be considered in the context of this uncertainty. This is done by a sensitivity analysis, whereby one or more of the input values is changed and the impact on the resultant economic criteria is calculated. For example, if the price of copper were 10% less than expected, how much would the net present value of the project be lowered? More complex analysis is possible by considering simultaneous changes to multiple variables. Probabilistic risk analysis and real option studies can be used to measure these impacts.

Michael D. Doggett

Computers in Mining

Computers are used at all stages of a mining operation to help geologists, engineers, designers, and management make informed decisions about the size and location of mineral deposits and how they can be profitably extracted. Computers have helped the mining industry be more systematic, make decisions based on quantifiable information, and implement procedures that are practical, are profitable, and meet stringent environmental safety requirements.

Exploration. Mineral deposits are located by examining surface samples, as well as samples taken from core drilling. Data collected from surface and drill samples are used to define the location of mineralization in three-dimensional space and also to define the elements and minerals composing the rock. Enormous amounts of data are collected, and computers provide a tool for analyzing the data in a timely manner.

The addition of a palmtop or laptop computer to a geologist's field tools is a recent development.

Portable computers have high computing capacity and are available in light, easily portable packages. These small computers contain disk storage media and software that the geologist needs to collect data directly in digital form. Data collected in the field is displayed on the computer screen, and analysis can be performed on site before the geologist leaves the area where the data were collected.

Palmtop computers are usually about the size of a large hand-held calculator and include a digital screen. Software on a palmtop computer displays in a form that prompts the geologist for input such as rock type and characteristics, mineralization, outcrop orientation, amount of fracturing, weathering, and other essential descriptive geological information. A geologist may also collect information from core drilling. As core is brought to the surface, geological information is captured. The geologist examines the rocks, and descriptions and measurements are entered directly into a computer. Direct computer entry bypasses errors introduced when a hand-written data sheet is converted to a digital format in a computer. Software for geological data collection is often designed so the geologist can enter codes for rock and other characteristics and minimize typing. Data from a palmtop computer are later transferred to a laptop, to a desktop, or onto a company computer network for installation into a database and production of statistics, maps, and other analysis such as reserve calculations. If a geologist is using a laptop computer, which has more data storage and computing capacity than a palmtop computer, the entire data collection, analysis, and mapping work may be done in the field very soon after or even during data collection.

The data collected are organized in computer databases and used throughout the life of a project. The data stored include geological characteristics, rock types, spatial locations, geochemical analyses, mineral assays and analyses, and any other pertinent information applicable to the type of deposit being evaluated. The digital database is readily available for additions and refinements, ongoing reserve calculations, and use in later analyses for economic evaluation, mine designs, mine plans, operating schedules, and reclamation plans. A mine database is often stored on a central network computer, so that all departments can access it for particular needs, such as short-term mine schedules, location of mine waste, and many other operational requirements.

When enough data are collected for a preliminary definition of a potential ore deposit, computer software helps the geologist interpret the data and calculate mineral quantities in a target area. Mineral reserve calculations are readily updated using mine-modeling software as more data are collected and added to the database. Continual revisions of the reserve estimates give the geologist clues about where additional information is needed. In the "past," weeks or months of laborious work was required to obtain reserves; and when more information was needed, it had to be collected during the next field season. Computer use has converted geological analysis from a static to a dynamic process. The power

of computer hardware (fast processors, large hard disks) and sophistication of modern graphically oriented software have provided exploration geologists the tools to keep pace with demand for information in a fast-paced industrial environment. *See* DATABASE MANAGEMENT SYSTEM.

The Internet is playing an increasingly important role in data transfer. For example, when a geologic map has been generated, the map and associated data can be sent to a private Web page where others can view it and instantly see the progress of the exploration efforts. Special Internet map server software provides an environment for the map user to look at a map or portions of a map at different scales and to include and overlay other information. For example, a composite map can be produced that shows a satellite image, location of drill holes, interpreted geological information, roads, power lines, political boundaries, and other features.

This nearly instantaneous access to information, and ability to evaluate it, provides a basis for evaluating the progress of an exploration program as the activities are happening rather than months later. Previously, geologists collected field information in the summer and spent the winter evaluating it. An ore body model often is required 6 months or more to complete using hand methods and additional months to adjust for new information. Computer hardware and software tools make the data collection and evaluation process a matter of days and weeks rather than months to complete an analysis of a prospect. *See* REMOTE SENSING.

Planning and operations. When geological reserves are available, the mining engineer begins the planning process to determine what part of a deposit can be mined economically and what type of mining system should be used: surface, underground, or some combination. The mine planner uses a desktop computer with specialized software to build a mathematical and graphical model of an open pit or underground mine design. Formerly, a single mine design often took up to 6 months to produce. An engineer with a computer and specialized mine planning software can produce a multitude of maps, designs, and plans in a few weeks. Each design or plan can be developed with a variety of design criteria, such as minimum grade, different metal prices, physical layouts, and a selection of economic conditions. The resulting surface or underground mine designs can be displayed in color and in three dimensions, and can be shaded to look almost like photographs when enough details are available. The graphics help the planner determine the workability of a design, and can also be used to show management and government regulatory agencies how a mine might look at different stages of development from startup to final reclamation. *See* CONTROL SYSTEMS.

The goal of an open pit mine designer is to find the most economic areas of a deposit, the areas where the value of the ore minerals exceeds the cost of removing waste material to expose valuable ore. Surface mine planning software is used to define ore and waste on each level of the proposed mine, as well as the final outline of the pit after it is mined.

Each amount of material in a volume of earth is assigned a net value based on the potential sale price of the ore that can be recovered, the cost of removing the material, and processing or dumping it.

An underground mine design is created using a computer-aided design (CAD) program to graphically lay out the mine openings and extraction areas. The program operator sees a three-dimensional display of the geological model on the monitor screen and can lay out shafts, drifts, entries, and other access to the ore, as well as define the areas of the ore body that will be removed. Ore values and costs of mining are calculated for a design. The design is evaluated, then changed as necessary, until a suitable result is attained that yields the highest return. *See* COMPUTER-AIDED DESIGN AND MANUFACTURING.

When the mine plans are complete, detailed economic evaluations are performed to determine whether the proposed mine is economical. If so, the next step is to prepare the information for mine permits, a long and complicated process in most countries. The types of software used to evaluate the geology and develop mine plans are also used to prepare documents for permitting. When permits have been obtained, development of a mine operation can proceed.

Surface operations. Computers are used for several aspects of guiding the progress of mining in open pit and strip mines. Surveyors lay out blast-hole patterns based on computer-generated designs. The surveyors use electronic surveying equipment to lay out individual blast holes, identify existing pit outlines, and other surface features and feed these data into a computer for further processing. Computer programs are also used to identify the quantity and type of explosive required to properly fragment the rock. Samples of blast-hole cuttings in an open pit mineral operation are taken during drilling and sent to an automated analysis lab to determine quantities of minerals in the sample. The results are entered into a database and used by the engineer to produce an ore/waste map for the next day's mining. Computers and software provide the ability to produce maps quickly and keep the operation moving while routing mined material to the proper location—ore to the mill, waste to the dump. Maps and ore models developed from the blast-hole information are compared to the model developed from the exploration drilling. The comparison helps the engineer and geologist make adjustments to the modeling method if necessary.

In an open pit, tracking the movement of mobile equipment such as trucks, shovels, and loaders is essential to keep the operation flowing smoothly. Computerized truck dispatching systems are installed in the larger mines to monitor where trucks and other mobile equipment are located at any time and where they travel to be loaded and dump the load. Computers on board equipment monitor activity and movement. Dispatching systems use beacon stations in the pit to sense vehicle movement or use global positioning systems (GPS) to track the location of equipment. The dispatching system provides a printed report

of all activity plus a graphic display showing where the equipment is at any time. Summary reports are available for management at any time if the dispatch data is stored on an in-company network. The dispatch system shows equipment operation hours and is sometimes connected with the maintenance system to help track availability, provide information for preventive maintenance, and record equipment downtime. The system may also collect personnel working hours.

Computer-controlled driverless equipment has been in development for many years, but is not yet widely accepted and used in the mining industry. Several prototype operations have been installed, and the technology is continually being improved. *See* CONTROL SYSTEMS.

Underground operations. Underground mining is a more complicated process than surface mining. Active mining areas are often isolated, and the mine shift bosses and foreman spend most of their time moving from one area to another to make sure extraction is progressing properly. Computer and radio equipment has improved communications underground and helps management keep material flowing smoothly from the active mining faces to the surface.

Sensors installed in an underground mine can detect airflow quantity, presence of gases such as methane or radon, location of equipment, operating condition of equipment, ground movement, and other conditions. The various sensors connected to a computer collect data and display graphs, numerical information, or live video. The information is then available for display on computers in underground offices and surface offices. Management can quickly see at any time the status of the operation and whether equipment is functioning properly, and can provide information to identify possible trouble spots before they become a crisis.

Computerized remote control and robotic equipment is used in some operations such as a low coal seam where it is very difficult or dangerous for a human to work in an area. Automated mining systems and equipment are still being developed and are not yet widely used in mines. *See* ROBOTICS.

Mineral processing. Computer process control is well developed in milling operations. These complex sensor/computer systems provide information to the mill operator about the condition of everything that is happening at any given time at any location in the mill. Process control systems are a dedicated system of computer hardware and specialized software. The process control system contains a central controller, sensors around the mill, one or more operator consoles, and video monitors for different parts of the process. The monitors show numbers such as motor horsepower and other measurable quantities, graphs showing the quantity of material being fed into a crusher, and actual videos when cameras are used to monitor particular parts of the operation. The operator has the ability to change equipment settings from the console without having to physically touch the equipment. *See* PROCESS CONTROL.

Active equipment maintenance in a milling oper-

ation is essential for efficient movement of material from the beginning to the end of the process. Properly maintained equipment keeps metal recovery levels high and contributes to a profitable operation. The computers and software used for fixed equipment maintenance, such as found in a mill, is often the same as that used for mobile equipment maintenance. *See* PROCESS CONTROL.

Mining systems. Simulations of all phases of the mining operation can be performed using computer software. Parts or all of a mine operation are simulated using mathematical and probabilistic models. Simulations commonly are run for planning haulage routes, designing a milling operation, evaluating an existing mill or mining operation to identify solutions to problems, and showing how material is actually moved from one place to another. Computer simulation offers the ability to try different mining scenarios at a very low cost. For example, a simulation can help define whether it is economic to use trucks to haul material out of a pit or to install an in-pit crusher and conveyor belt to move material.

Scheduling is an essential function for keeping production at the highest levels. Long-term and short-term production schedules are developed from the overall mine plan. Production targets, consisting of quantity and grade of material, are defined by the engineer, and then the areas that meet those qualities are selected from the mine plan and assigned for extraction to a calendar time. Because short-term scheduling is subject to change (for example, the blast holes may show a different grade than the original exploration drilling), computer programs permit mining engineers to dynamically change the outline for the next production period and calculate new projected grades and volumes. Another approach to automated scheduling is to use a graphical CAD technique where an area is defined and the ore quantity and grade are calculated from the mine's database.

Computerized maintenance and inventory systems are an essential part of modern mines. Maintenance systems produce reports about each piece of equipment (whether stationary or mobile) and when it is due for servicing or major overhaul. A history of each piece of equipment is readily available from the computer maintenance database that shows what parts were installed, any breakdowns that occurred, and when the next regular maintenance is due. The availability of each piece of equipment is reported, and replace or repair decisions can be made. The maintenance system is closely connected with parts and supplies inventory which helps operators identify when to order additional parts or consumables. The system also maintains historical records for parts reliability and equipment failure tracking.

Maintenance and inventory data are stored on networked computers, so it is readily available to the maintenance department personnel and management. Maintenance management systems are often tied to the mine accounting system as well.

Enterprise-wide software solutions are appearing in some of the larger mining operations. This software connects all parts of the mine operation,

including financial accounting, human resources, safety training, maintenance, inventory, facilities management, and material movement. Management can at any time access an enterprise system and see detailed activity reports or summaries of exactly what is happening in the operation at that time. Historical records of past performance help the manager identify changes in production levels and spot potential problem areas.

Mine reclamation. Geographic information systems (GIS) have become important throughout a mine operation, but are especially helpful in preparing maps showing a mine operation in relation to soils, vegetation, wildlife, surface hydrology, and other environmentally sensitive information. GIS combines a wide array of different types of information to produce maps showing how an area will appear during phases of the mining operation and to show final reclamation plans. One use is to show more efficient and effective ways of locating topsoil storage, so it is available in the right place when it is needed for final surface coverage. GIS can also be used for facilities management to keep track of existing infrastructure and future plans.

GIS is used during the exploration phase to combine satellite images with ground data and organize all the spatial data collected from a variety of sources. Surface geochemical data, geological data, geophysical data, assay results, and other spatial data can be posted on a map and analyzed for spatial trends. The information can be shown against a satellite photo background and can be viewed in perspective. See GEOGRAPHIC INFORMATION SYSTEMS. Betty Gibbs

Environmental Impact

Historically, mines have caused significant adverse environmental impact. Often, little consideration was given to the effect of the mine on the environment, and rarely was the land reclaimed. As a result, natural weathering processes have caused erosion of unreclaimed land, and reaction of certain rock types with rainwater, forming acid- and metal-laden runoff that may contaminate streams and ground water.

Contemporary mining operations have come a long way environmentally. Today, mining companies recognize that land reclamation and environmental protection are as much a part of mine planning as are the extraction and recovery of the mineral itself. As the practice of environmental assessment and the understanding of the impact of mining wastes on the environment have advanced, the adverse impact on the environment has been reduced.

Laws and regulations. To assure that mining operations act in an environmentally prudent manner, nearly every country has enacted laws and regulations requiring environmental protection. In the United States, laws such as the federal Surface Mining Control and Reclamation Act and various state reclamation acts specifically require mining operations to evaluate environmental effects and develop reclamation plans to mitigate impacts. A few federal laws that directly affect mining but are nonspecific to the industry include the Clean Air Act, the Clean Water Act, and the Resource Conservation and

Recovery Act. States also usually have programs similar to these federal laws.

In countries such as Peru, where economic development and public environmental awareness are advancing but are still in the formative process, the laws and regulations are also in the formative stage. These countries typically have an established front-end review and permitting process, but have limited enforcement capability due to limited financial support for the program. In underdeveloped countries where the political environment is unstable and there is very little economic development or public environmental awareness, only very general environmental protection laws may have been enacted.

Developing a mine usually requires a significant investment. Many mining companies must therefore secure mine development financing from international lending institutions. Although they have no environmental regulatory authority, nearly all of the major international lending institutions have specific guidelines for land reclamation and environmental protection. So even in underdeveloped countries where the government has little ability to administer and enforce its environmental protection laws, mine developers are still required to carry out environmental evaluations and reclamation programs in order to secure financing.

Environmental evaluations. Laws and regulations in nearly all countries, as well as most international lending institutions, require a mining company to assess the impacts that its development plans will have on the environment and to propose programs to limit those impacts before construction of the mine begins.

Pre-mine environmental conditions. The first step in evaluating the potential environmental effects of a mining project is to define the preexisting environmental conditions at the location where the mine will be developed. Baseline studies are typically commissioned to characterize the existing physical, biological, and human-interest resources in the area. Physical resource concerns include air quality and climatic and meteorological conditions; surface-water and ground-water quantity, quality, and flow characteristics; the nature, extent, and surrounding geologic media associated with the mineral deposit; and the agronomic capability, land uses, and erosion characteristics of the soils. Biological baselines characterize the presence, abundance, and diversity of plants and animals, including the aquatic species and insects, and the identification of any sensitive or protected plant or animal species. Human-interest resource baselines provide a profile of the population distribution, statistics, education, and income levels of the people near the project site; profile the existing infrastructure and services available; define the local economic and tax structure; define cultural and ethnic beliefs; and identify the presence of historical and archeological resources.

These studies provide a basis against which the impacts associated with mine development can be measured. The development of the environmental baseline data usually requires at least one year, or more, of data collection to address the seasonal

variation of certain resources. For this reason, most baseline studies are initiated several years before mine construction is anticipated.

Predicting environmental impacts. After the baseline conditions have been characterized, project design and development plans need to be analyzed relative to those baseline conditions to predict the effect that project development will have on the environment. This process requires an understanding of how the project designs will be implemented, operated, and maintained, as well as specific knowledge of the environmental resource being evaluated.

One of the most important factors that need to be evaluated in an impact assessment is how mined and processed materials and wastes will behave in the environment. Mining typically exposes to the atmosphere various rock types that have not been exposed for millions of years. Weathering processes can affect certain rock types, resulting in acid generation, metal leaching, and the release of elevated concentrations of pollutants to streams and ground-water systems. It is therefore extremely important to understand the geochemistry of the rock to be mined and the effect that the weathering process will have on the rocks.

Impact assessment is an iterative process in which mine plans are analyzed relative to each specific environmental resource. For example, the effect that mining will have on air quality is analyzed relative to the kinds of emissions that are expected to be generated from each source. Typical emission sources at a mine include exhaust systems from mobile equipment, stacks at processing operations, dust from rock crushing activities, and dust from roads and other surface disturbances. The impact of development is analyzed by understanding the air quality that exists before the mine is developed (the baseline condition), assessing all of the contributing emissions that are expected to be generated by mine development, and predicting what the air quality will be during the various phases of the mine operation. Similar predictions are then made for each of the other environmental resources, such as water, soils, vegetation, and people.

After the effects of project implementation have been predicted for each resource, it is necessary to further evaluate how those changes will fit with other reasonably foreseeable developments in the area. For example, if there are plans to develop other mines and mineral processing facilities over the next several years, the effect of each facility on the environment needs to be evaluated in order to anticipate the combined effect.

Reducing adverse environmental impacts. Most countries and lending institutions have specific limits on how much pollutant can be released from a mining or industrial operation. After the effect of project development has been predicted for each receiving resource, those predictions are compared to the specific requirements for that discharge. If the current plans fail to meet the required levels established for environmental protection, project planners must change their designs to provide a system that will assure compliance.

Mine designers can limit environmental impacts in many ways. Probably the most significant way to reduce the long-term environmental impact of a mining operation is by implementing an effective closure and reclamation plan. Some of the more common ways to control the discharge of pollutants include designing control systems for mining and processing wastes, water and wastewater treatment, and sedimentation control. Disturbances to sensitive plant and animal species, or to cultural or archeological sites, can often be avoided by changing the design location of certain facilities. Where such disturbances are unavoidable, translocation and data recovery programs can often be implemented to reduce the overall impact.

Environmental management systems. Environmental management and monitoring systems are excellent tools for limiting the environmental effects of mining to the extent possible. These systems are designed specifically to monitor the performance of the environmental controls and mitigation measures that were defined in the design stage, and to continuously provide improvements throughout the life of the project.

The environmental management system is a program, to be administered by mine personnel, that defines a systematic process for documenting environmental conditions at the site. It will typically include specific monitoring programs for air quality, meteorology, surface-water and ground-water quality and quantity, the success of revegetation testing, and wildlife presence. It will also clearly establish the lines of authority for implementation and contingency plans for upset conditions. However, because every mine is unique in its environmental effects, the environmental management program must be customized to address what is important for a specific site.

Systematic management and monitoring programs allow mine personnel to monitor the environmental conditions, evaluate their consistency with the predictions made in the impact assessment, and make timely modifications to environmental controls if the conditions vary from the original predictions or do not meet specific environmental requirements. Because these systems are continually being updated and improved, they also allow for new technologies to be easily integrated into the environmental protection program.

Closure and reclamation. Unlike historic mining practices, contemporary mining operations plan for the closure and reclamation of mined lands before the mine is developed. As the mine gets closer to the end of its operational life, the closure and reclamation plans that were defined conceptually in the planning and impact assessment phase must be more specifically defined so they can be implemented.

Closure activities are implemented to assure the long-term chemical stabilization of the site. Examples of closure activities include the rinsing or cleaning of rock piles or equipment that have been exposed to chemical reagents or reactions. The covers that may be required to isolate mine wastes from weathering processes would also be considered clo-

sure components if, without the covers, the wastes would release pollutants to the environment.

Land reclamation restores mine site disturbances to a safe, stable, and productive postmining condition. Reclamation programs are typically initiated after the closure activities are completed. Reclamation activities typically include precluding entry to mined-out workings, grading and shaping the land to a form that is stable over the long term, placing soils over the graded land disturbances, and planting vegetation to allow the final reclaimed surface to blend with the surrounding environment. *See* LAND RECLAMATION.

There are many examples of contemporary mines that have been developed, operated, closed, and reclaimed in an environmentally prudent manner. While it is impossible to mine without affecting the environment, it is possible to limit its environmental effects. With good up-front analysis and impact assessment, effective environmental management systems, and successful site closure and reclamation, the mining industry will continue to balance the need to protect our natural resources with the need to provide the raw materials demanded by our advancing society.

Barbara A. Filas

Restoration of Land

Land disturbance is frequently the most visible impact of mining on the environment. Pits and quarries from surface mining, waste rock dumps, low-grade ore leaching dumps or heaps, and tailings ponds for disposal of processing waste are prominent features in the landscape of a mining area. A variety of techniques are used to reduce or eliminate potential environmental hazards associated with these features.

The goal of reclamation is to return the land surface to as close to its premining condition as possible. When this is not possible, the goal becomes stabilization of the area and preparation of it for an acceptable postmining condition or alternative land use. The basic steps of reclamation usually include movement of materials to achieve a stable land contour, replacement or rebuilding of topsoil, and revegetation. The complexity of reclamation is related to the type of deposit, the mining method used, the type of processing done at the mine site, and the physical and chemical properties of the waste materials involved.

Surface mining. In surface mining, one of the first steps in mine development is removal of the topsoil layer. The topsoil is set aside so that it can be used to cover the disturbed area and be revegetated during reclamation. Shallow flat-lying deposits, such as those mined for phosphate in Florida, are amenable to area mining; that is, the overlying waste material and ore are removed sequentially in long narrow cuts. The waste rock from a cut can be used to fill the previous cut. After several cuts are completed, reclamation can be initiated and then continue throughout the life of the mine. The waste piles in the mined-out areas are recontoured by bulldozers and scrapers. Topsoil is replaced and vegetation is planted. In all types of reclamation, plant species are selected for the soil and climatic conditions on the

mine area. When fill materials are not available for the final cut, it may be contoured to form a lake or a wetland. The continuous reclamation process means that only a small portion of the surface area will be disturbed at a time.

Deeper mines. Deeper and more localized deposits, such as those mined for copper, iron, and gold, require the construction of large open pits. As production continues, the pits expand downward and outward, making reclamation impossible until all mining has ceased. Waste rock covering the ore-body is removed and placed in permanent waste dumps. The configuration of the waste dumps may be dictated by the topography. In steep mountainous regions, areas for dump construction may be limited; and dumps may be high, making reclamation more difficult. In flatter terrain, such as the desert areas of the southwestern United States, lower-profile dumps are constructed. They are generally more stable, blend in better with the surroundings, and are easier to revegetate. While continuous reclamation may not be possible under these conditions, individual dumps can be completed and reclaimed at various times during the life of a mine. The pit remaining when mining is completed presents a unique problem. Backfilling is usually neither practical nor economically feasible. However, the slope angle of the sides of the ultimate pit can be designed to give maximum stability. The steplike benches that are left on the sides of the pit can catch any loose rock and can be revegetated. Depending on the local hydrology, these pits may also fill with water and form lakes.

Quarries. Quarries and pits mined for construction materials represent yet another type of reclamation challenge. These mines are frequently located in populated areas. The visual impact can be reduced by constructing and landscaping berms of rock and soil around the perimeter of the property. This also helps suppress equipment noise that may be objectionable to neighbors. Deep quarries are treated in the same manner as open-pit metal mines. Shallower sand and gravel pits are often recontoured to make them valuable as areas for residential and commercial development, with the deeper pits left as lakes and ponds. *See* LAND RECLAMATION.

Leach dumps and heaps. In addition to dumps constructed of waste rock, special dumps and heaps are designed for the leaching of low-grade metallic ores. Leaching solutions containing acid or sodium cyanide are typically used for copper or gold, respectively. They are sprayed on the tops of the dumps. As the solutions percolate downward through the pile of rock, the valuable metals are dissolved and carried away in the solutions. The most important environmental concern is contamination of ground water if the solutions are not controlled. Leach dumps and heaps are designed to prevent solution loss. Dumps such as those used for copper leaching can be built on impervious bedrock with a system of channels and sumps constructed for solution collection. Dumps and heaps for leaching gold are constructed on a series of liners that prevent solution loss. The liners may be made of asphalt, synthetic materials,

or impervious clay. Leak-detection instruments are installed between the liners. Monitoring wells surrounding the leaching areas are a further precaution. Leaching operations must provide adequate storage reservoirs to contain solutions during periods of high rainfall or snowmelt.

Tailings ponds. Fine-grained waste-rock tailings from mineral-processing plants are stored in tailings ponds. Well-engineered dams constructed of the coarse waste materials are used to impound finer wastes. The design of tailings dams must provide long-term stability and safety. The tailings are transported to the ponds as slurries. As the solids settle from the slurry, the water is recovered and recycled to the mineral-processing operations. If excess water is generated as drainage from a tailings pond, it must be collected and, if necessary, treated before it is released to a stream. The level of treatment will depend on the chemistry of the water. In some instances, older tailings ponds have contaminated ground water. One method for correcting this problem is to drill wells to intercept the plume of contaminated water and pump the water to a treatment plant. Because of the fine particle size of tailings, dust control is also important. When a tailings pond is full, it must be reclaimed.

Underground mine subsidence. Most underground mining methods do not disturb large areas of the land surface. The underground workings may be back-filled with waste rock or tailings to prevent collapse, or they may be designed to remain open. However, some large-scale mining methods, such as block caving, are designed so that the openings collapse in a controlled manner. When the mine workings collapse, the land surface over the mining area will eventually subside. Depending on the depth and size of the workings, the extent and timing of the subsidence will vary. Although subsidence may take years to stabilize, reclamation of stable areas is possible.

Erosion and sedimentation. Storm-water runoff is the major cause of erosion and the resulting sediments that pollute rivers and streams. Surface water must also be controlled to prevent mine flooding and to reduce the potential for creating contaminated mine drainage. Streams in the vicinity of a mine can be diverted away from the active operations, although provisions may have to be made to accommodate fish. Retention ponds are designed with adequate capacity to hold water during a peak rainfall event and to release the water gradually after the event. Sediment basins are used to permit solids to settle to the bottom before the water is released. These basins are also used in placer mining to prevent the release of sediment into local streams. Revegetation provides another highly effective means of controlling erosion and sediment.

Treating contaminated mine water. The most common form of contaminated mine water is acid mine drainage. It is formed when water, air, sulfide minerals such as iron pyrite, and bacteria react to form sulfuric acid. This acid may in turn result in the release of

toxic metals into solution. When acid mine drainage is released into rivers and streams, it can damage water supplies, harm aquatic life, and ultimately render the waterway sterile. Several thousand miles of streams in the United States are contaminated by acid mine drainage from abandoned mines.

Eliminating contamination from acid mine drainage requires water treatment. Depending on the chemical elements involved, treatment processes range in complexity from simple neutralization with lime to more elaborate processes to remove toxic metals such as uranium, cadmium, or lead. Constructed wetlands are being used for treating acid mine drainage. The drainage flows through shallow ponds lined with crushed limestone and organic material such as compost or manure, and then cattails are planted. The metals are removed and become part of the organic substrate. Originally developed for treating coal mine drainage, several large-scale wetlands have been designed to treat metal mine drainage. See WETLANDS.

Dust control. Airborne dust can be generated in many parts of a mining operation. The simplest method of dust control is to spray water on mine roads, preventing trucks and other types of mobile equipment from raising clouds of dust. Chemicals and surfactants can also be used to improve control. In crushing plants, water sprays or dust collection systems and filters are used. After mine closure, revegetation reduces airborne dust. L. Michael Kaas

Bibliography. American Geological Institute, *Dictionary of Mining, Mineral, and Related Terms*, 2d ed., 1997; A. E. Annels (ed.), *Mineral Deposit Evaluation: A Practical Approach*, 1991; *Application of Computers and Operations Research in the Mineral Industry* (APCOM), various publishers of proceedings from conferences held since 1967; C. J. Bise, *Mining Engineering Analysis*, 2003; H. L. Hartman (ed.), *SME Mining Engineering Handbook*, 2 vols., 2d ed., 1992; H. L. Hartman et al., *Mine Ventilation and Air Conditioning*, 3d ed., 1997; H. L. Hartman and J. M. Mutmansky, *Introductory Mining Engineering*, 2d ed., 2002; C. C. Heald (ed.), *Cameron Hydraulic Data*, 19th ed., 2002; B. A. Kennedy (ed.), *Surface Mining*, 2d ed., 1990; J. J. Marcus (ed.), *Mining Environmental Handbook*, 1997; M. J. McPherson, *Subsurface Ventilation and Environmental Engineering*, 1993; C. J. Moon et al., *Introduction to Mineral Exploration*, 2d ed., 2006; National Institute for Occupational Safety and Health, *Criteria for a Recommended Standard: Occupational Exposure to Respirable Coal Mine Dust*, DHHS Publ. no. 95-106, 1995; R. V. Ramani, B. K. Mozumdar, and A. B. Samaddar (eds.), *Computers in Mineral Industry*, 1994; E. A. Ripley, R. E. Redman, and A. E. Crowder, *Environmental Effects of Mining*, 1996; I. C. Runge, *Mining Economics and Strategy*, 1998; M. Sengupta, *Mine Environmental Engineering*, vols. I and II, 1989; A. J. Sinclair and G. H. Blackwell, *Applied Mineral Inventory Estimation*, 2002; J. G. Stone and P. G. Dunn, *Ore Reserve Estimates in the Real World*, 3d ed., Soc. Econ. Geol. Spec. Publ. no. 3, 2002; T. F. Torries,

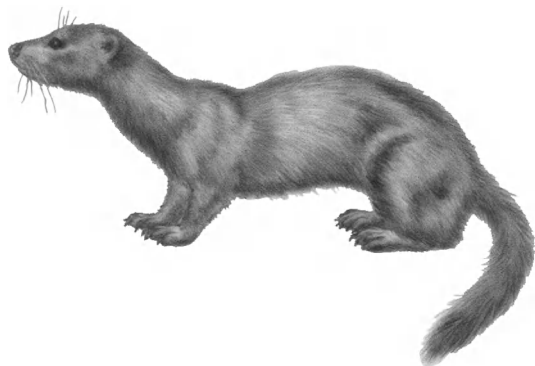
Evaluating Mineral Projects: Applications and Misconceptions, 1998; D. Wyllie and C. W. Mah, *Rock Slope Engineering: Civil and Mining*, 4th ed., 2004.

Mink

A semiaquatic, carnivorous mammal in the family Mustelidae. The American mink (*Mustela vison*) is found in Alaska and Canada and throughout most of the United States. The range extends southward in the western United States to northern New Mexico, northern Nevada, and central California. This species was deliberately introduced into the former Soviet Union, and escaped animals have established populations in France, Spain, Portugal, Germany, Poland, Ireland, Great Britain, and Scandinavia, as well as in Iceland.

The mink has a long, slender, weasel-like body and a tail that is bushy at the tip. It is short-legged and almost uniformly dark brown (see **illustration**). The underparts are only slightly paler than the back. Long, glistening guard hairs partially conceal the soft, luxurious underfur. A white patch is usually present on the chin and may extend to the throat and chest. The tail is less than half the length of the body. Males are considerably larger and heavier than females. Females possess eight mammae. Paired anal scent glands are present. Five toes are present on each foot, and the hindfeet may be slightly webbed. The sense of smell is exceptionally well developed, while the senses of hearing and sight are moderately developed. Mink can hear ultrasound within frequencies emitted by potential rodent prey. The dental formula is $I \frac{3}{3} C \frac{1}{1} Pm \frac{3}{3} M \frac{1}{2} \times 2$ for a total of 34 teeth. Adult mink are 530–660 mm (20–26 in.) in total length, including a 175–225-mm (7–9-in.) tail. They normally weigh 0.5–1.4 kg (1–3 lb).

The sea mink (*M. macrodon*) formerly occurred along the Atlantic coast from New Brunswick to Massachusetts but is now extinct. It is thought to have been considerably larger than the American mink. No complete specimen is known to exist. Descriptions are based on recorded observations and on numerous bone fragments and teeth found in Indian middens along the New England coast. This species apparently inhabited the rock crevices and ledges along the ocean shore. It became extinct about 1880,



American mink (*Mustela vison*).

possibly as a result of overhunting.

American mink are found mainly along the banks of streams, rivers, ponds, and lakes and in swamps and marshes. Forested, log-strewn, or brushy areas are preferred. Males may range several kilometers along a stream. Mink are mainly nocturnal, semiaquatic mammals and seem to be most active around dawn and dusk. They are excellent swimmers and divers. Although they spend a lot of time in the water, mink also travel much on land and on rare occasions climb small trees. Mink are active all year. They are normally solitary, except for family groups of mother and young. The mink may dig its own burrow, or it may appropriate one made by some other animal such as a muskrat or rabbit. Dens may also be located under logs or debris, in rocks, or in muskrat houses.

Mink feed primarily on mice, muskrats, rabbits, birds, frogs, fish, and crayfish. They are able to catch swimming fish. They have also been known to eat bats that were thought to have fallen from the ceiling of a cave. In northern areas, breeding normally occurs during February and March. Ovulation is induced by copulation. The single annual litter is produced in April or May after an average gestation of approximately 51 days (range 40–75 days) because of delayed implantation. Implantation occurs about 30 days before birth. Litters may contain from three to ten young, although the average is four or five per litter. The young are born in a nest of feathers, fur, bones, grass, and plant fibers. Newborn young have their eyes and ears sealed and are covered with fine, silvery-white hair. Within 2 weeks, pale reddish-gray hair replaces the white hair. The teeth begin erupting between 16 and 21 days. The eyes open at about 37 days. Weaning is at about 5 weeks, and the young leave the den when 6–8 weeks old. The young mink remain with the female until late summer or fall, at which time they disperse. Although the male may mate with several females, it usually stays with the last one and assists in caring for the young. Both males and females are capable of breeding during the breeding season following their birth. Mink have few enemies other than humans and their traps. Occasionally, one may be attacked by a fox, bobcat, or great horned owl. Captive ranch mink have lived 10 years, although the life-span in the wild is probably 3 or 4 years. Most of the mink fur used in commerce is produced on farms. Selective breeding has produced colors such as black, white, platinum, and blue, and descriptive trade names such as Aleutian, amber gold, blue iris, pastel, platinum, sapphire, and winter blue. See CARNIVORA; DENTITION; MARTEN; WEASEL.

Donald W. Linzey

Bibliography. G. A. Feldhamer, B. C. Thompson, and J. A. Chapman, *Wild Mammals of North America: Biology, Management, and Conservation*, 2d ed., Johns Hopkins University Press, 2003; R. M. Nowak, *Walker's Mammals of the World*, 6th ed., Johns Hopkins University Press, 1999; D. E. Wilson and S. Ruff (eds.), *The Smithsonian Book of North American Mammals*, Smithsonian Institution Press, 1999.

Miocene

The second subdivision of the Tertiary Period (Eocene, Miocene, and Pliocene) by Charles Lyell in 1833; the fourth in a more modern sevenfold subdivision (epochs) of the Cenozoic Era; and the first epoch of the Neogene Period (which includes in successive order the Miocene, Pliocene, Pleistocene, and Holocene). The Miocene represents the interval of time from the end of the Oligocene to the beginning of the Pliocene and the rocks (series) formed during this epoch. The Miocene was originally considered as a biostratigraphic (rather than temporal) entity, to unite rocks containing 20–40% extant molluscan species. Lyell based his concept of the Miocene primarily on the shallow marine sediments and associated molluscan faunas in the Superga Hill of northern Italy (near Turin), as well as other Piedmont localities and outcrops in southwestern France (Aquitaine Basin), Touraine, and the Vienna Basin. *See* CENOZOIC; HOLOCENE; PLEISTOCENE; PLIOCENE; OLIGOCENE; TERTIARY.

The Miocene spans the time interval between 23.8 and 5.32 million years ago (Ma) based on integrated astronomical and radioisotopic dating. The Miocene/Pliocene boundary is located in Sicily, just above a major unconformity separating the youngest late Miocene (Messinian) deposits (of the Great Terminal Miocene Salinity Crisis) and the overlying white chalks of the Zanclean. *See* UNCONFORMITY.

Subdivisions. While originally conceived as biostratigraphic units (that is, including rocks containing various broadly contemporaneous marine and terrestrial faunas), the epochs of Lyell (including the Miocene) have come to be considered as chronostratigraphic units (rocks having clearly specified, conceptually temporally isochronous and correlative boundaries) in which fossil faunas and floras are used for the purpose of characterization and correlation. Integrated studies using radiometric dating, paleomagnetic stratigraphy, stable-isotope (oxygen, carbon, strontium) studies together with (micro)paleontologic studies have provided a framework for regional and global correlation of the disjunct stratigraphies established in the marine and terrestrial biosphere (**Figs. 1 and 2**).

An essentially twofold standard subdivision of the lower Miocene is recognized: the Aquitanian and Burdigalian stages, having their type development in the Aquitaine Basin of southwestern France. The middle and upper Miocene have their typical devel-

opment in the tectonically active basins of Italy. The middle Miocene is represented by the Langhian and Serravallian stages and the upper Miocene by the Tortonian and Messinian stages. Marine sedimentation essentially ceased in the Anglo-Paris-Belgian basins (type areas for the Paleogene stages) in the mid-Oligocene, whereas in the Aquitaine Basin marine sedimentation ceased in the early middle Miocene (Langhian Stage). A comparable subdivision has been made for the relatively complete Miocene sequence which is developed in the Paratethys region of central and eastern Europe (Vienna Basin to the Urals).

Tectonics and volcanism. Major orogenic and volcanic events characterize the Miocene (**Fig. 3**). Plate-tectonic motions, originating in the Mesozoic, resulted in the gradual dismemberment of the Tethyan Ocean and the upthrusting of the Alpine-Himalayan orogenic belt in three major phases: the late Eocene (about 40 Ma) and the early (21–17 Ma) and mid-late Miocene (10–7 Ma). Along the eastern margins of the Pacific Ocean, the ocean crust was subducted under the North and South American continents, giving rise to major orogenic movements stretching from the Aleutians to Tierra del Fuego. The Andes range was thrust up during the later part of the Miocene. The Pacific Coast developed as a result of westward drift of North America over, and partial consumption by, the Farallon Plate and collision with the Farallon Ridge. Only two relatively minor plates remain as remnants of the Farallon plate: the Juan de Fuca and Cocos plates between Mexico and Alaska. The plate margin was bounded by transform faults rather than a subduction zone, and northwestern propagation of a major transform fault issuing from the Cocos plate formed the Gulf of California in the late Miocene, and its continued extension northward is familiar to residents of the west coast as the San Andreas Fault System. The latter was responsible for the formation of many of the off- and onshore basins of southern California, some of which contain prolific petroleum resources. Subduction of the Pacific plate at the Middle America Trench during the late Paleogene and Neogene resulted in arc magmatism and eventual uplift of the Central American Isthmus into a series of archipelagos in the late Miocene (about 7 Ma) and eventual fusion into a continuous land bridge in the early Pliocene (about 3 Ma) that resulted in the separation of the Atlantic and Pacific oceans and concomitant disruption in marine faunal communities as well as transcontinental migration of vertebrate animals in the Great American Faunal Interchange. *See* OROGENY; PLATE TECTONICS; SUBDUCTION ZONES; TRANSFORM FAULT.

The Mediterranean, Black, and Caspian seas are remnants of an ancient equatorial ocean called the Tethys that separated Africa and Europe and linked the Atlantic and Pacific oceans for over 100 million years. In the mid-Cenozoic the continued northward movement of the African plate resulted in its collision with Europe in the early Miocene (about 18–19 Ma) along a suture line represented today by the Taurus (Turkey) and Zagros (Iran and Iraq) mountain ranges; the formation of the Alpine, Dinaric, and

| | | |
|----------|------------|-------------|
| CENOZOIC | QUATERNARY | Holocene |
| | | Pleistocene |
| | TERTIARY | Pliocene |
| | | Miocene |
| | | Oligocene |
| | | Eocene |
| | Paleocene | |
| MESOZOIC | CRETACEOUS | |

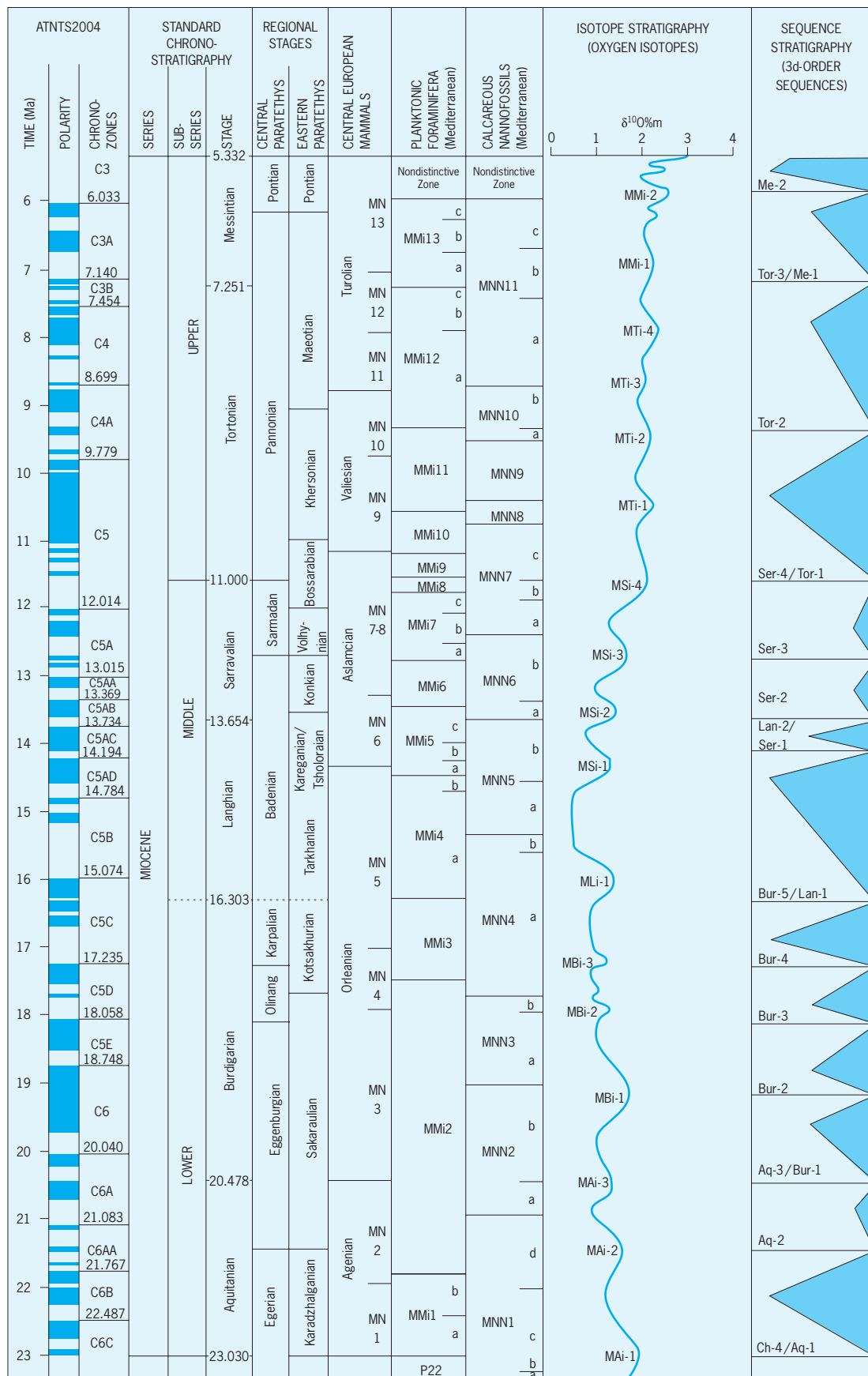


Fig. 1. Time scale of the Miocene.

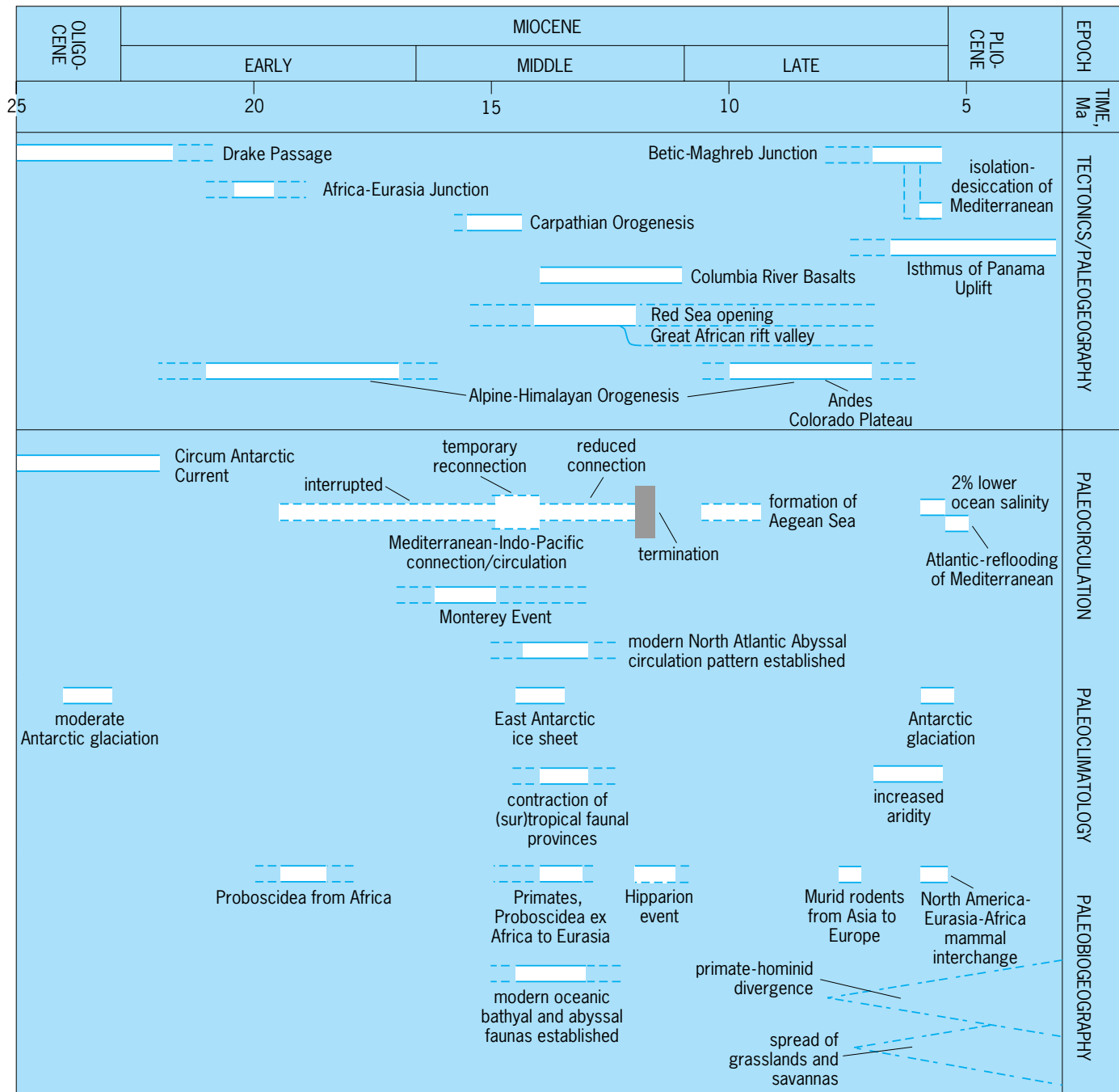


Fig. 2. Major events in the Miocene Epoch.

Hellenic mountain chain; and the eventual separation about 13–14 Ma of the Tethys Ocean into two inland seas: the Miocene (ancestral) Mediterranean (to the south) and the Paratethys (to the north). The Paratethys, which had existed as a shallow marine arm of the Tethys since the late Eocene–early Oligocene, underwent gradual reduction in salinity during the middle Miocene and local development of evaporite (salt) deposits reflecting continued compression, and eventually evolved into a series of both interconnected and separate brackish lakes. Concomitantly, early Miocene clockwise rotation of Africa away from Arabia widened the southern part of the Red Sea (which had opened up during late Oligocene time), and (at least intermittent) com-

munication with the Indian Ocean was maintained throughout the Miocene probably by way of the Afar Depression. The Gulf of Suez and Gulf of Aqaba/Eilat opened during the middle and late Miocene. The southward extension of this major rift zone into the heart of East Africa formed the Great African Rift Valley, an incipient mid-ocean ridge in the making. See CONTINENTS, EVOLUTION OF; MID-OCEANIC RIDGE; RIFT VALLEY.

The Colorado Plateau was uplifted from near sea level to nearly its present height of 5000–6000 ft (1500–1800 m) during the late Miocene and Pliocene, and the Colorado River began its down-cutting, resulting in Grand Canyon, a feature that exposes 2 billion years of geologic history. The Basin

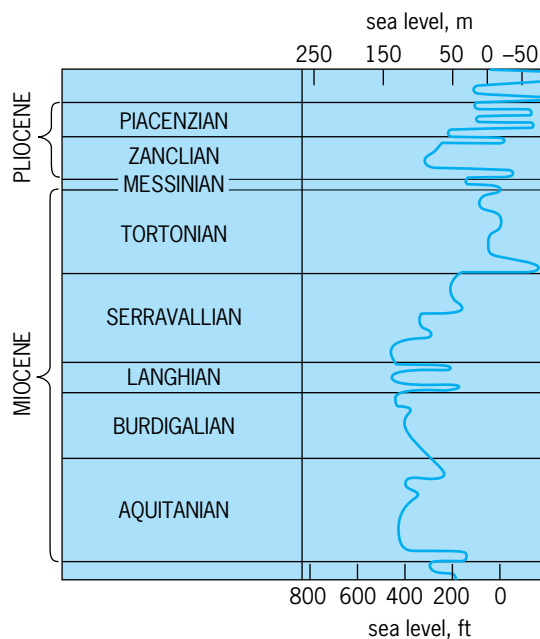


Fig. 3. Eustatic sea-level curve for the Neogene. The zero represents present sea level.

and Range Province of North America was subjected to early Miocene (rhyolite) volcanism, and block faulting and basaltic volcanism during the late Miocene have combined to give the region its characteristic linear topography of alternating mountains separated by broad valleys.

One of the most spectacular volcanic deposits can be found in the Pacific Northwest: the Columbia River basalts. Erupted during a 3–4-million-year span in the middle Miocene, these basalt flows extend over 50,000 mi² (200,000 km²) in eastern and central Washington and have a maximum thickness of over 8250 ft (2500 m). See BASALT; VOLCANOLOGY.

Resources. The Miocene is the source for major deposits of phosphates that are mined in North Carolina and Florida as well as extensive deposits of early-middle Miocene diatomaceous earth and phosphates in California which formed as a result of enhanced primary productivity in the eastern Pacific Ocean. Significant oil and gas are produced from Miocene rocks, particularly in the Gulf of Mexico and the west coast basins off California, northern South America, the Indonesian archipelago, and to a lesser extent the Arabian Peninsula and Gulf and northern Iraq and eastern Syria, among other places. See NATURAL GAS; PETROLEUM; PHOSPHATE MINERALS.

Ocean currents and climate. Ocean circulation essentially assumed its modern form during the Miocene as enhanced refrigeration in the form of growth of the Antarctic Ice Sheet plunged the Earth inexorably deeper into an icehouse state, although there were some details that were completed during the succeeding Pliocene and Pleistocene epochs. An ice cap has been present on Antarctica, at least intermittently, since at least the early Oligocene (about 34 Ma). The opening of the Drake Passage between South America and Antarctica took place during the latest Oligocene–early Miocene (about 25–23 Ma), allowing the unhindered circulation of ocean cur-

rents around the Antarctic continent. The development of the Circum-Antarctic Current thermally isolated high southern latitude waters and the continent of Antarctica from the warmer, low-latitude waters and resulted in the replacement of calcareous oozes (comprising planktonic foraminifera and calcareous nannoplankton) by biosiliceous oozes (diatoms and radiolarians). See GLACIAL EPOCH.

During the late early Miocene (about 16 Ma) a significant increase in biologic productivity occurred, and increased extraction and burial of organic carbon in oceanic sediments is interpreted to have lowered atmospheric levels of carbon dioxide and cooled global climates (reverse greenhouse). This has been called the Monterey Event because large amounts of organic matter were deposited around the perimeter of the northern Pacific and particularly in California, off which they are the main source rock for petroleum in the Monterey Formation. An alternative explanation for this event suggests that Early Miocene (about 21–17 Ma) Himalayan uplift caused increased chemical weathering and riverine nutrient flux to the sea, resulting in enhanced productivity and enhanced deposition of organic rich sediments. The latter theory suggests that drawdown of carbon dioxide (and attendant global cooling) is ultimately due to tectonics (and resulting silicate weathering) rather than organic carbon burial and that enhanced delivery of nutrients to the ocean, rather than stronger upwelling, is the driving force behind organic burial.

Global cooling did follow shortly after the Monterey Event. Some 2 million years later, during the Middle Miocene (about 14 Ma) major expansion and permanent emplacement of the East Antarctic Ice Sheet resulted in significant planetary cooling and constriction of tropical and subtropical belts and associated faunas. During the latest Miocene (about 7–5.5 Ma) another major expansion of the ice cap, comparable to, or somewhat less than, present-day volume, resulted in a global sea-level fall of some 230 ft (70 m), extensive glaciation as far away as Patagonia, and development of widespread aridity in midlatitudes of western North America, Australia, and Asia (India, Pakistan). This event coincided with, and is believed to have played an essential role in, the latest Miocene isolation of the Mediterranean Basin from the global ocean and its evaporation and transformation into a desert 2 mi (3 km) below sea level. The subsequent reconnection of the Mediterranean and Atlantic following a rise in sea level at the beginning of the Pliocene and the resulting southward flow of warm, salty surface water from the Mediterranean may have been responsible for the early Pliocene deglaciation of the West Antarctic Ice Sheet.

In the Northern Hemisphere, evidence suggests that ice was present in the polar regions as reflected in the form of glacial dropstones in North Atlantic marine sediments of late Miocene age (about 7 Ma), while tillites suggest the presence of intermittent glaciers at about 10 Ma (early late Miocene) in the Wrangell Mountains of Alaska.

It has been suggested that late middle and late

Miocene (about 10–7 Ma) renewed uplift elevated the Himalayan Plateau to a height where it was able to disrupt the global circulation system, modify the monsoonal airflow over southwestern Asia, and push the Earth climatic system across a delicate threshold which led to the development of permanent Northern Hemisphere glaciation in the mid-Pliocene (about 2.5 Ma). The globally significant expansion of savannah and grasses at this time and the increase in aridity in Africa and development of the arid-adapted flora of the Sahara, which appears to date to about 6–5 Ma (or slightly later), appear to be a consequence of the late Miocene Himalayan uplift. *See* PALEOCLIMATE; PALEOCEANOGRAPHY.

Sea-level changes. The predominantly regressive tendency of the Oligocene was replaced in the Miocene by an essentially transgressive early Miocene followed by a regressive middle and late Miocene phase (Fig. 3). A significant transgression occurred near the base of the Miocene, and a second led to the highest sea-level stand of the Miocene during the early middle Miocene. Two rapid sea-level falls in the early middle Miocene were followed by a gradual sea-level lowering during the late middle Miocene, which was punctuated in turn by a major, rapid fall near the middle/late Miocene boundary. Major, rapid sea-level falls bracket the terminal Miocene Mediterranean salinity crisis, followed by a rapid sea-level rise at the base of the Pliocene. Indeed, the boundaries of (at least the lower and lower middle) standard Miocene stage as delineated in Europe can be viewed as corresponding to, and reflecting, the major changes in sea level seen on the global sea-level curve.

Tectonic uplift along the Iberian (Spanish) and African continental margins is considered to have been the main cause of the closure of two marine seaways in the Betic and Rifian areas. Global sea-level lowering, associated with an interval of Antarctic glaciation with peak glacials at 5.79 and 5.75 Ma, may have aided in further isolation of the Mediterranean Basin; but it is not considered to have played a significant role in the initial isolation of the Mediterranean Basin from the global ocean inasmuch as these peak glacial stages postdate the initiation of the Messinian Salinity Crisis by ~200 kyr. Peak glacials at 5.79 and 5.75 Ma may have significantly lowered the global sea level, isolated the Sea of Japan, and temporarily turned into a freshwater lake.

Terminal Miocene salinity crisis. About 6 Ma the northwestern part of Africa collided with Europe, closing off the western end of the Mediterranean to the Atlantic. Together with a global lowering of sea level shortly thereafter of approximately 230 ft (70 m), caused by enhanced glaciation on Antarctica, the Mediterranean and Paratethys seas were isolated from the global ocean system and became large, inland lakes of (predominantly) interior drainage with parallel but distinct geologic histories as one of the most dramatic events in geologic history took place over the next half million years. While about 6% of the salt from the world's ocean was extracted, resulting in a lowering of ocean salinity by about 2%, Tethyan faunas emigrated out of the

Mediterranean to seek temporary safe haven in the Atlantic as the Mediterranean became hypersaline and deposited some 6600–10,000 ft (2–3 km) of salt, gypsum, and anhydrite over an area of 40,000 mi² (10⁶ km²) in a series of discontinuous hypersaline playa lakes and sabkhas (like the present-day Dead Sea or Great Salt Lake) over 10,000 ft (more than 3000 m) below sea level, eventually becoming a gigantic desert like Death Valley. The main difference between the two is that Death Valley is only slightly below sea level. This event has become known to geologists as the Messinian (or Terminal Miocene) Salinity Crisis. Since the Mediterranean salt deposits are over 1.2 mi (2 km) thick, and only 230 ft (70 m) of salt would be produced by a single isolation and evaporative event, intermittent (quasicontinuous) connection with the Atlantic Ocean must have been maintained, probably through either narrow passage (or both) in the Rifian Corridor (northwestern Morocco) or the Betic Straits (southwestern Spain). The Gibraltar Straits did not exist at this time. *See* ANHYDRITE; GYPSUM; HALITE; PLAYA.

The Messinian Stage of the Mediterranean consists of three distinct sedimentary deposits: a lower, normal marine unit exhibiting signs of gradual isolation from the global ocean (7.26–5.96 Ma); a middle unit consisting of two sets of evaporite deposits, the Lower and Upper Evaporite Series, deposited between 5.96 and 5.6 Ma and 5.52 and 5.45 Ma, respectively; and between the cessation of evaporite deposition in the Mediterranean (5.45 Ma) and the early Pliocene deposition of normal marine marls of the Trubi Formation (5.33 Ma)—that is, during the terminal Miocene (Messinian)—a series of nonmarine sediments characterized by a peculiar unique and indigenous megafauna (bivalves, or pelecypods, of the genus *Congerina* and carriids) and microfauna (ostracodes, particularly *Cythereis* and *Loxocoelha djajfarovi*) [Fig. 4]. These deposits, interpreted as having been formed in brackish-water lakes, have been named Lago Mare—the Italian equivalent of the French term applied for a long time to the middle and late Miocene (Paratethyan) Lac Mer. While the Mediterranean was undergoing evaporation, the Black Sea was transformed into a shallow brine lake (but in which no salt deposits have ever been found). The typical Lago Mare fauna of the upper part of the Mediterranean Messinian Stage has also been found in the Black Sea. The general interpretation of the transformation from evaporite deposition to the oligo-mesohaline conditions of the Lago Mare deposits was the results of a rearrangement of the drainage pattern of Europe, such that the Mediterranean captured the riverine runoff from the humid regions of central Europe and the Paratethyan (Black Sea) drainage via the Aegean region. The subsequent infilling of the Mediterranean “bathtub” by the Atlantic occurred at the beginning of the Pliocene Epoch; it was probably abrupt and catastrophic. However, the recent discovery of a fauna of marine euryhaline (95% gobies and trachinoids) and euryhaline and fully marine stenohaline (5% combtooth blennies, gray mullets, herrings, lanternfishes, porgies, pupfishes, silversides,

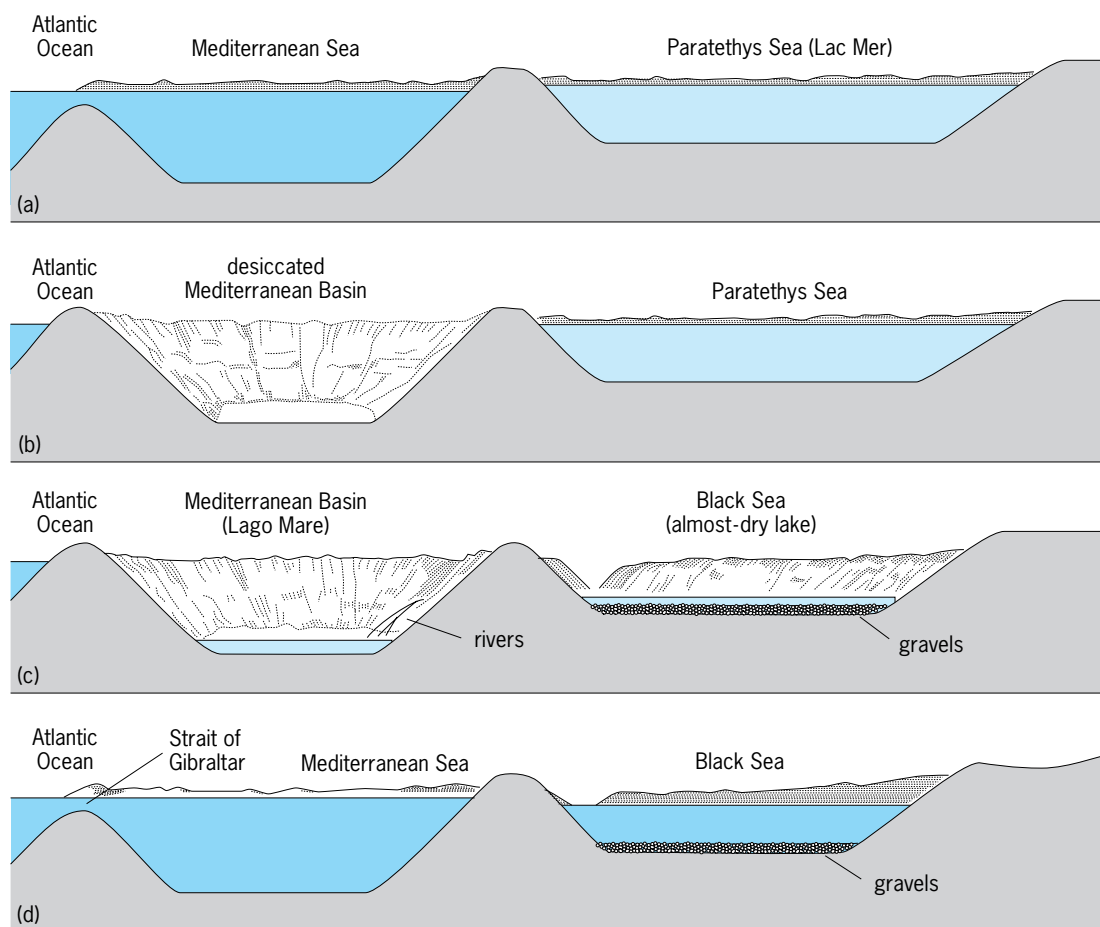


Fig. 4. Changes in Mediterranean Sea and contemporaneous development of the Black Sea. (a) Middle Miocene; about 15 Ma. (b) Late Miocene; about 6 Ma. (c) Late Miocene; about 5.5 Ma. (d) Pliocene; about 5 Ma.

toadfishes, viviparous brotulas) fishes belonging to at least 12 teleostean families in the upper part of Lago Mare deposits in Tuscany, Italy, has suggested a modification of long-held beliefs on the nature of the terminal Miocene depositional setting and history of the Lago Mare and of the (supposed) subsequent early Pliocene infilling of the Mediterranean “bathtub,” which is considered to have been probably abrupt and catastrophic. The discovery of a diverse marine fish fauna in the Lago Mare deposits indicates that normal marine conditions had been established at least during the upper part of the ~ 0.10 -kyr-long Lago Mare deposition and ~ 0.10 kyr before the Miocene/Pliocene boundary at 5.33 Ma. In this revised scenario, the Lago Mare brackish ostracode faunas are interpreted as local indicators of marginal basin conditions in a basin that had already undergone the reestablishment of normal marine conditions with the open ocean. The Miocene/Pliocene boundary at 5.33 Ma witnessed the further development of open marine conditions in the Mediterranean Basin with the formation of normal, deep-sea chalk deposits—the Trubi Marls—characterized by normal open ocean marine calcareous protists (planktonic foraminifera and calcareous nanoplankton), and was the results of ongoing tectonic activity inasmuch as the next significant glaciation is recorded at ~ 4.9 Ma (that is, early Pliocene).

Since the Tethyan Ocean had been cut off from the Indo-Pacific earlier in the Miocene, the coral reefs that thrived in the western Tethys (southern Spain) until the onset of the terminal Miocene evaporative stage never returned to the Mediterranean, and the Mediterranean today is characterized by its low diversity of fauna and low nutrient levels. During the latest Miocene (Messinian) interval of lowered sea level, major rivers around the Mediterranean cut gorges several thousand feet deep into granite in the Nile Valley, and comparable gorges, which have been traced down to the edge of the abyssal plain of the sea floor, have been documented in the Rhone Valley of southern France as far north (180 mi or 300 km from the coast) as Valence, in the Po Valley nearly up the Alps, on the islands of Corsica and Sardinia, and in the eastern (Syria, Israel) and southern (Libya, Algeria) margins of the Mediterranean Sea. *See BLACK SEA.*

Life. During the Miocene, life assumed much of its modern aspect. The spread of grasses and weeds throughout this epoch, but particularly in the late Miocene, and concomitant reduction in and thinning of forests reflected the global Neogene cooling as the Earth entered deeper into an ice house state. In this environment, snakes, frogs, and murids (rats, mice) expanded in diversity and habitat; songbirds reflect the expansion of seed-bearing herbs and, like frogs,

the concomitant diversification of insects, many of which are found entombed in middle Miocene amber from the Dominican Republic. Grazing animals (elephants, rodents, horses, camelids, and rhinos, for example) developed high crowned teeth to resist significant wear caused by silicon fragments in the developing grasses. Some animals assumed gigantic proportions such as *Baluchitherium*, a Eurasian rhino that stood 16 ft (5 m) at the shoulders, and the tallest camel known, a giraffelike form that was over 12 ft (3.5 m) tall.

In at least two adaptive radiations, early proboscideans (elephant family) and mastodonts migrated into Europe from Africa following the docking of Africa and Eurasia about 18 Ma. Over a dozen families of small mammals migrated to Africa, including insectivores and rodents, while larger migrants included ancestors of modern deer, cattle (bovids), antelope, pigs, and large forms such as chalicotheres, rhinos, and early giraffes. Felid-like carnivores (the cat family) originated in the Oligocene Epoch, ~35 Ma. Modern cat species originated in Asia in the late Miocene, ~11 Ma, and replaced the earlier, ancestral creodonts that originated in the Paleogene, with the divergence of the *Panthera* lineage ~10.8 Ma, followed shortly by the bay cat lineage at ~9.4 Ma. An early migration event at ~8.5–5.6 Ma brought an early progenitor of the caracal lineage to Africa, and a second migration at ~8.5–8.0 Ma saw the immigration of a common ancestor to five felid lineages (domestic cat, leopard cat, lynx, ocelot, and puma) to North America via the Bering Land Bridge. A concomitant immigration of Eurasian carnivores (mustelids, procyonids, ursids, and saber-toothed felids) is thought to have crossed from Eurasia to North America. During a migration event about 15 Ma, early apes spread out into the woodland and forests of Eurasia, while hyaenids and shrews infiltrated Africa.

In North America, grazing animals were among the earliest ruminants (camelids, horses), and these were joined shortly afterward by Eurasian migrants by way of the Bering Land Bridge, including over 50 genera of ungulates and large carnivores in the late Miocene. In the late Miocene (about 7 Ma), savannah grasslands underwent a significant areal expansion and major faunal changes occurred which coincided with the interval of the Terminal Miocene Salinity Crisis. Extinctions exceeded replacements, so that the main result was a diminution of faunal diversity by the early Pliocene. See MAMMALIA.

Placental ungulates reached South America in the early Paleocene, and by Miocene time they had evolved into a variety of forms and sizes. South America remained isolated from North America for over 100 million years and, with the exception of an apparently brief interval in the Paleocene when minor exchange took place between the two continents, little or no faunal interchange took place between the two until the latest Miocene–early Pliocene. However, parallel evolution took place among many of the terrestrial vertebrates, a good example being the South American Miocene forms *Thoatherium* and *Diadiaphorus*, which bore a strik-

ing resemblance to horses of North America. Large carnivorous marsupials were common, including wolflike boryhaenids.

The Neogene evolution of higher primates (anthropoids) was played out in the Southern Hemisphere primarily as late Paleogene cooling associated with modified tropical environments and climates forced the southward displacement of most primates. Old World monkeys evolved in Europe and Asia; New World monkeys evolved in South America. Hominoids [including hyalobatids (gibbons) and panids (African apes, or gorillas, and chimpanzees) and hominids (humans) evolved in Africa, while the orangutan evolved in Asia.

Two groups of apelike hominoids evolved during the Miocene: dryomorphs and ramapithecines (and their late Miocene Asian descendants, also known as sivapithecines). Dryomorphs were small primates that browsed in early Miocene forests and grasslands of East Africa. Ramapithecines and sivapithecines constitute an extinct group of predominantly Eurasian apes that probably evolved from an early Miocene African ancestor.

The relatively free interchange between Eurasia and Africa between 18 and 12 Ma appears to have come to an end in the late Miocene, after which (about 8 Ma) the hominoids of Eurasia and Africa appear to have followed separate and independent lines of evolution: the pongids in Asia, and the panids and hominoids leading eventually to the true hominids in (predominantly East) Africa. This scenario has been linked, in turn, with the development of the East African Rift (and its northward extension into the Red Sea and Gulf of Suez), which would have served as a geographic barrier allowing independent evolution toward forest (panid) and savannah (hominid) adapted forms. With the late Miocene change in climate (7–5 Ma) to cooler, drier conditions and the spread of open savannah and grasslands, monkeys came to dominate the African forest at the expense of dryomorphs. There is a gap in the terrestrial fossil record during this interval of time, and it is only in the early Pliocene (about 4 Ma) that the story of human evolution resumes with the discovery of the earliest true hominids (australopithecines) in East Africa, about 1 million years older than the australopithecine footprints of Laetoli and the skeletons of Lucy and other australopithecines at Hadar in Ethiopia at about 3 million years. Within the past few years, however, there have been reports of fossils of stem hominids in the latest Miocene (~7–5 Ma) in Africa, but these finds await further confirmation. See AUSTRALOPITHECINE; FOSSIL HUMANS.

In the marine realm, major radiation of mammals including walruses, seals, sea lions, and whales occurred during the early Miocene. Carnivorous sperm whales with large teeth were among the earliest representatives of the group; and baleen whales, long-nosed dolphins, and sharks, including the large gigantic white shark *Carcharodon*, were among distinctive early Miocene forms. On the sea floor, large bivalve mollusks of the scallop family thrived in the early Miocene, and a distinct horizon of large

pectenids occurs in lower Miocene rocks of Europe and North America and in corresponding levels in the deposits of the Paratethyan Sea in east-central Europe and at least as far to the east as Iran, attesting to an interval of global climatic amelioration. Among the protozoa, planktonic foraminifera experienced a major radiation in the early and middle Miocene following the drastic reduction in diversity during the middle and late Eocene, some 15–20 million years earlier. Mangroves and coral reefs flourished in a circum-equatorial belt spanning the Indo-Pacific and Caribbean regions, but the latter were eliminated from the Mediterranean during the terminal Miocene Salinity Crisis, never to return with early Pliocene flushing from the Atlantic. Siliceous oozes expanded and encircled Antarctica (and to a lesser extent the North Pacific) beginning in the early Miocene. This reflected the global cooling of the Neogene and the development of the Circum-Antarctic Current in the early Miocene, which isolated high southern latitudes and the Antarctic continent from the warmer ocean currents to the north for the remainder of the Neogene. *See* FORAMINIFERIDA; MANGROVE; MOLLUSCA; REEF.

A major biogeographic reorganization of tropical, shallow-water corals occurred over a several-million-year span in the early Miocene. About half of these shelf-edge-dwelling reef corals became regionally extinct in the Caribbean, with the majority surviving to the present day in the Indo-Pacific region. Upwelling of cold, nutrient-rich waters, reflected in contemporaneous circum-Caribbean phosphorite deposits, are thought to have been responsible for the local elimination of the corals. An essentially contemporaneous reduction in diversity (and regional extinction) of tropical, photosymbiont-bearing, larger benthic protozoa occurred as well, with a combination of local upwelling (Caribbean) and tectonic closure of ancient seaways (eastern Tethys) having probably played major roles. *See* EXTINCTION (BIOLOGY); GEOLOGIC TIME SCALE; PALEOECOLOGY; PALEONTOLOGY.

W. A. Berggren

Bibliography. G. Carnevale et al., Mare versus Lago-Mare fishes and the Mediterranean environment at the end of the Messinian Salinity Crisis, *J. Geol. Soc.*, 163(1):75–80, 2006; K. J. Hsu, When the Black Sea was drained, *Sci. Amer.*, 258(5):52064, 1978; K. J. Hsu, When the Mediterranean Sea dried up, *Sci. Amer.*, pp. 27–36, December 1972; W. Krijgsman et al., Revised astrochronology for the Ain el Beida section (Atlantic Morocco): No glacio-eustatic control for the onset of the Messinian Salinity Crisis, *Stratigraphy*, 1(1):87–101, 2004; S. R. Stanley, *Earth and Life Through Time*, 1986.

Mira

The first star recognized to have a periodic brightness variation. Mira (officially designated Omicron Ceti, in the constellation Cetus, the Whale) was dis-

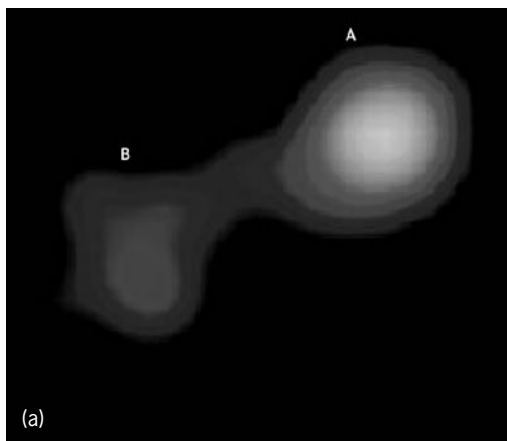
covered in 1596 by David Fabricius (or Fabricus), a clergyman and an amateur astronomer from Esens in East Friesland (now in Germany), while he was searching for Mercury. He mistook it for a nova because it later faded from view. However, he saw it reappear 14 years later. Not until 1638 was it recognized to be the first known variable star, when Johann Holwarda, from Franeker, also in Friesland (now in the Netherlands), rediscovered it, and in 1639 determined its period to be 11 months. In 1642, Johannes Hevelius of Danzig, who also observed the star, called it Mira, meaning The Wonderful. Mira is the prototype of an entire class of Mira-type pulsating long-period variables. Although it once resembled the Sun, Mira has evolved into a cool red giant star that is at the end of its life. *See* GIANT STAR; STELLAR EVOLUTION.

Contracting and expanding every 332 days, Mira typically varies in visual brightness from about magnitude 3.4 (brightest) to about 9.3 (faintest). However, an individual maximum can sometimes be as bright as second magnitude, while at other times the maximum may reach barely fifth magnitude. Mira is the brightest long-period variable because it is intrinsically bright and only 420 light-years from the Earth. Angular diameter measurements such as those from the Very Large Telescope (VLT) Interferometer show that the diameter of Mira varies from about 330 to 400 times the diameter of the Sun during its pulsational cycle, and has an average size larger than the orbit of Mars. *See* MAGNITUDE (ASTRONOMY).

Mira is also known as a symbiotic variable: a long-period variable star with a close hot compact companion star. The companion, VZ Ceti (or Mira B), forms a close visual double with Mira (or Mira A; the pair is often known as Mira AB). VZ Ceti was also discovered in 1959 by G. van Biesbroeck to be a variable star in its own right. It is a hotter, bluer, burned-out star, called a white dwarf, surrounded by material captured from Mira's wind. VZ Ceti varies in brightness between magnitude 9.5 and 12.0, and it is only about the size of the Earth. When VZ Ceti is bright, it may affect the apparent brightness of Mira, particularly when Mira is faint. The orbital period is about 498 years. *See* BINARY STAR; SYMBIOTIC STAR; WHITE DWARF STAR.

The *Chandra X-ray Observatory* has observed Mira and its companion (see **illustration**). The separation between Mira and its companion has been measured to be currently about 70 times more than that between the Earth and the Sun (about twice the distance between Pluto and the Sun, and equal to an angular size of 0.6 arcsecond).

The *Chandra* image shows both Mira A and its hot companion, Mira B. Mira A is losing matter from its upper atmosphere via a stellar wind. The image shows a faint bridge of this material between the two stars. Mira A is bright in this x-ray image because it was undergoing a flare; normally, red giants are very faint at x-ray wavelengths. Mira B is likely surrounded by a disk of the accreted material, which is likely the cause of its strange shape in this image.



Mira and its companion. (a) *Chandra* image, taken in x-ray wavebands, of Mira (labeled A) and its companion (B). (b) Artist's conception of the Mira AB system. (M. Karovska, Harvard-Smithsonian Center for Astrophysics; NASA)

At about 420 light-years from the Earth, Mira is the nearest wind-accreting binary system. See STAR; VARIABLE STAR. Arne Henden; Janet Akyüz Mattei

Bibliography. D. Hoffleit, History of the discovery of Mira stars, *J. Amer. Ass. Variable Star Observers*, 25:115-136, 1997; J. E. Isles, Mira's 400th Anniversary, *Sky Telesc.*, 91(2):72-73, February 1996; M. Karovska, High resolution observations of Mira, *J. Amer. Ass. Variable Star Observers*, 25:75-79, 1997; G. van Biesbroeck, Mira as a double star, *Publ. Astron. Soc. Pacific*, 71:462-465, 1959; H. C. Woodruff et al., Interferometric observations of the Mira star o Ceti with the VLTI/VINCI instrument in the near-infrared, *Astron. Astrophys.*, 421:703-714, 2004.

Mirage

A distorted scene of distant objects resulting from the passage of light through a nonuniform medium. In a uniform atmosphere light would travel in straight lines and distant scenes would appear without distortion. Since the density of the air changes with height and with temperature, the atmosphere is not uniform and light rays will follow curved paths. The direction of the curvature is to deflect a ray away from a warmer (less dense) layer of air. **Figure 1** shows some of the features of the most familiar desert or

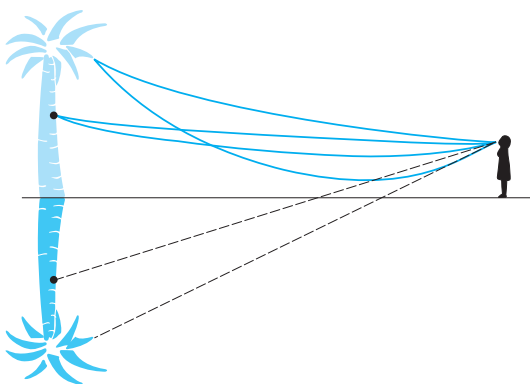


Fig. 1. Rays that give rise to the inverted image in the desert mirage. (After R. Greenler, *Rainbows, Halos, and Glories*, Cambridge University Press, 1980)

hot-road mirage. The ground is heated by the sun and in turn heats (and rarifies) the layer of air just above the ground. Then rays passing near the ground are bent (refracted) upward. The observer can see the top of the tree in the figure, for example, by looking in two directions. The strongly refracted rays produce an inverted view of the tree. The only other place in nature where such an effect can be seen is in reflection from a water surface, and it is natural to interpret the mirage as evidence of a body of water.



Fig. 2. Mirage over the frozen Arctic Ocean, resulting from an inversion layer. (From R. Greenler, *Rainbows, Halos, and Glories*, Cambridge University Press, 1980)

Sometimes a warm layer of air can be trapped between layers of cooler air overhead (a temperature inversion), resulting in light-ray paths that curve downward, so that distorted images of ships or land that normally would lie hidden below the horizon can be seen (**Fig. 2**). See METEOROLOGICAL OPTICS; REFRACTION OF WAVES; TEMPERATURE INVERSION. Robert Greenler

Mirror optics

The use of plane or curved reflecting surfaces for the purpose of reverting, directing, or forming images. The most familiar use of reflecting optical surfaces is for the examination of one's own reflected image in a flat or plane mirror. A single reflection in a flat mirror produces a virtual image which is reverted or reversed in appearance. The use of one or more

reflecting surfaces permits light or images to be directed around obstacles, with each successive reflection producing a reversal of the image. A curved mirror, either spherical or conic in form, will produce a real or virtual image in much the same manner as a lens, but generally with reduced aberrations. There will be no chromatic aberrations since the law of reflection is independent of the color or wavelength of the incident light. See ABERRATION (OPTICS); OPTICAL IMAGE.

An optical surface which specularly reflects the largest fraction of the incident light is called a reflecting surface. Such surfaces are commonly fabricated by polishing of glass, metal, or plastic substrates, and then coating the surface of the substrate with a thin layer of metal, which may be covered in addition by a single or multiple layers of thin dielectric films. The law of reflection states that the incident and reflected rays will lie in the plane containing the local normal to the reflecting surface and that the angle of the reflected ray from the normal will be equal to the angle of the incident ray from the normal. This law is a special case of the law of refraction in that the angles rather than the sines of the angles of incidence and reflection are equal. Formally, this relation is commonly used in calculations by setting the effective index of refraction prior to incidence on the surface. When this concept is introduced, all of the formulas relating to lenses are applicable to reflective optics. In this article, however, the imaging relations will be described in the most appropriate form for reflecting surfaces. See GEOMETRICAL OPTICS.

Plane mirrors. The formation of images in the plane mirrors is easily understood by applying the law of reflection. **Figure 1** illustrates the formation of the image of a point formed by a plane mirror. Each of the reflected rays appears to come from a point image located a distance behind the mirror equal to the distance of the object point in front of the mirror. In **Fig. 1**, the face of the observer can be considered as a set of points, each of which is imaged by the plane mirror. Since the observer is viewing the facial image from the object side of the mirror, the face will appear to be reversed left for right in

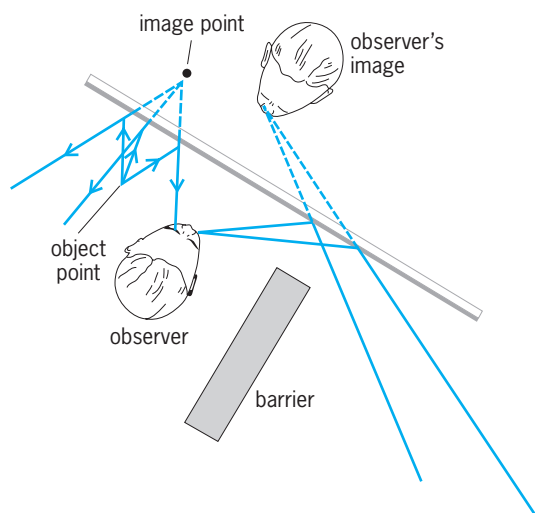


Fig. 1. Formation of images by a plane mirror.

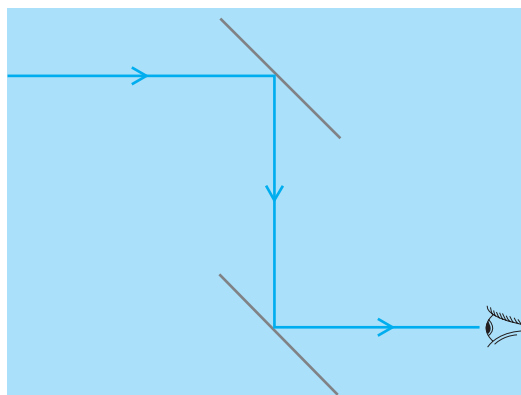


Fig. 2. Simple mirror periscope.

the virtual image formed by the mirror. Such a virtual image cannot, of course, be projected on a screen, but can be viewed by a lens, in this case the eyes of the observer. **Figure 1** also indicates the redirection of light by a plane mirror, in that a viewer who cannot observe the object point directly can observe the virtual image of the point formed by the mirror. A simple optical device which is based on this principle is the simple mirror periscope (**Fig. 2**), which uses two mirrors to permit viewing of scenes around an obstacle. In this case two reflections are present and provide an image which is correctly oriented, and not reverted, to the observer. The property of reversion in a complicated mirror system depends upon the location and view direction of the observer, as well as the number of reflections that take place and the orientation of the planes through which the light is directed. See PERISCOPE.

Prisms. These are solid-glass optical components that use reflection at the faces to provide redirection of the optical pencils passing through them. The advantage of the use of a prism is that the reflecting surfaces are maintained in accurate location with respect to each other by the integrity of the glass material making up the body of the prism. Difficulties with prisms are that very homogeneous glass is required since the light may make many passes through the prism, and that a prism is optically equivalent to insertion of a long block of glass into the imaging system. The insertion of such a glass block often results in a system which is mechanically shorter in space, but the aberration balance of the imaging system is changed, frequently requiring a redesign of the associated optical components in order to accommodate the increased glass path. It is obvious that the use of solid glass prisms may introduce more weight into the optical design than the equivalent metal mounts required for a similar arrangement of mirrors in air. In certain cases, the angle of incidence on a reflecting surface within the prism may exceed the critical angle of incidence, and no reflective coating may be required on such a surface. **Figure 3** shows some common types of optical prisms with reflecting surfaces. The applications of such prisms range from simple redirection of light to variable angle of rotation of the image passed through the prism and binocular combination of images. A special case of reflection is the use of a beam splitter which permits

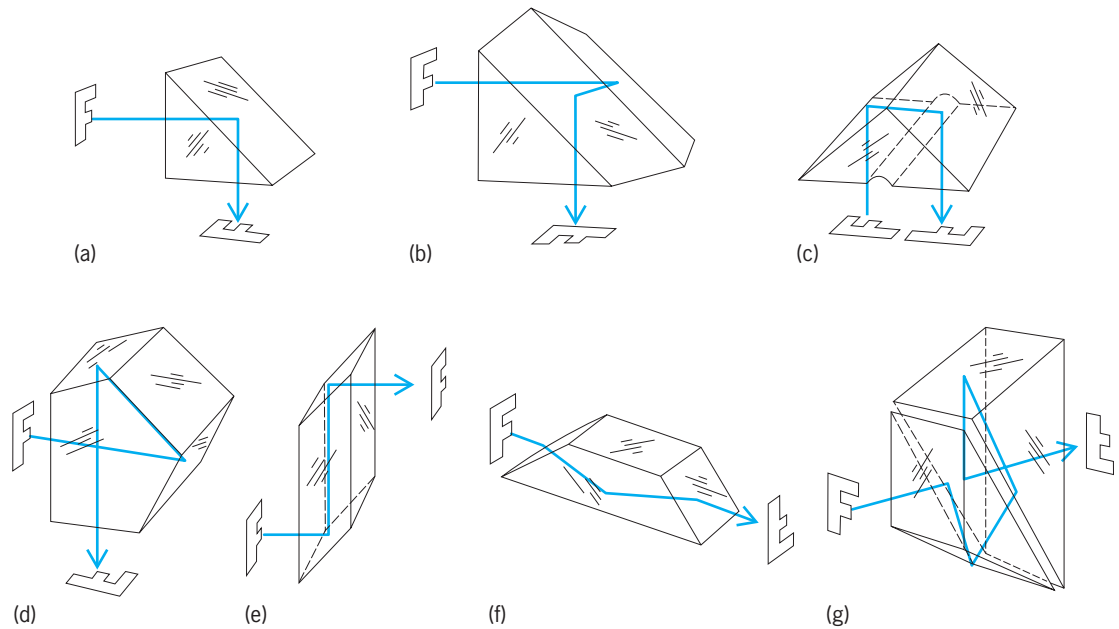


Fig. 3. Prism types. (a) Right-angle. (b) Amici roof. (c) Porro. (d) Pentaprism. (e) Rhomboid. (f) Dove. (g) Pechan.

the splitting of light, or the combining of two beams by the use of a surface which is partially reflecting and partially transmitting. See BINOCULARS; OPTICAL PRISM.

Spherical mirrors. These are reflecting components that are used in forming images. The optics of such mirrors are almost identical to the properties of lenses, with their ability to form real and virtual images. In the case of spherical mirrors, there is a reversal of the direction of light at the mirror so that real images are formed on the same side of the mirror as the object, while virtual images are viewed from the object side but appear to exist on the opposite side of the mirror. Both concave and convex spherical mirrors are commonly encountered. Only a virtual erect image of a real object will be formed by a convex mirror. Such mirrors are commonly used as wide-angle rearview mirrors in automobiles or on trucks. The image formed appears behind the mirror and is greatly compressed in space, with a demagnification dependent on the curvature of the mirror. A concave spherical mirror can form either real or virtual images. The virtual image will appear to the observer as erect and magnified. A common application is the magnifying shaving mirror frequently found in bathrooms. A real image will be inverted, as is the real image formed by a lens, and will actually appear in space between the observer and the mirror.

Figure 4 shows the formation of real and virtual images by a spherical mirror. The equation which applies to all of the image relations is given below.

$$\frac{1}{S'} + \frac{1}{S} = \frac{2}{R}$$

The distances S and S' are measured from the surface of the spherical mirror; when either is negative, a virtual image is formed behind the mirror. The constant R is the radius of curvature of the mirror. The

magnification of the image is the ratio of the image distance S' to the object distance S .

Conic mirrors. These are a special case of the spherical mirror with improved image quality. A spherical mirror will form an image which is not perfect, except for particular conjugate distances. The use of a mirror which has the shape of a rotated conic section, such as a parabola, ellipsoid, or hyperboloid, will form a perfect image for a particular set of object-image conjugate distances and will have reduced aberrations for some range of conjugate relations. Two of the most familiar applications for conic

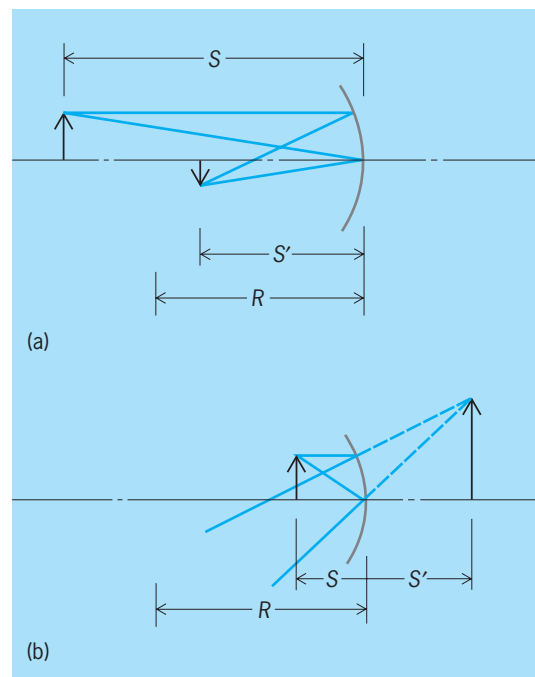


Fig. 4. Formation of images by spherical mirror. (a) Real image. (b) Virtual image.

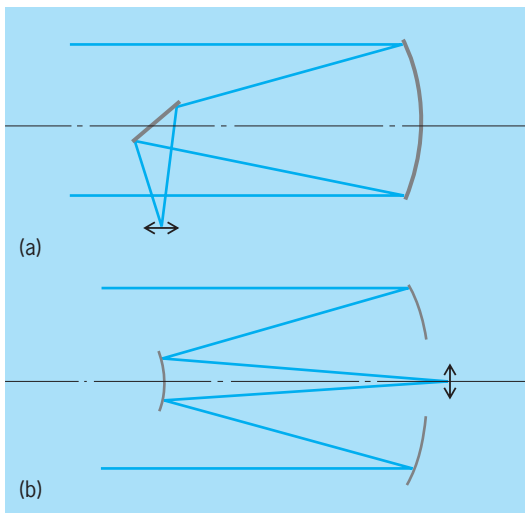


Fig. 5. Applications of conic mirrors. (a) Newtonian telescope. (b) Cassegrain telescope.

mirrors are shown in Fig. 5. Figure 5a shows the use of a paraboloid of revolution about the optical axis to form the image of an object at an infinite distance. In this drawing the image to be viewed by the observer at the eyepiece is relayed to the side of the telescope tube by a flat folding mirror in what is called a Newtonian form of a telescope. This demonstrates one of the difficulties that is found with the use of reflecting optical components to form real images; namely, that the image must often be relayed out of the incident path on the image-forming mirror, otherwise the observer will block some of the light from the object. Not all reflecting systems carry out this relaying in the same manner. The Cassegrain system uses a curved secondary mirror to achieve magnification of the final image while allowing the image to fall outside the telescope barrel.

Figure 5b shows the use of a paraboloid as the primary mirror with the image relayed to the final image location by a hyperbolic secondary mirror. Such a use of two mirrors permits the construction of a long-focal-length telescope within a relatively short space. This latter form, usually referred to as a Cassegrain telescope, serves as the principal type of modern reflecting astronomical telescope. One of the advantages of this type of telescope design is the freedom from chromatic aberration that would be present in a refracting telescope. See OPTICAL SURFACES.

Mirror coatings. The reflectivity of a mirror depends on the material used for coating the reflecting surface. The conventional coatings for glass mirror surfaces are silver or aluminum, which are vacuum-deposited or sputtered onto the surface. In some cases, chemical deposition will be used. Most mirrors intended for noncritical uses, such as looking glasses or wall mirrors, will have the reflecting metallic coating placed on the back side of the glass, thus using the glass to protect the coating from oxidation by the atmosphere. Mirrors for most critical or scientific uses require the use of front-surface reflectors, with the reflecting coating on the exposed front surface of the glass. In this case, a hard overcoat of a thin layer of silicon dioxide is frequently deposited over

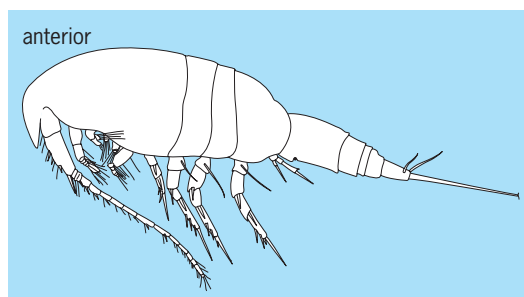
the metal to protect the delicate thin metal surface. The reflectivity of the mirror with respect to wavelength depends on the choice of the metal for the reflector and the material and thickness of material layers in the overcoat. In some cases, a fully dielectric stack will be used as a reflecting coating with special spectral selective properties to form a dichroic beam splitter, as in a color television camera, or as an infrared-transmitting "cold mirror" for a movie projector illumination system. See REFLECTION OF ELECTROMAGNETIC RADIATION.

R. R. Shannon

Misophrioida

A copepod order of free-living crustaceans made up of 34 species in 16 genera and 3 families. *Misophria pallida* was the first species described, by Axel Boeck in 1865, from a sample taken close to the sea floor in shallow water. A second genus and species, *Benthomisophria palliata*, was established by Georg Ossian Sars in 1909 for a bathypelagic species from the North Atlantic Ocean. The remaining species have been described since 1964 and reflect an increased interest by taxonomists in this order of copepods. Many of the recently described species have been recovered during the exploration of anchialine caves, which are coastal marine water bodies with subsurface rather than surface hydrologic connections to the sea. Anchialine caves are one category of crevicular habitat in coastal marine geological formations. The number of known species of Misophrioida is expected to increase significantly, as similar habitats, such as the interstices of coral reefs, are explored more systematically. See COPEPODA; CRUSTACEA.

Description. Adult females are 0.6–5.7 mm (0.02–0.23 in.) long and males 0.6–3.7 mm (0.02–0.15 in.); the length range includes small species living close to the sea floor and the larger, bathypelagic species. The cuticle of the adult body is distinctly ornamented. The second thoracic somite, bearing the first swimming leg, articulates with the first somite. However, a posterior extension of the cuticle of the first thoracic somite completely covers the second thoracic somite so that these two somites appear unarticulated (see **illustration**). The last two thoracic somites are associated with the posterior part of the body (the podoplean architecture). The internal anatomy of *B. palliata* includes a heart, but the naupliar (larval) eye, a normal feature of adult copepods,



Side view of typical misophrioidan.

is absent. The excretory system is an antennal gland, apparently retained from the naupliar phase of development, rather than a maxillary gland, which is the usual excretory system for postnaupliar stages of copepods. An embryo carried by a female of *M. pallida* hatched to a nauplius (a crustacean larval stage), which molted directly to the first juvenile copepodid. The addition of body somites and limb segments during the copepodid phase of development of *B. palliata* follows a pattern common for many copepods. The vertical distribution of copepodid stages of the bathypelagic *B. palliata* appears uniform between 2500 and 4000 m (8200 and 13,100 ft).

Phylogeny. The origin of the Misophrioida from either shallow marine waters directly over the sea floor or marine waters of anchialine caves continues to be debated. In phylogenetic analyses, the Misophrioida almost invariably are the least derived of the podoplean (the last two thoracic somites with the posterior part of the body) copepod orders, with the gymnoplean calanoids (only the last thoracic somite with the posterior part of the body) as their closest older order. This relationship is based on the presence in misophrioids of a heart, a podoplean body architecture, and a multisegmental configuration of several limb branches. However, the hypothesis has not been examined using the contemporary logic of phylogenetic systematics. *See* CALANOIDA.

Frank D. Ferrari

Bibliography. A. Boeck, Oversigt over de ved Norges Kyster jagttagne Copepoder henhørende til Calanidernes, Cyclopidernes og Harpacticidernes Familier, *Forbandl. Vidensk.-Selsk.*, Christiania, vol. 1865; G. A. Boxshall, On the anatomy of the misophrioid copepods, with special reference to *Benthomisophria palliata* Sars, *Phil. Trans. Roy. Soc. London*, ser. B, vol. 297, 1982; G. A. Boxshall and D. Jaume, On the origin of misophrioid copepods from anchialine caves, *Crustaceana*, vol. 72, 1999; G. A. Boxshall and D. Jaume, Discoveries of cave misophrioids (Crustacea: Copepoda) shed new light on the origin of anchialine faunas, *Zool. Anz.*, vol. 239, 2000; G. A. Boxshall and H. S. J. Roe, The life history and ecology of the aberrant bathypelagic genus *Benthomisophria* Sars, 1909 (Copepoda: Misophrioida), *Bull. Brit. Mus. Nat. Hist. (Zool.)*, vol. 38, 1980; R. Gurney, Notes on some Copepoda from Plymouth, *J. Mar. Biol. Ass. UK*, vol. 19, 1933; P. Martínez-Arbizu and D. Jaume, New hyperbenthic species of *Misophriopsis* and *Misophriella*, first record of misophrioid copepods (Crustacea) from Antarctic waters, *Helgoland Mar. Res.*, vol. 53, 1999; G. O. Sars, Note préliminaire sur trois formes remarquables de Copépodes, provenant des campagnes de S.A.S. le Prince Albert de Monaco, *Bull. Inst. Océanog.*, Monaco, vol. 147, 1909.

Mississippian

The fifth period of the Paleozoic Era. The Mississippian System (referring to rocks) or Period (referring to time during which these rocks were deposited) is employed in North America as the lower (or older)

| | | |
|-------------|---------------|---------------|
| CENOZOIC | QUATERNARY | |
| | TERTIARY | |
| MESOZOIC | CRETACEOUS | |
| | JURASSIC | |
| | TRIASSIC | |
| PALEOZOIC | PERMIAN | |
| | CARBONIFEROUS | PENNSYLVANIAN |
| | | MISSISSIPPIAN |
| | DEVONIAN | |
| | SILURIAN | |
| | ORDOVICIAN | |
| | CAMBRIAN | |
| PRECAMBRIAN | | |

subdivision of the Carboniferous, as used on other continents. The name Mississippian is derived from rock exposures on the banks of the Mississippi River between Illinois and Missouri.

The limits of the Mississippian Period are radiometrically dated. Its start (following the Devonian) is dated as 345–360 million years before the present (Ma). Its end (at the start of the next younger North American period, the Pennsylvanian) is dated as 320–325 Ma. The duration of the Mississippian is generally accepted as 40 million years (m.y.). Biochronologic dating within the Mississippian, based on a combination of conodont (phosphatic teeth of eel- or hagfish-like primitive fish), calcareous foraminiferan, and coral zones permits a relative time resolution to within about 1 m.y. *See* CARBONIFEROUS; PENNSYLVANIAN.

Subdivisions. The Mississippian is divided, in ascending order, into the Lower Mississippian, comprising the Kinderhookian and Osagean, and the Upper Mississippian, comprising the Meramecian and Chesterian. Kinderhookian, Osagean, Meramecian, and Chesterian are used in North America as series (for rocks) and as stages (for time). In Illinois, Valmeyeran is commonly used for the Osagean and Meramecian combined. The type localities for the Kinderhookian, Meramecian, and Chesterian are in the upper Mississippi Valley, whereas the type locality of the Osagean is in western Missouri.

Conodont zones are widely accepted as subdivisions of the Mississippian (Fig. 1) and are used for intercontinental correlation. The 10 conodont zones that subdivide the Kinderhookian and Osagean have been proposed as global standard zones, whereas the seven conodont zones that subdivide the Meramecian and Chesterian are used mainly in the United States. Correlated to these zones by detailed biostratigraphic studies are 15 foram (calcareous foraminiferan) zones and 13 coral zones (Fig. 1).

This combined, conodont-foram-coral biochronology provides the best tool for precise correlation within North America. See STRATIGRAPHY.

Lithofacies and paleogeography. During much of Mississippian time, the central North American craton (stable part of the continent) was the site of an extensive marine carbonate platform on which mainly limestones and some dolostones and evaporites were deposited. This platform extended either from the present Appalachian Mountains or Mississippi Valley to the present Great Basin, as exemplified by the reconstructed map of this area in mid-Mississippian (middle Osagean) time (Fig. 2). The paleoequator passed through Nevada and Wisconsin. Paralleling it to the south was the Transcontinental arch, composed of a peninsula and several islands (extending from Wisconsin Highlands to Zuni-Defiance Island) that separated the carbonate platform into halves. See DOLOMITE ROCK; LIMESTONE; SALINE EVAPORITES.

The Mississippian continent was bounded on its present east, south, and west sides by deep troughs and basins in which mainly siliciclastic sediments—siltstones, sandstones, and conglomerates—were deposited as debris produced by erosion of bordering orogenic highlands. In the east, these siliciclastic rocks were deposited as the Pocono clastic wedge and seaward Borden deltaic complex. In the west, the Antler highlands shed clastic sediments into an adjacent trough. More distal parts of the subsiding troughs, which were not overwhelmed by clastics, were the sites of sediment-starved basins, exemplified by the Deseret basin and Caballos-Ouachita trough (Fig. 2), in which mainly shales, cherts, and phosphorites were slowly deposited.

Most Mississippian limestones are bioclastic (composed of parts of organisms), except in the bordering troughs where they are largely calciclastic (turbiditic calcarenites, wackestones, and conglomerates), eroded from preexisting limestones, transported, and then redeposited.

The predominant Mississippian rock type on the carbonate platform is crinoidal limestone or encrinite formed from the skeletal remains of crinoids, blastoids, and other echinoderms. In fact, the Mississippian has been called the Age of Crinoids. The second most common carbonate rock is oolite, formed by aggregates of ooids (sand-size spheres composed of chemically precipitated concentric grains). Another important rock type is very fine grained limestone termed micrite because of its microscopic grains. Evaporitic rocks, mainly anhydrite and some halite, were also deposited in peritidal environments of the Madison shelf (Fig. 2) and Williston basin.

In Kinderhookian and Osagean times, reeflike mudmounds were deposited on the slopes between the carbonate platform and adjacent basins. Fine-grained sediments in these mudmounds were held in place by organisms which, although not preserved, were probably mats of algae or bacteria. Crinoids growing on these mudmounds contributed coarser debris to the sediments. See ANHYDRITE; CRINOIDEA; DEPOSITIONAL SYSTEMS AND ENVIRONMENTS; ECHINODERMATA; HALITE; MARINE SEDIMENTS; OOLITE.

| CONODONT ZONES | FORAM ZONES | CORAL ZONES | TRANSGRESSIONS → REGRESSIONS ← | NA S | EU S | |
|---|--|-------------|-----------------------------------|------------|-----------------|---------------|
| <i>primus</i> | 19 | VI | ← | CHESTERIAN | NAMURIAN (part) | |
| <i>muricatus</i> | | | | | | |
| <i>unicornis</i> | | | | | | |
| <i>naviculus</i> | 18 | V B | → | | | |
| | 17 | | | | | |
| <i>bilineatus</i> – Upper <i>Cavusgnathus</i> | 16s | V A | | | | |
| | 16i | | | | | |
| Lower <i>Cavusgnathus</i> | 15 | IV | | ← | MERAMECIAN | VISEAN |
| | 14 | III D | | | | |
| <i>homopunctatus</i> – Upper <i>texanus</i> | 13 | | III C | | | |
| | <i>mehli</i> – Lower <i>texanus</i> | 12 | III B | | | |
| 11 | | III A | | | | |
| 10 | | | | | | |
| <i>anchoralis</i> – <i>latus</i> | 9 | II B | → | OSAGEAN | TOURNAISIAN | |
| Upper <i>typicus</i> | 8 | | | | | |
| Lower <i>typicus</i> | 7 | II A | | | | |
| <i>isosticha</i> – Upper <i>crenulata</i> | Pre-7 | I C | | ← | | KINDERHOOKIAN |
| | | I B | | | | |
| Lower <i>crenulata</i> | I A | | | | | |
| <i>sandbergi</i> | | | | | | |
| Upper <i>duplicata</i> | | | | | | |
| Lower <i>duplicata</i> | | | | | | |
| <i>sulcata</i> | | | → | | | |

Fig. 1. Conodont-foram-coral biochronology, sea-level changes, and rock series subdivisions of the Mississippian in North America. NA S, North American Series; EU S, European Stages. Length and weight of arrows show relative scale of the major North American sea-level rises (transgressions) and falls (regressions).

Sandstones and siltstones predominate in the Borden deltaic complex (Fig. 2) and in the younger Humbug deltaic complex, which filled the former Deseret basin in Meramecian time. Coarse-to-fine conglomerates composed of boulders, cobbles, and pebbles of deeper marine, older Paleozoic rocks that had been folded and elevated to form the Antler highlands intertongue with shales and argillites in the adjacent flysch trough. See SANDSTONE; SEDIMENTARY ROCKS.

In the Deseret basin, fine siliceous rocks termed

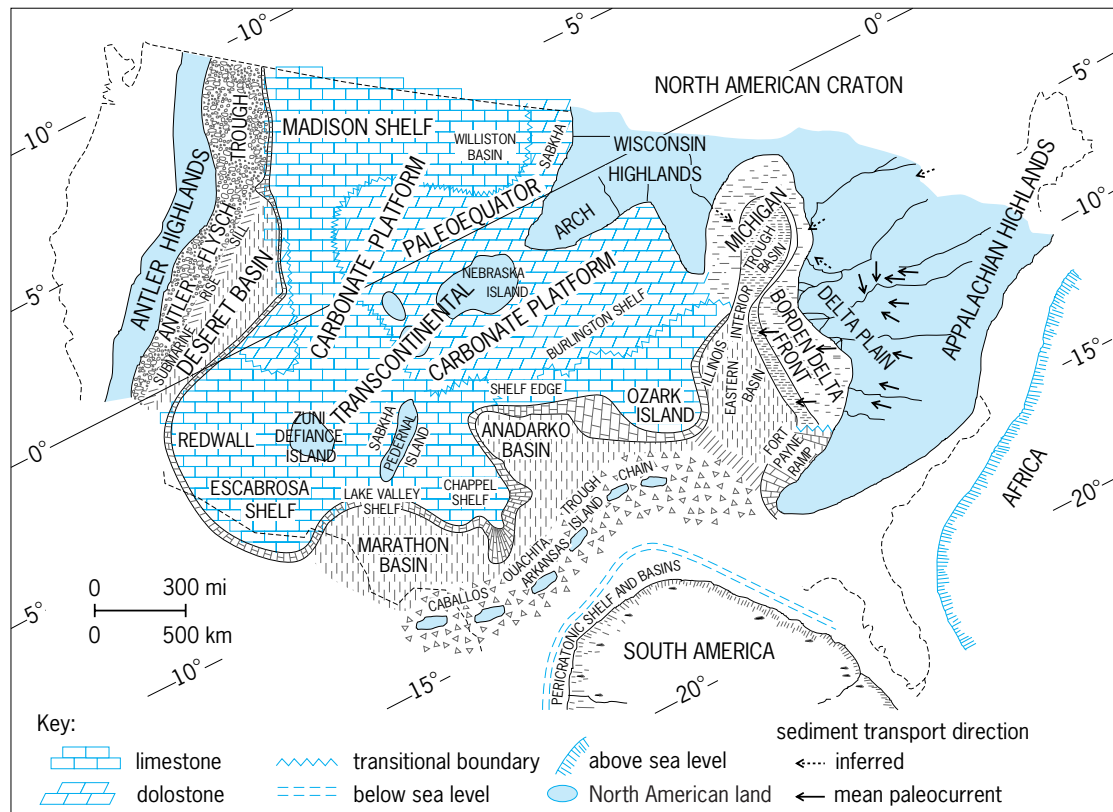


Fig. 2. Mid-Mississippian (middle Osagean) paleogeographic and lithofacies map of conterminous United States showing the widespread carbonate platform, continental spine (Transcontinental arch), bounding highlands and basins, and position of paleoequator. (After R. C. Gutschick and C. A. Sandberg, *Mississippian continental margins of the conterminous United States*, Soc. Econ. Paleont. Mineral. Spec. Publ., no. 33:79–96, 1983)

cherts are black, well bedded, and formed mainly of preserved and dissolved radiolarian tests. The cherts are commonly phosphatic and are interbedded with thicker beds of phosphatic shale and siltstone and thinner beds of oolitic and peloidal phosphorite. Thicker, lighter-colored siliceous rocks in the Caballos-Ouachita trough are termed novaculites.

Terrestrial Mississippian rocks, particularly in the Pocono and Antler clastic wedges, contain thin coal beds. These coal beds are economically insignificant in contrast to those in Pennsylvanian rocks. See COAL.

Paleoceanography. The North American craton was covered by shallow, warm, tropical epicontinental seas that had maximum depths of only about 60 m (200 ft) at the shelf edge. A chain of islands occupied the Transcontinental arch between the Madison Sea on the west and the Eastern Interior Sea on the east (Fig. 1). The end of the arch plunged southwestward into the Redwall-Escabrosa Sea. Depths in the Antler trough far exceeded 500 m (1640 ft), whereas those in the Caballos-Ouachita and Eastern Interior troughs exceeded 300 m (1000 ft). See PALEOCEANOGRAPHY.

Paleotectonics. During the Mississippian, the North America continent was subjected to three plate-tectonic events—the Acadian, Antler, and proto-Ouachita orogenies—that created orogenic highlands and inboard, deep foreland troughs. These structural features influenced sedimentation on its

present-day eastern, southern, and western margins. First, the European and African continental plates collided on the east beginning in Middle Devonian time, producing the Acadian orogeny. Next, the decoupled proto-Pacific oceanic plate collided on the west beginning in latest Devonian time, producing the Antler orogeny. At about the same time, the South American plate began moving toward North America, producing the proto-Ouachita orogeny. This movement created the Caballos-Arkansas island chain (Fig. 1) as a tectonic forebulge during the Mississippian, although actual continental collision did not take place until Pennsylvanian time. See OROGENY; PLATE TECTONICS; SEDIMENTOLOGY.

Sea-level changes. Mississippian North America was subjected to a number of sea-level rises, associated with transgressions, and sea-level falls, associated with regressions (Fig. 1). The two major rises, which were probably eustatic (referring to worldwide change of sea level), took place during the late Kinderhookian lower *crenulata* zone and at the start of the middle Osagean *ancboralis-latus* zone. These sea-level rises caused progradation or seaward migration of carbonate platforms as organisms produced buildups that maintained their niches relative to sea level. The rises also caused stratification of the water column in deeper basins, so that bottom conditions became deficient to lacking in oxygen. See ANOXIC ZONES; BASIN.

A major sea-level fall occurred at the start of

the middle Meramecian lower *Cavusgnathus* zone (Fig. 1). This fall caused sediments to bypass the carbonate platforms and to be deposited offshore as proximal deltaic deposits and distal prodeltaic deposits. Later, with increased sea-level fall, the carbonate platforms were subaerially exposed and subjected to karstification and cave formation. See CAVE; DELTA; KARST TOPOGRAPHY.

The Mississippian, which began with a moderate transgression that continued from latest Devonian time, ended with a moderate regression during the *muricatus* and *primus* zones.

Conodont biofacies and zonation. Conodonts are the phosphatic teeth of primitive fishes that had soft bodies, which rarely are preserved in rocks. These hard parts remained on the sea bottom after the remains of the conodont-bearing animal had decayed, and were generally subjected to postmortem sorting by currents. However, conodonts were only occasionally transported outside their original, broad paleotectonic setting. Different conodont faunas of the same age, which lived in various paleoenvironmental settings, are termed conodont biofacies. Contemporaneous dating of conodont biofacies is enabled by the few conodonts that lived at the fringes of, or were transported just outside, their preferred habitats.

Conodont zonation has become the major biochronologic tool for dating Mississippian rocks. Conodonts lived in most marine environments and many were distributed worldwide. Most genera and species were rapidly evolving and short-ranging, especially in warm tropical waters. Most other organisms evolved slowly, had restricted habitats, were not commonly preserved, or became disarticulated after death and hence difficult to identify.

The Late Devonian and Early Mississippian were the mid-Paleozoic heyday of conodont evolution, but the number and diversity of taxa began to decline gradually beginning with the extreme sea-level fall at the start of the Meramecian lower *Cavusgnathus* zone. The conodont zonation of the Kinderhookian and Osagean is considered a global standard against which other zonations are measured (Fig. 1). However, Meramecian and Chesterian conodont zones are less firmly established and represent longer time intervals. See CONODONT.

Life. Crinoids were probably the most abundant biota in Mississippian seas, but are only uncommonly used for correlation because most specific identifications require study of calyxes, which generally disarticulated after death. Crinoid and similar blastoid columnals are major constituents of Mississippian carbonate rocks. See CRINOIDEA.

Corals are probably the most widely preserved Mississippian megafossil group, and they are one of the most useful biochronologic tools for constructing biostratigraphic zones for carbonate-platform rocks. The earliest Mississippian is characterized mainly by solitary corals and tubular colonial corals known as *Syringopora*, which had survived the late Frasnian mass extinction. Other colonial corals began a gradual return late in the Kinderhookian. The Late Mississippian was a heyday for reef-building colonial corals

and large solitary corals. See CORALLINALES; REEF.

Other commonly occurring Mississippian megafossils are the brachiopods, mollusks, bryozoans, and ostracodes. Thick-valved brachiopods, such as spiriferids and productids, are found in abundance in carbonate-platform and upper-slope rocks; thin-valved brachiopods occur in basinal rocks. Among the mollusks, cephalopods, particularly the ammonoids, are most important for biochronologic dating and correlation. Gastropods (snails) and pelecypods (bivalves) occur in most biofacies, but taxa are long-ranging and not generally useful for correlation. Fenestrate (meshlike) and ramose (branched) bryozoans are found in shallower-water rocks. Ostracodes are ubiquitous in all rock types, and deeper-water species locally provide excellent biostratigraphic tools. See BRACHIOPODA; BRYOZOA; CEPHALOPODA; MOLLUSCA; OSTRACODA.

Small, phosphatic fish remains (ichthyoliths), including teeth, scales, bones, and dermal denticles, are found in almost all marine Mississippian rocks.

Radiolarians and calcareous and agglutinated foraminiferans are common marine Mississippian microfossils. Pelagic radiolarians are useful in dating deep-water rocks. Calcareous forams are important in dating most carbonate-platform and upper-slope rocks. Agglutinated forams (those that build their shells from detritus) provide a biofacies tool. Sponge spicules are common in basinal rocks. Worm jaws (scolecodonts) occur in some fine-grained rocks. See FORAMINIFERIDA.

Trace fossils (ichnofossils), including grazing traces, trails, burrows, and fecal pellets, are common in most rocks. Their occurrences are controlled by biofacies, and hence they are useful as depth indicators. Tracks of early amphibians are scarce. See TRACE FOSSILS.

Among marine plants, bacterial and algal mats were important in binding sea-bottom sediments, but their remains may not be preserved. However, calcareous algae, calcispheres, algal stromatolites, and trochiliscid charophytes occur commonly in carbonate-platform rocks.

Forests flourished during the Mississippian, and tree trunks, plant stems, roots, and spores occur commonly in terrestrial and peritidal rocks, particularly in coal beds. *Lepidodendron* trunks and *Stigmaria* roots are among the best-known plant remains. See GEOLOGIC TIME SCALE; PALEONTOLOGY; PALEOZOIC.

Charles A. Sandberg

Bibliography. E. Atherton, C. Collinson, and J. A. Lineback, Mississippian System: Handbook of Illinois Stratigraphy, *Ill. St. Geol. Surv. Bull.*, 95:123-163, 1975; L. C. Craig and C. W. Connor (coords.), Paleotectonic investigations of the Mississippian System in the United States, *USGS Prof. Pap.*, no. 1010, 1979; R. C. Gutschick and C. A. Sandberg, Mississippian continental margins of the conterminous United States, *Soc. Econ. Paleont. Mineral. Spec. Publ.*, 33:79-96, 1983; F. G. Poole and C. A. Sandberg, Mississippian paleogeography and conodont biostratigraphy of the western United States, *Pacific Section, Soc. Econ. Paleont. Mineral. Spec. Publ.*, 67:107-136, 1991.

Mistletoe

The name given to several species of the mistletoe family (Loranthaceae). The true mistletoe of Europe is *Viscum album*, and among the early nations this was an important ceremonial plant, which probably accounts for the origin of the custom of kissing under the mistletoe. In the United States, the common representative of the group is *Phoradendron flavescens* (see **illus.**). All of the mistletoes are green hemipar-



Mistletoe (*Phoradendron flavescens*).

asites; that is, they obtain water and minerals from the host plant but manufacture their own food. See SANTALALES. Perry D. Strausbaugh; Earl L. Core

Mitochondria

Specialized compartments (organelles) of all eukaryotic (nucleated) cells that use oxygen (**Fig. 1**). Often called the powerhouses of the cell, mitochondria are responsible for energy generation by the process of oxidative phosphorylation. In this process, electrons produced during the oxidation of simple organic compounds are passed along a chain of four membrane-bound enzymes (the electron transport or respiratory chain; **Fig. 2**), finally reacting with and reducing molecular oxygen to water. The movement of the electrons releases energy that is used to build a gradient of protons across the membrane in which the electron transport chain is situated. Like a stream of water that drives the turbines in a hydroelectric plant, these protons flow back through adenosine triphosphate (ATP) synthase, a membrane-bound enzyme that acts as a molecular turbine. Rotation of part of ATP synthase results in storage of energy in the form of ATP, the universal energy currency of the cell.

Besides their role in energy generation, mitochondria house numerous enzymes that carry out steps essential to metabolism. Defects in mitochondrial assembly or function generally have serious consequences for survival of the cell. In humans, mitochondrial dysfunction is the underlying cause of a wide range of degenerative diseases, with energy-demanding cells such as those of the central nervous

and endocrine systems, heart, muscle, and kidney being most severely affected.

Structure. Mitochondria are bounded by two concentric membranes referred to as the outer and the inner. This creates two distinct compartments, the matrix and the intermembrane space.

The outer membrane consists of a bilayer containing about 80% lipid. It is freely permeable to molecules smaller than about 5000 daltons, a feature that is attributable to the presence of many copies of a channel-forming protein called porin.

The inner membrane is also a lipid bilayer. It is extremely rich in protein (about 75%) and is impermeable to even the smallest of ions. This latter property may be due to the presence of a relatively high proportion of cardiolipin (10%) in the lipid bilayer. The inner membrane contains the enzymes of the electron transport chain and the ATP synthase, together with a set of transporter proteins that regulate the movement of metabolites in and out of the matrix space. Mitochondria of cells that depend on a high level of ATP production are usually extensively folded to produce structures called cristae (**Fig. 1**). These greatly increase the surface area of the inner membrane, allowing many more copies of the enzymes of oxidative phosphorylation. Cristae are generally believed to be baffle-shaped structures in the inner membrane that allow free communication of the intermembrane and intracristal spaces (**Fig. 1**). In reality, however, as shown by transmission electron microscopic tomography (**Colorplate 1**), cristae may be connected to the peripheral inner membrane by narrow tubular connections that may considerably restrict diffusion of molecules between intermembrane and intracristal spaces.

The intermembrane space contains enzymes

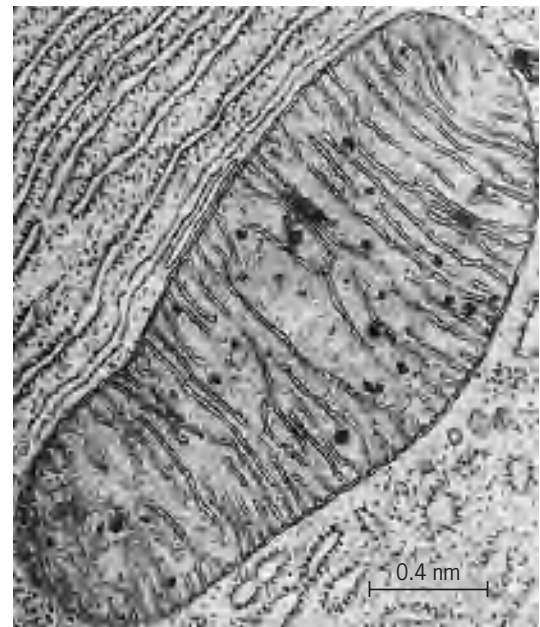


Fig. 1. Electron micrograph of a thin section through the pancreas of a bat, showing a typical mitochondrion in profile. Note how the cristae are formed by extensive folding of the inner membrane. (Courtesy of K. R. Porter)

capable of using some of the ATP that is transported out of the matrix to phosphorylate other nucleotides.

The matrix space is packed with a hundred or so water-soluble proteins that form a sort of semisolid gel. They include enzymes of the tricarboxylic acid (Krebs) cycle and enzymes required for the oxidation of pyruvate and fatty acids, comprising steps in the biosynthesis or degradation of amino acids, nucleotides, and steroids.

Dynamics. Textbooks often depict mitochondria as rod- or sausage-shaped structures with diameters of 0.5–1.0 micrometer and lengths of 1–10 μm , depending on cell type (Fig. 1). This convention, based largely on older electron micrographs, fails to bring out the highly dynamic character of the organelle. Studies that have employed confocal laser microscopy and time-lapse photography show mitochondria to be anything but static. Within minutes, they can change shape, grow in length, branch, divide, or fuse, often moving about in the cytoplasm as they do so. In many types of cell, mitochondria accumulate at sites of ATP utilization, for example between adjacent myofibrils in heart muscle cells, or wrapped around the flagellum of a sperm cell. In others, they may cluster about other organelles or structures, thereby facilitating the exchange of metabolic intermediates or ions (Colorplate 2). An association between mitochondria and microtubules has also frequently been observed. This may play a role in determining the way in which mitochondria are distributed in the cell and in which they segregate during cell division.

Energy conservation. Most cells derive their energy from the coupled processes of electron transport and oxidative phosphorylation in mitochondria. Energy production is fueled mainly by pyruvate, produced during glycolysis, or by fatty acids. Both compounds are selectively transported into the mitochondrial matrix where, after conversion to two-carbon units in the form of acetyl coenzyme A, they are fed into the citric acid cycle for further oxidative breakdown. Operation of the cycle converts the two-carbon units to CO_2 and H_2O . High-energy electrons extracted during this process are temporarily held by the coenzymes NADH and FADH_2 before being passed to molecular oxygen via the enzymes of the respiratory chain. The energy that is released as these electrons pass from one enzyme to the next is used to pump protons across the inner membrane from the matrix space into the intermembrane space (Fig. 2).

The pumping of protons creates an electrochemical gradient across the inner membrane. This gradient represents a form of stored energy, which can be used to perform different kinds of work. Commonly, as mentioned above, backflow of the protons down the gradient is used to drive synthesis of ATP through the action of the enzyme ATP synthase. Other energy-requiring processes can also be driven, including the transport of metabolites or ions against their concentration gradients by coupling their movement to the backflow of protons.

The overall process of oxidative phosphorylation

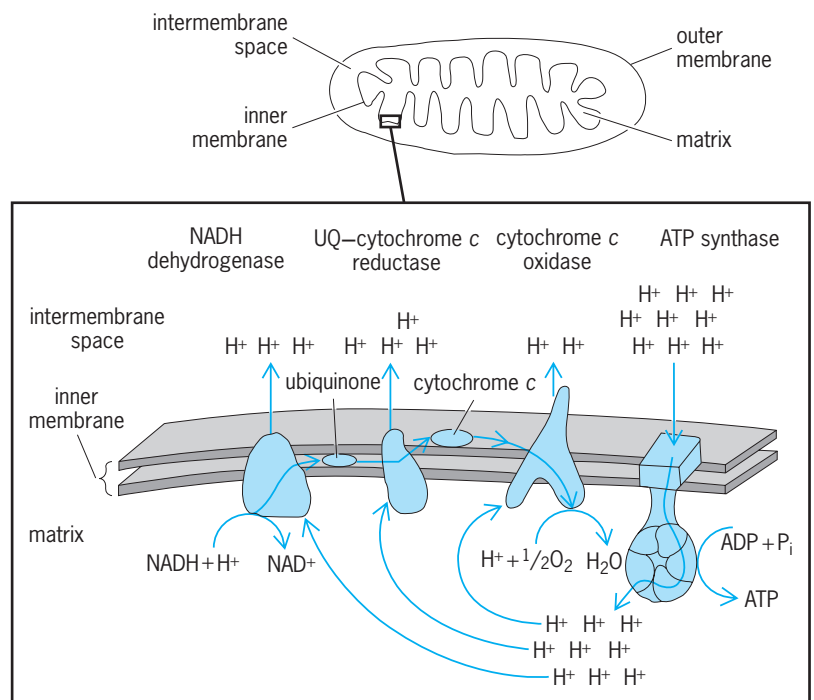


Fig. 2. Mitochondrial inner membrane, showing the pathway of electrons through the electron transport chain and the use of a proton gradient to drive ATP synthesis. Electrons pass from one enzyme to another; each enzyme is one of a series of carriers that are embedded in the membrane. Each carrier in the chain has a progressively higher affinity for electrons, so that these tumble onward, losing energy. Final transfer is to oxygen, which has the highest affinity of all. At two points in the chain, mobile carriers, ubiquinol and cytochrome c, mediate the passage of the electrons. Energy released during electron transfer is used by the enzymes of the chain to pump protons from the matrix into the intermembrane space. (After B. Alberts et al., *Molecular Biology of the Cell*, Garland, 1983)

by mitochondria is remarkably efficient, with about 80% of the energy theoretically available from oxidation of carbohydrates and fats being stored in the form of ATP. Compare this with the 10–20% efficiency of manufactured energy-conversion devices such as the internal combustion engine or electric motor. Lower efficiencies are sometimes found, however, and the mitochondria of brown fat cells are an interesting example. The inner membrane of these mitochondria contains a protein that allows protons to leak down the electrochemical gradient into the matrix. The energy stored by oxidative phosphorylation is thus wasted in the form of heat, but for a reason—such cells are usually found in newborn infants and hibernating animals, where they produce warmth necessary for survival. See ADIPOSE TISSUE; CITRIC ACID CYCLE.

Electron transport chain. Understanding the mechanisms of electron transport and oxidative phosphorylation has been increased dramatically by the recent elucidation of the three-dimensional (3D) structures of three of the enzymes involved (b_c_1 complex, cytochrome c oxidase, and ATP synthase). This was a technical tour de force. First, analysis of the 3D structure required prior crystallization of the enzymes, which was no easy task in this case because of their tendency to aggregate and become insoluble after their release from the membrane. Second, the complexity of the structures required that the analytical techniques used be pushed to their limits.

Cytochrome bc_1 complex. Cytochrome bc_1 (ubiquinol cytochrome c oxidoreductase; complex III) of mammalian mitochondria contains 11 subunits, but only 3 of these carry the redox centers necessary for electron transport (**Colorplate 3**). These are cytochrome b , the so-called Rieske iron-sulfur protein, and cytochrome c_1 . Both the iron-sulfur protein and cytochrome c_1 are anchored in the inner membrane by their C-terminal tails and possess headpieces that extend into the intermembrane space. The headpiece of the iron-sulfur protein carries an iron-sulfur (Fe_2S_2) cluster, while that of cytochrome c_1 contains a c -type heme group. Cytochrome b [encoded by mitochondrial (mtDNA)] contains eight helical regions, which are all contained within the lipid bilayer of the inner membrane. It also contains two heme groups, each of which forms an active site in electron transport. Cytochrome b delivers electrons to the iron-sulfur protein, whose headpiece rotates through about 60° to make contact with that of cytochrome c_1 . A key component of the bc_1 complex is not a protein at all, but a low-molecular-weight lipid-soluble and highly mobile electron carrier called ubiquinol. In a complicated mechanism that has been designated the Q-cycle, electrons are donated to cytochrome b one at a time. Operation of the Q-cycle drives protons across the inner membrane, providing part of the proton-motive force that is later used to synthesize ATP. As occasional by-products, operation of the Q-cycle also produces highly reactive semiquinol intermediates that can react with molecular oxygen to form various type of reactive oxygen species (ROS). These can cause extensive damage to proteins, lipids, and deoxyribonu-

cleic acid (DNA) and as such are probably responsible for many of the pathological developments in cells that contain dysfunctional respiratory chains.

Cytochrome c oxidase. The second respiratory enzyme for which detailed 3D structural information is available is cytochrome c oxidase (complex IV; **Colorplate 4**). Like the bc_1 complex, cytochrome c oxidase acts both as an electron carrier and as a proton pump. In mammalian cells, it consists of 13 subunits, of which only 3 (Cox1, 2, and 3) have clearly defined roles in enzyme activity. These 3 subunits are all encoded by mtDNA. The active site of the enzyme is situated in Cox1, which is buried in the inner membrane, spanning it several times with a series of alpha helices. The substrate for the enzyme is cytochrome c , a water-soluble protein with a c -type heme group. The electrons it donates are used by the enzyme to reduce molecular oxygen to two water molecules. Cytochrome c is loosely attached to the outer surface of the inner membrane and shuttles electrons between cytochrome c_1 and cytochrome c oxidase. Electrons received from reduced cytochrome c are picked up by Cox2, which contains a dinuclear, mixed valence copper center (Cu_A). Interaction between the two proteins is possible because Cox2, like cytochrome c_1 , is anchored in the inner membrane only by its C-terminal part. The remainder of the protein forms a flexible headpiece that extends into the intermembrane space. The electrons are then passed to the subunit I, the functional core of the enzyme, first to cytochrome a (a low-spin heme) and then to the bimetallic cytochrome aa_3/Cu_B site. One of the Cu_B ligands is a histidine residue which forms an adduct with a neighboring tyrosine. It has been suggested that free-radical generation by this adduct may play a role in the reduction of molecular oxygen.

Each atom of oxygen is reduced to a molecule of water by acceptance of two electrons and two protons. At the same time, two protons are translocated across the enzyme from the matrix to the intermembrane space. Both protons that are consumed in the reaction, and those that are pumped reach the active center of the enzyme through hydrophilic crevices in the enzyme, possibly by passing along chains of water molecules. Beyond the active site, the mechanism of proton movement is less clear, since this region of the enzyme is strongly hydrophobic.

ATP synthase. Energy stored in the form of a proton gradient across the mitochondrial inner membrane can be converted to ATP by the action of the turbine-like enzyme F_1F_0 -ATP synthase (complex V). The enzyme consists of two parts. F_0 is the proton channel. It is embedded in the membrane. F_1 is the catalytic part. It protrudes out of the F_0 complex into the matrix space and consists of five different subunits (α , β , γ , δ , and ϵ) in a ratio of 3:3:1:1:1. Examination of the crystal structure of mitochondrial F_1 (**Fig. 3**) reveals that synthesis of ATP is achieved by rotation of a central stalk consisting of γ and ϵ subunits inside a globular head of α and β subunits. Structural studies on both the mitochondrial and *Escherichia coli* ATP synthases suggest that the head is fixed relative to

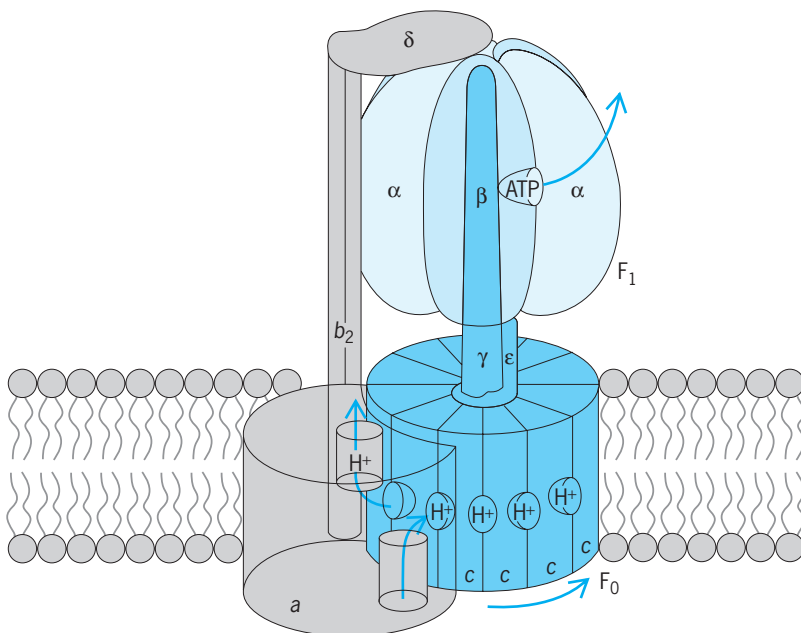


Fig. 3. Three-dimensional structural model of the mitochondrial ATP synthase based in part on x-ray crystallography of bovine F_1 ATPase and of the *Escherichia coli* F_0 . In the latter, the a , b , and c subunits correspond to the mitochondrial Atp6, b , and Atp9 subunits. The bovine F_1F_0 ATP synthase probably consists of some 16 different proteins and has a molecular weight of greater than 500 kDa. (From P. D. Boyer, *What makes ATP synthase spin?*, *Nature*, 402: 247-248, 1999)

the membrane by the stator. This is a complex of a membrane-embedded subunit (a subunit in *E. coli*), together with the β and δ subunits that link to F_0 . Gated movement of protons through a complex in F_0 produces structural changes which cause the parts to rotate (rather like the blades in a real turbine). This rotation induces the rotation of the central γ - ϵ stalk, which in turn induces conformational changes in the α - β head, thereby driving the conversion of bound adenosine diphosphate (ADP) and inorganic phosphate to ATP and its subsequent release. Elegant studies in which the torque of isolated F_1 has been measured during rotation caused by reversal of the normal enzyme reaction (that is, hydrolysis of ATP) suggest that the ATP synthase operates with almost 100% efficiency.

Mitochondrial genetic systems. Both mitochondria and chloroplasts contain DNA and the machinery necessary to express the information stored there. In both cases, the DNAs are relatively small and simple compared with DNA in the nucleus. While chloroplast DNA tends to be very similar in size in all organisms examined, mtDNA varies widely in complexity, from 16–18 kilobases in metazoa to upward of 2000 kilobases in some higher plants. Much of this variation can be explained by differences in organization: compact in metazoa; expanded in the other organisms, with large intergenic regions and introns in many genes. In the case of higher-plant mtDNAs, some extra sequences appear to have been picked up from chloroplast and nuclear DNAs. MtDNA is generally circular, although some linear exceptions are known among the yeasts, algae, and protozoa.

The information content of most mtDNA is limited. This means that most of the several hundred proteins found in these organelles are encoded by genes located in nuclear DNA. These proteins are synthesized in the cytosol and subsequently transported specifically to the respective organelles.

The contributions of the two genetic systems are usually closely coordinated, so that cells synthesize organelles of more or less constant composition. However, isolated mitochondria and chloroplasts are capable of synthesizing both RNA and protein *in vitro* and *in vivo*; both are capable of continuing for a short time with synthesis even after nuclear RNA synthesis or after cytosolic protein synthesis has been blocked.

Both types of activity have provided a means of identifying the products made by each genetic system. Thus the compound cycloheximide inhibits protein synthesis on cytosolic ribosomes but does not affect protein synthesis by either mitochondria or chloroplasts. In contrast, antibiotics such as chloramphenicol, tetracycline, and the macrolides inhibit organellar protein synthesis, but not cytosolic. This inhibition underlies the toxic effects often observed during long-term administration of these antibiotics to patients.

With some minor variations, all mtDNAs contain genes for proteins belonging to enzymes required for electron transport and oxidative phosphorylation. In humans, they include 7 subunits of the NADH dehydrogenase, 3 subunits of cytochrome *c* oxidase,

1 subunit of the cytochrome *bc*₁ complex (ubiquinol cytochrome *c* reductase), and 2 subunits of the ATP synthase. Genes for distinctive, organellar ribosomal (rRNAs) and transfer RNAs (tRNAs) are also present. Curiously, while some mtDNAs encode a complete set of the latter, others (including plants and trypanosomes) do not, and prevailing evidence suggests that the missing tRNA species have to be specifically imported from the cytoplasm.

The first mitochondrial genome to be completely sequenced was human mtDNA. With a length of 16,569 base pairs, this is one of the most compact of all mtDNAs, and it displays a number of unusual features:

1. Nearly every nucleotide appears to be part of a coding sequence; there is little or no room for regulatory sequences.

2. The genome contains only one major promoter region. Transcripts initiated within one part of this region extend over the full length of both strands (symmetric transcription). Others, initiated a short distance upstream from the first site, display a high tendency to terminate immediately downstream of the genes encoding the two rRNAs. This process of attenuation ensures a higher rate of synthesis of rRNAs relative to other transcripts.

3. Individual rRNAs, tRNAs, and mRNAs are generated by processing, signaled by tRNA sequences which are interspersed between the other genes.

4. DNA sequences of most protein-coding genes lack complete translational stop codons. Full-stop codons are generated by addition of a poly (A) tail.

5. The genetic code differs from the universal genetic code in three important respects: UGA (stop) encodes Trp, AUA (Ile) encodes Met, and AGA/AGG (Arg) encodes stop.

6. Only 22 genes for tRNAs are present. This number is sufficient for mitochondrial protein synthesis, but only because restrictions on codon-anticodon recognition operative in other genetic systems appear to be relaxed.

Inheritance. Many organisms, including humans, show uniparental inheritance of mitochondrial genes because one parent contributes more cytoplasm to the zygote than the other. In humans, it is the egg cell provided by the mother that contributes the cytoplasm. Human mitochondrial genes are thus inherited maternally.

Protein import. Only a small fraction of the total number of mitochondrial proteins is synthesized within the organelle itself. The remainder are specified by nuclear genes and are imported into the growing mitochondrion after their synthesis in the cytoplasm. These proteins are directed to the organelle by specific targeting sequences contained within them. In most cases, these sequences are located at the N-terminus of the newly synthesized protein and are removed after import by a specific protease. A smaller group of imported proteins contains noncleaved targeting sequences that are less well characterized. For both groups, import is initiated by interaction of precursor proteins with receptor complexes located in the outer membrane. Subsequent

steps include transfer to specific translocation complexes of the inner membranes at sites at which inner and outer membranes make transient contact, sorting to one of the intramitochondrial compartments and folding to a native conformation. Both unfolding and refolding of import proteins are carried out by a special class of proteins known as chaperones. Import of all proteins requires ATP and, with the exception of outer membrane proteins, is dependent on an electric potential across the inner membrane.

Diseases. Recent years have seen a growing interest in human diseases that result from mitochondrial dysfunction. A number of these result from mutations in mtDNA. Others are linked to nuclear genes, whose mutation disturbs oxidative phosphorylation or impairs mitochondrial assembly. Mutations in mtDNA are remarkably frequent and lead to a wide range of degenerative, mainly neuromuscular diseases. Most of these diseases are maternally inherited, but some appear to be spontaneous, possibly resulting from error-prone replication of mtDNA. One of the first mitochondrial diseases to be characterized was Leber's hereditary optic neuropathy (LHON), a maternally inherited form of sudden onset of blindness caused by death of the optic nerve. LHON patients possess a missense mutation in one of the mtDNA-encoded subunits of the NADH dehydrogenase (often ND1, ND4, or ND6). Single-base changes in mtDNA-encoded tRNA genes are responsible for the syndromes known as myoclonic epilepsy and ragged-red fibers (MERRF) and mitochondrial myopathy, encephalopathy, lactic acidosis and stroke-like (MELAS) episodes. In other myopathies, which include chronic external ophthalmoplegia (CEOP), Kearns-Sayres syndrome (KSS), and Pearson marrow/pancreas syndrome, large-scale deletions of segments of mtDNA are responsible for the disease. The sites of such deletions are often bounded by short repeated sequences in mtDNA, and they may arise as a result of the process of slippage-mispairing during mtDNA replication.

A striking feature of mtDNA-related diseases is the enormous diversity in clinical presentation. This diversity is attributable to two main factors: (1) Heterogeneity in the mtDNA population. Most human tissues contain many thousands of mtDNA molecules per cell. The severity of clinical symptoms depends on the number of mutated molecules present. (2) Dependence of a particular cell type on mitochondrial function (mainly ATP production). Cells with a high requirement for mitochondrially generated ATP (central nervous system, endocrine system, heart, muscle, kidney) are more severely affected than cells with alternative sources of ATP (such as liver).

Apoptosis. Besides their role in metabolism and energy-linked processes, mitochondria have recently been identified as important players in the initiation of apoptosis (programmed cell death). On one hand, the mitochondrial outer membrane houses a number of members of the Bcl-2 family of apoptosis regulatory proteins. On the other hand, release of certain mitochondrial proteins from the intermembrane space is instrumental in activating specialized

proteases called caspases. These catalyze a degradative cascade in the cytoplasm that eventually ends in cell death. Cytochrome *c* is just such a caspase-activating protein. Details of how its release is triggered are lacking, but a plausible chain of events includes opening of pores in the outer membrane mitochondrial permeability transition pores, swelling of the inner membrane, and subsequent collapse of the electric potential across the inner membrane. Activation of the mitochondrial permeability transition pores can be induced by excessive uptake of calcium (Ca^{2+}), exposure to reactive oxygen species, or decline in energetic capacity. Since decline in energetic capacity and increase in oxidative stress are common consequences of many mitochondrial diseases, pore opening and consequent triggering of apoptosis is likely to contribute to the pathophysiology of the disease.

Evolution. According to the endosymbiont hypothesis originally formulated by L. Margulis in 1968, mitochondria are proposed to have originated from an aerobic bacterium at a time of increasing oxygen levels in a world inhabited by proto-eukaryotes, the ancestors of present-day eukaryotic cells. These primitive cells survived by engulfing and digesting other organisms. The bacterium is likely to have been an α -proteobacterium similar to present-day *Paracoccus*. Having escaped digestion after engulfment, it is proposed to have developed a stable, symbiotic relationship with its host by providing respiration-derived ATP in exchange for metabolizable substrates and physical protection. As time passed, its genome was drastically reduced by loss of genetic information needed only for independent growth and by transfer of many genes needed for respiration to the host nucleus. The result of this process of reductive evolution is the modern mitochondrion, its DNA encoding a handful of genes that failed to be transferred before divergence of the organellar and nucleocytoplasmic genetic systems made further genetic exchange impossible.

The endosymbiont hypothesis has up to now been widely accepted. It received strong support from molecular sequence data published in the period 1975–1995. These showed that both mitochondrial genes and nuclear genes encoding mitochondrial proteins are indeed of α -proteobacterial origin, while many other genes in the eukaryotic nucleus turned out to be of archaeobacterial origin.

However, subsequently published sequences, in many cases derived from genomic sequence projects, now reveal that nuclear genomes carry many genes of bacterial origin that have nothing to do with mitochondrial functions. Additionally, primitive amitochondrial eukaryotes, such as the microsporidia, which are considered to be the direct descendants of the ancient proto-eukaryotes, also carry bacterial genes in their nuclear genomes. Plausible alternatives, but still controversial explanations, for both features are offered by what have been termed the hydrogen hypothesis (put forward by W. Martin and M. Müller in early 1998) and the syntrophy and oxygen toxicity hypothesis (both put

forward by S. Andersson and C. Kurland in 1999). Like the endosymbiont hypothesis of Margulis, each of these models proposes that mitochondria indeed have their origins in symbiosis between an α -proteobacterium and a member of the Archaea, and that mitochondrial genome evolution has involved both gene loss and transfer. They each differ, however, in the suggested identity of the driving force behind symbiosis. The endosymbiosis hypothesis proposes that this was respiration-generated ATP; the hydrogen hypothesis proposes the production of hydrogen; the syntrophy hypothesis proposes the utilization of methane; the oxygen toxicity (ox-tox) hypothesis proposes the local removal of oxygen from the intracellular environment of the primitive anaerobic host.

Mammalian mtDNA evolution. Mammalian mtDNAs accumulate mutations at high rates and evolve correspondingly fast (up to 12–15 times faster than single-copy genes in nuclear DNA and up to 100 times faster for rRNA and tRNA genes). This behavior reflects both a high incidence of mutations and a high probability of their fixation. The first is probably related to oxidative damage to mtDNA by oxygen free radicals produced as by-products of electron transfer through the respiratory chain. The second has been attributed to the lack of efficient DNA repair (mitochondria lack nucleotide-excision repair) and to a relatively high tolerance of many mitochondrial gene products to mutational change.

The rapid rate of sequence evolution of mammalian mtDNAs makes these genomes highly sensitive indicators of recent evolutionary relationships. Unlike their nuclear counterparts, mtDNAs do not undergo recombination during sexual transmission and are strictly maternally inherited. Sequence changes in mtDNA therefore provide a clear record of the history of the female lineages through which this DNA has been transmitted. In 1992 the geneticist Allan Wilson compared a short section of these sequences in 182 individuals. His results allowed the construction of an evolutionary tree that indicates that the common ancestor of all surviving human mtDNA types lived in Africa between 140,000 and 290,000 years ago. This ancestor is often misleadingly referred to as Mother Eve, the suggestion being that Eve corresponds to a single woman from whom all living humans have descended. This is incorrect. The root of the tree simply indicates the point in time at which the first mutation occurred in mtDNA. All other ancestral types simply failed to survive. Additionally, the vast majority of human genes are encoded by chromosomes in the nucleus. A large number of individuals, both men and women, are thus likely to have contributed to the genetic makeup of the individual whose mtDNA gave rise to the many variants extant today. *See* BIOLOGICAL OXIDATION; CELL (BIOLOGY); ENZYME.

Les A. Grivell

Bibliography. G. Attardi and A. Chomyn (eds.), *Mitochondrial Biogenesis and Genetics*, vol. 264 of *Methods in Enzymology*, Academic Press, 1996; V. Darley-Usmar and A. H. V. Schapira (eds.), *Mitochondria, DNA, Proteins and Disease*, Portland Press

Res. Monog. Ser., no. 5, 1994; I. E. Scheffler, *Mitochondria*, Wiley, 1999; A. Tzagoloff, *Mitochondria*, Plenum, 1982.

Mitosis

The series of visible changes that occur in the nucleus and chromosomes of non-gamete-producing plant and animal cells as they divide. During mitosis, the replicated genes, packaged within the nucleus as chromosomes, are precisely distributed into two genetically identical daughter nuclei (**Fig. 1**). The series of events that prepares the cell for mitosis is known as the cell cycle. When viewed in the context of the cell cycle, the definition of mitosis is often expanded to include cytokinesis, the process by which the cell cytoplasm is partitioned during cell division. Although the continuity of heredity by cell division was first predicted by R. Virchow in 1858, the central role of mitosis was not fully understood until the mid-1900s, when the deoxyribonucleic acid (DNA) within chromosomes was proven to contain the hereditary blueprints for life. *See* CELL CYCLE; DEOXYRIBONUCLEIC ACID (DNA).

Mitotic spindle. Chromosome segregation is mediated in all nonbacterial cells (that is, eukaryotes) by the transient formation of a complex structure known as the mitotic spindle (**Fig. 2**). During mitosis in most higher plants and animals, the nuclear membrane surrounding the replicated chromosomes breaks down, and the spindle is formed in the region previously occupied by the nucleus (open mitosis). In lower organisms, including some protozoa and fungi, the spindle is formed and functions entirely within the nucleus which remains intact throughout the process (closed mitosis).

All spindles are bipolar structures, having two ends or poles. In animal cells, each spindle pole contains an organelle, the centrosome (**Fig. 3**), onto which the spindle focuses. As a result, the spindle in animal cells looks like a football. The polar regions of plant spindles lack centrosomes and, as a result, are much broader. In animals the bipolar nature of the spindle is established by the separation of the centrosomes, which is critical for a successful mitosis; the presence of only one pole produces a monopolar spindle in which chromosome segregation is inhibited. The presence of more than two poles produces multipolar spindles which distribute the chromosomes unequally among three or more nuclei. Centrosomes are duplicated during interphase near the time that the DNA is replicated, but then act as a single functional unit until the onset of mitosis. In plants, and during meiosis in some animals, the two spindle poles are organized by the chromosomes and by molecular motors that order randomly nucleated microtubules into parallel bundles. *See* CENTROSOME; PLANT CELL.

Microtubules. Microtubules are the primary structural components of the mitotic spindle and are required for chromosome motion (**Fig. 3**). These 25-nanometer-diameter, hollow, tubelike structures are

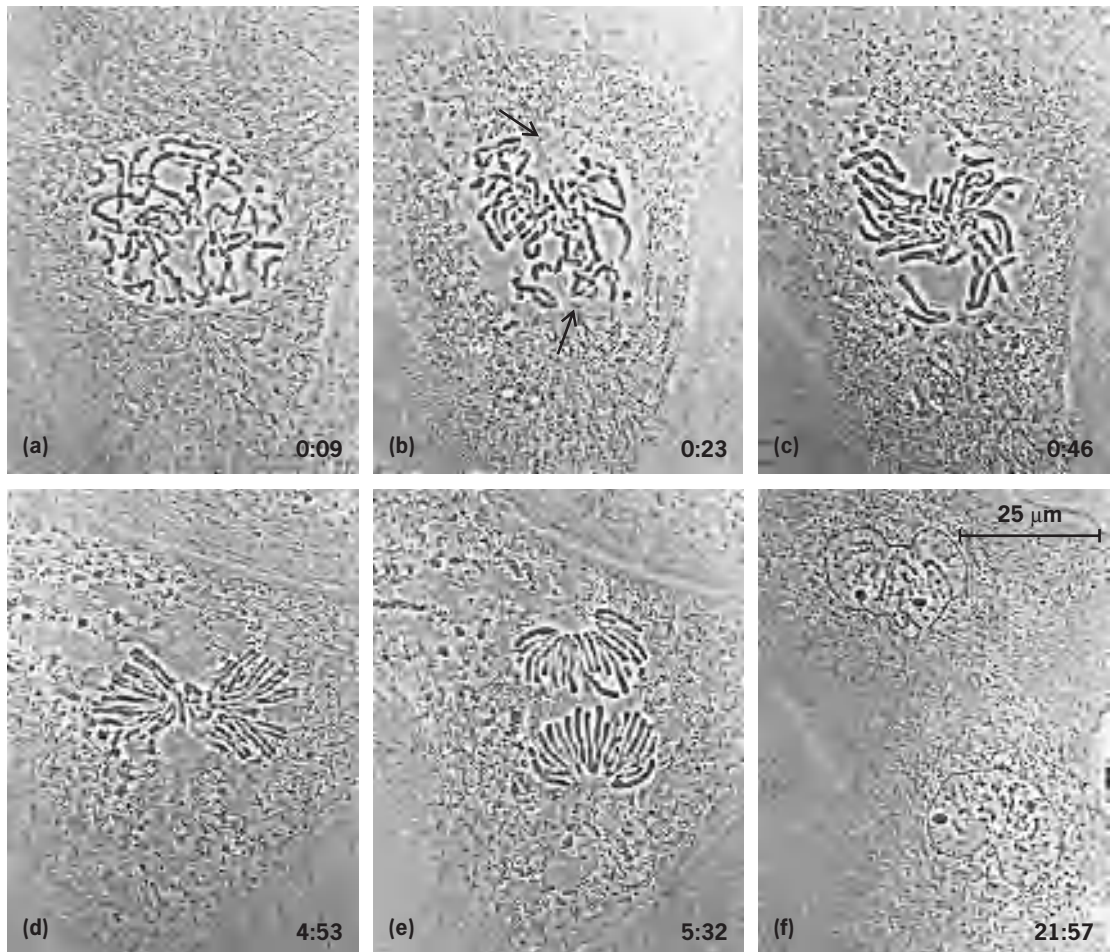


Fig. 1. Selected phase-contrast light micrographs showing changes in chromosome position during mitosis in a living newt lung epithelial cell (elapsed time in hours and minutes). (a) Late prophase. (b) Prometaphase. (c) Mid-prometaphase. (d) Metaphase. (e) Anaphase. (f) Telophase.

formed from the polymerization of protein subunit dimers composed of alpha and beta tubulin. Microtubules are polarized structures that grow much faster at one end than at the other; the fast-growing end is referred to as the positive (+) end, the slow-growing end as the negative (−) end. During interphase, microtubules are distributed throughout the cytoplasm, where they serve to maintain cell shape and also function as polarized roadways for transporting organelles and cell products. As the cell enters mitosis, the cytoplasmic microtubule network is disassembled and replaced by the mitotic spindle. The microtubules in animal cells originate from the centrosome that, like the chromosomes, was inherited during the previous mitosis where it functioned as a spindle pole. The centrosome contains a unique type of microtubule protein (gamma tubulin) that is involved in seeding microtubule assembly. Once growth is initiated, the positive ends of centrosomal microtubules grow away from the centrosome, while the negative ends either remain associated with the centrosome or are shed from it. At any one time, many microtubules within the cell are growing or shrinking as subunits are added or removed from their positive ends. Changes in the parameters that

modulate this dynamically unstable behavior of microtubule ends lead to changes in the numbers and average length of microtubules.

The motion associated with microtubules is mediated by several families of molecular motors which bind to and move along the wall of the microtubule. These motors include kinesins that move cargo along microtubules toward their positive ends away from the centrosome, and cytoplasmic dynein (and also some members of the kinesin family) that transports cargo along individual microtubules toward their centrosomal or negative ends. *See* CYTOSKELETON.

Kinetochores. As mitosis begins, each replicated chromosome consists of two identical sister chromatids that are joined along their length. In most cells, chromosomes possess a unique region of highly condensed chromatin (DNA plus protein), known as the centromere, which forms an obvious constriction on the chromosome, referred to as the primary constriction. Spindle microtubules attach to a small specialized structure on the surface of the centromere known as the kinetochore (**Fig. 4**). Fragments of chromosomes lacking a kinetochore do not move poleward; it is always the kinetochore that leads in the poleward motion of the chromosome.

Kinetochores in most plant cells are shaped like a ball, whereas those in animal cells resemble a plate. The centromere region of each replicated chromosome contains two sister kinetochores, one attached to each chromatid, that lie on opposite sides of the primary constriction.

Regulation of mitosis. Cells that are in the cell cycle contain an inactive kinase, known as $p34^{cdc2}$. Once chromosome and centrosome replication is completed at the end of S phase, two cyclin proteins (A and B) are synthesized that rapidly bind to and activate $p34^{cdc2}$. This activated enzyme is called cyclin-dependent kinase 1 (CDK1), formerly known as mitosis or maturation promoting factor (MPF). CDK1 then drives the cell into mitosis by phosphorylating specific proteins involved in the interphase/mitosis transition. In vertebrates, cyclin A is synthesized before cyclin B; and once active CDK1/cyclin A accumulates in the nucleus, it initiates the early events of mitosis including chromosome condensation. The subsequent accumulation of active CDK1/cyclin B then commits the cell to mitosis, and the cell will stay in mitosis as long as CDK1/cyclin B activity remains high. See CELL CYCLE.

Stages. Once initiated, mitosis is a continuous process that, depending on the temperature and organism, requires several minutes (*Drosophila*, yeast) to many hours (newts) to complete. Traditionally it has been subdivided into five consecutive stages that are distinguished primarily by chromosome structure, position, and behavior. These stages are prophase, prometaphase, metaphase, anaphase, and telophase.

Prophase. The first visible sign of an impending mitosis occurs within the nucleus as the DNA-containing chromatin begins to condense into chromosomes. This condensation is due to the phosphorylation of chromatin-associated proteins, including histone H1 and H3, by CDK1/cyclin A and the aurora B kinase within the nucleus. Up to a point, this is a reversible process, and cells in early-to-mid prophase can be induced to decondense their chromosomes and return to interphase by a variety of treatments that stress the cell and/or damage the DNA. However, once active CDK1/cyclin B accumulates in the nucleus at late prophase, the cell becomes irreversibly committed to the division process.

By late prophase the individual chromosomes are well defined and the nucleoli begin to dissipate. Near this time the cytoplasmic network of microtubules breaks down, and is replaced in animal cells by two radial astral arrays of microtubules growing from the two spindle poles (centrosomes). When compared to interphase, the microtubules associated with mitotic centrosomes contain more but significantly shorter microtubules. The functional changes that occur in the centrosome as the cell enters mitosis are mediated by the activity of CDK1 and other kinases that modify both centrosome chemistry and the proteins that control microtubule dynamic instability.

As the asters form, they separate and move toward opposite sides of the nucleus. During this sep-

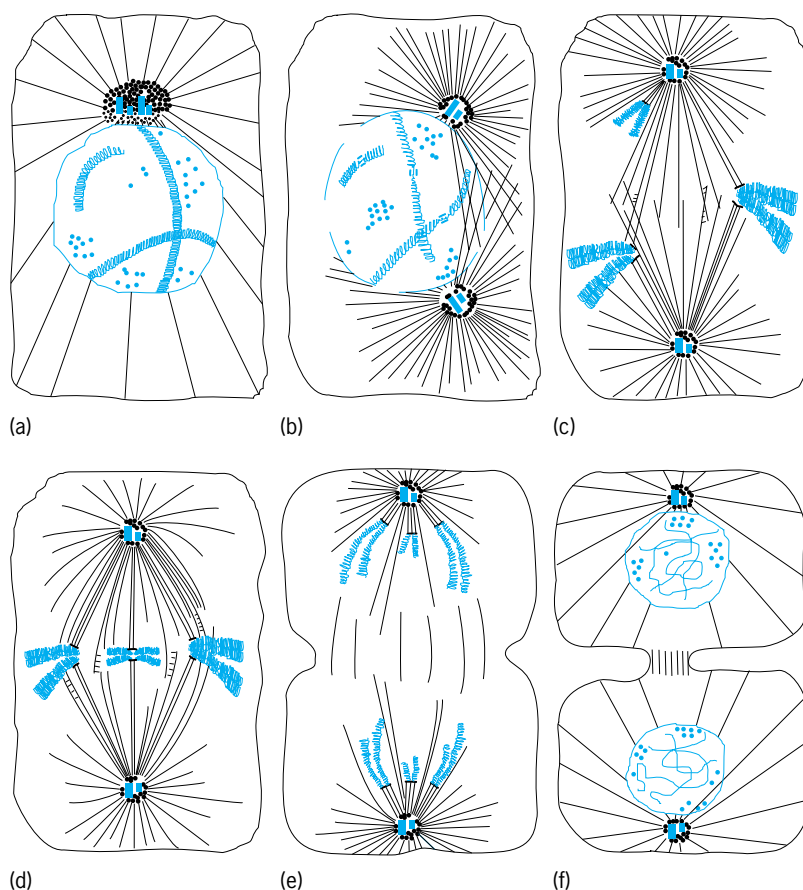


Fig. 2. Changes in microtubule distribution and chromosome position as an animal cell enters into and proceeds through mitosis. (a) Mid-prophase: chromosomes are condensing within the nucleus. (b) Late prophase/early prometaphase: centrosomes separate; the cytoplasmic microtubule complex is replaced by two radial astral microtubule arrays; microtubules in each aster grow and shorten at their ends away from the centrosome; the nuclear envelope breaks down; kinetochore-containing primary constrictions are sometimes visible on the well-condensed chromosomes. (c) Mid-prometaphase: the kinetochores and chromosomes interact with the asters to form the spindle. (d) Metaphase: all of the chromosomes are aligned on the spindle equator; sister kinetochores are attached to opposite poles by kinetochore fibers. (e) Anaphase: the sister chromatids separate and move toward their respective spindle poles (anaphase A); at the same time the spindle poles move farther apart (anaphase B). (f) Telophase: the two groups of sister chromosomes have become two well-separated sister nuclei; the cytoplasm of the cell divides (cytokinesis).

aration the microtubule arrays from each aster exhibit considerable overlap, and within this overlap the microtubules from opposing asters are of opposite polarity. In some lower organisms, including yeast and diatoms, the spindle poles are pushed apart by the action of microtubule plus end motors that bind microtubules of opposite polarity within the region of overlap, and then slide them away from one another. However, in higher animal cells, including vertebrates, the asters continue to move apart even when their microtubule arrays no longer overlap. Thus, it is likely that the separation of spindle poles in higher animals occurs via multiple mechanisms, including one that involves pulling.

Prometaphase. The sudden activation of CDK1/cyclin B during late prophase results in the rapid phosphorylation of various nuclear and cytoplasmic proteins. The most obvious consequence of this activity is that the nuclear envelope surrounding the chromosomes suddenly swells and dissolves, as many

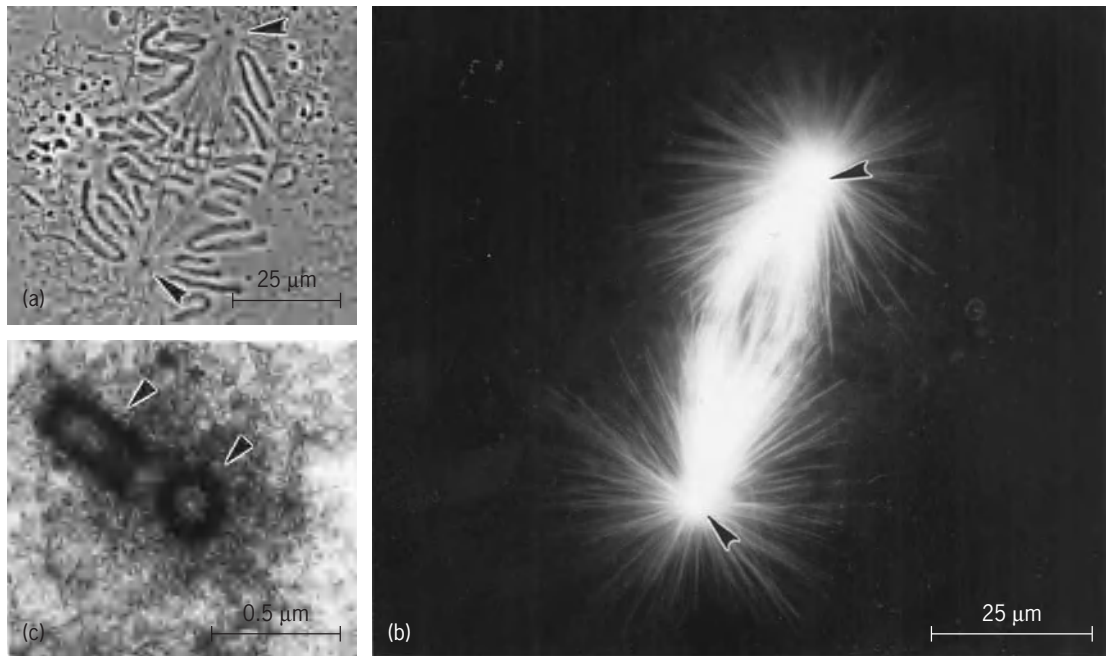


Fig. 3. Phase-contrast light micrograph of a newt lung cell in mid-prometaphase. (a) The two centrosomes or spindle poles (arrowheads) and some spindle fibers are visible, as are a number of mono-oriented chromosomes. (b) The same cell photographed by fluorescence microscopy. Most of the microtubules are concentrated within the spindle between the poles. (c) High-voltage electron micrograph of a thick section through two orthogonally arrayed centrioles (arrows) and surrounding dense material, cut from a mitotic centrosome depleted of its associated microtubule by cold treatment. (Parts a and b from L. Cassimeris, C. L. Rieder, and E. D. Salmon, *Microtubule assembly and kinetochore directional instability in vertebrate monopolar spindles: Implications for the mechanism of chromosome congression*, *J. Cell Sci.*, 107:285–297, 1994)

of its constituent proteins become phosphorylated and soluble. At the same time, CDK1 activity also induces other membrane systems within the cell, including the Golgi apparatus and the endoplasmic reticulum, to fragment into numerous small vesicles that then become randomly distributed.

The breakdown of the nuclear envelope initiates the prometaphase stage of mitosis. During this stage the chromosomes attach to the separating centrosomes as their kinetochores capture dynamically unstable astral microtubules that are randomly probing the cytoplasm. Once an astral microtubule is captured by a kinetochore, it becomes known as a kinetochore microtubule, and its positive end becomes firmly anchored within the kinetochore. A kinetochore begins to move as soon as it encounters its first microtubule. However, as a rule, kinetochores continue to capture as many astral microtubules as their surface area allows and, throughout the capture process, they induce these microtubules to bundle into a kinetochore fiber. In vertebrates, kinetochore fibers rarely form at the same time on sister kinetochores. Instead, the tendency is for a chromosome to attach first to the closest spindle pole as the kinetochore facing that pole acquires microtubules. Because only one of its kinetochores is attached to the spindle, such chromosomes are mono-oriented, and they rapidly move toward the pole to which they have become attached. Over time the opposing sister kinetochore on a mono-oriented chromosome, which lacks an attachment to microtubules, must become associated with the positive ends of microtubules generated from the opposing half-spindle (that is, the

far pole). This occurs by chance when it captures one or more microtubule ends growing from the far pole. To increase the chances of this happening in a reasonable time, the unattached kinetochore often becomes laterally associated with the surface of adjacent microtubules residing in the same half-spindle as the mono-oriented chromosome. Once this lateral association is established, microtubule positive-end centromere-associated protein E (CENP-E) motors on the surface of the unattached kinetochore transport it, and its associated chromosome, along the surface of the microtubules toward the spindle equator (that is, away from the closer pole). As the kinetochore approaches the equator, its chances of encountering a positive microtubule end growing from the opposing half-spindle dramatically increase. Once both sister kinetochores are attached to the spindle, so that each kinetochore is attached to different and opposing poles, the chromosome is considered to be bioriented. The failure of a chromosome to acquire this proper biorientation leads to the production of aneuploid cells with unequal numbers of chromosomes, with corresponding detrimental consequences to the organism.

Once a chromosome has achieved biorientation status, it undergoes a complex series of motions, termed congression, that ultimately position its centromere on the spindle equator midway between the poles. In order for congression to occur, the kinetochore fiber microtubules that terminate in the sister kinetochores must elongate and shorten in a coordinated fashion. For example, as a bioriented chromosome moves toward the equator, those kinetochore

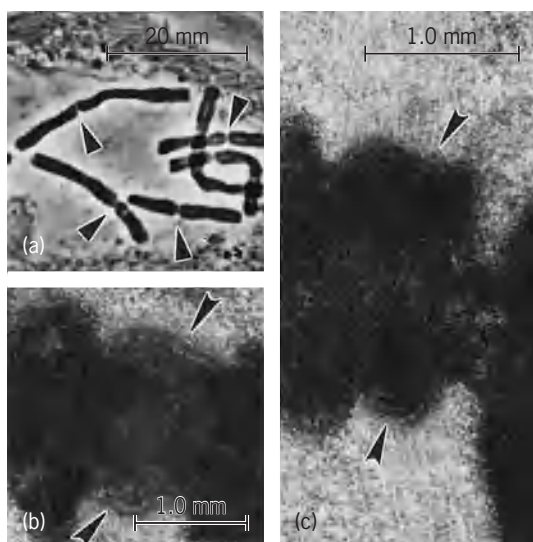


Fig. 4. Kinetochores. (a) Phase-contrast light micrograph of six unattached chromosomes in a living mitotic newt lung cell. The centromere-containing primary constrictions (arrowheads), and the replicated nature of the chromosomes, are clearly visible. (b) High-voltage electron micrograph of a thick section cut through the primary constriction of an unattached rat kangaroo chromosome. Two platelike sister kinetochores (arrowheads) are positioned on opposite sides of the primary constriction. (c) Same as b except that each sister kinetochore is attached to opposing spindle poles by a bundle of microtubules. (Part b from C. L. Rieder and G. G. Borisy, *The attachment of kinetochores to the pro-metaphase spindle in PtK₁ cells*, *Chromosoma*, 82:693–716, 1981)

microtubules that attach the chromosome to the proximal pole must elongate, while those that attach it to the distal pole must shorten. In vertebrates, the elongation and shortening of kinetochore fiber microtubules during chromosome motion occur primarily by the addition and deletion of tubulin subunits into microtubules at the kinetochore, but shortening also occurs to a lesser extent from subunit loss at the microtubule minus end associated with the pole.

Mono-oriented chromosomes exhibit constant-velocity, oscillatory motions toward and away from their associated spindle pole. The tendency of the only attached kinetochore on a mono-oriented chromosome to autonomously switch between poleward and away from the pole motion means that the motility of kinetochores is directionally unstable. When the unattached kinetochore of a mono-oriented chromosome attaches to the distal pole, it begins to move toward that pole. In order for this congression motion to continue, its sister kinetochore, attached to the closer pole, must switch into a prolonged movement away from the pole state of motion. The factors that coordinate the motile behavior of sister kinetochores, to allow for congression, remain unknown.

Metaphase. When the last chromosome becomes positioned on the spindle equator, the cell is considered to be in metaphase of mitosis. By this time, many of the astral microtubules have been sequestered into the spindle, which is now a fully ma-

ture structure, and cyclin A has been degraded. The metaphase spindle consists of two half-spindles, tethered to one another by the sister kinetochores on the chromosomes and by interactions between their associated overlapping (antipolar) microtubule arrays. Each half-spindle contains the same number of kinetochore fiber microtubule bundles interspersed among more numerous microtubules growing from the pole or free within the spindle.

Anaphase. The sudden and largely synchronous separation (disjunction) of sister chromatids initiates the anaphase stage of mitosis. This disjunction is not dependent on forces generated by the spindle as evidenced by the fact that it also occurs in some cells when spindle formation is inhibited by drugs. In all but embryonic cells, entry into anaphase is controlled by a cell-cycle checkpoint that delays chromatid disjunction until all of the kinetochores have achieved a stable attachment to the spindle. In these cells, unattached or weakly attached kinetochores trigger a signal transduction cascade that delays anaphase by inhibiting the activity of a macromolecular assembly known as the anaphase promoting complex (APC). These large, spindle-associated complexes selectively target several proteins involved in the metaphase/anaphase transition, including securin that maintains the cohesion between the sister chromatids and cyclin B, for proteolytic destruction by ubiquitinating them. This “kinetochore attachment” or “spindle assembly” checkpoint ensures that anaphase does not normally start until all of the chromosomes have achieved a bipolar orientation, which is the minimum requirement for equal chromosome segregation.

Once the chromatids disjoin in non-drug-treated cells, they slowly ($1\text{--}2\ \mu\text{m}/\text{min}$) move toward their respective poles in a process referred to as anaphase A. During this time the once highly organized spindle begins to disassemble. After anaphase A is under way, the poles themselves begin to move farther apart in a process known as anaphase B.

Telophase. The activation of APCs at the metaphase/anaphase transition not only leads to chromatid disjunction but also inactivates the mitotic (CDK1) kinase by destroying its associated cyclin B. As CDK1 activity falls, the various proteins that were phosphorylated during the initial stages of mitosis become dephosphorylated, and pathways are initiated that ultimately return the dividing cell to interphase. During this telophase stage of mitosis, the groups of separated sister chromosomes, which are now positioned near their respective poles, begin to swell and stick to one another. As this occurs, a nuclear envelope is deposited on the surface of the decondensing chromatin, and a Golgi apparatus reforms in association with each of the two centrosomes.

During telophase, new microtubule-based structures, known in animals as stem bodies and in plants as phragmoplasts, also form between the now-separated daughter nuclei. The phragmoplasts are responsible for directing the construction of a new cell wall that ultimately partitions the dividing plant cell into two separate (but connected) entities. In

animals the stem bodies act as a template and catalyst for forming and stabilizing the cytokinetic furrow, which constricts the cell into two independent daughters. At the end of telophase the centrosome inherited by each daughter cell nucleates another interphase complex of cytoplasmic microtubules. *See* CYTOKINESIS.

Chromosome movement. In some cells a kinetochore (and thus its associated chromosome) moves toward a spindle pole because the fiber to which it is attached is pulled toward the pole. This “traction fiber” mechanism accounts for 100% of chromosome poleward motion in some systems, including insect spermatocytes, various oocytes, and perhaps plant cells. It remains to be resolved whether the pulling force for this motion is produced along the entire length of the fiber by microtubule plus end motors anchored in the spindle matrix, or only at its end in the spindle pole. For poleward motion to occur, however, the kinetochore microtubules must shorten as they slide poleward, and this shortening clearly occurs by subunit loss at the microtubule minus ends at the pole. In contrast to a traction fiber mechanism, during poleward chromosome motion in vertebrate somatic cells only 25% of kinetochore fiber shortening can be attributed to subunit loss at the spindle pole. The other 75% occurs by subunit loss at the microtubule plus ends within the kinetochore. Immunological studies show that kinetochores in vertebrates contain cytoplasmic dynein, CENP-E, and other microtubule-based motor molecules, and that during prometaphase they can move rapidly along the surface of a microtubule toward a centrosome (spindle pole). It remains to be determined if the kinetochore-based force for poleward chromosome motion in these cells is produced by microtubule minus-end molecular motors associated with the kinetochore or simply from the disassembly of microtubule minus ends within the kinetochore. Regardless, in humans the kinetochore can be viewed as a complex biological machine that uses its associated kinetochore fiber microtubules to generate the force for chromosome poleward motion while allowing the plus ends of these microtubules to grow and shorten within its confines.

The mechanism which ultimately positions the sister kinetochores of a bioriented chromosome stably on the spindle equator, halfway between the two spindle poles, remains to be solved. For this to occur, the behavior of sister kinetochores must somehow be coordinated, so that when one is moving toward its associated pole, the other is moving away from its pole. At the moment the best guess for how this occurs is that kinetochores somehow “sense” their position and the spindle, and regulate their behavior accordingly. *See* CELL (BIOLOGY); CELL DIVISION; CELL NUCLEUS.

Conly L. Rieder

Bibliography. D. Mazia, Mitosis and the physiology of cell division, pp. 77–412 in *The Cell*, vol. III, ed. by J. Brachet and A. E. Mirsky, Academic Press, New York, 1961; C. L. Rieder and A. Khodjakov, Mitosis through the microscope: Advances in seeing inside live dividing cells, *Science*, 300:91–96, 2003;

S. Gadde and R. Heald, Mechanisms and molecules of the mitotic spindle, *Curr. Biol.*, 14:R797–R805, 2004; B. A. A. Weaver and D. W. Cleveland, Decoding the links between mitosis, cancer and chemotherapy: The mitotic checkpoint, adaptation, and cell death, *Cancer Cell*, 8:7–12, 2005.

Mitteniales

An order of the true mosses (subclass Bryidae) found in Australia and New Zealand. The order consists of a single species, *Mittenia plumula*, which is adapted for growth in caves or cavelike places. Branches of the persistent protonema consist of spherical cells which reflect light from a backing of chloroplasts, thus providing a glow. The stems are simple and erect, and the leaves are 2–4-ranked and oblong-lingulate and blunt or apiculate. The midrib ends above the midleaf, and the cells are short and smooth. The sporophytes are terminal, the setae elongate, and the capsules erect and oblong-cylindric. The opercula are conic-subulate. The peristome consists of 16 filiform teeth, inrolled when moist and recurved when dry, and 32 smooth, hyaline, nodulose segments attached to a short membrane. The calyptrae are mitrate and smooth.

The order is remarkably similar to the Schistostegales of North Temperate distribution in protonema and habitat, but the other features of gametophyte and sporophyte are quite different. A distinctive feature is the double peristome with 32 segments derived in part from the outermost cell walls of the endothecium (that is, the inner portion of the embryonic capsule). All other mosses of the subclass Bryidae have peristomes entirely of amphithecial origin. *See* BRYIDAE; BRYOPHYTA; BRYOPSIDA; SCHISTOSTEGALES.

Howard Crum

Mixer

A device with two or more signal inputs and one common output. The two primary classes are linear (additive) and nonlinear (multiplicative) mixers. Linear mixers are used to add or blend together two or more signals, nonlinear mixers mainly to shift the spectrum (center frequency) of one signal by the frequency of a second signal.

Linear mixer. Linear mixing is the process of combining signals additively, such as the summing of audio signals in a recording studio. This operation can be accomplished passively by simply using a resistive summing network (**Fig. 1a**). Although this approach appears very economical, there is a loss in signal strength and an interaction of the signal amplitudes as the gains are adjusted. These problems are demonstrated by the equation below for the output signal, e_o . The numerator contains the proper linear

$$e_o = \frac{R_L \left(\frac{e_1}{R_1} + \frac{e_2}{R_2} + \cdots + \frac{e_N}{R_N} \right)}{1 + R_L \left(\frac{1}{R_1} + \frac{1}{R_2} + \cdots + \frac{1}{R_N} \right)}$$

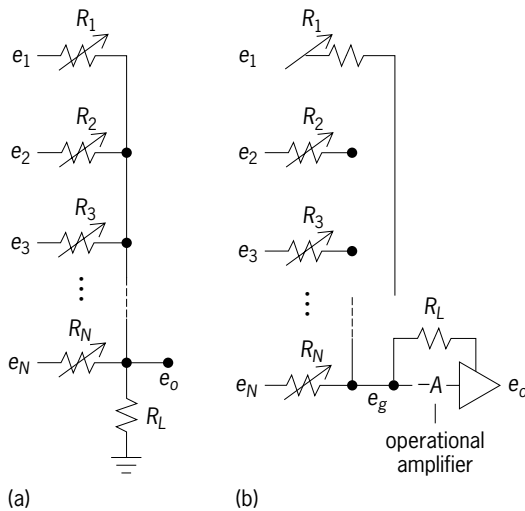


Fig. 1. Linear mixers. (a) Passive linear mixer based on a resistive summing network. (b) Active linear mixer with an operational amplifier.

mixing terms. The i th signal, e_i , is weighted by R_i/R_i , where R_i is the corresponding input resistance and R_L is the load resistance, and the result of all weighted signals is summed. The denominator contains all of the problems. It provides the overall attenuation of the output, which is further complicated by being a function of the individual R_i settings; one adjustment changes all the gain values.

Inexpensive integrated circuits have solved both problems and have revolutionized this application dramatically. Operational amplifiers of reasonably high quality that will eliminate the adjustment interactions and also provide gain are readily available. The input signals are summed into the virtual ground summing node (signal e_g) at the input of the operational amplifier (Fig. 1b). The troublesome denominator in the above equation is replaced by unity, so that interaction and attenuation are eliminated. There is a sign change in the output, but that is a small drawback compared to the advantage of having the virtual ground provided by the operational amplifier. See AMPLIFIER; INTEGRATED CIRCUITS; LINEARITY; OPERATIONAL AMPLIFIER.

Nonlinear mixer. Perhaps the most familiar application of nonlinear mixers is in radio and television receivers, but they are widely used in such applications as amplitude modulation (AM) and demodulation, frequency demodulation, phase detection, frequency multiplication, and single-sideband (SSB) generation. The incoming information to a receiver has been transmitted and received at a frequency (say, ω_i) far too high to permit efficient amplification and processing. Therefore the signal is translated or frequency-shifted or heterodyned by a mixer to a lower frequency (ω_o), known as the intermediate frequency (IF), where amplification and processing are performed efficiently by an IF processor, sometimes referred to as the IF strip. See AMPLITUDE-MODULATION DETECTOR; AMPLITUDE MODULATOR; FREQUENCY-MODULATION DETECTOR; FREQUENCY MODULATOR; FREQUENCY MULTIPLIER;

PHASE-ANGLE MEASUREMENT; RADIO RECEIVER; SINGLE SIDE BAND; TELEVISION RECEIVER.

If the input signal has frequency ω_i , and the adjustable (tuned) local oscillator signal into the mixer has frequency $\omega_i - \omega_o$, then it follows from trigonometry that the mixer output, which is the product of the input and oscillator signals, is the sum of two terms. One of these terms is at a frequency equal to the sum of the input and oscillator frequencies, that is, $2\omega_i - \omega_o$, and since ω_i is much larger than ω_o , this frequency is double that of the input; therefore, this term can be completely filtered out. The frequency of the remaining signal is at the difference of the input and oscillator frequencies, namely ω_o , the IF-channel frequency. See HETERODYNE PRINCIPLE.

A second application of a nonlinear mixer is frequency synthesis, where a stable but not easily changed signal at a high frequency, ω_b , is made tunable by mixing it with an easily tunable signal at a low frequency, ω_v , which, perhaps, can be varied in precise increments of any size. As before, the output is the sum of two terms, in this case at frequencies $\omega_b + \omega_v$ and $\omega_b - \omega_v$. The utility of the method is limited by the ability to filter or separate one frequency term from another (the frequency separation of the two terms is only $2\omega_v$), thereby determining the minimum practical value of ω_v for the application.

A mixer is an integral part of an AM-radio integrated circuit which contains virtually all AM-radio functions except filters. A particular type of mixer, the quadrature detector, is included in the frequency-modulation (FM)-radio integrated circuit.

A circuit commonly used to implement the multiplication of two signals to give the product necessary for nonlinear-mixer applications consists of a set of three differential amplifiers (Fig. 2). Two of these differential amplifiers (consisting of transistor pairs Q_1 - Q_2 and Q_3 - Q_4 , respectively) are driven out of phase with one another and into saturation by the first input signal. The output points are at the nodes of the cross-coupled collectors (Q_1 - Q_3 and Q_2 - Q_4). The linear differential drive by the second input signal to the remaining transistor pair (Q_3 - Q_4) results in the output being the second signal, square-wave-modulated (or demodulated) by the first signal. The

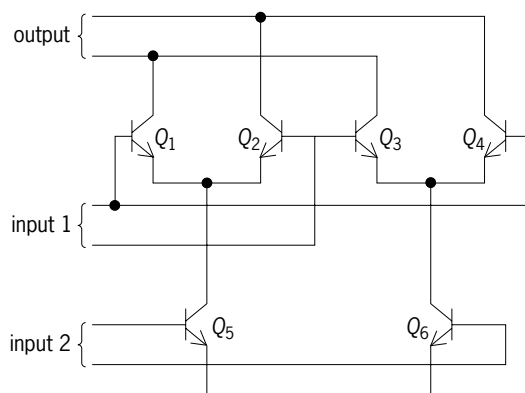


Fig. 2. Balanced modulator (nonlinear mixer). Only signal paths are shown. Biasing must be added.

consequence of this very efficient but nonlinear operation is that the output spectrum contains the desired spectrum, plus images (or aliases or spectral replicas) centered at odd harmonics of the first input signal. These images are removed by subsequent filtering. See DIFFERENTIAL AMPLIFIER. Stanley A. White

Bibliography. A. V. Oppenheim and A. S. Willsky, *Signals and Systems*, 2d ed., 1996; M. S. Roden, *Analog and Digital Communication Systems*, 4th ed., 1995; M. Schwartz, *Information Transmission, Modulation, and Noise*, 4th ed., 1990.

Mixing

A common operation to effect distribution, intermingling, and homogeneity of matter. Actually the operation is called agitation, with the term "mixing" being applicable when the goal is blending, that is, homogeneity. Other processes, such as reaction, mass transfer (includes solubility and crystallization), heat transfer, and dispersion, are also promoted by agitation. The type, extent, and intensity of agitation determine both the rates and adequacy of a particular process result. The agitation is accomplished by a variety of equipment.

Most liquid mixing is done by rotating impellers in vertical cylindrical vessels. A typical impeller-type liquid mixer with a variety of features is shown in Fig. 1. The internal features, including the vessel it-

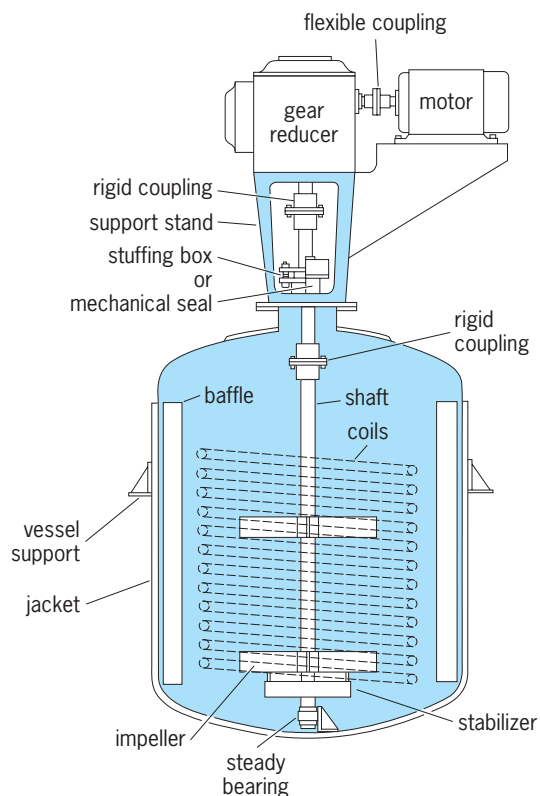


Fig. 1. Typical impeller-type liquid mixer. (After V. W. Uhl and J. B. Gray, *Mixing: Theory and Practice*, vol. 2, Academic Press, 1967)

A classification of operations requiring mechanical agitation and mixing

| Physical criteria | General kinds of operations | Chemical and mass transfer criteria |
|------------------------|--|-------------------------------------|
| Pumping, circulation | Fluid motion | Heat transfer |
| Blending Emulsions | Miscible liquids Immiscible liquids | Reactions Extraction, reaction |
| Dispersions | Liquid-gas systems | Absorption, stripping, reaction |
| Dispersion, suspension | Liquid-solid systems | Dissolving, crystallization |

self, are considered as a whole, that is, as the agitated system. The forces applied by the impeller develop overall circulation or bulk flow. Superimposed on this flow pattern, there is molecular diffusion, and if turbulence is present, also turbulent eddies. These provide micromixing. Solids, granular to powder, are mixed in a variety of contrivances.

Fluid mixing. The range of industrial processes for which mechanical agitation is used is conveyed by the table. The processes may be batch or continuous. Often several operations are conducted in concert; for example, in the hydrogenation of vegetable oils, the catalyst is suspended, a reaction takes place, and heat is transferred to the jacket wall. The performance criteria for different operations vary, and although several agitation schemes often will accomplish a given process result, certain systems can be shown to be preferable. Here the governing factors generally are the type, size, and speed of the impeller as well as major internals such as baffles.

Fluid motion, both large-scale (bulk circulation) and small-scale (turbulent eddies), is required in turbulent flow. The bulk circulation results when the fluid stream is discharged by the impeller. Turbulence is generated mostly by the velocity discontinuities adjacent to the stream of fluid flowing from the impeller, but also by boundary-and-form separation effects; turbulence spreads throughout the bulk flow and, although attenuated, is carried to all parts of the vessel. It is recognized that some mixing operations should require relatively large bulk or mass flows, whereas others need a high intensity of turbulence (termed shear in this connection). From this it follows that there should be an optimum ratio of flow to shear for a given process result, for example, high flow for blending of miscibles and suspension of solids, and high shear for gas absorption (mass transfer) and dispersion.

Energy must be supplied to produce fluid motion. The power imposed by the mixing impeller is proportional to the flow multiplied by the head developed, which is equated to turbulence (shear). The ratio of the power supplied for flow to that for turbulence can be varied for the same power input and type of impeller; this variation is demonstrated by Fig. 2. Impeller shape also has an effect which can be significant, with a rotating disk representing the extreme case for high turbulence or shear.

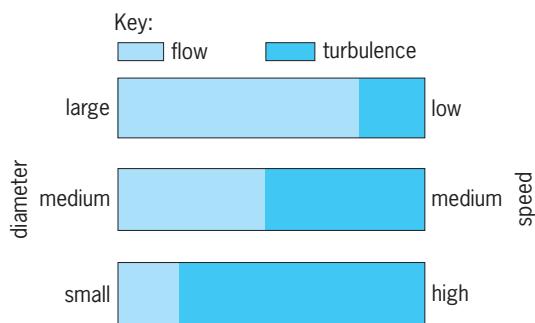


Fig. 2. Effect of impeller size and speed on flow and turbulence (shear) at constant power.

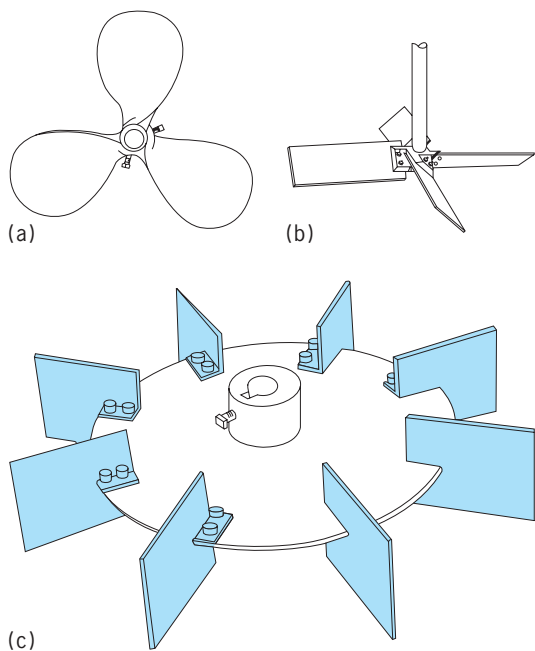


Fig. 3. Common impellers. (a) Marine-type mixing propeller. (b) Axial-flow (pitched-blade) turbine. (c) Flat-blade (Rushton) turbine. (After R. H. Perry and C. H. Chilton, eds., *Chemical Engineers' Handbook*, 5th ed., McGraw-Hill, 1973)

In all bench-scale and pilot-plant work where mixing is important, the type of fluid motion in terms of ratio of flow to shear should be found. This becomes a basis for scale-up in design.

Mixing vessels and impellers. Circulation or bulk flow depends on the shape of the container; the fittings it contains; the type, size, and position of the impeller; and the properties of the fluid. Lateral (radial) and vertical flow currents usually produce the best agitation; these currents must penetrate to all portions of the fluid.

For the laboratory, whenever matters are mixed, cylindrical containers such as beakers should be used. Vertical baffles should be provided for impellers which are centrally located; if a propeller is used in an off-center position, the baffles may be omitted. Round-bottomed flasks are unsatisfactory unless modified by creases blown in the side; this form can be purchased.

The most common and useful types of impellers are the marine-type propeller (Fig. 3a), the axial-

flow (pitched-blade) turbine (Fig. 3b), the flat-blade (Rushton) turbine (Fig. 3c), and the flat paddle (generally two vertical blades relatively large compared to the other types). Any of these impellers centrally positioned produces rotating fluid motion (Fig. 4) with a vortex around which liquid swirls. This motion often results in separation or stratification rather than intermingling. Relatively little power can be applied; a dearth of turbulence and of vertical and lateral flow motion results. Inserting projections into the body of the fluid stops the rotary motion, and the vortex disappears. Thus agitation is improved. Such projections at the side of the tank are called baffles (Fig. 5). The propeller and axial-flow turbine with tank-wall baffles (Fig. 5a) will generate an axial-flow pattern; the paddle and Rushton turbine produce radial flow (Fig. 5b). These flow patterns are conducive to good agitation.

Swirl can be obviated in laboratory work by using the propeller off-center. Good vertical and lateral flow motion without swirl and surface vortex can be attained by discharging the propeller downward and positioning as in Fig. 6; the exact position is critical but can be found by trial.

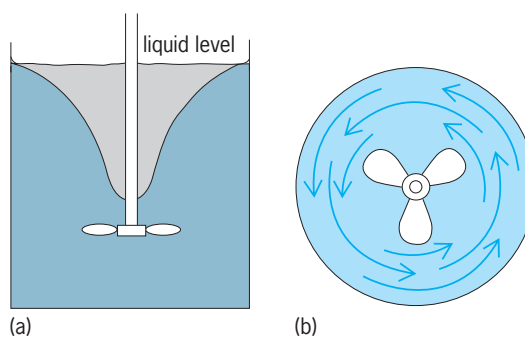


Fig. 4. Swirling flow pattern for impeller of any shape, without baffles. (a) Side view. (b) Bottom view.

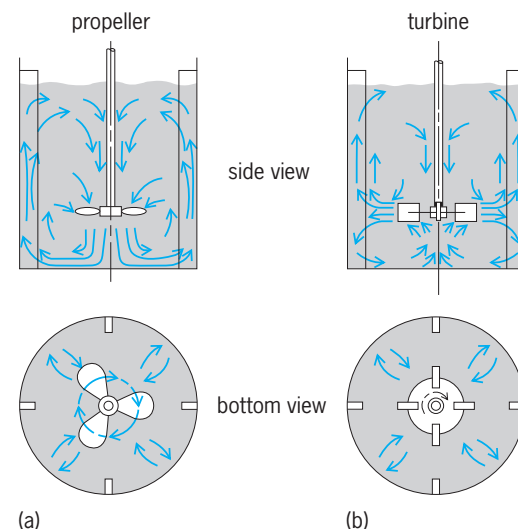


Fig. 5. Flow patterns with baffles. (a) Flow pattern for propeller with baffles at tank wall. (b) Flow pattern for turbine with baffles at tank wall.

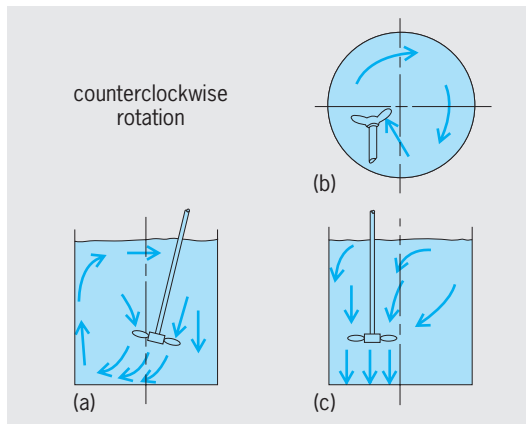


Fig. 6. Flow pattern for top-entering, off-center propeller without baffles: (a) front, (b) top, and (c) side views.

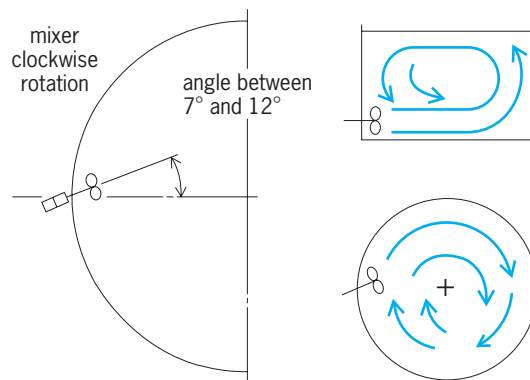


Fig. 7. Side-entering propeller mixer position. With proper propeller position, no vortex will result.

Liquids in large tanks, for example, 3×10^6 gal (11,370 m³), are often blended by a side-entering propeller. **Figure 7** shows the correct position for a right-hand, clockwise-rotation propeller; proper angular positioning is critical.

The behavior of low-viscosity liquids has been described above and is shown in Fig. 4. High-viscosity liquids, those above 1000 centipoises, will have much less rotary flow; without baffles the motion will approximate that in Fig. 5.

Power and flow. In liquids of low viscosity, power imposed by an impeller in a baffled tank is proportional to the cube of its speed, the fifth power of its diameter, and the density. The fluid regime would be turbulent. For constant power and varying diameter of a given type of impeller, the corresponding speed can be found from the equation $N_r = (1/D_r)^{5/3}$, where N_r is the ratio of speeds and D_r is the ratio of corresponding impeller diameters.

In liquids of high viscosity, the power imposed by the impeller is proportional to the viscosity of the liquid, the square of the speed, and the cube of the diameter. Here the fluid regime is laminar. See VISCOSITY.

The discharge rate of flow from impellers is proportional to the speed and to the cube of the diameter. Clearly, for the same power there will be greater flow from the large-diameter, low-speed im-

PELLER than from a smaller-diameter, higher-speed one (Fig. 2).

High-viscosity liquids. These liquids and pastes often require different techniques from those used for mixing low-viscosity liquids. Special apparatus is necessary to provide for wiping, stretching, and squeezing, because turbulence cannot be generated in such fluids to provide for the small-scale mass transfer necessary to cause interpenetration of substances. There are few quantitative data available which describe the performance of the various types of equipment. An example of typical equipment is shown in Fig. 8.

Continuous processing. For continuous pipeline blending, a baffled mixing cell with one or two impellers on the shaft is available. Centrifugal pumps sometimes provide in-line agitation as an auxiliary service. For continuous two-phase contacting, a column containing compartment separation plates, baffles, and impellers is used.

Solids mixing. Solids of different density and size are mixed in tumblers (a double cone turning end on end) or with agitators (a helical ribbon rotating in a horizontal trough). The duration of mixing is an important additional variable because classification and separation often occur after attainment of the desired distribution if the operation is carried on too long.

Equipment. Portable mixers generally range in size from fractional horsepower to 3 hp (2240 W); they are designed for use in open tanks up to 3000 gal (11.4 m³) capacity and are clamped to the tank shell. Direct-drive units usually run at 3600, 1750, or 1150 rpm or at variable speeds. Gear-drive units run at about 420 rpm or at variable speeds.

Light-duty, permanently mounted mixers range from fractional horsepower to 3 hp, and are used on open or closed tanks up to 3000 gal (11,400 L) capacity. Heavy-duty, permanently mounted, top-entering mixers are made up to 500 hp (373,000 W). They are normally designed for speeds of 45–200 rpm, made



Fig. 8. Ribbon turbine with double spiral, for high-viscosity liquids.

with independent mounting for the mixer shaft, and connected to the drive by a flexible coupling to protect the speed-reduction gearing. See UNIT OPERATIONS. Vincent W. Uhl

Advances, trends, and challenges. The understanding of mixing of fluids and solids has benefited by major advances in computations and chaos theory. These insights help to explain why some designs work and others do not, and also suggest novel designs.

The objective of fluid mixing is to produce the maximum amount of contact or interfacial area between fluids in the minimum amount of time or with the least amount of energy. High interfacial area aids molecular diffusion, ultimately resulting in a homogeneous mixture. The key to effective mixing lies in stretching and folding fluid elements. Stretching and folding is synonymous with chaos and results in an exponential increase of interfacial area with time. Experiments have shown that it is relatively straightforward to produce chaotic flows. Mixing in chaotic flows does not require turbulent motion, and the concepts apply even to very viscous liquids, where turbulence would be difficult to produce. Chaos can be achieved by time modulation of the flow field or by periodic changes in geometry, as in static mixers. Chaos, however, is rarely complete and, in general, chaotic systems contain both chaotic regions, where mixing is effective, as well as regions of poorly mixed fluid (called islands, regular regions, or tubes). The objective of good mixing is the elimination of islands.

Chaotic mixing occurs in impeller-type mixers, in continuous duct-type flows such as static mixers, and even applies to mixing in very small scale systems, such as in microfluidic applications. One example of a duct-type chaotic flow is the so-called partitioned-pipe mixer (Fig. 9). This flow consists of a pipe partitioned with a sequence of orthogonally placed rectangular plates. The cross-sectional motion is induced through rotation of the pipe with respect to the assembly of plates. Two streams are injected at the top. In Fig. 9a, both streams are in chaotic regions and mix well. In Fig. 9b, one of the streams is injected into an island and remains unmixed.

Similar results have been uncovered in stirred tanks where mixing is less efficient than once thought, with only a fraction of the energy input actually participating in mixing. This area has benefited from computational flow codes where rapid advances have taken place. Significant advances have occurred also in areas where the objective is the creation of fine dispersions of drops or particles, such as the blending of immiscible polymers, composites, and nano-composites, and in the mixing of fluids with complex rheological characteristics. Mixing effects in reactive systems are critical in the case of multiple reactions, when reactions are fast, or in viscous media, when mixing is slow compared to the rate of reaction. Improperly designed systems lead to slowing of desired reactions and sometimes even halting before completion, increase of undesired reactions, and reduction of product selectivity.

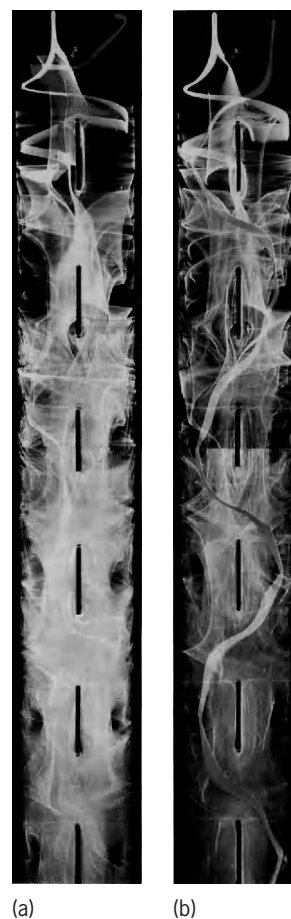


Fig. 9. Coexistence of (a) chaotic and (b) nonchaotic mixing in a duct-type flow.

Mixing of granular solids is pervasive in industry and natural processes. In contrast to mixing of liquids, mixing of granular materials received disproportionately little attention until recently. Computation and chaos theory concepts have aided in understanding. The simplest case of mixing of dry solids has benefited by simulation via particle dynamics (also called discrete element methods), where each particle is simulated using contact force models that capture friction and inelasticity. Presently it is difficult to simulate mixing with the number of particles that are contained in industrial mixers, such as a double-cone blender and bladed mixers, but laboratory experiments are within reach and simulations with 10^4 – 10^5 particles are routine. Simulations with 1–2 million particles are now possible.

Mixing of solids is more complicated than mixing of liquids. Mixing happens only when there is flow (there is no molecular diffusion). Also, granular solids can segregate or become unmixed. Unmixing happens if particles differ in size or density (and in practice they often differ in both). In this case, there may be differential motion of particle types, with the end result of magnifying variations in composition and, in the worst case scenario, nearly complete unmixing (Fig. 10). There are numerous processes where a nonuniform composition can have serious consequences. In pharmaceuticals



Fig. 10. Unmixing of particles of two different sizes in a rotating tumbler operating in a time periodic mode.

poor mixing often requires batches of products to be either remixed or disposed. Theory reveals that mixing of solids is highly sensitive to mixer filling. Improper filling may leave entire regions in the mixer unmixed. In general, low fill levels perform better. The shape of the mixer is of critical importance; noncylindrical shapes, such as an equilateral triangle, may mix better than a cylinder, the most common shape. Recent advances have taken place for slurries, where particles are immersed in a liquid, and for powders, where cohesion between particles is important.

Julio M. Ottino

Bibliography. N. Harnby, M. F. Edwards, and A. W. Nienow, *Mixing in the Process Industries*, 2d ed., 1997; G. Metcalfe et al., Avalanche mixing of granular materials, *Nature*, 374:39–41, 1995; J. M. Ottino, The mixing of fluids, *Sci. Amer.*, 260:56–67, 1989; R. H. Perry and D. Green (eds.), *Perry's Chemical Engineer's Handbook*, 7th ed., 1997; T. Shinbrot and F. J. Muzzio, Nonequilibrium patterns in granular mixing and segregation, *Phys. Today*, pp. 25–30, March 2000; V. W. Uhl and J. B. Gray, *Mixing: Theory and Practice*, vol. 1, 1966, vol. 2, 1967, vol. 3, 1986; J. J. Ulbrecht and J. B. Gray (eds.), *Mixing of Liquids by Mechanical Agitation*, 1985.

Mizar

The second-magnitude star that marks the bend in the handle of the Big Dipper, ζ Ursae Majoris. Mizar was the first telescopic visual binary to be discovered (1650), the first ever to be photographed (1857), and the first star discovered to be a spectroscopic binary (1889). Its distance to the Sun is 24 parsecs (7.5×10^{14} km or 4.7×10^{14} mi). See BINARY STAR; CONSTELLATION; URSA MAJOR.

It is one of the finest examples of a multiple star system. Mizar (the Horse) forms a wide visual binary with the 4th-magnitude star Alcor (the Rider), at a separation of $12'$. The two stars have nearly the same

proper motion, and the pair is visible to the naked eye. Mizar itself is a close visual pair, with bluish-white components of apparent magnitude 2.27 and 3.95, approximately $14''$ apart. The spectral types are A2Vp and Am, respectively. The spectrum of the primary star, Mizar A, shows periodic doubling of the absorption lines due to the Doppler effect, caused by the motion of two stars around their common center of mass. The orbital period of this spectroscopic binary is 20.5 days, and the stars have been resolved with an interferometer, the mean angular separation being only $0.012''$. By combining the elements of the spectroscopic orbit with those from the astrometric orbit, the masses of the two stars have been derived, and are both about 2.5 times that of the Sun. Mizar B is also a spectroscopic binary with a period of 176 days. See DOPPLER EFFECT; SPECTRAL TYPE.

Mizar is a member of the group of stars known as the Ursa Major moving group, which has an estimated age of 3×10^8 years. David W. Latham

Mobile communications

Radio communication in which one or both ends of the communication path are movable. The term "mobile" refers to movement of the radio rather than association with a vehicle (for example, hand-held portable radios are included in the definition). The movement need not occur during the communications. The Federal Communications Commission (FCC) licenses and regulates non-federal-government radio activity in the United States, while the National Telecommunications and Information Administration (NTIA) oversees federal-government uses. Other countries have similar agencies. International coordination is afforded through the International Telecommunication Union (ITU) and international treaty. Historically, the term "mobile radio" indicated systems where the user terminal was located in a vehicle. Current usage includes "outdoor" transportable devices that may not be actually moving during a communication session, as would be the case for a person using a laptop at a park bench.

Private land mobile services. Users who lease or purchase radio equipment for personal use fall into this category. Examples are public safety, special emergency, industrial, land transportation, and location radio services. Spectrum over a wide range of radio frequency bands is allocated; for example, low band (30–50 MHz), high band (150–174 MHz), and 900-MHz band (896–901 MHz, paired with 935–940 MHz). Dispatch is the normal mode of operation; that is, all members of the group hear all communications. To accomplish this, high-power, high-site, base station repeaters are generally used so that a single site covers the entire area of interest. Coverage radius varies with frequency band, local terrain, and permissible power levels, but values on the order of 20 mi (32 km) are commonplace. Where areas to be covered are even larger (for example, statewide police systems) or where coverage reliability must be greater than that possible from a single site

(for example, for ambulance communications), multiple sites can simulcast the communications. Current technology allows for data exchanges, vehicle location, and secure digitized voice. See RADIO SPECTRUM ALLOCATIONS.

Specialized mobile radio (SMR) is a type of mobile radio service in which individual users with business interests are licensed to operate their mobile, portables, and control stations on channel pairs repeated by specialized mobile radio base stations. Full interconnection to the public switched telephone network (PSTN) is possible. To boost the spectrum efficiency of specialized mobile radios relative to shared repeaters already in use, the FCC required that channels be trunked. Trunking in the context of radio systems means not only sharing equipment but sharing frequencies as well. Trunking channels means that when a user wishes to place a call it can be served by any one of the channel pairs that is available. The peak capacity that is obtained through trunking is no different than that obtained with single-channel operation, but the average grade of service is much improved. For example, with conventional operation, if each channel is busy half the time, the probability of being blocked from service is 50%. However, when five channels are trunked, a new call request will be blocked only if all channels are simultaneously busy. The probability of this is only about 7% according to Erlang-B traffic theory.

Paging service. Although paging is primarily a one-way radio system, two-way operation with such functions as page acknowledgment and short message reply is available. Typical paging operation is as follows. Via the public switched telephone network, the land party calls a general telephone number for the paging system to which the desired party subscribes. When the call is answered, a number unique to the subscriber is entered. Paging control equipment routes the call to appropriate paging transmitters (multiple transmitters are generally required to assure high reliability over large coverage areas). The transmitters simulcast coded radio signals that are recognized only by the desired individual's paging receiver. The user is alerted by a series of beeps or by mechanical vibration.

Some types of paging receivers display digits and letters (alphanumeric displays) that allow the calling party's number and a brief message to be displayed,

and a message operator becomes unnecessary. Since display of paging messages involves little information, thousands of users can share a paging channel, thus making the system extremely spectrum efficient. Other types of paging receivers provide for brief voice messages following the alert (tone and voice). See RADIO PAGING SYSTEMS.

Public land mobile service. Mobile telephone service (MTS) in the United States initially involved wireline and radio common carriers. The earliest mobile telephone systems operated in push-to-talk mode with calls handled by special operators. The advent of improved mobile telephone service (IMTS) allowed automatic, duplex operation, with customers able to dial to and receive calls from anywhere in the world. Custom features such as call waiting, three-way calling, speed calling, and voice mail became available, as did data information exchange and facsimile transmission when modems were attached to the radios.

Trunked duplex operation (also known as frequency-division duplexing or FDD) was facilitated by assigning the radio channels in pairs, sufficiently separated in frequency such that the transmitted signal did not desensitize the receiver. Thus, conversations in the land-to-mobile direction and the mobile-to-land direction could occur simultaneously. Customers could subscribe for service in one area and travel to other areas and continue service, provided that the companies involved had previously agreed to honor such operation (termed roaming).

Cellular telecommunications service. Until the development of cellular mobile radio, the total number of mobile telephone users was severely limited by the modest number of radio channels allocated. Internationally, many different standards for cellular are in use (Table 1). While they differ in frequency, channel bandwidth, and control signaling, all these first-generation systems use analog frequency modulation (FM) supporting a single conversation per channel. See FREQUENCY MODULATION.

Cellular technology allows hundreds of thousands of users to be handled in a single metropolitan area. Rather than link into the telephone system from a single high-power, high site that covers the entire metropolitan area (Fig. 1a), users are linked via many low-power, low sites. A single low site, of course, can cover only a limited area, termed a cell, but

TABLE 1. Analog cellular radio systems

| Name* | Frequency, MHz | | Peak FM deviation, kHz | Channel bandwidth, kHz |
|--------|----------------|----------------|------------------------|------------------------|
| | Mobile-to-land | Land-to-mobile | | |
| AMPS | 824-849 | 869-894 | 12 | 30 |
| JTACS | 870-885 | 925-940 | 5 | 25 |
| NMT450 | 453-457.5 | 463-467.5 | 4.7 | 25 |
| NMT900 | 890-915 | 935-960 | 4.7 | 25 |
| TACS | 890-905 | 935-950 | 9.5 | 25 |
| ETACS | 872-888 | 917-933 | 9.5 | 25 |

*AMPS = Advanced Mobile Phone System; JTACS = Japanese Total Access Communication System; NMT = Nordic Mobile Telephone (number following denotes band); TACS = Total Access Communication System; ETACS = Extended TACS.

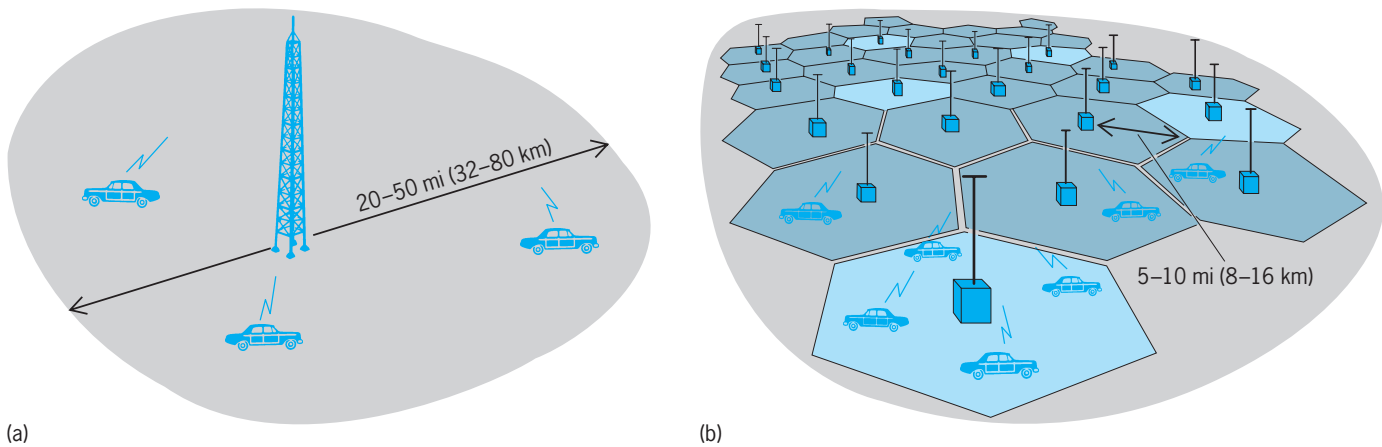


Fig. 1. Comparison of conventional and cellular mobile radio. (a) Conventional mobile radio covers a broad geographical region from a single, high site using high-powered base stations. (b) Cellular mobile radio divides the service area into cells with low-power, low-antenna base stations, allowing channels to be reused by cells that are sufficiently separated.

many low sites taken together can cover the entire metropolitan area (Fig. 1b). Spectrum efficiency stems from reusing the same frequency at all sites that are sufficiently separated. To further limit interference caused by such frequency reuse, each cell may be divided into sectors, and directive antenna patterns may be used.

An attractive feature of cellular radio is the ability to vary the size of cells in accordance with user density; hence, cell size can increase away from city centers. To sustain the reuse pattern with mixed cell sizes, power levels are tailored to produce comparable signal levels at all cell boundaries. Also, as more customers are added, radio channels can be created to serve them by constructing new base stations (hence, new cells) in geographical locations between existing cells. This concept is called cell splitting (Fig. 2). Geographical coverage of the system can be expanded as well by constructing new base stations on the periphery of the existing system and assigning frequencies consistent with the original reuse pattern.

Automatic, continuous coverage as users move across cell boundaries is provided by the cell hand-

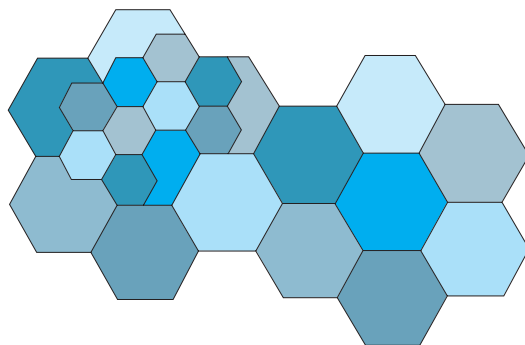


Fig. 2. Cell splitting in a cellular mobile radio system. In this example, nine cells have been created out of parts of four original cells. The additional capacity can be provided without increasing the available radio spectrum, because low-power transmitters in nearby cells can use the same frequencies without unacceptable interference.

off feature of cellular (also termed handover and automatic link transfer). Calls in need of handoff are recognized by monitoring call quality and comparing it to some required threshold. Handoff control procedures for first-generation analog frequency-modulation systems are also in operation.

Development of mobile radio services. Following more than a decade of trials and standardization, the first commercial cellular service in the United States began in Chicago in 1983, joining Japanese and Nordic systems that became commercial in 1979 and 1985 respectively. Initially, 40 MHz of spectrum was allocated to cellular service, and in 1986 this was extended to 50 MHz, in the 824–849 and 869–894 MHz bands. There are at least two licensed service providers in each of 305 metropolitan statistical areas and 428 rural statistical areas in the United States.

As acceptance of cellular technology has exceeded original expectations, concerns about the ability of the technology to support the large number of users has arisen. Practical limits on the cell-splitting concept occur due to difficulty in controlling interference as cell size shrinks, to capital investment requirements for continued building of new cells, and to local zoning issues. System modifications using split or offset channels have been developed to boost the effective number of simultaneous conversations possible per cell. For example, Narrow-band Advanced Mobile Phone Service (NAMPS) provides three channels in the bandwidth of a single 30-kHz channel of the Advanced Mobile Phone System (AMPS), the original United States analog cellular standard. Likewise, the narrow-band version of the Japanese analog cellular, the Japanese Narrow-band Total Access Communications System (JTACS), provides two channels in the bandwidth of a single 25-kHz Japanese Total Access Communication System (JTACS) channel. Along with narrowing the channel bandwidth, these modified systems generally have improved handoff by including mobile participation in the process. Data operation has likewise improved with the introduction of cellular digital

TABLE 2. Digital cellular radio systems

| Name* | Frequency, MHz | | Channel bandwidth | Conversations per channel | Multiplex method |
|-------|----------------|----------------|-------------------|---------------------------|------------------|
| | Mobile-to-land | Land-to-mobile | | | |
| GSM | 890–915 | 935–960 | 200 kHz | 8 | TDMA |
| IS-54 | 824–849 | 869–894 | 30 kHz | 3 | TDMA |
| IS-95 | 824–849 | 869–894 | 1.25 MHz | 20–60 | CDMA |
| JDC | 810–826 | 940–956 | 25 kHz | 3 | TDMA |
| MIRS | 806–824 | 851–869 | 25 kHz | 6 | TDMA |

*GSM = Global System for Mobile Communications; IS = Interim Standard; JDC = Japan Digital Cellular; MIRS = Motorola Integrated Radio System.

packet data (CDPD). This allows data traffic to be handled in short pieces or packets during idle periods on the voice servers, a major improvement over circuit-switched data, which require a voice server continuously even though most of the time no real exchange of data is occurring.

2G digital cellular service. Despite the improvements in analog cellular operation, schemes that digitally encode speech and suitably modulate the carrier have been adopted in second-generation (2G) cellular systems (Table 2). Such schemes handle data traffic as well as voice traffic. See MODULATION.

European digital technology (GSM, or Global System for Mobile Communications) uses 124 carriers of 200 kHz each over 25 MHz of spectrum allocated for each direction of communication (890–915 MHz and 935–960 MHz). Each 200-kHz carrier uses time-division multiple access (TDMA) with eight user time slots per carrier; hence, eight simultaneous conversations can be supported. See MULTIPLEXING AND MULTIPLE ACCESS.

Several standards for digital cellular in the United States and Canada exist:

1. Interim Standard 54 (IS-54) specifies a 30-kHz carrier using time-division multiple access with three user time slots per carrier. The existing cellular spectrum allocation is used, with analog FM carriers being gradually displaced by digital channels. This required the initial use of dual-mode radios.

2. Interim Standard 95 (IS-95) spreads all signals simultaneously over the same bandwidth and separates them via code-division multiplex access (CDMA). With the help of precise power to control near-far interference problems (that is, the tendency of nearby users to obliterate base station reception of distant, weak users), the same bandwidth can be reused in each and every cell. This is the means of achieving high spectral efficiency. Speech is encoded at a variable rate according to voice activity. Again, dual-mode radios were initially required.

3. MIRS is the result of a 1990 waiver request to the FCC to operate a cellular system in the frequency bands allocated to specialized mobile radio. Such systems are called enhanced SMRs (ESMRs) and provide six TDMA time slots in 25-kHz channels.

Second-generation digital cellular systems have attempted to improve handoffs by using more meaningful measurements of quality such as bit error rate (BER) and by having the subscriber units assist in the handoff process. IS-95 systems even introduce

the concept of soft handoff, wherein the instantaneously best-received signal from a number of cell sites is used.

Cordless telephones. Cordless telephones provide a wireless extension to the public switched telephone network over an extremely limited range. Many units rely on mature analog radio technology that provides only nominal security and almost no interference protection.

In Europe and many Asian countries, there are enhanced cordless telephone services that use digital modulation. The first of these services, known as CT-2, for second-generation cordless telephone, uses frequency-division multiple access (FDMA) and time-division duplexing (TDD); each conversation uses a selected frequency exclusively (FDMA), but only one frequency is used and each direction of communication takes turns using it (TDD). By comparison, analog cordless telephones are also FDMA but use frequency-division duplexing (FDD) in which each direction of communication uses a different radio channel.

Digital European Cordless Telephone (DECT) is another (TDD) system that was developed primarily for in-building business application. In contrast to CT-2, DECT uses TDMA on each radio channel and can support up to 12 user voice channels on a given radio-frequency carrier. DECT uses a dynamic channel allocation (DCA) scheme in which a radio can change carrier frequencies every TDMA frame, if necessary, to avoid interference. In contrast, CT-2 and analog cordless telephones use the same frequency for the duration of a call and thus have no interference avoidance mechanism.

Another TDMA/TDD cordless telephone system, the Personal Handyphone System (PHS), developed in Japan, supports up to four voice circuits in each radio-frequency channel but uses much less bandwidth than DECT; channel spacing for PHS is 300 kHz, while the spacing between DECT carriers is 1.728 MHz.

Personal communication systems (PCS). The great demand for cellular phones and related wireless services has been addressed to some extent by the addition of new spectrum, by the introduction of narrow-band and digital cellular systems, and by cell splitting where practical. Acknowledging that these techniques for increasing capacity would quickly be exhausted, most countries have allocated additional spectrum for mobile and portable communications.

Since these frequency bands have much more spectrum than those previously allocated to cellular service, a greater variety of services are possible.

A spectrum of 120 MHz in the 1850–1910 and 1930–1990 MHz bands was allocated for licensed personal communication system (PCS) operation in the United States, and 20 MHz of spectrum in the 1920–1930 MHz band for unlicensed operation, split evenly between voice (isochronous) and data (asynchronous) applications. The spectrum allocation for licensed PCS operation is divided into six frequency blocks, three of which contain 30 MHz of spectrum, and the other three 10 MHz. It is thus possible in a given region to have as many as six competing service providers, in addition to the two 900-MHz cellular service providers.

The unlicensed PCS band was authorized to spur innovation and to encourage rapid development of service by allowing PCS equipment to be deployed without going through a licensing process. However, equipment designed to operate in this band must comply with a spectrum etiquette developed by the industry. This etiquette is a set of rules and requirements that facilitates simultaneous use of multiple systems in a common portion of the spectrum.

3G systems. In response to the ever-increasing demand for higher capacity and new services, the ITU defined the standards for third-generation systems, also known as International Mobile Telecommunications 2000 (IMT-2000). The ITU specified minimum bit rates of 144 kilobits per second (kbps) in mobile (outdoor) environments and 2 megabits per second (Mbps) in fixed (indoor) environments.

The IMT-2000 initiative gave birth to the Third-Generation Partnership Project (3GPP) for the evolution of GSM networks and the Third-Generation Partnership Project-2 (3GPP2) for the evolution of CDMA networks. The bands 1885–2025 MHz and 2110–2200 MHz are intended for use, on a worldwide basis, by administrations wishing to implement IMT-2000.

The 3GPP is a global initiative under the auspices of the European Telecommunications Standards Institute (ETSI). The 3GPP scope is to generate a global technical specification for a third-generation mobile system based on evolved GSM core networks and their corresponding radio technologies in both FDD and TDD modes. The scope grew to include the evolution of the specifications for the General Packet Radio Service (GPRS) and EDGE (Enhanced Data Rates for GSM Evolution). A permanent project support group called the Mobile Competence Centre (MCC) has been established to ensure the efficient day-to-day running of 3GPP. The MCC is based at the headquarters of the European Telecommunications Standards Institute (ETSI) in Sophia Antipolis, France. A similar partnership project, 3GPP2, was created to define global specifications for a third-generation mobile system based on the evolution of the ANSI/TIA/EIA-41 core network (developed by the American National Standards Institute, the Telecommunications Industry Association, and the Electronic Industries Alliance) and the corre-

sponding radio access technology (Code-Division Multiple Access-2000 or CDMA2000). Members of both 3GPP and 3GPP2 include ARIB (Association of Radio Industries and Businesses, Japan), CCSA (China Communications Standards Association, China), TTA (Telecommunications Technology Association, Korea), and TTC (Telecommunications Technology Committee, Japan).

In 1999, the ITU approved five radio interfaces for IMT-2000 standards as part of the ITU-R M.1457 recommendation. These are:

1. CDMA Direct Spread, also known as UTRA-FDD (Universal Terrestrial Radio Access-Frequency Division Duplexing) or UMTS (Universal Mobile Telecommunications System), including W-CDMA (Wideband CDMA). [The Japanese service provider NTT DoCoMo is using an ARIB-standardized W-CDMA solution.] UMTS was developed by 3GPP. While there are some differences, the terms UMTS and W-CDMA are used to refer to both the ITU and the ARIB W-CDMA standards.

2. IMT-2000 CDMA Multi-carrier, also known as CDMA2000-3xRTT (3 Times Radio Transmission Technology), developed by 3GPP2. IMT-2000 CDMA2000 includes 1xRTT (1 Times Radio Transmission Technology) components, like CDMA2000 1xRTT EV-DO (Evolution Data Optimized).

3. IMT-2000 CDMA TDD, also known as UTRA TDD and TD-SCDMA (Time Division-Synchronous Code-Division Multiple Access). TD-SCDMA is developed in China and supported by the TD-SCDMA Forum.

4. IMT-2000 TDMA Single Carrier, also known as UWC (Universal Wireless Communications)-136 (EDGE), supported by UWCC (Universal Wireless Communications Forum).

5. IMT-2000 DECT, supported by DECT Forum.

There were estimated to be about 2 billion mobile subscribers in March 2005, about one-third the population of the world, with just over 50% subscribers utilizing 2G technologies, a negligible fraction still using first-generation (analog) technology, and just under half utilizing either enhanced digital technologies, known as 2.5G, or full ITU-approved 3G technologies. There were reportedly about 256 million CDMA subscribers in March 2005. The first commercial CDMA2000 system was deployed in October 2000. TD-SCDMA was not yet in commercial deployment as of 2006. DECT is generally considered to be better suited for in-building and enterprise systems.

Two key members of the global family of third-generation systems identified by the ITU are the Universal Mobile Telecommunication System (UMTS) from 3GPP, intended for new spectrum, and the Code-Division Multiple Access-2000 (CDMA2000) from 3GPP2, also known as CDMA2000 1xRTT, intended for deployment by operators within their existing spectrum previously allocated for CDMA2000. Both systems can serve multiple users simultaneously by assigning a distinct code to each user (code-division multiple access). Each code is a unique sequence of pseudorandom sign reversals applied to each user symbol, thereby spreading the user

TABLE 3. Third-generation mobile communication standards

| | CDMA2000 | W-CDMA (UMTS) | W-CDMA (ARIB) |
|----------------------------|-------------------------|---|---|
| Frequency band* | Any | 1920–1980 MHz uplink, 2110–2170 MHz downlink | 1920–1980 MHz uplink, 2110–2170 MHz downlink |
| Minimum bandwidth required | 2×1.25 MHz | 2×5 MHz | 2×5 MHz |
| Reuse factor | 1 | 1 | 1 |
| Chip rate [†] | 1.2288 Mcps | 3.84 Mcps | 4.096 Mcps |
| Maximum user data rate | 144–230 kbps | 384 kbps | 2 Mbps |
| Frame duration | 5, 10, 20, 40, or 80 ms | 10 ms | 10 ms |
| Spreading factors* | 8, ..., 128 | 4, ..., 256 uplink, 4, ..., 512 downlink | 1, ..., 256 |

*Uplink = transmissions from the mobile terminal to the base station; downlink = transmissions from the base station to the mobile terminal.
[†]Mcps = megachips per second (1 megachip = 2^{20} chips).

symbol across a given number of code chips to obtain a desired spreading factor. Spreading each user symbol with a unique code provides protection against interference from adjacent transmissions, making it possible for all transmissions to use the same frequency with complete reuse. To increase efficiency, multiple user symbols are grouped into frames. The main features of the CDMA2000 and W-CDMA third-generation standards are summarized in Table 3. See SPREAD SPECTRUM COMMUNICATION.

3G extensions for data. With the increased interest in Internet usage in particular and data applications in general, extensions of the 3G standards have been developed to increase performance by taking advantage of the distinct nature of data as compared to voice traffic. For instance, introducing delay can be very annoying in conversational speech, and thus it is maintained under strict control in voice-optimized systems. In contrast, data packets can be subject to variable delays of up to several hundred milliseconds to exploit better channel conditions with no noticeable degradation in the user experience. The data-optimized evolutions of the third-generation standards are CDMA2000 1xEV-DO (Evolution Data Optimized) and High Speed Down Link Packet Access (HSDPA) for W-CDMA. HSDPA was designed to support simultaneous voice and data traffic in the same 5-MHz channel. HSDPA achieves a peak rate of 14.4 Mbps and, in realistic deployments, it is expected to deliver on average rates in the 400–700 kbps range. 1xEV-DO, first deployed in 2003, devotes a complete 1.25-MHz channel to data users only. This design philosophy is not being strictly followed in that an emerging new service is the delivery of voice in data format, using IP (Internet Protocol) packets (Voice over IP or VoIP). In another departure from previous usage, in CDMA2000 1xEV-DO, the multiple CDMA codes are used to increase the bit rate delivered, while communicating with one user at a time (in the downlink, that is, in transmissions from the base station to mobile terminals), instead of using these codes to separate transmissions to multiple users, as is the case in CDMA2000. With this change it is possible to schedule preferentially transmissions to users that experience good channel conditions, in what is known as proportional fair scheduling. This scheduling technique results in throughput improvements, also referred to as mu-

tiuser diversity. CDMA2000 1xEV-DO delivers a peak rate of 2.4 Mbps and an average rate in the 300–500 kbps range. Further evolution to 1xEV-DO, known as 1xEV-DO RevA, should achieve faster packet set-up, reduced latency, and improved data rates. See VOICE OVER IP.

Smart antenna technologies for 4G systems. With the deployment of third-generation systems well underway, system designers and standard organizations have turned their attention to the evolution of systems “beyond 3G” and to 4G proper. While there is no universal definition of what will constitute a true 4G system, it is possible to extrapolate from the previous generational transitions that a key element will be a significant increase (that is, a tripling or at a least doubling) in system capacity, throughput, and average bit rates. It is expected that such increases will be made possible in large part by advanced antenna technologies.

Antenna diversity. Perhaps one of the best-known multiple-antenna techniques is antenna diversity, where multiple antennas are used to maximize the chances that the radio channel condition to at least one of the antennas will be satisfactory. Clearly, the more antennas, the higher this chance will be. In the technique of selection diversity, the signal from the best antenna is selected. The optimum approach, known as maximal ratio combining, is to perform a sum of all received signals, weighting each in proportion to its signal-to-noise ratio, thus emphasizing the better-quality signals.

This technique, using two antennas, is already in use at the base station receiver for uplink traffic (that is, transmissions from mobile terminals to the base station). It is expected that future base stations may use four-antenna diversity and handsets, and mobile terminals will also employ multiple diversity antennas. While less effective due to overall power limitations, diversity antennas may also be used in transmission.

Beam-forming antennas and future enhancements. Wireless systems can benefit greatly from using antennas that focus radio energy in the geographical direction of interest. In the past, the use of such beam-forming antennas was limited to fixed point-to-point installations. Modern electronics make it possible to coordinate the signals from a set of antenna elements to generate, and even steer, a beam in the

desired direction, by properly phasing the signals associated with each antenna element. Thus, beam-forming techniques can be used in cellular systems. The beam can be fixed or electronically steered to track user movements. The beam can be configured in reception as well and in transmission. Most benefit will be observed when the signal directions are constrained to a narrow angle. Thus, beam-forming techniques are, in general, better suited for base station sites high above the surrounding clutter, rather than at the user end, where signals may appear to be arriving from multiple objects reflecting and diffracting the signals in almost all directions.

Spatial-division multiplexing. With sufficient antenna elements, it is possible to form more than a single beam. In this case, multiple messages intended for users in different directions can be associated with correspondingly many beams, generated by the appropriate beam-forming weights for multiple beams in each of the desired directions. By judiciously choosing users that are placed in direction separated by at least the beam width, it is possible to deliver simultaneous messages to all these users without interference. Thus, a system capacity increase factor equal to the number of simultaneous beams may be achieved.

MIMO systems. Link capacity can be greatly increased with a promising technique using multiple antennas at the transmitter and the receiver, known as multiple-input, multiple-output (MIMO). In MIMO systems, the transmitter sends multiple simultaneous messages in the same channel bandwidth without increasing the transmitted power. In a scattering environment, the receiver can process the signals at multiple receive antennas and separate the multiple transmitted messages, thereby creating multiple parallel virtual channels. Hence the link capacity is increased in direct proportion to the number of the multiple virtual channels. In this case, most of the processing is carried out at the receiver. For cost reasons, it is attractive in the downlink to shift the processing burden to the base station transmitter. In this case, the terminal feeds back the propagation channel estimates to the base station. With this information, the base station can compute generalized beam-forming weights to ensure that, at the receiver antenna locations, the multiple messages are received without any significant interference between the multiple transmitted messages. In this case, it is even possible to physically separate the receiving antennas and use this technique for the case in which the (multiple-antenna) base station is sending multiple messages to separate, single-antenna users. Furthermore, with this technique, essentially no changes are needed at the mobile user terminal, making it possible to use unmodified legacy terminals. This technique is sometimes also described as generalized beamforming.

Satellite telephony. There are many scenarios where mobile communications services are needed, yet no network infrastructure, such as base station transceivers, is available. In this case, satellite-based mobile services are extremely valuable. There are

three types of satellite-based mobile systems, depending on the orbit in which the satellite base transponder is placed. The first type of system utilizes geostationary Earth orbit (GEO) satellites, where the orbit is chosen such that the satellite appears stationary with respect to Earth. This simplifies the Earth station system design. However, geostationary orbits are at such a large distance from Earth (about seven Earth radii) that large transmission delays are introduced. More recently, lower orbits have been used to reduce delay. These are used in medium Earth orbit (MEO) and low Earth orbit (LEO) types of systems. In these cases, the satellite is in movement with respect to the Earth station location. Hence, the system must provide for user terminal handoffs, as the serving satellite disappears below the horizon and another satellite appears above the horizon. In these systems, many satellites are needed to ensure coverage at all times. See COMMUNICATIONS SATELLITE.

Reinaldo A. Valenzuela

Bibliography. J. D. Gibson (ed.), *The Mobile Communications Handbook*, 2d ed., CRC Press, 1999; T. S. Rappaport, *Wireless Communications: Principles and Practice*, 2d ed., Prentice Hall, 2002; R. Steele and L. Hanzo (eds.), *Mobile Radio Communications*, 2d ed., Wiley, 1999.

Mode of vibration

A characteristic manner in which vibration occurs. In a freely vibrating system, oscillation is restricted to certain characteristic patterns of motion at certain characteristic frequencies; these motions are called normal modes of vibration. An ideal string, for example, can vibrate as a whole with a characteristic frequency defined in the equation below, where L

$$f = \frac{1}{2L} \sqrt{\frac{T}{m}}$$

is the length of string between rigid supports, T the tension, and m the mass per unit length of the string. Displacements of different parts of the string are governed by a characteristic shape function. Specifically, the motion of any part of the string is proportional to $\sin(\pi x/L) \sin(2\pi ft)$, where x is the distance of the part of the string from a fixed end and t is the time. This simplest kind of vibration is the first, or fundamental, mode of vibration of the string; its frequency is the fundamental frequency. All parts of the string vibrate with the same frequency, and move outward from equilibrium at the same time.

The string is also capable of vibrating in two segments, one of which goes outward from equilibrium in a positive direction at the same time as the other part goes outward in a negative direction, and conversely. Again, the motion of any part of the string can be described by the product of a space function by a sinusoidal function of time: $\sin(2\pi x/L) \sin(2 \times 2\pi ft)$. All parts of the string move together as a sinusoidal function of the time and at the same frequency; the space function governs the

motion in opposite directions. The frequency of the second mode of vibration is twice that of the first mode. Similarly, modes of higher order have frequencies that are integral multiples of the fundamental frequency.

Because the frequencies are in the ratios 1:2:3 . . . , the modes of vibration of an ideal string are properly called harmonics. Not all vibrating bodies have harmonic modes of vibration, however. The ideal drumhead, for example, vibrates freely with frequencies in the ratios 1:1.59:2.14:2.30 In fact, most real systems vibrating freely have modes of vibration whose frequencies are not exactly in the ratios of integers. See HARMONIC (PERIODIC PHENOMENA); VIBRATION.

Robert W. Young

Model theory

The body of knowledge that concerns the fundamental nature, function, development, and use of formal models in science and technology. In its most general sense, a model is a proxy. A model is one entity used to represent some other entity for some well-defined purpose. Examples of models include: (1) An idea (mental model), such as the internalized model of a person's relationships with the environment, used to guide behavior. (2) A picture or drawing (iconic model), such as a map used to record geological data, or a solids model used to design a machine component. (3) A verbal or written description (linguistic model), such as the protocol for a biological experiment or the transcript of a medical operation, used to guide and improve procedures. (4) A physical object (scale model, analog model, or prototype), such as a model airfoil used in the wind-tunnel testing of a new aircraft design, or an electronic circuit used to simulate the neural activity of the brain. (5) A system of equations and logical expressions (mathematical model or computer simulation), such as the mass- and energy-balance equations that predict the end products of a chemical reaction, or a computer program that simulates the flight of a space vehicle. Models are developed and used to help hypothesize, define, explore, understand, simulate, predict, design, or communicate some aspect of the original entity for which the model is a substitute.

Formal models are a mainstay of every scientific and technological discipline. Social and management scientists also make extensive use of models. Indeed, the theory of models and modeling cannot be divorced from broader philosophical issues that concern the origins, nature, methods, and limits of human knowledge (epistemology) and the means of rational inquiry (logic and the scientific method). Philosophical notions of causality are also central to modeling. See LOGIC; SCIENTIFIC METHODS.

Advantages of models. Models are so pervasive because a model is usually more accessible to study than the system modeled. Models typically are less costly and less time-consuming to construct and test. Changes in the structure of a model are easier to implement, and changes in the behavior of a model

are easier to isolate, understand, and communicate to others. A model can be used to achieve insight when direct experimentation with the actual system is too dangerous, disruptive, or demanding. A model can be used to answer questions about a system that has not yet been observed or built, or even one that cannot be observed or built with present technologies.

Role of model theory. Every discipline develops its own models and its own approach and techniques for studying these models. The specific models developed in different disciplines may differ in subject, form, and intended use. However, basic concepts such as model description, validation, simplification, and simulation are not unique to any particular discipline. Model theory seeks a formal logical and axiomatic understanding of the underlying concepts that are common to all modeling endeavors.

Mathematical models. General and mathematical systems theory have stimulated many of the important developments in model theory. This article outlines the concepts of model theory as these relate to mathematical and computer simulation models of systems. Mathematical models are particularly useful, because of the large body of mathematical theory and technique that exists for the study of logical expressions and the solution of equations. The power and accessibility of digital computers have increased the use and importance of mathematical models and computer simulation in all branches of modern science and technology. A great variety of programming languages and applications software are now available for modeling, computational analysis, and system simulation. Many applications that formerly relied on other types of models now also use mathematical models and computer implementations extensively. See DIGITAL COMPUTER; SIMULATION; SYSTEMS ANALYSIS; SYSTEMS ENGINEERING.

Systems as models of systems. A system is a collection of entities that interact. Every system has a boundary. Entities not included within the boundary are known as the environment of the system.

A system is characterized by a set of attributes, that is, a set of quantities which assume values. Static attributes have fixed values and are called the parameters of the system. Dynamic attributes can assume different values at different times and at different points in space and are known as the descriptor variables of the system.

Three kinds of descriptor variables are input variables, output variables, and state variables (Fig. 1). Input variables represent the action of the environment on the system. They may be represented by

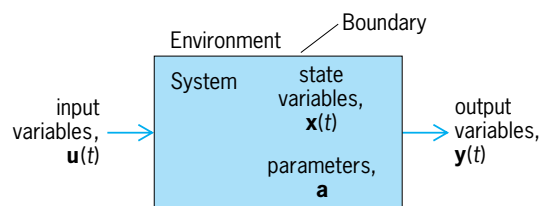


Fig. 1. Conceptual model of a system.

Eq. (1), which indicates that there are m input vari-

$$\mathbf{u}(t) = \{u_1(t), u_2(t), \dots, u_m(t)\} \in \mathbf{U} \quad (1)$$

ables, u_1, u_2, \dots, u_m ; that each depends on time t ; that collectively they form an m -component vector \mathbf{u} ; and that they are members of a range set \mathbf{U} . Output variables, Eq. (2), represent the action

$$\mathbf{y}(t) = \{y_1(t), y_2(t), \dots, y_q(t)\} \in \mathbf{Y} \quad (2)$$

of the system on the environment. State variables, Eq. (3), represent the internal condition of the sys-

$$\mathbf{x}(t) = \{x_1(t), x_2(t), \dots, x_n(t)\} \in \mathbf{X} \quad (3)$$

tem. Changes in the input precipitate changes in the state, which in turn are observed as changes in the output. Changes in the output are assumed to have no direct effect on inputs. See CONTROL SYSTEMS; LINEAR SYSTEM ANALYSIS.

While the state of a system is unique, the state variables that describe this internal condition are not. Instead, the state variables are any (minimal) set of variables that has the following two properties:

1. Knowledge of the values of the state variables $\mathbf{x}(t_i)$ at any instant of time t_i is sufficient to determine uniquely the values of any possible set of output variables $\mathbf{y}(t_i)$ at that same instant of time. That is, there exists an output function $\lambda: \mathbf{X} \rightarrow \mathbf{Y}$ defined for all t_i in the time base that maps current states into current outputs.

2. Knowledge of the values of the state variables $\mathbf{x}(t_i)$ at the time t_i , together with knowledge of values of the inputs $\mathbf{u}(t)$ over the time interval $[t_i, t_j]$, $t_i \leq t \leq t_j$, is sufficient to determine uniquely the values of the state variables $\mathbf{x}(t_j)$ at time t_j . That is, there exists a state transition function $\delta: (\mathbf{X}, \mathbf{U}) \rightarrow \mathbf{X}$ defined for all $[t_i, t_j]$ in the time base that maps current states and inputs into future states. (This is known as the Markov property.)

The pattern of changes exhibited by the state (or output) variables in response to any particular input is known as a state (or output) trajectory of the system. The collection of all possible state (or output) trajectories, corresponding to all possible inputs, is known as the state (or output) behavior of the system. The behavior of the system is generated by the functions δ and λ , which depend on the system parameters and possibly on time and spatial location as well.

The relationships among the input, state, and output variables of a system that give rise to its behavior are known as the structure of the system. A system, therefore, can be completely defined by the algebraic structure $\langle \mathbf{U}, \mathbf{X}, \mathbf{Y}, \delta, \lambda, t \rangle$, where \mathbf{U}, \mathbf{X} , and \mathbf{Y} are understood to be the range sets of the input, state, and output variables.

If a system S has the structure $\langle \mathbf{U}, \mathbf{X}, \mathbf{Y}, \delta, \lambda, t \rangle$, then a model of S is just some other system S' with structure $\langle \mathbf{U}', \mathbf{X}', \mathbf{Y}', \delta', \lambda', t' \rangle$. The system S' is used as a substitute for S for some specific purpose. Model theory deals with the nature and adequacy of the various

possible relationships between these two systems for the purpose at hand.

Modeling elements. Comprising five basic elements, a formal theory of modeling and simulation that builds on the ideas of systems was proposed.

Real system. This is the system modeled. It is simply a source of observable data, in the form of input/output pairs $(\mathbf{u}(t), \mathbf{y}(t))$. Typically, there are no other clues available to determine its structure.

Base model. This is the investigator's image or mental model of the real system. It is a system that is capable (at least hypothetically) of accounting for the complete behavior of the real system. If the real system is highly complex, the base model also is highly complex. The cost, time, and difficulty of realizing the base model explicitly most often are prohibitive, unwarranted, or impossible. As a practical matter, therefore, the structure of the base model is at best partially known to the investigator.

Experiment frame. This is the set of limited circumstances under which the real system is to be observed and understood for the purpose of the modeling exercise. It is a restricted subset of the observed output behaviors, as indicated in Eq. (4).

$$(\mathbf{u}(t), \mathbf{y}(t))_e \subseteq (\mathbf{u}(t), \mathbf{y}(t)) \quad (4)$$

Lumped model. This is the concept most often associated with the term model. It is a system which is capable of accounting for the output behavior of the real system, under the experiment frame of interest. It is an explicit simplification and partial realization of the base model. Its structure is completely known to the investigator.

Computer. This is the means by which the behavior of the lumped model is generated. The computer is not necessarily a digital computer. For simple models, it may represent the explicit analytical solution to the model equations, worked out by hand. For more complex models, however, the computer may need to generate individual trajectories step by step, based on instructions provided by the model. This step-by-step process is what is most often associated with the concept of simulation, and is usually conducted by using a digital computer.

Modeling relationships. Model theory addresses the relationships among modeling elements. Three modeling relationships are validation, simplification, and simulation.

Validation. This concerns the relationship between models and the real system. The objective of validation is to ensure that a model matches the system modeled, so that the conclusions drawn about the model are reasonable conclusions about the real system as well. A base model is valid to the extent that it faithfully reproduces the behavior of the real system in all experiment frames. On the other hand, a lumped model is valid to the extent that it faithfully matches the real system under the experimental frame for which it is defined. There can be many different lumped models that are valid, and a lumped model can be valid in one experiment frame and not another.

Model validation is a deep and difficult issue, and there are many different levels and interpretations of validity. For example, a model has face validity simply if it is accepted as reasonable for its intended purpose by people who are knowledgeable about the system under study. A model is replicatively valid if its trajectories match the real system input/output data used in its development. A model is predictively valid if its trajectories also match experimental data not used in its development. A model is structurally valid if it is behaviorally valid and also matches the structure of the base model, that is, if it generates behavior within its experiment frame in substantially the same way that the real system is believed to generate this behavior. There are also many other forms of validity.

Simplification. This concerns the relationships between a base model and its associated lumped models. The objective of simplification is to achieve the most efficient and effective lumped model that is valid within the experiment frame for which it is defined. Simplification can be achieved in many ways. These include dropping relatively insignificant descriptor variables and their associated structures, replacing deterministic variables and structures with random variables and their generating functions, coarsening the range set of descriptive variables, and aggregating descriptor variables and structures into larger blocks. Many of the formal ideas associated with simplification, such as isomorphism and homomorphism, concern the preservation of structural similarities between mathematical systems.

Simulation. This concerns the relationship between models and the computer. The objective of simulation is to ensure that the computer faithfully reproduces the behavior implied by the model. The behav-

ior of a lumped model must be distinguished from the correctness of its computer implementations or solutions, in the same way that the behavior of a real system must be distinguished from the validity of its models. While a valid model may have been developed, it is also necessary to have a correct simulation. Otherwise, the model solution cannot be used to draw conclusions about the real system. Formal ideas associated with simulation include the completeness, consistency, and ambiguity of the computer implementation. The process of matching a simulation to its lumped model sometimes is known as verification. Many of the same techniques used to validate models are also used to verify simulations.

Example. An example that illustrates the concepts of model theory involves the population changes of a region over time. Models of population dynamics are used to help assess future needs for housing, retail and commercial space, and schools and other public services, as well as to forecast revenues that will be available for these needs. The real system in this case might be a county, city, state, or nation, where census data are a source of information about past population trends. The base model comprises all that is known about how and why populations change, including factors that determine natural changes in the population (such as births, deaths, and aging) and influence migration among regions (such as economic and social conditions) [Fig. 2].

An experiment frame can be constructed to seek a very general understanding of population changes. The system boundary in this case corresponds to the geographic boundary of the region. A simple lumped model appropriate to this experiment frame might involve the following descriptor variables: total net migration into the region (the input), total regional

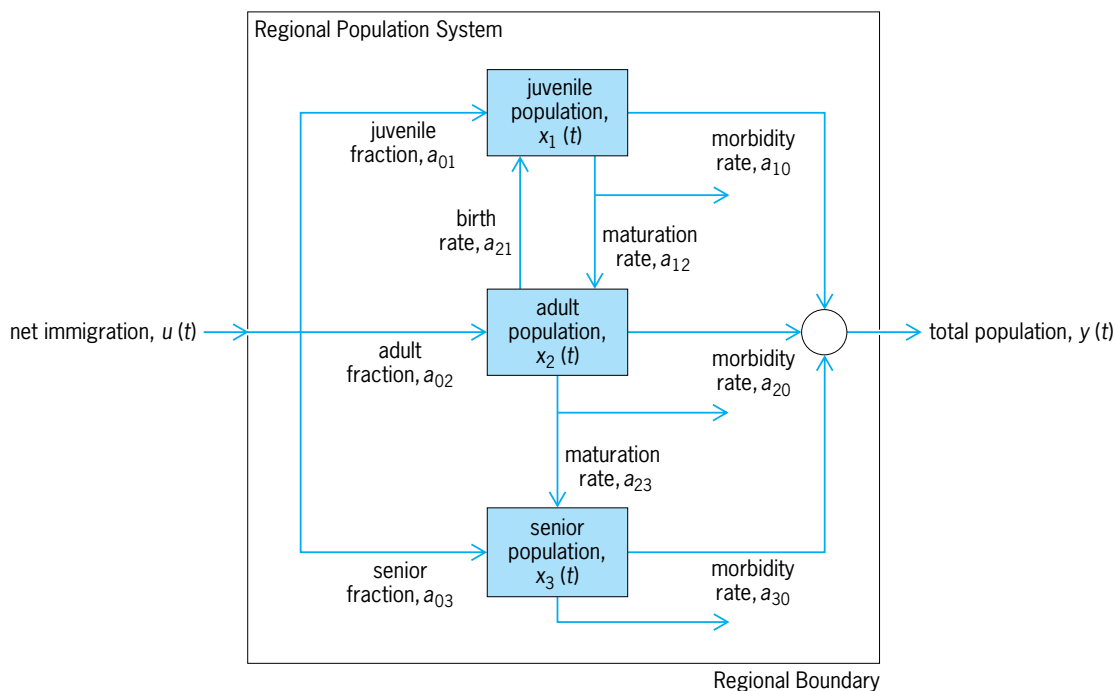


Fig. 2. Model of a regional population system.

population (the output), and the three state variables juvenile population, adult population, and senior population. The interactions among all these variables over time can be expressed in mathematical terms by a set of difference equations that represent a theory about how the state of the system (the values of the state variables) at a given time is determined by the state at some preceding time. The parameters of this model, whose values are determined from census data, reflect survival of the population within each age cohort, maturation from younger cohorts to older cohorts, births, and the age distribution of migrants. The output mapping defines the total population at any time to be the sum of all the population cohorts (juvenile, adult, and senior populations) at that time.

The model equations can be solved analytically, or a digital computer can be used to simulate the evolution of the state and output trajectories. Comparison of the model trajectories against actual census data would then provide a test of the validity of the model.

K. Preston White, Jr.

Bibliography. G. Gordon, *System Simulation*, 2d ed., 1978; N. A. Kheir (ed.), *Systems Modeling and Computer Simulation*, 2d ed., 1995; M. Kutz (ed.), *Mechanical Engineers' Handbook*, 2d ed., 1998; A. M. Law and W. D. Kelton, *Simulation Modeling and Analysis*, 3d ed., 1999; M. G. Singh (ed.), *Systems and Control Encyclopedia*, 1988; B. P. Zeigler, *Theory of Modelling and Simulation*, 1976, reprint 1984.

Modeling languages

Graphical or textual languages that are used to build models by describing the qualities with which the models are concerned. A model represents something and includes some but not all qualities of what it represents. For example, an architect's model of a new development area concentrates on the physical appearance, while an aircraft designer's wind-tunnel model concentrates on aerodynamic qualities. When models are used in systems and software engineering, their important qualities are structure, behavior, communication, and requirements.

Need for models. Systems and software engineering are both characterized by a combination of large information content with a need for different people with different backgrounds and interests (system stakeholders) to understand at least some system aspects. During the development of systems, developers can see only a small part of the complete system at a time. They will then need models to assist them in understanding the context of their own work and the requirements on it. Models are useful prior to system development to study requirements and design alternatives, and to ensure that all stakeholders have a common understanding of why a system should be built or modified. They are also useful during and after development of a system to assist developers and users to navigate the system's structure and to understand the system's behavior. Finally,

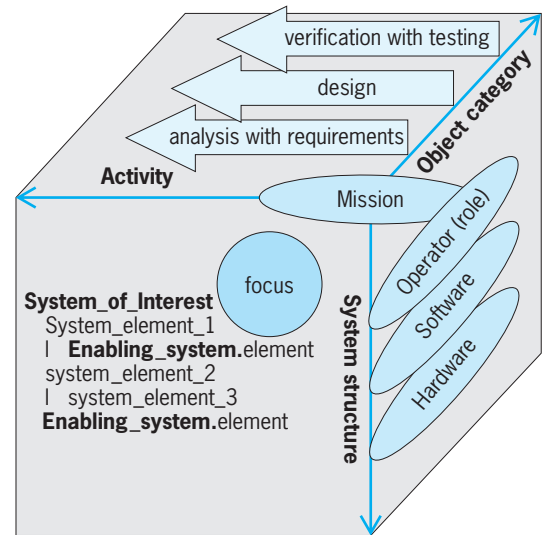


Fig. 1. Systems engineering work space.

models are essential for communication between different participants in projects and as a “corporate project memory.” See MODEL THEORY.

Systems and software engineering. Systems engineers work with a collection of operator roles (people), software modules, and hardware modules, which are combined to complete one or more missions. Thus, they work in a systems engineering work space with three dimensions (Fig. 1): the activity dimension (what they work with), the object category dimension (what kind of object they work with), and the structure dimension (where they work in the system). See OBJECT-ORIENTED PROGRAMMING.

It is normally not humanly possible to retain an overall understanding of the system together with a deep insight into the system's details, so there is a need to focus somewhere in this space. A complete system model will then be useful to navigate the system and to ensure that the part that is focused on is consistent with its environment.

Obviously, software is a subset of systems engineering. However, since building the right software requires a correct understanding of the software's environment, it is advisable to include this environment in software engineering models. See SOFTWARE ENGINEERING; SYSTEMS ENGINEERING.

Unified Modeling Language diagrams. The Unified Modeling Language (UML™) has emerged as a standard for software engineering. This standard includes a set of diagrams:

Use Case diagram. This diagram shows the interaction between a user and a software system. This is basically a dependency relation where the user depends on the system.

Class diagram. This diagram is basically an enhanced entity-relationship diagram where the entities are classes with names, attributes, and actions. A “software class” can be seen as a “template” from which instances are created to implement the software. The diagram includes relations such as generalization, association, and aggregation. The diagram can

be used for conceptual modeling, for specification, and for documenting a software solution.

Sequence diagram. Message sequence charts have a long tradition for visualizing the interaction between concurrent processes communicating by way of messages. These charts are included in the UML as Sequence diagrams.

Collaboration diagram. This diagram resembles a block diagram with its boxes and arrows. It basically contains the same information as the Sequence diagram.

Package diagram. This diagram shows dependencies between major system components.

State diagram. This diagram visualizes behavior.

Activity diagram. This diagram is a development from flow charts.

Deployment diagram. This diagram shows the distribution of software objects on hardware nodes.

Component diagram. This diagram shows dependencies between components and can also be drawn to show aggregation (components included in other components).

The qualities of a software system that are primarily modeled by the UML are requirements, communication, structure, and behavior. The table shows how the diagrams support different qualities.

Extension of UML to systems engineering. At present there is a trend in the systems engineering community toward extended use of the UML for systems engineering. This trend is based on the fact that the UML views software as composed of modules called “classes” or “objects” characterized by their having (1) an offered interface, which contains services offered by the object, (2) an internal behavior, and (3) a required interface, with services required from other objects to complete the internal behavior.

This view is not limited to software, but can be extended to systems engineering when one includes not only software objects but also hardware and operator objects. An example of this extended use is a UML Component diagram (Fig. 2) which models a simple system to control the windows in a car.

Figure 2 shows that the car driver requires services from a switch bank and a button to control the car’s windows. Further, these hardware control devices depend on software in the car’s central computer and in the car’s local computers, one in each door. The software in the local computer in the door depends

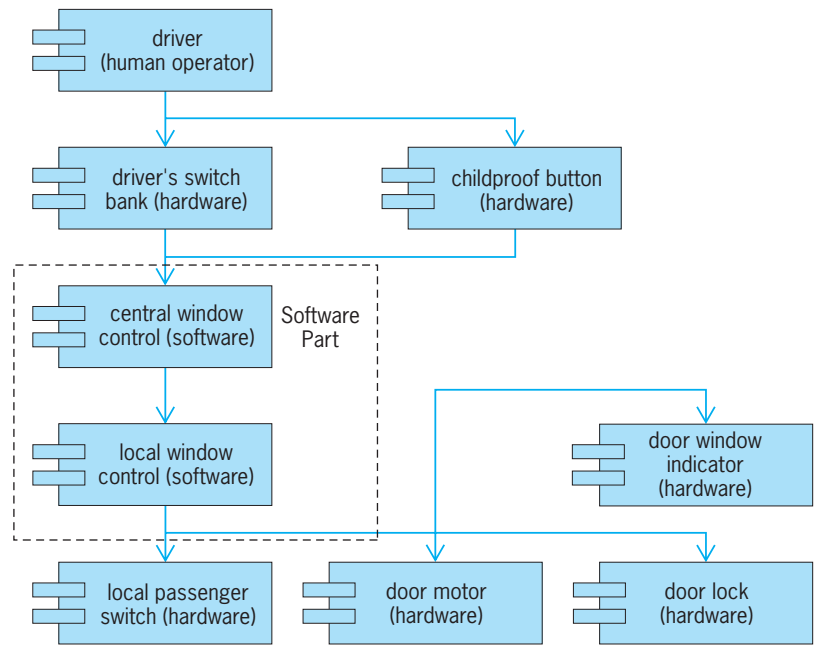


Fig. 2. Component diagram for car window control. The small rectangles signify that this is a component diagram.

in turn on a set of hardware devices in the door to complete the driver’s commands.

Other modeling languages. While the UML is the emerging standard for modeling languages, there are several alternatives, including mathematical formal languages, petri nets, and formalized natural language.

Mathematical formal languages are used in so-called formal methods, with the main advantage being that these languages can be used to create formal specifications, which can then be checked automatically for compliance with a completed software system. The main disadvantage is that the mathematical language used can be difficult to understand for some stakeholders.

Petri nets give a graphical model of how concurrent processes communicate and influence each other through the transmission of “tokens.”

A formalized natural language is obtained by simplifying a programming language for modeling use while keeping the “reserved words” and including variables of defined types. An example with part of the behavior for the software module “central window control” in Fig. 2 is:

| Usage of UML diagrams | | | | |
|-----------------------|---------------------|-----------|---------------|----------|
| Diagram | Qualities supported | | | |
| | Requirements | Structure | Communication | Behavior |
| Use Case | X | — | — | — |
| Class | — | X | — | — |
| Sequence | — | — | X | — |
| Collaboration | — | — | X | — |
| Package | — | X | — | — |
| State | — | — | — | X |
| Activity | — | — | — | X |
| Deployment | — | X | — | — |
| Component | — | X | — | — |

```

begin
  case direction_up is
    when true ==>
      send up_message (window_concerned)
      -- send message to move window
      -- upwards to the local control
      -- software for the concerned window in
      -- the car (driver;
      -- other_front, rear_left or rear_right)
    when false ==>
      send down_message (window_concerned)
    end case
  end
end
    
```

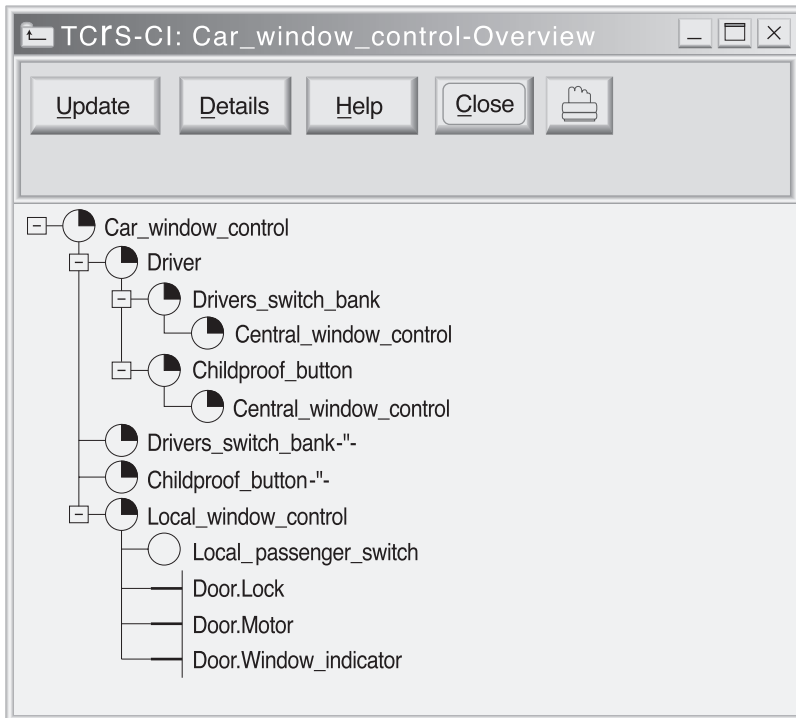


Fig. 3. Example of deep dependency structure.

See NATURAL LANGUAGE PROCESSING; PROGRAMMING LANGUAGES.

Tailoring of standards. The table shows how various diagrams in the UML address different aspects of modeling, and also shows that some aspects are covered by multiple diagrams. The obvious conclusion is that the standard needs tailoring prior to efficient use. Such tailoring should result in a set of modeling languages, some of which may be taken from within the UML, possibly modified, and some of which may be taken from outside the UML. Using a good set of modeling languages can be a blessing for a project, while careless selection, for example through uncritical acceptance of a “proven toolkit,” may result in confusion.

An example of tailoring concerns a systems engineering project with the following requirements on the modeling languages used:

1. The system’s missions shall be clearly seen.
2. Objects of categories Operator, Software, and Hardware shall be distinguished in the model.
3. The system’s “deep dependency structure” shall be clearly seen.
4. System structure and behavior shall be modeled.
5. Stakeholders, without previous systems or software engineering experience, shall be able to understand the models immediately.

It may seem that the UML cannot satisfy much of these requirements, but tailoring can help. The Component diagram in Fig. 2 really shows a dependency structure with objects of different categories. The diagram can be modified by keeping only the objects, ordered according to how they depend on each other, and introducing a new “Mission” object

at the top to state the mission “Car window control.” The result is the diagram in Fig. 3, which is known as the deep dependency structure of the system.

This modified Component diagram has only one component (object) on each line, and each component depends on the components below and one step indented. The “little clocks” indicate completion status for the components, and the quotation marks indicate repeated parts of the diagram, when an object appears a second time. The result is that a way has been found to satisfy the requirements for structural models.

The next step is to construct behavioral models, where the UML diagrams have difficulties with ignorant stakeholders. A possible solution is then to select a text-based modeling language, as described above. Thus, it is possible to find a limited set of modeling languages that can be used for communication between stakeholders with minimum education.

Ingmar Ogren

Bibliography. E. Best, R. Devillers, and M. Koutny, *Petri Net Algebra*, Springer-Verlag, 2001; M. Fowler, *UML Distilled*, 3d ed., Addison-Wesley, 2004; C. Larman, *Applying UML and Patterns*, 3d ed., Prentice Hall, 2005; I. Ogren, On principles for model-based systems engineering, *Sys. Eng.*, 13(1):38–49, 2000; I. Ogren, Possible tailoring of the UML for systems engineering purposes, *Sys. Eng.*, 3(4):212–224, 2000.

Modem

A device that converts the digital signals produced by terminals and computers into the analog signals that telephone circuits are designed to carry. Despite the availability of several all-digital transmission networks, the analog telephone network remains the most readily available facility for voice and data transmission. Since terminals and computers transmit data using digital signaling, whereas telephone circuits are designed to transmit analog signals used to convey human speech, a device is required to convert from one to the other in order to transmit data over telephone circuits. The term modem is a contraction of the two main functions of such a unit, modulation and demodulation. The device is also called a data set. See INTEGRATED SERVICES DIGITAL NETWORK (ISDN); MODULATION.

In its most basic form a modem consists of a power supply, transmitter, and receiver. The power supply provides the voltage necessary to operate the modem’s circuitry. The transmitter section contains a modulator as well as filtering, wave-shaping, and signal control circuitry that converts digital pulses (often input as a direct-current signal with one level representing a digital one and another level a digital zero) into analog, wave-shaped signals that can be transmitted over a telephone circuit. The receiver section contains a demodulator and associated circuitry that is used to reverse the modulation process by converting the received analog signals back into a series of digital pulses (Fig. 1). See DEMODULATOR;

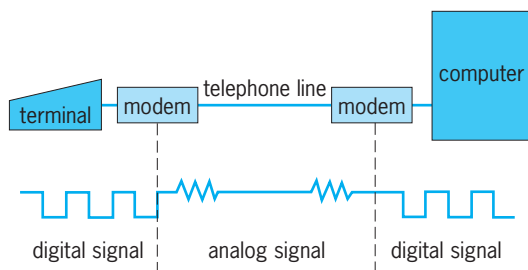


Fig. 1. Signal conversion performed by modems. A modem converts a digital signal to an analog tone (modulation) and reconverts the analog tone into its original digital signal (demodulation).

ELECTRIC FILTER; ELECTRONIC POWER SUPPLY; MODULATOR; WAVE-SHAPING CIRCUITS.

Handshaking. When a modem initiates a transmission sequence, it generates a tone known as a carrier signal. The carrier signal is simply a continuously repeating signal, which by itself conveys no information. However, it serves to inform the device at the opposite end of the telephone circuit that a connection between two modems has occurred. The modem that receives the carrier tone then generates its own carrier (at a different frequency from that of the originating device), which, when recognized by the original sending modem, indicates that a connection in both directions has been established. This process, commonly referred to as handshaking, is for full-duplex modems that are capable of transmitting and receiving data simultaneously. Half-duplex modems that can only send or only receive data at any particular instant switch their carrier signal off and on to reverse the direction of data flow. See CARRIER (COMMUNICATIONS).

Modulation process. The modulation process alters the characteristics of the repeating carrier signal and results in the impression of information onto the carrier. Normally, the carrier is a sine wave. The carrier's characteristics that can be altered include its amplitude for amplitude modulation (AM), its frequency for frequency modulation (FM), and its phase angle for phase modulation (PM or ϕM). In addition, a combination of amplitude and phase modulation is commonly employed to transmit high-speed data. See AMPLITUDE MODULATION; FREQUENCY MODULATION; PHASE MODULATION.

Amplitude modulation. Amplitude modulation involves varying the magnitude of the carrier signal from a zero level to represent a binary zero to a fixed peak-to-peak voltage level to represent a binary one (Fig. 2a). Amplitude modulation by itself is normally used for very low data rates. However, it is also employed in conjunction with phase modulation to obtain a method of modulating high-speed digital data source.

Frequency modulation. Frequency modulation refers to how frequently a signal repeats itself at a given amplitude. In frequency-shift keying (FSK), the transmitter shifts from one frequency to another as the input digital data changes from a binary one to a binary zero or from a zero to a one. This technique is used pri-

marily by low-speed modems operating at data rates up to 300 bits per second in a full-duplex mode and up to 1200 bits per second in a half-duplex mode (Fig. 2b).

Phase modulation. Phase is the position of a waveform of a signal with respect to the origination of the carrier cycle. Thus, phase modulation is the process of varying the carrier signal with respect to the origination of its cycle (Fig. 2c). Several forms of phase modulation are used in modems, including single- and multiple-bit phase-shift keying (PSK) and the combination of amplitude and multiple-bit phase-shift keying.

In single-bit phase-shift keying, the transmitter simply shifts the phase of the signal to represent each bit entering the modem. Thus, a binary one might be represented by a 90° phase change, while a zero bit could be represented by a 270° phase change; this technique is known as two-phase modulation.

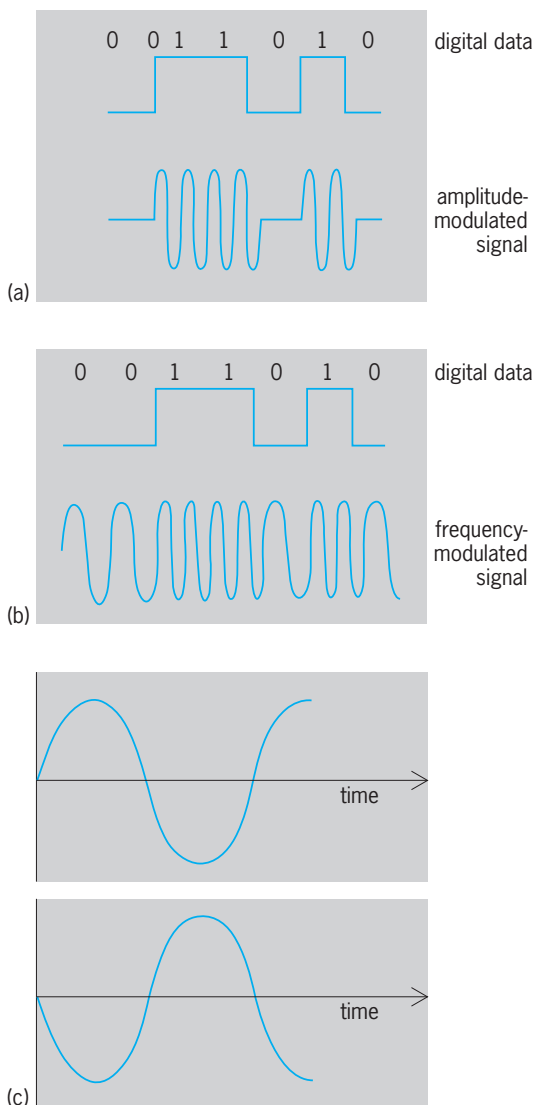


Fig. 2. Use of modulation to encode a digital data stream into analog signals. (a) Amplitude modulation. (b) Frequency-shift keying, a type of frequency modulation. (c) Phase modulation. The bottom wave is 180° out of phase with a normal sine wave illustrated at the top.

In multiple-bit phase-shift keying, two or more bits are grouped together and represented by one phase shift in a signal. Most modems that operate between 600 and 4800 bits per second use multiple-bit phase-shift keying modulation because this technique permits more information to be conveyed in each signal change than single-bit phase-shift keying.

Combined modulation techniques. One of the most frequently used combined modulation methods involves both amplitude and phase modulation. This technique, known as quadrature amplitude modulation (QAM), enables four bits to be represented by one amplitude and phase change. For example, if the modem varies its signal 2400 times a second and each signal represents four bits, then the modem has a data rate of 9600 bits per second.

The first implementation of QAM involved a combination of phase and amplitude modulation, in which 12 values of phase and 3 values of amplitude were employed to produce 16 possible signal states (Fig. 3). Most 9600 bit-per-second modems adhere to the Consultative Committee for International Telephone and Telegraph (CCITT) V.29 standard. The V.29 modem uses a 1700-Hz carrier that is varied in both phase and amplitude, resulting in 16 combinations of 8 phase angles and 4 amplitudes.

High-speed modems. The key to the ability to transfer information at high data rates (up to 115.2 kilobits per second) resulted from the use of echo cancellation, trellis coding, and data compression.

Echo cancellation. Echo cancellation permits a modem to operate in a full-duplex transmission mode on a two-line circuit, such as the switched telephone network. Under echo cancellation a modem establishes two high-speed channels in opposite directions. Through the use of digital-signal-processing circuitry, the modem's receiver uses the shape of the modem's transmitter signal to cancel out the effect of its own transmitted signal,

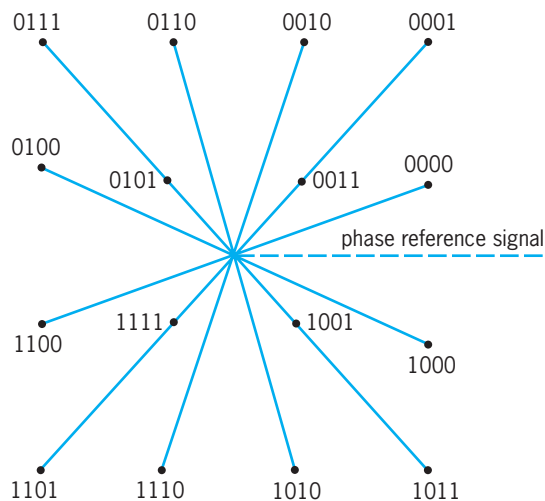


Fig. 3. Early version of quadrature amplitude modulation that produced 16 signal states from a combination of 12 phases (represented by angles of lines with respect to the phase reference signal) and 3 amplitude levels (represented by radial distances of points from the center).

enabling the modem to distinguish its sending signal from the signal being received.

The first modem to use echo cancellation to obtain full-duplex transmission on the switched telephone network at 9600 bits per second was the V.32 device. That modem was followed by the V.32bis (bis meaning the most updated version of the specification), which can operate at up to 14,400 bits per second, and the V.34 modem, which operates at data rates up to 28,800 bits per second. The key to the ability of each of those modems to transfer data at high data rates with extremely low error rates is their use of trellis coding.

Trellis coding. Trellis coding provides a higher tolerance to noise and other transmission impairments. When conventional QAM modems are used, a line impairment displaces the received signal from its appropriate location in the signal constellation. The modem's receiver is designed to select the signal point in the constellation closest to what it received. Hence, an error occurs when line impairments are large enough to cause the received point to be closer to a signal point that is different from the one transmitted.

Under trellis coding the modem's transmitter converts the serial data stream into groups of bits, adding one or more redundant bits to each group based upon the type of trellis coding employed at a particular modem operating rate. For example, a V.32bis modem will convert the serial data stream to be transmitted into 6-bit symbol groups and encode 2 of the 6 bits by using a binary convolutional encoding scheme which adds a code bit to the 2 selected input bits. See INFORMATION THEORY.

The redundancy introduced by the encoder means that only certain sequences of signal points are valid. Thus, if an impairment causes a signal point to be shifted, the receiver will compare the observed point to all valid points and select the valid signal point closest to the observed signal. As a result, a modem using trellis coding is only half as susceptible as a conventional QAM modem to an equal amount of noise power, and the resulting error rate is approximately two to three orders of magnitude below that of a QAM modem.

The V.34 modem employs a very complex form of trellis coding, using a 64-state, four-dimensional scheme to make more efficient use of the constellation space, while achieving an error rate below that of conventional 9600-bit-per-second QAM modems.

Data compression. Data-compression technology gained widespread acceptance with the promulgation of the V.42bis standard which defines the use of a modified version of the Lempel-Ziv method. Lempel-Ziv compression is a string-based compression technique in which strings are stored in a dictionary and numeric codes assigned to each string are transmitted in place of the string. Under the V.42bis standard, the algorithm used to develop a dictionary of strings, the number of entries in the dictionary, and its dynamic update are defined.

On the average, a fourfold reduction in data can be expected through the use of V.42bis compression.

This means, for example, that a V.34 modem with V.42bis compression, which transmits data at 28,800 bits per second, can achieve a data throughput of $28,800 \times 4$ or 115.2 kilobits per second. See DATA COMMUNICATIONS; DATA COMPRESSION; ELECTRICAL COMMUNICATIONS.

Gilbert Held

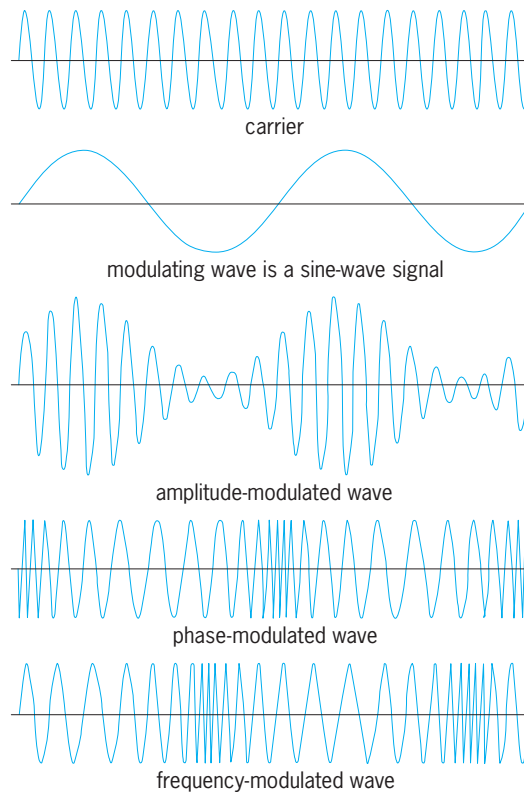
Bibliography. G. Held, *The Complete Modem Reference*, 3d ed., 1996; G. Held, *Data Communications Networking Devices*, 4th ed., 1998; G. Held, *Data and Image Compression: Tools and Techniques*, 4th ed., 1996; G. Held and R. Sarch, *Data Communications: A Comprehensive Approach*, 3d ed., 1995.

Modulation

A technique employed in telecommunications transmission systems whereby an electromagnetic signal (the modulating signal) is encoded into one or more of the characteristics of another signal (the carrier signal) to produce a third signal (the modulated signal), whose properties are matched to the characteristics of the medium over which it is to be transmitted. The encoding preserves the original modulating signal in that it can be recovered from the modulated signal at the receiver by the process of demodulation. The main purpose of modulation is to overcome any inherent incompatibilities between the electromagnetic properties of the modulating signal and those of the transmission medium. Of primary importance in this respect is the spectral distribution of power in the modulating signal relative to the passband of the medium. Modulation provides the means for shifting the power of the modulating signal to a part of the frequency spectrum where the medium's transmission characteristics, such as its attenuation, interference, and noise level, are favorable. See ELECTROMAGNETIC WAVE TRANSMISSION; RADIO-WAVE PROPAGATION.

Two forms of modulation are generally distinguished, although they have many properties in common: If the modulating signal's amplitude varies continuously with time, it is said to be an analog signal and the modulation is referred to as analog. In the case where the modulating signal may vary its amplitude only between a finite number of values and the change may occur only at discrete moments in time, the modulating signal is said to be a digital signal and the modulation is referred to as digital.

In most applications of modulation the carrier signal is a sine wave, which is completely characterized by its amplitude, its frequency, and its phase relative to some point in time. Modulating the carrier then amounts to varying one or more of these parameters in direct proportion to the amplitude of the modulating signal. In analog modulation systems, varying the amplitude, frequency, or phase of the carrier signal results in amplitude modulation (AM), frequency modulation (FM), or phase modulation (PM), respectively. Since the frequency of a sine wave expressed in radians per second equals the derivative of its phase, frequency modulation and phase modulation are sometimes subsumed under



Amplitude, phase, and frequency modulation of a sine-wave carrier by a sine-wave signal. (After H. S. Black, *Modulation Theory*, Van Nostrand, 1953)

the general term "angle modulation" or "exponential modulation." The **illustration** shows an example of an unmodulated sine-wave carrier signal and the signal resulting from modulating its amplitude, phase, or frequency with the amplitude of an analog modulating signal, which is also taken to be a sine wave.

Digital modulation. If the modulating signal is digital, the modulation is termed amplitude-shift keying (ASK), frequency-shift keying (FSK), or phase-shift keying (PSK), since in this case the discrete amplitudes of the digital signal can be said to shift the parameter of the carrier signal between a finite number of values. For a modulating signal with only two amplitudes, "binary" is sometimes added before these terms. In a particular form of binary ASK, known as on-off keying (OOK), the carrier signal is turned on or off in accordance with the two amplitudes of the modulating signal.

Digital modulating signals with more than two amplitudes are sometimes encoded into both the amplitude and phase of the carrier signal. For example, if the amplitude of the modulating signal can vary between four different values, each such value can be encoded as a combination of one of two amplitudes and one of two phases of the carrier signal.

Pulse modulation. In certain applications of modulation the carrier signal, rather than being a sine wave, consists of a sequence of electromagnetic pulses of constant amplitude and time duration, which occur at regular points in time. Changing one

or the other of these parameters gives rise to three modulation schemes known as pulse-position modulation (PPM), pulse-duration modulation (PDM), and pulse-amplitude modulation (PAM), in which the time of occurrence of a pulse relative to its nominal occurrence, the time duration of a pulse, or its amplitude are determined by the amplitude of the modulating signal. PAM and PPM are thus in many respects analogous to amplitude modulation and phase modulation. Another form of this type of modulation is pulse-frequency modulation (PFM), in which the rate at which the pulses are transmitted is varied in accordance with the modulating signal. PFM is analogous to frequency modulation of a carrier signal, with its instantaneous frequency being a continuous function of the modulating signal. *See PULSE MODULATION.*

Amplitude modulation. In this form of modulation, the amplitude of the sinusoidal carrier signal is varied in direct proportion to the amplitude of the modulating signal, with the frequency of the carrier signal held constant. If the modulating signal has a bandwidth of B hertz, so that the power in the modulating signal is distributed over the frequency range 0 to B hertz, the spectrum of the resulting modulated signal occupies the range from $f + B$ hertz to $f - B$ hertz, where f is the frequency of the carrier signal. The modulating signal power has therefore been shifted to a part of the spectrum determined by the frequency of the carrier signal, and the bandwidth of the modulated signal is twice the bandwidth of the modulating signal. The part of the spectrum from f to $f + B$ hertz is known as the upper sideband, and the part from f to $f - B$ hertz is known as the lower sideband. Either sideband contains all the information required to recover the original modulating signal. *See AMPLITUDE MODULATOR; SIDEBAND.*

Three special forms of amplitude modulation are frequently employed. In double-sideband, suppressed-carrier modulation (DSB-SC), the power contained in the unmodulated carrier is eliminated prior to transmission of the modulated signal. In single-sideband modulation (SSB), only one of the sidebands is transmitted. In vestigial-sideband modulation (VSB), one complete sideband and a fraction of the other sideband are transmitted. *See SINGLE SIDEBAND.*

Demodulation of an AM signal, which requires the translation of the spectrum of the modulating signal to its original position, can take two different forms. Coherent demodulation, also known as synchronous detection, involves the multiplication of the modulated signal with a precise copy of the carrier signal and subsequent removal of any high-frequency power by filtering. Noncoherent demodulation, also known as envelope detection, involves a nonlinear operation such as the squaring of the modulated signal and subsequent low-pass filtering. *See AMPLITUDE MODULATION; AMPLITUDE-MODULATION DETECTOR.*

Angle modulation. Angle modulation rests on the definition of a generalized sinusoidal signal $s(t) = A \cos \theta(t)$, where $\theta(t)$ is its generalized angle, and the time derivative $\theta'(t) = \omega(t)$ is its instantaneous

angular frequency. Two forms of angle modulation can be considered.

In FM, the instantaneous angular frequency $\omega(t)$ is varied in direct proportion to the modulating signal $m(t)$, so that $\omega(t) = \omega_c + km(t)$. Here ω_c is the angular frequency of the unmodulated carrier and k is a constant. If m_p denotes the peak value of $m(t)$, then the maximum deviation of $\Delta\omega$ of ω_c produced by $m(t)$ is given by $\Delta\omega = km_p$.

In PM, the modulating signal $m(t)$ is used to vary the instantaneous phase of the carrier so that $\theta(t) = \omega_c t + km(t)$. Since in this case $\omega(t) = \theta'(t) = \omega_c + km'(t)$, it can be said that modulating the phase of a carrier with $m(t)$ is equivalent to modulating the frequency of the carrier with the time derivative of $m(t)$. The maximum frequency deviation is then $\Delta\omega = km'_p$, where m'_p is the maximum value of $m'(t)$.

For both FM and PM, the modulated signal $s(t)$ has constant amplitude A and constant average power of $A^2/2$, independent of $m(t)$. Its bandwidth W is in theory infinite but for practical purposes is usually calculated from Carson's rule, according to which $W = 2(\Delta f + B)$. Here B is the bandwidth of $m(t)$ or $m'(t)$, and $\Delta f = \Delta\omega/(2\pi)$ is the maximum change in the frequency of the carrier signal produced by $m(t)$ in the case of FM or by $m'(t)$ in the case of PM. Equivalently, by defining $\beta = \Delta f/B$ as the deviation ratio, Carson's Rule can be stated as $W = 2B(\beta + 1)$.

Two forms of FM and PM are normally considered: If β is very much less than 1, the modulation is said to be narrow-band and the bandwidth of $s(t)$ equals $2B$, independent of the peak value of $m(t)$ or $m'(t)$. If β is very much larger than 1, the modulation is referred to as wide-band. The bandwidth of $s(t)$ is then equal to $2\Delta f$. For FM the value of Δf is determined by the peak value of $m'(t)$. Since the latter depends strongly on the spectral distribution of power in $m(t)$, the bandwidth of wide-band PM differs significantly from that of wide-band FM in that it depends on both the amplitude and spectral distribution of $m(t)$.

FM demodulation involves the generation of a signal whose amplitude is linearly proportional to the frequency of the modulated signal. Devices that produce such a response are known as discriminators. Among the most prominent are slope detectors, balanced discriminators, especially the Foster-Seeley discriminator, zero-crossing detectors, and phase-lock loops. *See FREQUENCY-MODULATION DETECTOR.*

Because of the intimate relationship between phase and frequency modulation, any of the FM demodulators can also be employed for the demodulation of a phase-modulated signal, with the additional step of integrating the output of the frequency demodulator. *See FREQUENCY MODULATION; PHASE MODULATION; PHASE-MODULATION DETECTOR.*

Quadrature amplitude modulation (QAM). This is an example of a multilevel digital modulation technique in which the distinct amplitudes of a digital modulating signal are encoded into combinations of phases and amplitudes of the carrier signal. A popular form of this scheme is 16-QAM, in which the carrier signal assumes 12 phases 30° apart, and three amplitudes. Carrier signals with phases of 45° , 135° , 225° , or 315°

assume two of the amplitudes, and the remaining eight phases are assigned the third amplitude. The resulting 16 states of the carrier signal are related to 16 distinct amplitudes of the modulating signal, each of which may represent a group of four binary digits. The bandwidth of any QAM signal is always twice the bandwidth of the modulating signal. Demodulation of a QAM signal involves the measurement at the receiver of both the phase and amplitude of the received signal.

Optical modulation. As in the modulation of electrical carrier signals, the amplitude, frequency, or phase of an optical sinusoidal carrier signal, typically generated by a light-emitting diode (LED) or a semiconductor laser, can be changed in accordance with the amplitude of the modulating signal. The modulating signal can also be used to change the carrier signal's polarization. For binary digital modulating signals the most common form of modulation employed is on-off keying, in which the carrier signal is shifted in power between two values. The ratio of these power levels is known as the extinction ratio. See LASER; LIGHT-EMITTING DIODE; POLARIZATION OF WAVES.

In so-called direct modulation schemes the injection current of the laser, which determines the frequency and power of the output signal, is varied directly in accordance with the modulating signal. This method may result in undesirable frequency shifts and other types of signal distortion. Much of this can be avoided through the use of external modulation schemes. Here the amplitude of the laser output is kept at a constant level, but is subsequently varied by inserting an attenuator between the laser and the fiber-optic transmission system whose level of attenuation is controlled by the modulating signal amplitude. External modulators are also used in effecting phase or frequency changes of the carrier signal for phase-modulation and frequency-modulation systems.

Several different methods for the demodulation of optical signals have been implemented. In direct-detection receivers, mostly used in conjunction with on-off keying, a photodetector converts received light energy into an electrical current or voltage, which is then compared to a threshold. Signal values above or below the threshold are decoded as one or the other amplitude of the original modulating signal. Several types of semiconductor photodiodes have been employed as photodetectors. In coherent receivers the received light, together with a monochromatic light signal generated by a local light source, is applied to the photodetector. The latter produces an output signal whose frequency is the difference in the frequency of the two input signals. Comparing this signal to a threshold completes the demodulation process. See OPTICAL COMMUNICATIONS; OPTICAL FIBERS; PHOTODIODE.

Performance of modulation systems. The quality of an analog modulation-demodulation system is normally determined by the ratio of signal power to noise power at the input to the demodulator and is measured by the same ratio at the output of the de-

modulator. For binary digital modulation systems the measure of quality is the bit error rate, the number of incorrect binary digits relative to the total number of binary digits at the output of the demodulator. In general, the performance of a modulation system depends critically on the amount of transmitted power, the bandwidth, the level of attenuation and signal distortion of the channel, the amount of interference present in the received signal, and the method of demodulation. See ELECTRICAL COMMUNICATIONS.

Hermann J. Helgert

Bibliography. P. E. Green, Jr., *Fiber Optic Networks*, Prentice Hall, 1993; F. Halsall, *Data Communications, Computer Networks and Open Systems*, 4th ed., Addison-Wesley, 1996; B. P. Lathi, *Modern Digital and Analog Communication Systems*, 3d ed., Oxford University Press, 1998; M. Schwartz, *Information Transmission, Modulation, and Noise*, 4th ed., McGraw-Hill, 1990.

Modulator

A device that combines an electrical information signal with a periodic electrical carrier signal for efficient transmission. The modification of a usually sinusoidal electrical carrier signal (or simply carrier) for the purpose of transmitting information which is carried on a second electrical modulating signal or information signal is called modulation. The device or circuit that performs the modulation is called a modulator. The information may be from any of a variety of sources such as sensor data, speech signals, or image data. The modulator varies the carrier in amplitude, frequency, phase, or some combination in order to imbed the information for efficient transmission. During the process of modulation, information is shifted in frequency from baseband to the carrier frequency for efficient transmission. See CARRIER (COMMUNICATIONS).

The primary classes of modulation are amplitude modulation (AM), frequency modulation (FM), and phase modulation (PM). Frequency and phase modulation are highly amplitude-noise resistant because the amplitudes of the received signal are forced to a constant value.

The information signals usually are of analog origin but may be converted to digital format for more exact processing. Transmission is always analog or continuous. Robust and versatile digital signal modulation of any type may be implemented directly with microprocessors or digital components (such as multipliers and adders) by mechanizing the modulation algorithms directly. The precisely controlled modulated digital signal is transformed into an analog signal by a digital-to-analog converter (DAC). Finally, by means of a linear amplifier, its power is increased to a level required for successful radio-frequency (RF) transmission. Because linear RF amplifiers are notoriously inefficient, an alternative lower-cost approach to AM and quadrature amplitude modulation (QAM) is used to first generate the baseband analog signals and then impress them on the RF carriers as described below.

High-quality data resolution is sacrificed in exchange for economy. See DIGITAL-TO-ANALOG CONVERTER.

Amplitude modulator. Amplitude modulation is subdivided into the classes double-sideband transmitted carrier (variously called DSTC, AMTC, or “standard” AM); the more complex but power-efficient double-sideband suppressed carrier (DSSC); the even more power- and bandwidth-efficient (and complex) single-sideband suppressed carrier (SSBSC or simply SSB); vestigial sideband (VSB), in which one sideband is only partially removed, leaving only a vestige of itself; and QAM, in which two AM signals are transmitted simultaneously on orthogonal carriers (that is, 90° out of phase with one another), which consequentially stay out of one another’s way while sharing a common spectrum. See AMPLITUDE MODULATION; SINGLE SIDEBAND.

A sinusoidal carrier signal is given by Eq. (1),

$$c(t) = \cos(2\pi f_c t) \quad (1)$$

where f_c is the carrier frequency in hertz. It is useful to consider the special case where this signal is amplitude-modulated by a sinusoidal information signal given by Eq. (2), where $A \geq 0$ and $B \geq 0$. If $A \geq$

$$s(t) = A + B \cdot \cos(2\pi f_i t) \quad (2)$$

B , the information may be recovered by simple envelope detection. Amplitude modulation is a multiplication process, where the transmitted signal is given by Eq. (3). The arguments of the trigonometric func-

$$\begin{aligned} T(t) &= c(t) \cdot s(t) = A \cdot \cos(2\pi f_c t) \\ &+ B \cdot \cos(2\pi f_i t) \cdot \cos(2\pi f_c t) = A \cdot \cos(2\pi f_c t) \\ &+ \frac{1}{2}B \cdot \cos[2\pi(f_c + f_i)t] \\ &+ \frac{1}{2}B \cdot \cos[2\pi(f_c - f_i)t] \end{aligned} \quad (3)$$

tions show that a standard AM or DSTC modulated signal has transmitted components at frequencies f_c , $f_c + f_i$, and $f_c - f_i$ with respective amplitudes of A , $\frac{1}{2}B$, and $\frac{1}{2}B$ (Fig. 1). One can show that this conclusion is also valid for general bandwidth-limited information signals. See DEMODULATOR.

If $A = B$, half of the transmitted power is in the carrier, which contains no information. If we set $A = 0$, the carrier is suppressed, transmitted power is reduced by 50% compared to DSTC, and a DSSC signal results. If the maximum information frequency contained in $s(t)$ is f_{\max} , then the bandwidth of the transmitted double-sideband signals is $2f_{\max}$. Trans-

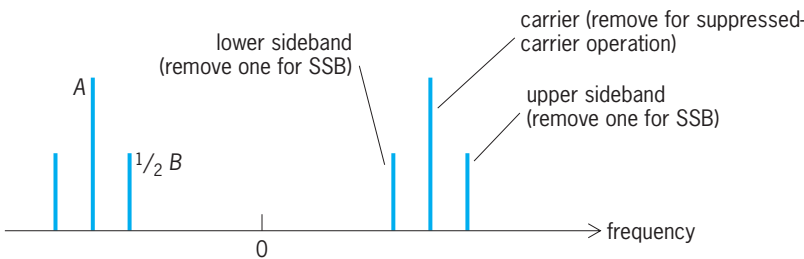


Fig. 1. Amplitude modulation (AM) spectrum.

mitting power may be halved again, and the bandwidth halved as well, by use of SSB, that is, by sending only one sideband after removing the other by linear band-pass filtering or by actively canceling the unwanted sideband. A little signal manipulation yields Eq. (4). By choosing the plus or minus sign, we can

$$\begin{aligned} \cos(2\pi f_c t) \cdot \cos(2\pi f_i t) -/+ \sin(2\pi f_c t) \cdot \sin(2\pi f_i t) \\ = \cos[2\pi(f_c +/- f_i)t] \end{aligned} \quad (4)$$

elect to generate and to transmit either the upper (+) or lower (–) sideband (Fig. 2). See ELECTRIC FILTER.

The multiplication function called for in simple and inexpensive analog amplitude modulation is achieved economically by varying, via the modulating signal, the applied power-supply voltage to the output RF amplifier that follows the carrier-frequency oscillator. This can be done by simply transformer-coupling the modulating signal voltage into the amplifier’s power-supply output line. See AMPLITUDE MODULATOR.

Quadrature amplitude modulator. A simple quadrature amplitude modulator can be implemented by inserting a second signal path from the reference carrier-frequency oscillator output consisting of a 90° phase shifter followed by a second linear amplifier whose power-supply output line is varied by a second transformer-coupled modulating signal voltage. A QAM signal is formed by passively summing the outputs of the two amplifiers delivering amplitude-modulated sine and cosine (quadrature) carriers.

By using analog or digital signals to amplitude-modulate each quadrature carrier, a QAM transmitter can generate an information-bearing carrier varying in both amplitude and phase. For digital signals, a so-called M -ary (that is, each data symbol can be represented by one of M possible states in the complex plane) digital QAM signal occupies the same bandwidth as a PSK signal (discussed below), but the power efficiency is greater. An M -ary digital QAM signal is generated by segmenting the input-signal data into M -bit lengths. These are divided into two sets, A and B, of $M/2$ bits each. A common segmentation approach is to steer the odd-numbered bits of the input-signal data into the A set and the even-numbered bits into the B set. This pair of $M/2$ -bit words is modified by complementing the most significant bit of each word and appending a “1” at the end to provide a pair of $M/2 + 1$ -bit words that will be treated as two’s complement signed numbers, passed through a pair of $M/2 + 1$ -bit digital-to-analog converters, and then multiplied (modulated) by $\cos(2\pi f_c t)$ for the A set and by $\sin(2\pi f_c t)$ for the B set. These two sets of analog information are then summed for transmission. See NUMBERING SYSTEMS.

This seemingly arcane exercise provides a rectangular M -by- M array of evenly spaced points in the complex plane, centered at the origin, that represent the M^2 possible data states. The case $M = 2$ is, of course, binary. The case $M = 4$ is robust and commonly used. As M increases, both the channel capacity and the sensitivity to noise similarly increase.

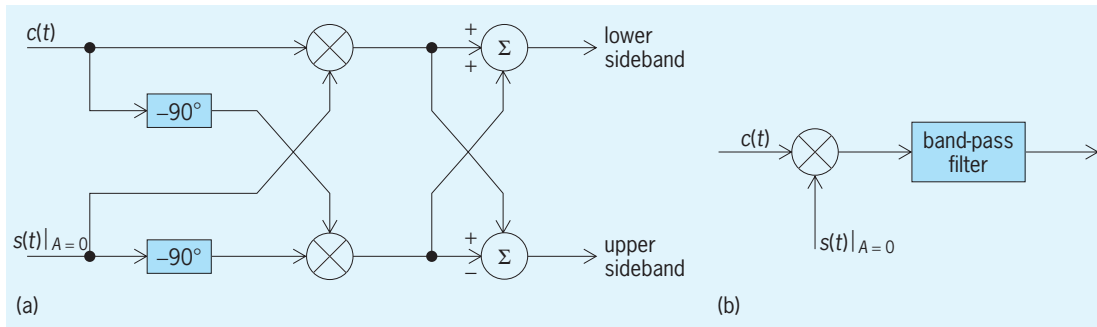


Fig. 2. Single-sideband generation by (a) phase shifting and (b) band-pass filtering.

Frequency modulator. Amplitude-modulated signals are degraded by additive noise, usually encountered along their propagation paths. This problem is overcome by transmitting information on constant-amplitude carriers and using frequency or phase modulation. The received signals are clipped to a constant amplitude, thereby removing the additive noise.

In frequency modulation, the constant-amplitude signal to be transmitted, given by Eq. (5), is varied in

$$T = \cos \left(2\pi \int_0^1 f_i(\tau) d\tau \right) \quad (5)$$

frequency. The instantaneous frequency is f_i where $f_i = f_c + k_{vf} \cdot s(t)$, f_c is the carrier frequency, $s(t)$ is the information signal, and k_{vf} is the voltage-to-frequency modulation index. If $|s(t)| \leq 1$, k_{vf} is also the peak deviation of the instantaneous frequency from f_c . Equation (5) demonstrates that a time-varying frequency must be integrated with respect to time, not just multiplied by time, in order to obtain phase. The Carson bandwidth (CBW), an approximation to the bandwidth within which virtually all energy lies, is $CBW = 2(k_{vf} + f_{max})$, where f_{max} is the highest frequency present in $s(t)$. See FREQUENCY MODULATION.

In a frequency modulator, the operating frequency of the carrier-frequency oscillator can be determined by the reactance of a tuned LC (inductor-capacitor) circuit. The oscillator frequency is determined by $1/\sqrt{LC}$, so varying either element in the oscillator's tuning circuit varies the frequency. A voltage-controlled capacitor called a varactor can be used in the circuit. The modulating voltage is applied directly to the varactor in order to produce an FM signal. See VARACTOR.

Transmission of digital data by frequency modulation is called frequency-shift keying (FSK). The frequency of the transmitted signal is determined by the digitized information signal. M-ary FSK employs M distinct frequencies. See FREQUENCY MODULATOR.

Phase modulator. In phase modulation, the signal to be transmitted is varied in phase according to the information signal, as given by Eq. (6), where k_{vp}

$$T(t) = \cos[2\pi f_c t + k_{vp} \cdot s(t)] \quad (6)$$

is a voltage-to-phase scaling constant. This can be

expanded as Eq. (7). If $|s(t)| \leq 1$, k_{vp} is the maxi-

$$T(t) = \cos(2\pi f_c t) \cdot \cos[k_{vp} \cdot s(t)] - \sin(2\pi f_c t) \cdot \sin[k_{vp} \cdot s(t)] \quad (7)$$

mum phase excursion due to the modulating signal. If k_{vp} is small, we have narrow-band phase modulation and the transmitted signal is approximated by Eq. (8), which can be mechanized (Fig. 3). Again,

$$T(t) = \cos(2\pi f_c t) - k_{vp} \cdot s(t) \cdot \sin(2\pi f_c t) \quad (8)$$

the CBW = $2(k_{pf} + f_{max})$. See PHASE MODULATION.

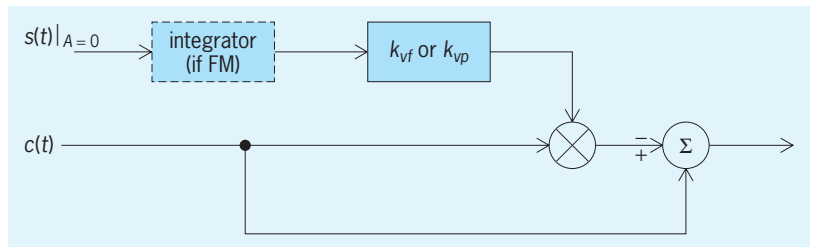


Fig. 3. Narrow-band frequency or phase modulator.

Transmission of digital data by phase-angle modulation is called phase-shift keying (PSK). PSK employs a single frequency. The phase of the transmitted signal is determined by the digitized information signal. Each data symbol in an M-ary PSK system may be represented by a point on a unit circle because the amplitude is held constant. The points are separated in angle (phase) by $360^\circ/M$. See MODULATION; PHASE MODULATOR. Stanley A. White

Bibliography. G. A. Breed (ed.), *Wireless Communications Handbook*, Cardiff Publishing, Englewood, CO, 1992; S. Gibilisco, *Handbook of Radio and Wireless Technology*, McGraw-Hill, New York, 1999; J. G. Proakis, *Digital Communications*, 4th ed., McGraw-Hill, New York, 2001; T. S. Rappaport, *Wireless Communications*, 2d ed., Prentice Hall, Upper Saddle River, NJ, 2002; M. S. Roden, *Analog and Digital Communications Systems*, 4th ed., Prentice Hall, Englewood Cliffs, NJ, 1995; M. E. Van Valkenburg and W. M. Middleton (eds.), *Reference Data for Engineers: Radio, Electronics, Computer and Communications*, 9th ed., Butterworth-Heinemann, Woburn, MA, 2002.

Mohair

The long, lustrous hair of the Angora goat, which originated in the area around Ankara (Angora), Turkey. This plateau, about 3000 ft (900 m) high, has a dry climate with extremes in temperature. The Angora goat probably originated before Biblical times, as references to the use of goat's hair are found in the Bible. Angora goats were brought into the United States from Turkey in 1849 and taken westward after the Civil War, especially to Texas and California.

Production and value. About half of the world mohair production is produced in the United States, followed by Turkey, South Africa, and Lesotho (Basutoland). Production in other parts of the world is insignificant. Generally, over half of all mohair produced is exported for processing and manufacturing. Principal importers are the European countries, the United Kingdom, and Japan.

Properties and uses. Mohair is a smooth, strong, durable, and resilient fiber. It enhances softness and luster in fabrics. Mohair absorbs dye evenly and brilliantly, retains color well, and permits unusual decorative effects. It is mainly used as an apparel fiber but may be used in upholstery, draperies, wigs, hairpieces, and rugs. Leather produced from the skin is useful for gloves, purses, and novelties.

Mohair fleeces may be classified by the different types of locks. The ringlet type is represented by long, tight locks or curls (Fig. 1). These fleeces have good length but may lack density, resulting in dirt penetration. Flat, wavy locks may be associated with light fleeces (Fig. 2). The web type resembles both the flat and ringlet (Fig. 3). It may be flat near the skin with a ringlet near the tip.

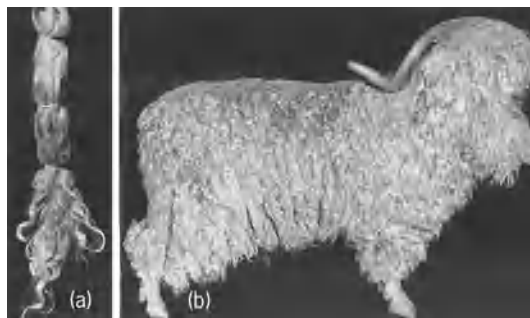


Fig. 1. Ringlet-type Angora: (a) lock from (b) buck. (From M. Camp, *Texas Sheep and Goat Raisers' Magazine*)

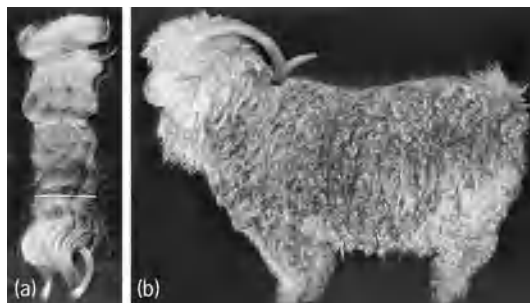


Fig. 2. Flat lock-type Angora: (a) lock from (b) buck. (From M. Camp, *Texas Sheep and Goat Raisers' Magazine*)

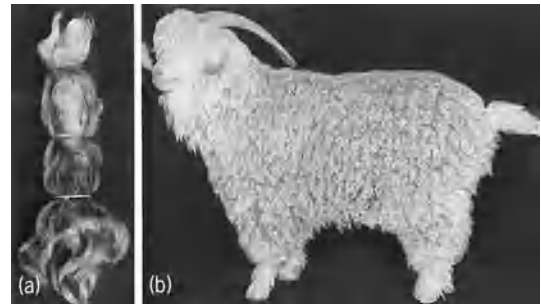


Fig. 3. Intermediate-type Angora: (a) lock from (b) buck. (From M. Camp, *Texas Sheep and Goat Raisers' Magazine*)

Growth characteristics. Mohair fiber follicles characteristically develop in groups, with connective tissue separating the groups. Three primary or first-appearing follicles are often found in each group. These are associated with an accessory sebaceous gland, a suboriferous gland, and an arrector muscle. Later-developing or secondary follicles generally lack accessory structures, although many have sebaceous glands. The number of secondary fibers per group varies from about 16 to 30. Medullated or hollow fibers are more likely to arise from primary follicles, but generally make up less than 1% of the fibers.

Structure. The mohair fiber is similar to wool in microscopic appearance, with epidermal scales about 20 micrometers long which barely overlap. The mohair fiber has a smooth appearance and is slightly wavy, with less than one crimp per inch. The mohair cortex is made up largely of ortho cells, while wool with its greater numbers of crimps has a cortex of both ortho and para cells. Mohair fibers are circular in cross section and range 10–90 μm in diameter; kid mohair ranges 10–40 μm in diameter. As goats grow older, up to about 8 years, mohair coarsens. Mohair grows at the rate of about 1 in. a month, resulting in a 6-in. (15-cm) staple for a 6-month clip. Large, chalky, brittle fibers known as kemp sometimes occur and have an undesirable effect on spinnability and dyeability. See WOOL.

Breeding practices. Angora goats may be improved by selecting for mohair and kid production. Weight of fleece, which is moderately heritable, deserves emphasis in selection. Individual fleece weights, especially for males, should be recorded. Length of staple is important as a quality factor and for its relationship to fleece weight. Fine mohair is preferred. Kemp or colored fibers are undesirable. Fleece density and good covering over the body and belly are preferred. Animals with a tendency to shed should be culled. Size, vigor, and fertility are important in both sexes. Body conformation is important since many goats are sold for meat. Clair E. Terrill

Moho (Mohorovičić discontinuity)

The level in the Earth where the velocity of sonic waves first increases rapidly or discontinuously to a value between 7.6 and 8.6 km/s (4.7 and 5.3 mi/s).

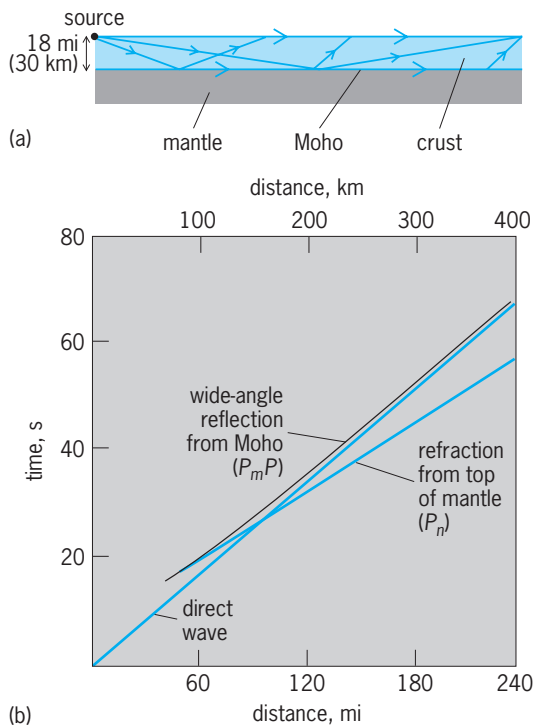


Fig. 1. Characteristics of waves traveling along the Moho. (a) Schematic cross section of the continental crust and upper mantle showing typical wave paths from a source of direct waves traveling near the surface of the crust, refracted waves traveling along the Moho (P_n), and wide-angle reflections from the Moho (P_mP). (b) Graph showing arrival times for various waves depicted in a. (After M. H. P. Bott, *The Interior of the Earth: Its Structure, Constitution and Evolution*, 2d ed., Elsevier, 1982)

A. Mohorovičić discovered this boundary while investigating seismograms of the Zagreb (now capital of Croatia) earthquake of October 8, 1909. He recognized that low-velocity waves traveling directly from the earthquake source were overtaken at large distances by refracted waves traveling through the deeper, high-velocity layer (Fig. 1). Modern determinations of the depth and nature of the Moho are commonly made in seismic refraction studies that use artificial seismic sources, such as explosions, rather than earthquakes. This method allows identification of the wave traveling in the high-velocity medium (P_n) and a wide-angle reflection (P_mP) from the boundary (Fig. 1). The Moho is generally assumed to mark the boundary between the crust and mantle, although this need not always be the case. Despite drilling attempts in the early 1960s, the Moho has not been directly sampled. Knowledge of the oceanic and continental Moho is based on interpretation of geophysical data, as well as geological interpretation of ophiolites (believed to be sections of oceanic crust and mantle uplifted onto land). See OPHIOLITE.

Formation. The oceanic Moho forms at spreading centers where partially molten material, which has the composition of basalt, both erupts and intrudes to form the oceanic crust, with mafic cumulates and sills forming the lower crust. The higher-density residual ultramafic material cools to form the

mantle beneath the crust. Much of the continental crust was formed in the early history of the Earth as low-density, silica-rich materials segregated from the mantle and moved to high levels. Because direct observations of the continental Moho are not available, the mechanics of this segregation process remain enigmatic. However, the continental Moho that formed during early crust formation may not be observable, because the continental Moho is a dynamic feature that can be modified by tectonic and volcanic processes. See EARTH CRUST; EARTH INTERIOR.

Continental Moho. Wide-angle and near-vertical reflections from the continental Moho suggest that it is not a first-order discontinuity but a complex zone several kilometers thick composed of interlayered crustal and mantle rock types. A band of strong reflections from the Moho is reported from a few regions, and in others the Moho appears to correspond to a transition from a highly reflective lower crust to a nonreflective upper mantle.

Moho depth, as determined by refraction experiments, varies considerably under the continents as evidenced by the Moho depth (crustal thickness) map of the United States (Fig. 2). Worldwide, Moho depths average between 35 and 40 km (22 and 24 mi) and range from 20 to 70 km (12 to 35 mi). These pronounced variations are generally related to tectonic processes. For instance, the Moho is shallow where extension creates thin crust such as in the Basin and Range Province in the western United States. The Moho is deep beneath fold and thrust belts where the crust is thickened by continental collision (for example, the Himalayas and Alps), and under continental magmatic arcs where volcanism and plutonism add new crustal material (for example, the Andes). These observations suggest that the Moho is a dynamic feature that changes its position and characteristics with time in response to tectonism. Certain regions, however, do not fit these generalities. The Moho under the Great Plains, for instance, is deeper than under the adjacent Rocky Mountain region. Significant Moho depth variations in Precambrian shield areas do not correlate with topography but with old tectonic features, indicating that the Moho in these regions represents an unmodified discontinuity established during the Precambrian. See GEODYNAMICS; VOLCANOLOGY.

The petrologic nature of the continental Moho is poorly known. Deep continental crust above the Moho is probably composed of a heterogeneous mixture of various metamorphic rocks and igneous rocks with sonic velocities less than 7.5 km/s (4.5 mi/s). Velocities of P_n range from 7.6 to 8.4 km/s (4.7 to 5.0 mi/s), values typical of peridotites, dunites, and eclogites. All these rock types are observed as xenoliths in diatremes, suggesting they are present in the uppermost mantle. In addition to variable upper-mantle composition, significant variation of pressure and temperature at the Moho under different continental regions can cause these lateral variations in velocity. There is evidence that upper-mantle velocities in several regions vary with propagation direction. This anisotropic behavior suggests that, in at least

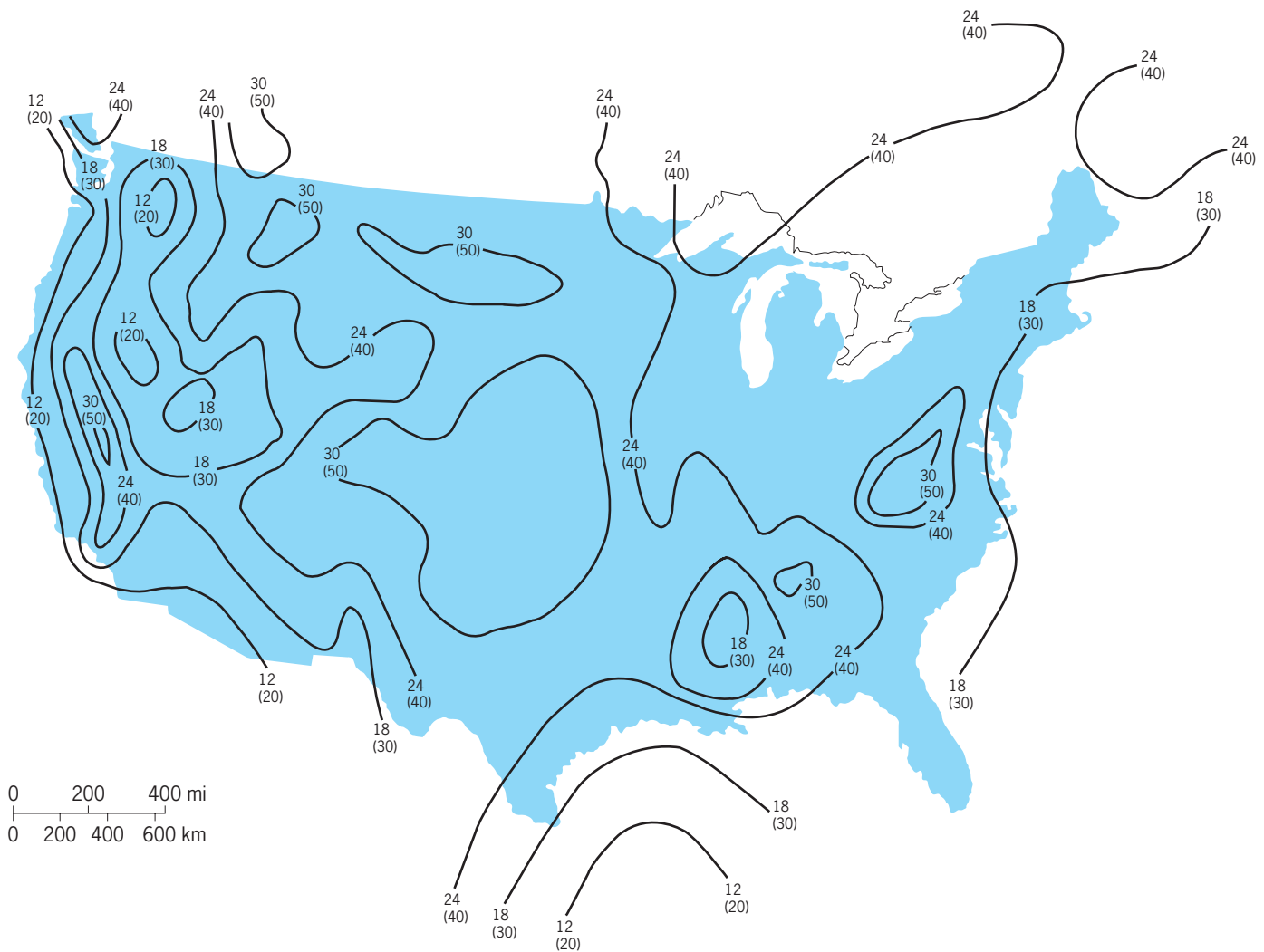


Fig. 2. Contour map of Moho depth in the United States. Contour values are given in miles (kilometers). (After R. J. Allenby and C. C. Schnetzler, *United States Crustal Thickness, Tectonophysics*, 93:13–31, 1983)

some regions, the subcontinental mantle is predominantly composed of peridotites with strong olivine alignment. The continental Moho in some places is probably a compositional boundary separating silica-rich igneous and metamorphic rocks of the crust from denser rock types, such as peridotites, within the mantle.

Results from a Canadian experiment known as LITHOPROBE suggest that four competing hypotheses for the origin of the continental Moho may apply in different regions and may not be mutually exclusive. These are a relict Moho (preserved from an original oceanic Moho); magmatic underplating (from sills); metamorphic front (phase transition); and regional décollement (structural detachment). See DUNITE; ECLOGITE; OLIVINE; PERIDOTITE; XENOLITH.

David M. Fountain; Maya Tolstoy

Oceanic Moho. While the definition of Moho in the continents is based solely on seismic refraction observations, the term has been utilized to describe a range of observations pertaining to the oceanic environment. In the strictest sense, Moho refers to the transition boundary where material velocities, as determined from seismic refraction methods, exceed

5 mi/s (8 km/s). Seismic studies conducted throughout the oceans since around 1950 reveal that the arrival of the Moho signal is best modeled as originating from a region (Moho transition zone) of relatively rapid velocity increase from about 4.5 mi/s (7.2 km/s; mafic lower crust) to about 5 mi/s (8 km/s; ultramafic upper mantle). On average, this transition begins at a depth of about 4.0 mi (6.5 km) beneath the top of the oceanic crust (ignoring variable thicknesses of sediments that may also be present), though considerable variability is observed, particularly in the slower-spreading environment. From geological studies of ocean crust and ophiolites (ancient oceanic sections subsequently emplaced on continents), oceanic crust is understood to originate from partial melting of mantle that upwells and decompresses in response to sea-floor spreading. The eruption and intrusion of this melt, and the formation of Moho, occurs in a very narrow zone at the axis of sea-floor spreading. Different mantle conditions can lead to different volumes of melting, accounting for most of the observed variations in crustal thickness (depth to the Moho). Very slow spreading centers, while rare, are expected to form somewhat thinner

crust, largely reflecting the magma-starved nature of these areas. The Moho is a proxy for the transition from crustal materials (for example, basalts, diabase dikes, or gabbros) formed by mafic melts extracted from the mantle to the ultramafic residual mantle (for example, peridotite) that has remained at depth. See BASALT; DOLERITE; GABBRO; PETROLOGY; PLATE TECTONICS.

The Moho transition zone can also be imaged by seismic reflection techniques. However, the reflection Moho may originate from the top of a transition zone and thus does not necessarily occur at the same depth as the true (refraction) Moho. Also, seismic reflection imaging is influenced by structural perturbations within the crust. For example, while refraction studies conducted around the relatively fast-spreading East Pacific Rise reveal Moho signal arrivals at a near-constant depth everywhere, reflection studies in the same area are unable to image the Moho in the vicinity of the spreading axis itself because of the presence of a low-velocity zone associated with a small crustal magma chamber perched atop very hot crust. Similarly, seismic reflection studies conducted in the crust formed at the relatively slow-spreading Mid-Atlantic Ridge environment rarely reveal a reflection Moho, even though refraction determinations of depth to Moho, that is, depth to residual mantle, are commonly indistinguishable from those observed in oceanic crust formed elsewhere. The reason is that slower sea-floor spreading is accompanied by a significantly greater degree of extensional faulting within the crust, as revealed by topographic roughness of the young crust, prominent intracrustal reflectors observed in older oceanic crust, and occurrences of hydrated residual mantle (serpentinized peridotite) in slow-spreading environments, particularly near fracture zones. During hydration (serpentinization), the intrinsic velocity of the rock decreases significantly such that the serpentinized peridotite no longer retains a Moho velocity. In such settings, the depth to the reflection Moho can change with time, depending on the depth to which hydration persists. See MAGMA.

Results from reflection work provide evidence for frozen magma lenses at and below the Moho supporting earlier refraction, compliance, and ophiolite work suggesting that lenses of gabbroic material at the base of the crust play a significant role in formation of the lower crust. It is thought that the presence or absence of these gabbroic lenses may explain some of the variability in Moho reflectivity within areas of similar topographic relief.

A usage encountered in many texts is petrologic Moho. While quite a departure from the original geophysical derivation of the term Moho, this term has been introduced to distinguish between the depth to the reflection Moho (originating within a Moho transition zone) and the depth to the true residual mantle (beneath any such transition zone); it is often used in descriptions of ophiolites. See IGNEOUS ROCKS; MARINE GEOLOGY; MID-OCEANIC RIDGE; SERPENTINE.

Carolyn Z. Mutter; Maya Tolstoy

Bibliography. D. L. Anderson, *Theory of the Earth*, 1989; K. Benn, A. Nicholas, and I. Reuber, Mantle-

crust transition zone and origin of wehrlitic magmas: Evidence from the Oman ophiolite, *Tectonophysics*, 151:75–85, 1988; W. E. Bonini and R. R. Bonini, Andrija Mohorovii: Seventy years ago an earthquake shook Zagreb, *EOS Trans. Amer. Geophys. Union*, 60(41):699–701, 1979; M. H. P. Bott, *The Interior of the Earth: Its Structure, Constitution and Evolution*, 2d ed., 1982; W. C. Crawford and S. C. Webb, Variations in the distribution of magma in the lower crust and at the Moho beneath the East Pacific Rise at 9°–10° N., *Earth Planet. Sci. Lett.*, 203:117–130, 2002; D. W. Eaton, Multi-genetic origin of the continental Moho: Insights from LITHOPROBE, *Terra Nova*, 18:34–43, 2006; C. M. Jarchow and G. A. Thompson, The nature of the Mohorovii discontinuity, *Annu. Rev. Earth Planet. Sci.*, 17:475–506, 1989; R. Meissner, T. Weaver, and E. R. Flüh, The Moho in Europe: Implications for crustal development, *Ann. Geophys.*, 5B:357–364, 1987; J. C. Mutter and J. A. Karson, Structural processes at slow spreading ridges, *Science*, 257:627–634, 1992; M. R. Nedimovic et al., Frozen magma lenses below the oceanic crust, *Nature*, 436:1149–1152, 2005; E. E. Vera et al., The structure of 0- to 0.2-m.y.-old oceanic crust at 9°N on the East Pacific Rise from expanded spread profiles, *J. Geophys. Res.*, 95:15529–15556, 1990.

Moiré pattern

When one family of curves is superposed on another family of curves, a new family called the moiré pattern appears. A familiar example of moiré is the pattern one sees on looking through the folds of a nylon curtain. When the curtain is moved slightly, the moiré pattern is observed to move wildly about. This effect raises at once the possibility of using the moiré effect to measure minute displacements. Moreover, the moiré patterns in the curtain are reminiscent of the complex patterns of waves seen at the shore of a lake. This suggests again that moiré might be used to describe in graphical form certain physical phenomena.

To produce moiré patterns, the lines of the overlapping figures must cross at an angle of less than about 45°. The moiré lines are then the locus of points of intersection. **Figure 1** illustrates the case of two identical figures of simple gratings of alternate black and white bars of equal spacing. When the figures are crossed at 90°, a checkerboard pattern with no moiré effect is seen. At crossing angles of less than 45°, however, one sees a moiré pattern of equispaced lines, the moiré fringes. The spacing of the fringes increases with decreasing crossing angle. This provides one with a simple method for measuring extremely small angles (down to 1 second of arc). As the angle of crossing approaches zero, the moiré fringes approach 90° with respect to the original figures.

Even when the spacing of the original figures are far below the resolution of the eye, the moiré fringes will still be readily seen. When two diffraction gratings of the transmitting type are superimposed and

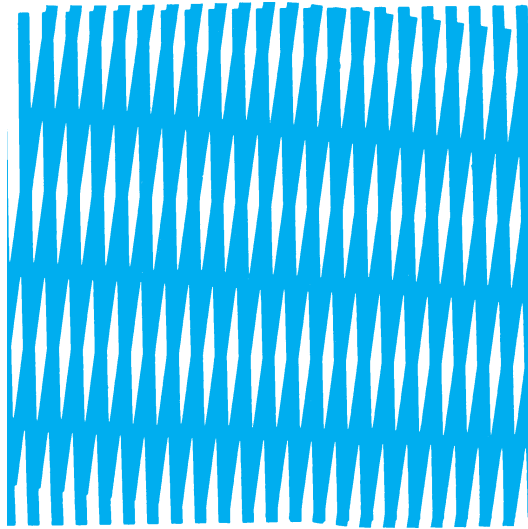


Fig. 1. Two simple gratings crossed at a small angle.

a distant light source viewed through them, moiré fringes can be seen by the unaided eye. This experiment, first performed by Lord Rayleigh in 1874, provides a means of checking the fidelity of a replica of a diffraction grating. If the repeat spacing of one grating differs slightly from that of the other, a beat pattern will be observed when the lines of the two figures are crossed at zero angle. The beat spacing is inversely proportional to the difference in spacings of the two gratings. Thus with nearly identical gratings a slight relative displacement will cause many fringes to pass by. This beat effect is the same as that seen when one is moving past two repetitive structures, such as the railings of a bridge; the separation between the posts of the railing closer to the observer appears to be slightly greater than that of the farther railing. As a consequence, one sees a beat when a post of the nearer railing gets in step (or in phase) with an open space of the farther railing.

The moiré-beat method, using diffraction gratings, has been applied industrially to automotive precision machinery, the fringes being detected by a photocell and the electric pulses so produced fed into a computer. See DIFFRACTION GRATING.

Industrial applications. Moiré techniques are used in the stress analysis of metals. One procedure is to produce photographically a copy of a grating on the surface of the material under examination and to view it through the master grating. In the unstressed condition no pattern is observed (that is, the moiré fringes are at infinity). When the sample is subjected to elongation or shear, a moiré pattern appears that can be readily interpreted in terms of the strain distribution. See PHOTOELASTICITY.

The degree of flatness of a surface can be determined by the moiré technique. The shadow of a grating on the surface serves as the second grating. If the surface is not perfectly flat, the moiré pattern observed no longer consists of parallel lines, and from the pattern a contour map of the surface is obtained. In this manner the examination of large optical surfaces can be carried out.

Any lens alters the apparent spacings of a grating. If now this altered image is viewed through another grating, the moiré pattern obtained is determined by the focal length of the lens. The aberrations of the lens become apparent, since the fringes are no longer straight and equidistant. When a periodic grating is viewed through a refractive index gradient (for example, sugar molecules diffusing into water), the spacings of the grating are modified. On overlaying this figure with another periodic grating, a moiré pattern is obtained which gives directly the refractive index gradient curve. See ABERRATION (OPTICS).

Phenomena in physics. Many concepts in physics may be demonstrated with a moiré kit, which consists of families of curves printed on transparent plastic. A simple grating of black and white bars can be regarded as the representation of a plane periodic wave, whose wavelength is given by the length of the periodic element of the grating. The moiré pattern of Fig. 1 represents the interference (or superposition) of two plane monochromatic waves. This occurs, for example, in x-ray diffraction; the interfringe distance corresponds to the reflecting planes of atoms, and Bragg's law is satisfied. The figure also represents other situations in which two plane waves are crossing, such as can occur with reflection at a seawall, in holography, or in a microwave cavity. The rays (perpendicular to the wavefront) satisfy the condition of reflection, namely, that the angle of incidence equals the angle of reflection. See X-RAY DIFFRACTION.

Cylindrical or spherical waves are represented by a figure consisting of equispaced concentric circles. Superposing two such figures gives a moiré pattern (Fig. 2) consisting of hyperbolas, and in the central portion a family of ellipses. The hyperbolas give the location of the fringes seen in Thomas Young's famous experiment, in which he illuminated two fine slits with a single source of light. The family of ellipses represents the standing-wave modes in an elliptical cavity. It would represent, for example, the sound patterns in a so-called whispering gallery,

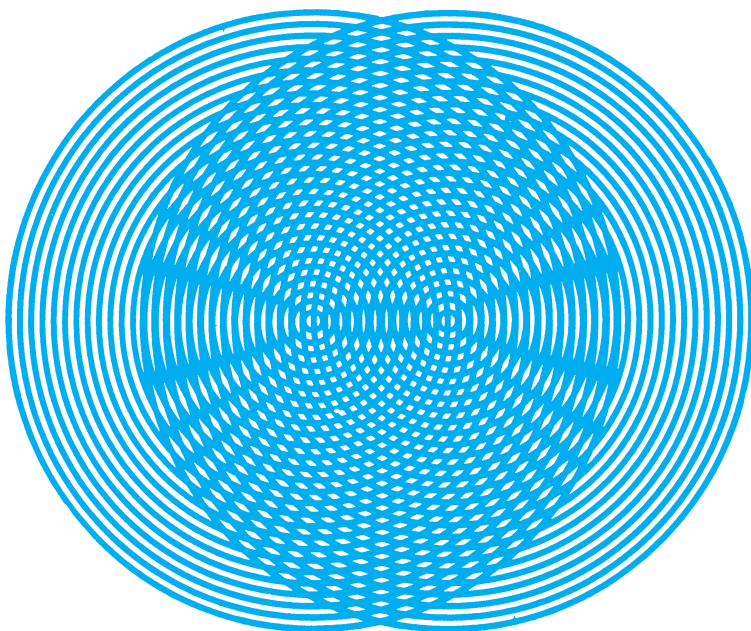


Fig. 2. Superposition of two families of equispaced concentric circles.

where one speaks at one focus of the room and the voice is heard clearly at the other focus.

Diffraction of light by a straightedge can be represented by the moiré pattern produced by the superposition of a straight-line grating and a circular grating; each point of the straightedge becomes a source of cylindrical waves. The moiré pattern is the interference pattern, or hologram, of a point scatterer (a speck of dust, for example). Indeed, the hologram obtained for more complicated objects may be regarded as the moiré pattern produced by the superposition of many such simple figures. See HOLOGRAPHY.

The moiré technique can also be applied to problems of fields and flows. Thus in hydrodynamics the flow of liquid from a source may be represented by a figure of radiating lines; the greater the source strength, the smaller will be the angular openings. As no positive or negative radial direction is indicated, the figure can also be used to represent a sink. Two such figures when superposed give a moiré pattern, which is the resultant flow of liquid for a source near a sink. This is also the field for an electric dipole, a positive charge close to a negative charge. When a simple parallel-line grating is superposed on a radial figure, one obtains the flow pattern for a source of liquid in a moving stream. This result, which is ordinarily difficult to calculate, is readily obtained and in graphic form by the moiré method. See HYDRODYNAMICS.

The vortex, or circular motion, of fluids is represented by circles in which the spacing is greater as one goes out from the center. Superposition of two such figures gives the resulting flow for two vortex centers of opposite sense (Fig. 3). The moiré pattern also represents the lines of the same voltage (the equipotentials) looking down the ends of two wires which are oppositely charged. It should be noticed that the pattern is turned 90° from that for the radial figures. By superposing this circular figure with other figures, one can obtain the flow of air about an airplane wing, for example. See AERODYNAMIC WAVE DRAG; VORTEX.

Mathematical solutions. The moiré technique offers a graphical method for the representation and solution of problems in many branches of mathematics. For example, the problems just described are essen-

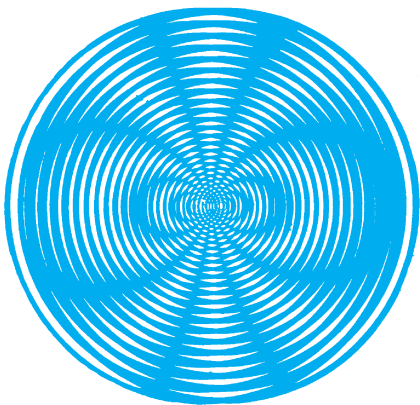


Fig. 3. Moiré pattern produced by superposition of two figures representing vortex or circular motion.

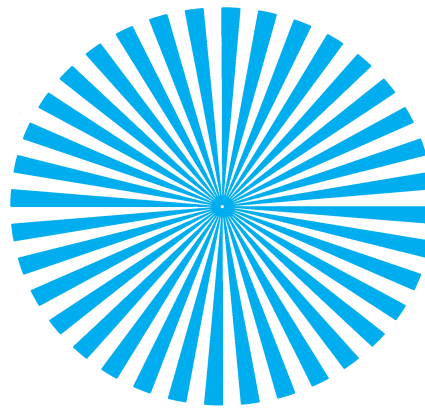


Fig. 4. Radial pattern of 10° openings.

tially the solutions of Laplace's differential equation in two dimensions. Consider a simpler case, namely, conic sections. A family of curves represents a contour map of some surface. Thus the simple grating represents the level lines for a plane; the more inclined the plane, the closer are the lines. The figure of equispaced circles is the contour map of a cone. Combining a cone map with that of a plane gives the conic section. Superposing a simple grating with a circular grating of the same spacing gives a family of parabolas as the moiré pattern. If the spacings of the straight-line grating are smaller (that is, the inclination of the plane is higher), one obtains hyperbolas; if the spacings are larger, one gets ellipses.

Psychological phenomena. The moiré phenomenon was known to the Chinese in ancient times. Indeed, they produced a luxurious fabric, moiré, or watered silk. It is made by doubling over a glossy, corded silk cloth and pressing the facing surfaces together under steam pressure. When the fabric is unfolded, the ribbed silk with its embossed image superposed creates a moiré pattern. In very recent years moiré has been developed into an art form, and constructions consisting of superposed but separated screens have been exhibited in art museums. The effect on the viewer is one of dynamic motion, because any slight movement of the viewer (even breathing) causes the pattern to move or even change its character.

There are a number of visual psychological phenomena associated with moiré. For example, rapid back and forth movement of a single figure of repetitive elements coarse enough to be resolved by the eye produces a fleeting image. Even when a figure, such as Fig. 4, is stationary, the unconscious, rapid, oscillatory movements of the eye give a sparkling character to the figure. The transient moiré pattern produced by the superposition of the immediate image and the transient, displaced afterimage is probably the basis for the liveliness of many examples of so-called optical art (Op art). See PERCEPTION; VISION.

Gerald Oster

Bibliography. O. Kafri and I. Glatt, *The Physics of Moiré Metrology*, 1990; K. Patorski, *Handbook of Moiré Fringe Technique*, 1993; D. Post, B. Han, and P. Ifju, *High-Sensitivity Moiré: Experimental Analysis for Mechanics and Materials*, 1997.

Moisture-content measurement

Measurement of the ratio or percentage of water present in a gas, a liquid, or a solid (granular or powdered) material. Nearly all materials contain free water, the relative amount being dependent upon the physical and chemical properties of the material. The primary purpose of determining and maintaining moisture contents within specified limits can usually be traced to economic factors, trade practices, or legal requirements.

Moisture content has a number of synonymous terms, many of which are specific to certain industries, types of product, or material. The water content in solid, granular, or liquid materials is usually referred to as moisture content on either the wet or dry basis; the wet basis is common to most industries. Specifically, moisture content on the wet basis refers to the quantity of water per unit weight or volume of the wet material. A weight basis is preferred. The textile industry uses the dry basis for moisture content of textile fibers. Often referred to as regain moisture content, the dry basis or regain refers to the quantity of water in a material expressed as a percentage of the weight of the bone-dry (thoroughly dried) material. The relationship between the wet and dry moisture-content basis is shown in Fig. 1.

The moisture content in air is referred to as humidity, either absolute or relative. Absolute humidity is the number of pounds of water vapor associated with 1 lb of dry air, also called humidity. Relative humidity (RH) is the ratio, usually expressed as a percentage, of the partial pressure of water vapor in the actual atmosphere to the vapor pressure of water at the prevailing temperature. Relative humidity is customarily reported by the U.S. Weather Bureau because it essentially describes the degree of saturation of the air. However, air which is saturated (100% RH) at 50°F (10°C) is quite dry (19% RH) when heated to 100°F (38°C). A changing basis of this type is not convenient for many purposes such as computations used in air-conditioning, combustion, or chemical processing; therefore absolute units, such as dew point or grains of water per pound of dry air, are more acceptable. Dew point is the temperature at which a given mixture of air and water vapor is

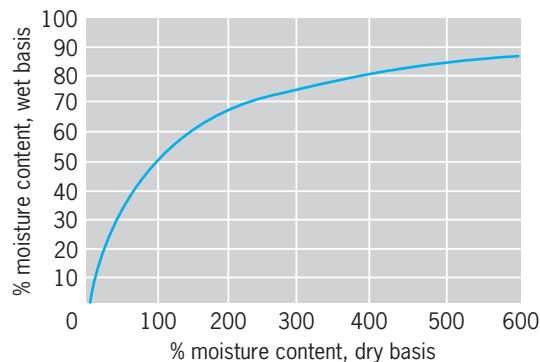


Fig. 1. Dry versus wet moisture-content basis.

saturated with water vapor. See DEW POINT; HUMIDITY.

Gases

The measurement of water content in gases and mixtures of air and gases is important in industry. A number of commercially manufactured instruments are available for these measurements; their principles of operation include condensation, used in dew- or fog-point indicators; dimensional change, used by hygrometers; thermodynamic equilibrium, used by wet-bulb psychrometers; and absorption methods, which serve as the basic principle for gravimetric and electric conductivity or dielectric types.

The importance of humidity in relation to personal comfort is well known. As a result the air-conditioning industry has grown to considerable proportions by producing equipment to maintain comfortable conditions of temperature and humidity. Considerable industrial air conditioning is also done for process reasons. The textile industry makes wide use of humidity control in rooms for weaving, carding, spinning, and other processes, because the amount of moisture held or absorbed by the textile fibers affects these operations. Paper manufacturers face problems similar to those of the textile industry, as do certain chemical, plastic, and allied processing industries in which control of humidity is important for product quality.

Control of humidity is important in the preservation of materials, especially those which are hygroscopic, and in the storage of food products. In many instances, humidities must be maintained at high levels, as in the storage of apples and vegetables, whereas humidity is maintained at low levels in the storage of dried milks, eggs, and similar products.

Human comfort is affected by high humidities, because the air is so close to its saturation content that it cannot absorb moisture from the surface of the skin and thus cool the individual by evaporation. The higher the temperature of the air, the greater the amount of moisture it can hold. A rule of thumb for human discomfort is that any combination of temperature and relative humidity totaling 130 or higher is uncomfortable. See HUMIDITY CONTROL.

Psychrometers. A psychrometer is a device for measuring moisture content of air or gases by means of two thermometers. One thermometer bulb is covered with a wick and maintained wet (the wet bulb); the other bulb is exposed directly to the air or gas (the dry bulb). The evaporation of water from the moistened wick of the wet bulb produces a lowering of its temperature, and by observation of the difference in temperature between the two bulbs the absolute or relative humidity can be determined. For accurate results, the gas or air must have a velocity of 15–20 ft/s (4.5–6 m/s) past the wet bulb. Psychrometric charts or tables are used with the readings obtained from the two thermometers to determine the moisture content of the air or gas. Instruments utilizing this principle are often called wet- and dry-bulb thermometers. See PSYCHROMETER; PSYCHROMETRICS.

Hygrometers. Hygrometers measure humidity by the change in dimensions of a hygroscopic material, such as human hair, organic membranes, wood, and plastics. Their most dependable range of operation is from 15 or 20% to 85 or 90% relative humidity at temperatures of 0 to approximately 160°F (−18 to 71°C). Stability of these instruments is better when they are not subjected to extremes of temperature or humidity. There is considerable time lag in the system response to changing humidity conditions when operated at low temperatures. Accuracies of $\pm 3\%$ relative humidity can be expected at normal room temperatures.

Salt conductivity type. Another type of hygrometer utilizes changes in the electrical conductivity of a hygroscopic salt as its operating principle. The conductivity of a solution is dependent upon its concentration (amount of water if the salt content is constant) and its temperature. See HYGROMETER.

Gravimetric type. The change in weight of an absorbing material can be used to measure its moisture content, a method known as gravimetric hygrometry. The measurement of moisture content in gases is obtained by passing a known volume of the gas through a suitable desiccant, such as phosphorus pentoxide, silica gel, or similar material, and observing its change in weight. This method is considered a primary standard and is often used in the exact calibration of instruments. It is necessary to make certain that all of the gas or atmosphere has been in contact with the desiccant for a sufficient time to ensure complete absorption of the water vapor. Exact initial and final weighings are also necessary. **Figure 2** shows the basic principle of this instrument.

Hygroscopic materials may also be employed directly to determine changes in water vapor in air. The Aldrich regain indicator makes use of a loose ball of cotton attached on one arm of a sensitive balance; the other arm of the balance serves as a pointer, as in **Fig. 3**. It is important that the sensing material change weight only with change in relative humidity of the surrounding atmosphere.

Dew-point indicators and recorders. When water vapor is cooled, a temperature is reached at which the phase changes to a liquid or solid. This temperature is known as the dew point. The classical method of determining the dew point consists of slowly cooling a polished metal surface until condensation takes place; the temperature at which the first droplet appears is taken as the dew point. The manually operated dew cup is one of the simplest forms of dew-

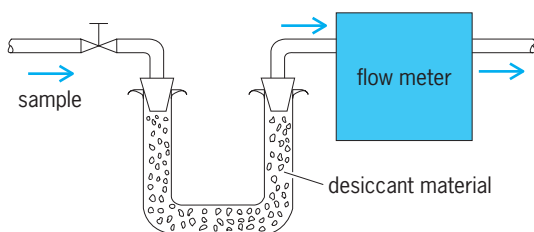


Fig. 2. Gravimetric absorption method.

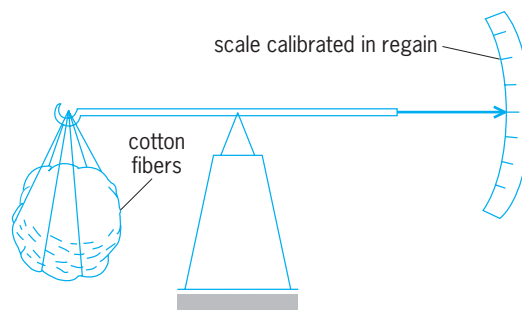


Fig. 3. Aldrich regain indicator.

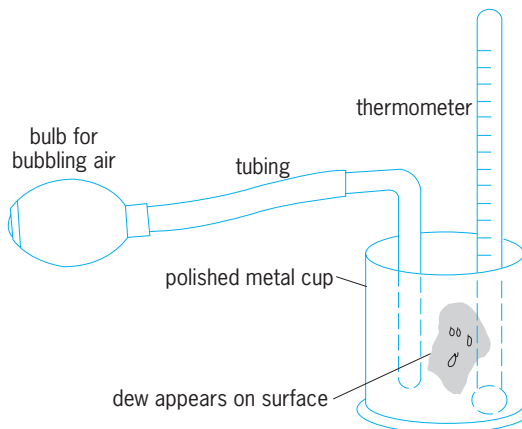


Fig. 4. Manually operated dew-point unit.

point apparatus. This consists of a polished metal cup containing ether or another volatile liquid into which is placed an accurate mercury-in-glass thermometer. Air is bubbled through the volatile liquid, lowering its temperature and the temperature of the metal cup until dew (condensation) forms on the cup (**Fig. 4**).

These instruments are capable of high accuracy, provided correct techniques are employed. This technique is widely used for measuring water vapor in flue gases, gasoline vapors, furnace gases, compressed gases, and others. More refined instruments utilize refrigeration and closed systems to contain the sample while the condensation is viewed through inspection windows. Others automatically record the condensation point by means of photocells, which alternately control heat or refrigeration to the target area. An adaptation of the dew-point technique makes use of the Wilson cloud chamber principle. The gas sample is compressed and then vented to atmosphere, producing cooling by adiabatic expansion. By repeated trials a pressure ratio can be obtained at which a cloud or fog forms. Dew point is computed from the ratio of the initial to final pressure.

Miscellaneous methods. Several alternate methods exist for determining the amount of water vapor present in gases or air.

The difference in thermal conductivity of dry air and air with water vapor may be determined by measuring the difference in the electrical resistance (temperature) of a hot wire sealed in a small cell. This

method is affected by changes in gas composition. Usually a bridge circuit is used, with a hot-wire cell containing dry air as the reference and a cell with the sample to be tested as the unknown. See BRIDGE CIRCUIT.

Spectroscopic methods and index of refraction have been used experimentally, as has the measurement of pressure or volume after the absorption of the water vapor from a sample. See SPECTROSCOPY.

Liquids and Solids

The rapid development of instruments for the measurement and control of moisture in liquids and solids was due in a large part to the great need that exists in many processes where the control of a precise moisture is critical. The desirability of a specific moisture content in a product during its preliminary manufacturing process is often required. In general, however, the rigid control of moisture content occurs most frequently in the final product to assure its quality and the fulfillment of legal or trade practices for the individual product.

Instrument types. Instruments suitable for the measurement of moisture content may be classified as periodic and continuous. In general, only those instruments offering continuous measurement are practical for the automatic control of moisture content in a product. The periodic instrument types are generally automated versions of conventional laboratory moisture-analysis procedures. The speed of response for the periodic (intermittent) instruments is typically 2 min or longer, often 15–20 min, making this type of instrument impractical for automatic control. The initial cost of the periodic or intermittent sampling instrument is usually less than for continuous types.

Moisture measuring instruments may also be classified by operating principle. Those instruments employing electrical conductivity (either dc or ac), absorption of electromagnetic energy (radio-frequency regions), electrical capacitance (dielectric constant change), and infrared energy radiations are more readily adapted to continuous measurements inasmuch as the response of these instruments to moisture changes is very fast. Those instruments employing automatic oven drying, chemical titrations (Karl Fischer technique), equilibrium hygrometric methods, distillation methods, and so forth are usually of the intermittent, or periodic, type; the time for analysis is usually 2 min or longer. Most of the more popular methods are outlined in the following discussion.

Electrical conductivity. These methods are based on the relationship between dc resistance and moisture content for such materials as wood, textiles, paper, grain, and similar products. Specific resistance plotted against moisture content results in an approximate straight line up to the moisture saturation point. Beyond the saturation point, where all of the cells and intermediate spaces are saturated with free water, conductivity methods are not reliable. This point varies from approximately 12 to 25% moisture content, depending on the type of product.

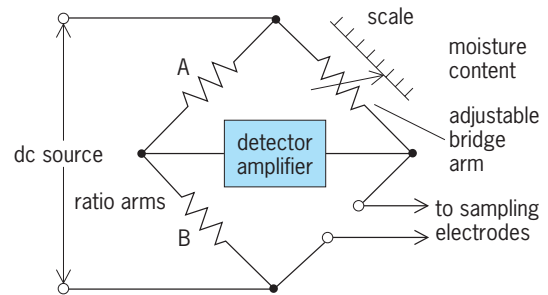


Fig. 5. Basic bridge circuit for conductivity method of moisture-content measurement.

The sample under test is applied to suitable electrodes in the form of needle points for penetrating into woods, plaster, and similar products or flat plates for sheet materials. Granular or fibrous materials may make use of electrodes in the form of a cup or clamp arrangement to confine the material to a fixed volume. The electrodes and sample under test make up one arm of a Wheatstone bridge (Fig. 5). The high sensitivity required of the detector dictates the use of electronic amplifiers. The range of resistance values corresponding to normal moisture contents varies from less than 1 to 10,000 megohms or higher, depending upon material, electrode design, and moisture content. Increasing moisture content results in decreasing resistance values. See RESISTANCE MEASUREMENT.

Electrical capacitance. These methods are based on the principle of the change in dielectric constant between dry and moist conditions of a material. The dielectric constant of most vegetable organic materials is 2–5 when dry. Water has a dielectric constant of 80; therefore the addition of small amounts of moisture to these materials causes a considerable increase in the dielectric constant. The material being measured forms part of a capacitance bridge circuit, which has radio-frequency power applied from an electronic oscillator (Fig. 6). Electronic detectors measure bridge unbalance or frequency change, depending on the method employed.

Electrode design varies with the type of material under test. Parallel-plate types are used for sheet materials, whereas cylindrical electrodes are usually adapted for liquids or powders. Modifications of the parallel-plate capacitor into a rectangular enclosure are used with granular, fibrous, or powdered materials. Range of moisture contents measurable is from

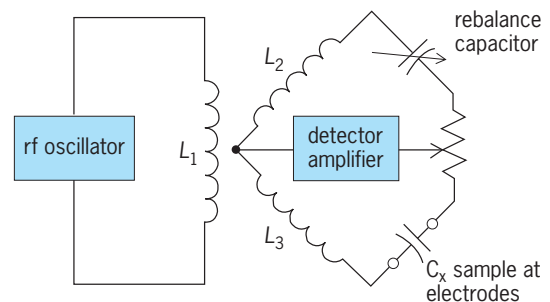


Fig. 6. Basic bridge circuit for capacitance method of moisture-content measurement.

2 or 3% to 15 or 20%, varying with the product. See CAPACITANCE; PERMITTIVITY.

Radio-frequency absorption. There are instruments that utilize the attenuation of electromagnetic energy when passed through the material. The radio-frequency types are operated at frequencies below 10 MHz and, for best results, are used with products composed of polar materials. The radio-frequency energy is passed through the polar material and the water molecules absorb some of the energy as molecular motion. See ELECTROMAGNETIC RADIATION.

Figure 7 shows the basic block diagram of a typical radio-frequency moisture instrument. These instruments must be calibrated to the particular product under test. By means of suitable electrodes for the sampling technique the instrument can be applied to solid or sheet materials; granular materials are placed into cell-type electrodes. The test results will be affected by any polar material in the sample; however, water of hydration in a product will not be detected. These instruments are mostly available as periodic types but can be adapted for continuous measurements on many products. Moisture content is dependent on the material and ranges from 0.1 to 60%; for certain materials the range will be quite narrow.

Microwave absorption. Instruments employing very high-frequency electromagnetic energy are in limited use for both grab samples and continuous measurement applications. The frequency of the electromagnetic energy is in the region of 1 gigahertz and higher. The 2.45-GHz S band, 8.9–10.68-GHz X band, and 20.3–22.3-GHz K band frequencies have all been used for microwave moisture-analysis instruments. See MICROWAVE.

The principle of operation is based upon the fact that the water molecule greatly attenuates the transmitted signal with respect to other molecules in the material in the S and X band frequencies. In the case of the K band microwave frequencies, the water molecule produces molecular resonance. There are no other molecules that respond to this particular resonant frequency, making this frequency most specific to the moisture (free water) in paper products. The wavelength of the K band frequencies are approximately 1.35–1.5 cm long; at these very short wavelengths, the electromagnetic energy may be guided (transmitted) by means of metallic pipes, usually of rectangular shape, called waveguides. Thus the output of a microwave oscillator can be coupled to a horn-shaped opening on the end of the waveguide, propagated through the sample material, and collected by means of an input or receiver horn. The output and input horn are physically separated so that a noncontacting (nondestructive) continuous measurement of the material is possible. The basic block diagram of a microwave moisture instrument is shown in **Fig. 8**. The detection of energy absorption can be made by attenuation (loss) or by phase shift methods, or both measurements may be made simultaneously. Moisture content ranges of less than 1% to 70% or higher can be made. At low moisture contents the S and X bands require a minimum mass of

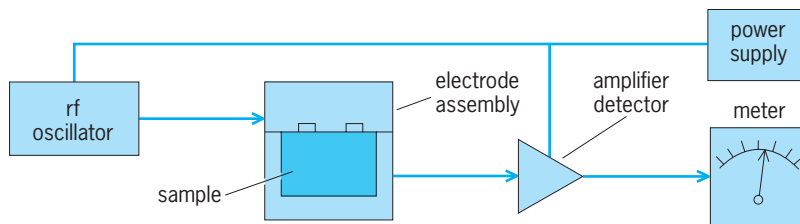


Fig. 7. Radio-frequency-absorption moisture indicator.

material for practical operation. Accuracies of these instruments are within $\pm 0.5\%$ of indication up to approximately 15% moisture content. See MICROWAVE SPECTROSCOPY.

Infrared absorption. These instruments operate on the principle of absorption of infrared radiation when passed through the sample material. The water molecule becomes resonant at certain infrared frequencies and thus the amount of energy absorbed by the water absorption band is a measure of the moisture content.

One instrument, used for measuring the moisture content of sheet paper, makes use of infrared energy at two different wavelengths by measuring the differential absorption (**Fig. 9**). Infrared radiation is passed through a rotating disk chopper containing two filters, one filter passing a wavelength of 1.94 micrometers (which covers the water absorption band) and the second filter passing a band of 1.80 micrometers (which is not significantly affected by moisture content). The paper sheet is exposed to the dual-frequency narrow-band radiation and the reflected radiation is collected in an integrating sphere, where it activates a lead sulfide detector. The signals are amplified and the ratio of the two narrow-band signals is read out directly as moisture content. This type of measuring system is quite immune to the effect of paper variables such as basis weight, coatings, density, and composition. The useful range is from 1 to 12% and can be used as either contacting or noncontacting. See INFRARED SPECTROSCOPY.

An infrared instrument useful for gas and liquid measurements operates on the principle of selective energy absorption of the sample when compared to a reference material having little or no absorption. The instrument contains two infrared sources (**Fig. 10**). One source is directed through a cell containing the sample material, and the second source is directed through the reference cell, which in gas analysis is

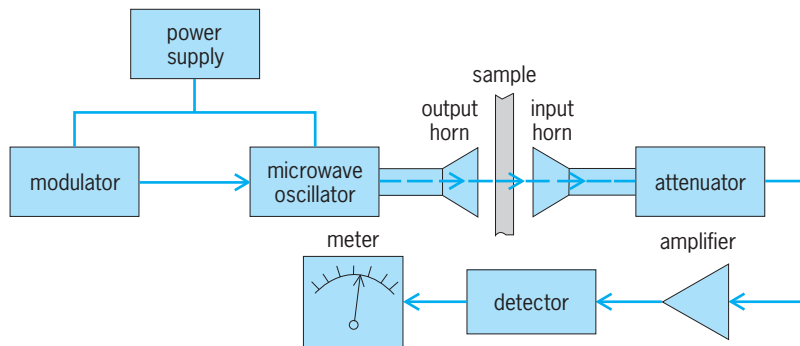


Fig. 8. Microwave-absorption moisture indicator.

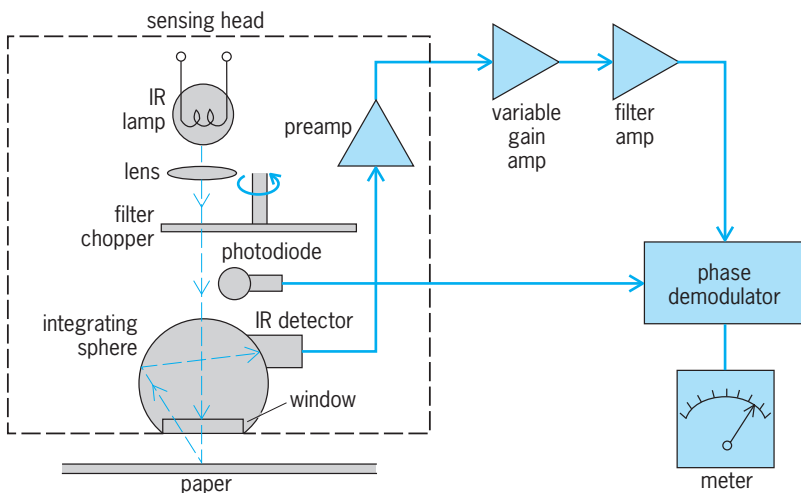


Fig. 9. Infrared-backscatter moisture instrument.

filled with an inert gas, usually air. A detector unit is present which consists of two cavities separated by a metallic diaphragm that acts as a variable capacitor in respect to a fixed plate within the cavity. The operation consists of chopping the infrared energy to the sample and reference cells at 10 Hz. The gas or fluid in the sample cell absorbs the wavelengths which are characteristic of the material and proportional to its concentration. The reference cell does not absorb any appreciable infrared energy. The detector, which has two equal volumes separated by a diaphragm, is filled with gas or vapor of the component of interest. The infrared energy passes through the cells and into the detector, which causes the temperature of gas in the detector to rise; however, if some of the energy has already been absorbed by the sample cell, the sample side of the detector receives less energy than the reference half of the detector. The temperature rise of the gas in the detector causes the pressure to increase and, if the temperature is unequal, the pressure increase is unequal, causing the diaphragm to move in a direction that will tend to equalize the two pressures. This in turn causes the capacitance existing between the diaphragm and its sensing plate to change; this change is used to modulate an oscillator.

The chopper alternately blocks and then passes the energy from each infrared source. Each time the energy is blocked, the pressures equalize in the detector compartments and again the energy is permitted to enter the detector cavities. The modulating frequency is 10 Hz and the magnitude of the capacitance change is directly related to the amount of energy absorbed in the sample cell. The signal from the oscillator is amplified and demodulated for meter indication or applied to a recorder. The range of moisture in gases is from 1000 ppm to 100% water vapor, and in fluids from 6 to 100% water. Response speed is 0.5 s for 90% of full-scale output.

Equilibrium. The equilibrium moisture content of the air at the surface of a material is representative of the moisture within the material. This is particularly true of hygroscopic granular or fibrous materials. Humidity-measuring instruments are used by inserting the measuring element into the material.

Moisture content is determined from data of relative humidity plotted against moisture content for the material under test.

Absorption. The water content of liquid organic compounds can be determined by the use of spectrophotometers. The sample is placed in a suitable cell and monochromatic infrared radiation is passed through it. Water vapor absorbs the infrared radiation, but the material absorbs little radiant energy. Thus the radiation varies with the water content. The radiation is focused on a sensitive thermocouple and the resulting voltage amplified. The indication of a sensitive galvanometer can be calibrated in terms of moisture content. This method has been used for accurately detecting 0–10 ppm of water in Freon-12.

Chemical methods. A widely used chemical titration method is the Karl Fischer technique. The method offers extreme sensitivity with good accuracy and covers a wide range of materials. Small samples can be analyzed readily. End points can be determined either visually or electrometrically. The Karl Fischer reagent is added in small increments to a glass flask containing the sample until the color changes from yellow to brown or a change in potential is observed at the end point. Dark-colored solutions require electrical end-point measurement. This method does not differentiate between free and combined water.

Various other methods involve chemical reactions with the free water in liquids or solids by (1) the evolution of free acidic or basic compounds, (2) the evolution of an inert gas, or (3) the formation of an insoluble precipitate. In general, these methods fail to distinguish active hydrogen from water in many materials and are reliable only under laboratory control.

Distillation. A representative sample of the material can be placed in a suitable flask with an excess of liquid which usually has a boiling point higher than water but is immiscible with water. Heat is applied to the flask, and the water and some of the liquid are distilled off. The combined vapors are condensed and collected. The water is separated and

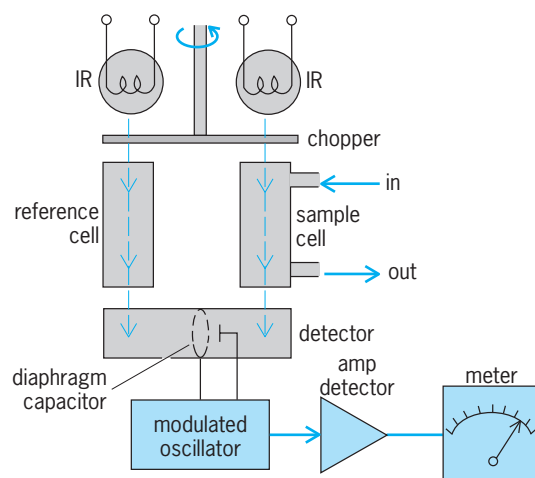


Fig. 10. Infrared gas or liquid analyzer.

measured volumetrically to calculate the moisture content. A variety of laboratory glass apparatus is available for distillation of liquids which are either lighter or heavier than water. The measuring tubes are etched with graduations in cubic centimeters corresponding to grams of water. The moisture content is calculated from the weight of the original sample and the weight of water. This method is slow, and the accuracy depends upon the care exercised and the apparatus used. The method was widely used in the analysis of grains but has largely given way to electrical or oven-drying methods. *See* DISTILLATION.

Oven drying. The oldest and most common analytical method of determining moisture consists of heating the sample to ensure complete drying. Moisture is calculated on the basis of loss in weight between original and dried sample. The method is applicable to many solids and some liquids, and does not require unusual operator skill. Semiautomatic drying and weighing ovens are now available; the moisture content is indicated directly by the weighing scale built into the oven. Problems involve making certain that the material has not lost some of its weight by loss of volatile products other than water; that the material is completely dry at the time of final weighing, and that accurate weighings are made. The oven method is widely used and often serves as the primary standard for calibration of electrical and other indirect methods. Lee E. Cuckler

Bibliography. J. Carmody and J. Istiburek, *Moisture Control Handbook: Principles and Practices for Residential and Small Commercial Buildings*, 1996; D. M. Considine (ed.), *Process/Industrial Instruments and Control Handbook*, 5th ed., 1999; D. M. Considine and S. D. Ross (eds.), *Handbook of Applied Instrumentation*, 1964, reprint 1984; Instrument Society of America, *Moisture and Humidity*, 1985; A. Pande, *Handbook of Moisture Determination and Control: Principles, Techniques, Applications*, vols. 1-4, 1974-1975.

Mole (chemistry)

A unit (symbolized mol) used to measure the amount of material in a chemical sample. Gram-molecular weight is an older name for the mole and is seldom used. The mole is defined by international agreement as the amount of substance (chemical amount) of a chemical system that contains as many molecules or entities as there are atoms in 12 g of carbon-12 (^{12}C). When the mole is used, the elementary entities need not be molecules, but they must always be specified. They may be atoms, molecules, ions, electrons, or specified groups of such particles.

Quantitative measurements in chemistry are concerned with the amount of material in a given sample. Three obvious ways of measuring the amount are to measure the mass m of the sample (using a balance), to measure the volume V (using a measuring cylinder), or to count the number of molecules in the sample, N . Although it is more difficult to devise

an experiment to count molecules, this third way of measuring amount is of special interest to chemists because molecules react in simple rational proportions (for example, one molecule of A may react with one, or two, or three molecules of B, and so forth). For example, in calculating how much air should be provided to completely burn a given amount of octane petroleum fuel to carbon dioxide and water, it is necessary to determine from the molecular formulas and the equation for the reaction how many molecules of oxygen (O_2) react with each molecule of octane, and to count the number of molecules of oxygen in a sample of air and the number of molecules of octane in a sample of the fuel.

However, to count molecules is difficult experimentally and inconvenient in practice because the numbers are so large. For any chemical, a mass of 1 kilogram of the sample contains a large number of molecules, of the order 10^{23} - 10^{24} . The mole is defined so that 1 mole of any substance always contains the same number of molecules. This number is large, approximately 6.02×10^{23} , and is known as Avogadro's number. The mole is a more convenient unit in which to measure the amount of a chemical than counting the number of molecules, and it has the same advantages. *See* AVOGADRO'S NUMBER.

The International System of Units (SI) is the internationally accepted system of quantities and units. The mole is the SI unit of the amount of substance (chemical amount), just as the kilogram is the SI unit of mass, and the meter is the SI unit of length. Although amount of substance is the approved name for this quantity, in practice it is not widely used; the name commonly used is number of moles.

The amount of substance (chemical amount) of a sample, n , may be determined in practice by one of three methods.

The value of n (the amount of substance) may be determined from the mass m by dividing by the molar mass M of the sample, as in Eq. (1). If m is

$$n = \frac{m}{M} \quad (1)$$

expressed in g and M in g/mol, then the value of n will be obtained in mol.

For a gas, the value of n may be determined from the volume V , pressure p , and absolute temperature T by using the ideal gas equation (2), where R is the

$$n = \frac{pV}{RT} \quad (2)$$

gas constant ($R = 8.3145 \text{ J K}^{-1} \text{ mol}^{-1}$). If pV is expressed in $(\text{N m}^{-2}) \times (\text{m}^3) = \text{J}$, and RT in J mol^{-1} , then pV/RT gives the value of n in mol. From the ideal gas equation it is determined that at a pressure of 1 atm (= 101,325 pascals) and a temperature of 0°C (= 273.15 K) the volume occupied by 1 mol of an ideal gas, $V = nRT/p$, is 22.414 liters (= $22.414 \times 10^{-3} \text{ m}^3$); and at a pressure of 1 bar (= 10^5 Pa) the volume is 22.711 liters (1 liter = 10^{-3} m^3). *See* AVOGADRO'S LAW; GAS.

For a solution, the amount of solute (or the amount concentration of solution) is frequently determined

by titration: if ν_A molecules of A react with ν_B molecules of B in the titration, then at the end point the amount of A used (n_A) is related to the amount of B (n_B) by Eq. (3), so that if one is known the other

$$n_A = \frac{\nu_A}{\nu_B} n_B \quad (3)$$

may be determined. See TITRATION.

The best experimental value for the number of atoms in 12 g of ^{12}C is $6.0221377(36) \times 10^{23}$, where the number in parentheses is the uncertainty in the least significant digits. It follows that one mole of any substance contains this number of molecules (or other entities). This leads to the best estimate of the value of the Avogadro constant, N_A , which is $N_A = 6.0221377(36) \times 10^{23} \text{ mol}^{-1}$.

The concentration of a solution may be recorded as (mass of solute)/(volume of solution), in units gram/liter; or as (chemical amount of solute)/(volume of solution) in units mol/liter. Because of the proportionality of chemical amount to number of molecules, the latter is the more useful measure of concentration and is generally used in chemistry and biochemistry. The word concentration without a qualifying adjective is generally taken to mean amount concentration, that is, amount-of-substance concentration. See CONCENTRATION SCALES.

Ian. M. Mills

Mole (zoology)

A mammal belonging to the order Soricomorpha (previously Insectivora), family Talpidae (the true moles). There are 39 species in 17 genera of talpids distributed on all continents except Australia. See INSECTIVORA; MAMMALIA.

Description. Moles are small, subterranean mammals that are highly specialized for burrowing. The eyes and ears are much reduced, and they have a long pointed snout and short soft fur that can lie either way. The powerful forelimbs are adapted for digging, having very powerful muscles and very large strong claws. Moles have sharp pointed teeth; the dental formula for the star-nosed mole (*Condylura cristata*), hairy-tailed mole (*Parascalops breweri*), and the western moles (*Scapanus*) is I 3/3 C 1/1 Pm 4/4 M 3/3 for a total of 44 teeth. That of the eastern mole (*Scalopus aquaticus*) is I 3/2 C 1/0 P 3/3 M 3/3 = 36, and that of the shrew-mole (*Neurotrichus gibbsii*) is I 3/3 C 1/1 Pm 2/2 M 3/3 = 36. The mole's body is stout and cylindrical with a short neck. See DENTITION.

Moles are solitary animals and rarely come above-ground. Earthworms are a staple food of many moles, and they are often stored for later use. The moles paralyze the worms by biting them. There is usually one litter per year, which is consistent with their underground mode of existence; that is, they have few predators. The young approach the size of the adults and begin to fend for themselves when about 4 weeks old. The testes of the males become huge in the mating season.

Molehills are formed of excess dirt pushed up and onto the surface when runways are being dug. Molehills are especially common in wet weather, when the moles are digging new foraging burrows. Foraging burrows are usually temporary and are often visible from above because the soil is pushed up as a ridge. Permanent runs are usually deeper and leave no telltale ridges above the burrows. Moles can turn around in the burrows, and they can run backward in them nearly as fast as they can run forward.

North American moles. The main North American moles—*Condylura* (1 species), *Scalopus* (1), *Parascalops* (1), and *Scapanus* (3)—fall into a related group, the subfamily Scalopininae, separate from the Old World moles, but this group also includes one species from China, *Scapanulus oweni*. The rest of the moles occur mostly in the Old World and are in the subfamily Talpinae, although one genus and species from this group, *Neurotrichus gibbsii*, occurs on the west coast of North America.

The eastern mole (*Scalopus aquaticus*) has the largest range of any of the North American moles: from the Atlantic coast west to southern Minnesota, through Nebraska, and to West Texas. It has a very short, nearly naked tail. The main food of this species, like most moles, is earthworms, but they eat numerous insect larvae, especially "grubworms" (scarabaeid or Junebug larvae), many other invertebrates, and some plant material. They enter ant nests and feed heavily on ant eggs, larvae, and adults. They are active for about 4 or 5 hours, then sleep for a similar time. These moles move rapidly through deeper underground tunnels to a foraging area, then they move rapidly back and forth thus disturbing the soil, which probably causes earthworms and other prey to move and be detected. Mating occurs early in the spring; after a gestation of 30–40 days the single litter of three to seven is born.

The hairy-tailed mole (*Parascalops breweri*) occurs mostly in woodlands from southeast Canada through much of New England and New York. This species has a short, hairy tail. It may come above-ground more than most moles, and may even forage some on the surface, which helps to account for the number caught by cats and other predators. There are many foraging burrows near the surface; permanent burrows are deeper and fewer. Mating is in late March or April, and the young (usually four or five) are produced in late April or May.

Three species of *Scapanus* occur on the west coast of North America—*S. townsendii* and *S. orarius* from southern British Columbia to northwestern California, and *S. latimanus* from southern Oregon to northern Baja California. The tail is short and fairly thick. These moles are often found in forests. Other small mammals, such as mice and shrews, often use their burrows. Nest chambers are lined with dry plant material.

The star-nosed mole of North America (*Condylura cristata*) is unusual in that it is gregarious, is a good swimmer, and prefers wet, mucky areas. It has a long, hairy tail. Encircling the snout is a highly developed sensory organ of 22 projections, or tentacles. These

tentacles were long thought to be sensory, but recent behavioral and experimental evidence suggests that the tentacles are used to detect electrical current. The moles dip their snouts into the water before entering and scan under water with the tentacles. They direct their bites toward areas with strong electrical output, the clitellum (“muscle”) and reproductive parts of earthworms. The burrows often end at the water. It is the only North American mole that occurs in mucky situations. Molehills in muckland in the northeastern United States are sure to be those of *Condylura*, although this species sometimes ranges into lawns from associated ponds.

Neurotrichus gibbsii is called the shrew-mole and is closely related to the Japanese shrew-mole (*Urotrichus*). It is much smaller and thinner than other moles. The forefeet are much smaller in *Neurotrichus* than in other talpids, but they do have long curved foreclaws. The tail is about one-third the body length. *Neurotrichus* occurs on the west coast of North America from southern British Columbia to northern California, primarily in deep woodlands. Like other moles, it feeds heavily on earthworms, slugs, snails, centipedes, and insects and their larvae, and has been known to eat salamanders. It is a good swimmer. It climbs low bushes, apparently in search of insects. There may be several litters per year consisting of one to four. Young have been seen in all months except December and January.

Eurasian moles. The desmans—*Desmana moschata* of Russia and the Ukraine and *Galemys pyrenaicus* of the Pyrenees Mountains in France and the northern part of the Iberian peninsula—are, like *Condylura*, long-tailed and highly aquatic. Both have very long, flexible snouts, useful for exploration, and large hindfeet useful for swimming. The Russian desman has a widened flat snout, with grooves above and below. Scent glands at the base of the tail give this animal a musky odor. The hindfeet are webbed, and there are stiff hairs along the edges of the feet. Both of these characters give greater surface area, which is helpful in swimming. This species lives along freshwater streams and ponds and feeds on aquatic foods—fish, crustaceans, insects, and worms. Its den is in a chamber above the water line. A tunnel up to 6 ft (1.8 m) long opens into the water below the surface. The Pyrenean desman is found along swift-flowing streams and is a strong swimmer. Unlike many moles, there are three annual peaks of breeding in this species.

The genus *Talpa* of Eurasia contains nine species. The European mole (*T. europaea*; see **illustration**), found in northern and central Asia as well as Europe, is a rather typical representative of the group and spends most of its life underground, emerging in the spring to gather leaves for its nest and to look for water. The tunnels range from just below the surface to depths of nearly 3 ft (0.9 m). There is usually one litter per year of two to seven. Gestation takes about 28 days.

Other types. Other moles are *Scapanulus* (1 species), *Scaptonyx* (1), and *Uropsilus* (4) of



European common mole (*Talpa europaea*).

China; *Parascaptor* (1) of Burma; and *Euroscaptor* (6), *Mogera* (5), *Scaptochirus* (1), and *Dymecodon* (1) of Asia.

In addition, there are two families of molelike mammals that are not talpids. They are the African golden moles, Chrysochloridae, and the marsupial moles, Notoryctidae. These two families occur in Africa and Australia, respectively, parts of the world where talpids do not occur. There they occupy similar niches to the talpids, which means they behave somewhat similarly and have evolved to fit similar conditions.

John O. Whitaker, Jr.

Bibliography. M. L. Gorman and R. D. Stone, The natural history of moles, Comstock Publishing Associates, a division of Cornell University Press, Ithaca, 1990; E. Gould, W. McShea, and T. Grand, Function of the star in the star-nosed mole, *J. Mammal.*, 74:108–116, 1993; G. D. Hartman and T. L. Yates, Moles, Talpidae, pp. 30–55 in G. A. Feldhamer, J. A. Chapman, and B. C. Thompson (eds.), *Wild Mammals of North America: Biology, Management, and Conservation*, 2d ed., Johns Hopkins University Press, Baltimore, 2003; R. Hutterer, Order Soricomorpha, pp. 220–311 in *Mammal Species of the World: A Taxonomic and Geographic Reference*, ed. by D. E. Wilson and D. M. Reeder, 2005; R. M. Nowak, *Walker's Mammals of the World*, 3 vols., Johns Hopkins University Press, Baltimore, 1999; J. O. Whitaker, Jr., and W. J. Hamilton, Jr., *Mammals of the Eastern United States*, Cornell University Press, Ithaca, 1998; D. E. Wilson and S. Ruff, *The Smithsonian Book of North American Mammals*, 1999.

Molecular adhesion

The tendency of dissimilar solids or liquids to cling together as a result of the interatomic forces which they exert upon each other across their common interface. Some factors affecting molecular adhesion are the physical state of the materials, the composition and the topology of the surfaces in contact, the temperature, and the presence of foreign materials such as adsorbed gases on one or both surfaces.

Mutual forces of attraction, electrical in origin, exist between virtually all atoms and molecules. It is these forces which, under proper conditions of temperature and pressure, compel isolated atoms or molecules to condense into solids and liquids. On average, these forces are approximately equally strong in all directions and result in an atom or molecule

achieving its lowest possible energy, that which is thermodynamically favored, when it is surrounded on all sides by other atoms. The presence of a real surface breaks the three-dimensional continuity of the substance, leaving atoms at the surface with an unfulfilled capacity for bonding. It is this capacity for bonding that is responsible for the phenomena of surface tension, adsorption of gases, and adhesion. See ADSORPTION; COHESION (PHYSICS); INTERMOLECULAR FORCES; SURFACE TENSION.

In principle, the strength of the adhesive bond between two substances may be quite large—larger, in fact, than the cohesive strength of the weaker material. In practice, however, many factors act to diminish the strength of adhesion. Probably the most significant are incomplete contact of the surfaces due to their microtopography and the presence of adsorbed gases or other surface contaminants. These may diminish the surface's capacity to further bond or simply replace one of the original surfaces with an overlayer which is significantly weaker.

One technological application in which molecular adhesion plays an especially significant role is the deposition by evaporation or sputtering of thin films on substrates of dissimilar composition. A particularly important example is the use of such films to form conducting paths—microscopic wires only a few tens of nanometers thick—in integrated circuits. These films must both adhere well to the underlying substrate and produce contacts with appropriate electrical properties. At present the understanding of adhesion is not sufficiently detailed to be able to predict such properties in advance of experiments. See INTEGRATED CIRCUITS; SPUTTERING.

Enhancement has been observed in the degree of adhesion between evaporated layers and the underlying medium as a result of irradiation by heavy ions with energies of a few megaelectronvolts. In these experiments, enhanced adhesion has been observed between metal films and metal, semiconductor, and insulating substrates. Materials as dissimilar as gold and poly(tetrafluoroethylene) have been joined, and it has also been observed that particle irradiation transforms electrically rectifying interfaces into ohmic ones. Further research in ion-beam enhanced adhesion promises not only new technological advances but also improved understanding of molecular adhesion at the most fundamental level. See ADHESIVE; INTERFACE OF PHASES; SURFACE TENSION.

Robert A. Weller

Bibliography. K. W. Allen (ed.), *Adhesion*, vols. 1–15, 1977–1991; L. H. Lee, *Fundamentals of Adhesion*, 1991; D. E. Packman (ed.), *Handbook of Adhesion*, 1993; R. L. Patrick (ed.), *Treatise on Adhesion and Adhesives*, vols. 1–7, 1967–1991; M. Prutton, *Introduction to Adhesion*, 1994.

Molecular anthropology

The study of primate phylogeny and human evolution through the genetic information in the deoxyribonucleic acid (DNA) of genomes and in the proteins that genes encode. The first studies in molecu-

lar anthropology used immunological and biochemical methods to obtain information from proteins on the degrees of genetic similarity of humans and other primates. These results not only placed chimpanzees and gorillas closest to humans rather than to orangutans but also indicated that the very close kinship between chimpanzees and gorillas was not any closer than the relation of each to humans. Subsequent studies that extracted genetic information directly from DNA extended this original finding. Indeed, the accumulating comparative DNA sequence data provide mounting evidence that the closest genetic kinship is between chimpanzees and humans rather than chimpanzees and gorillas.

DNA data. The results gathered in DNA studies of primate phylogeny challenge the traditional anthropological view that humans are very different from all other animals. DNA results show that, genetically, humans are only slightly remodeled apes. Humans share with their closest relatives, the chimpanzees and bonobos (pygmy chimpanzees), more than 98.3% identity in typical noncoding DNA and probably about 99.5% identity in the active coding sequences of functional nuclear genes. Humans share about 98.0% identity in nuclear genomic DNA with gorillas, 96.5% identity with orangutans, and 95.0% identity with the most distant ape relatives, the gibbons and siamangs. Apes and humans share with the other branch of catarrhines, the Old World monkeys, about 92% identity in nuclear genomic DNA, and with the platyrrhines, the New World monkeys, about 87% identity. Even with nonanthropoid primates, the tarsiers and strepsirhines (lemurs and loriforms, such as bushbabies), the anthropoids (platyrrhines and catarrhines) share a DNA identity in the range of 71–76%. New information on the social lives and high intelligence of chimpanzees and other apes also challenges the traditional view that humans are very different from other animals. Chimpanzees use tools and have material culture, are highly social, and may even have the mental capacity for abstract thought and symbolic communication, for example, for learning rudimentary forms of language. See DEOXYRIBONUCLEIC ACID (DNA).

Traditional primate classifications provide a poor guide to the course of primate phylogeny. These classifications use the vague concept of grades of evolutionary advancement to place the smaller-brained primates in the suborder Prosimii (the primitive grade) and the larger-brained primates in the suborder Anthropoidea (the advanced grade). Moreover, on viewing humans as the most advanced primates, traditional primate classifications have humans as the sole living members of family Hominidae, while the great apes of Africa (chimpanzees, bonobos, gorillas) and Asia (orangutans) are members of subfamily Ponginae of family Pongidae, despite overwhelming evidence that the African great apes share a more recent common ancestry with humans rather than with orangutans. This is the traditional anthropocentric view of the place of *Homo sapiens* in the order Primates. In contrast, a strictly objective view based on molecular evidence, but also congruent morphological evidence from both living and fossil primates,

not only places apes with humans within the family Hominidae but also within this family places chimpanzees and bonobos with humans in the genus *Homo*. See APES; FOSSIL HUMANS; MAMMALIA.

Primate phylogeny. The course of primate phylogeny, as reconstructed from the molecular and fossil evidence, may be sketched as follows. During the Paleocene geological epoch of 65–57 Ma (Mega annum, or million of years before the present), the ancestors of the modern primates divided first into strepsirhines and haplorhines and then, within the haplorhines, into the tarsiiiforms and basal anthropoids. In the early Eocene epoch at about 50 Ma, the strepsirhines divided into lemuriforms and loriforms; and in the middle to later Eocene epoch at about 40 Ma, the basal anthropoids divided into platyrrhines and catarrhines. In the late Oligocene epoch at about 28–25 Ma, while strepsirhines and platyrrhines were dividing into familial clades, the basal catarrhines divided into cercopithecids (Old World monkeys) and hominids (apes and humans).

In the early Miocene epoch at about 23–22 Ma, familial clades divided into subfamilial clades, which by 20–15 Ma divided into tribal clades. Thus, because the last common ancestor of all living apes and humans lived at about 18 Ma, the living hominids are grouped together as subfamily Homininae. This phylogenetic classification not only has each taxon represented by a monophyletic group or clade but also indicates by gradations in taxonomic rank how far back in time is found the last common ancestor of the living members of the clade.

The phylogenetic classification of living hominids is as follows [each age (in Ma units) shown for a taxon is the age of the last common ancestor of all living members of that taxon]:

Family Hominidae

Subfamily Homininae (18 Ma)

Tribe Hylobatini

Subtribe Hylobatina (8 Ma)

Symphalangus syndactylus: siamang

Hylobates lar: white-handed gibbon

Tribe Hominini (14 Ma)

Subtribe Pongina

Pongo pygmaeus: Borneo orangutan

Subtribe Hominina (7 Ma)

Gorilla gorilla: gorilla

Homo (6 Ma)

H. (Pan) (3 Ma)

H. (P) troglodytes: chimpanzee

H. (P) paniscus: bonobo (pygmy chimpanzee)

H. (Homo) sapiens: humankind

After Homininae divided into tribes Hylobatini and Hominini during the late Miocene epoch at about 8 Ma, subtribal hylobatans of the hylobatin clade divided into *Hylobates* (gibbons) and *Symphalangus* (siamangs). The hominin clade in the middle Miocene epoch at about 14 Ma divided into subtribes Pongina, from which *Pongo* (orangutans) evolved, and Hominina. The latter, in the late Miocene epoch at about 7 Ma, divided into *Gorilla* and *Homo*, and

at about 6 Ma *Homo* divided into the subgenera *Homo (Homo)*, out of which humankind evolved, and *Homo (Pan)*, which in the Pliocene epoch at about 3 Ma separated into *H. (Pan) troglodytes* (chimpanzees) and *H. (Pan) paniscus* (bonobos, or pygmy chimpanzees). Other subtribal clades such as those of New World monkeys as well as those of Old World monkeys also divided into genera during the late Miocene epoch at about 10–7 Ma. Then genera divided into subgenera or species groups in the late Miocene to early Pliocene epoch at about 6–4 Ma. Thus, classifying humans, chimpanzees, and bonobos as members of the same genus is equivalent to how other primate clades at the same age are classified.

Ongoing evolution. After the divergence of *Homo (Homo)* from *Homo (Pan)*, humankind's emergence was marked by mutations (such as DNA sequence changes) that spread to fixation in the ancestors of all modern humans. These mutations are the human-specific factors which distinguish the human species genetically from all other species. To obtain a comprehensive catalog of these mutations, genomes of a chimpanzee and a bonobo (closest relatives to humans) and a gorilla (next closest relative) need to be sequenced and compared to the sequence of a human genome. Scientists who call for a Human Genome Evolution Project as a companion to the Human Genome Project envision that the overall goal would be to identify and analyze the functions of human-specific genetic factors involved in the evolution of unique features of human anatomy (for example, bipedal locomotion, greatly enlarged brain) and behavior (such as speech, higher-order cognitive functions).

Ongoing evolution involves mutations that have not spread to fixation, either because they occurred too recently or because natural selection has maintained a polymorphic state. These mutations occur at frequencies that occasionally differ from one human group to another. They account for the genetic diversity found in the human species. Extensive comparative data now exist on the genetic diversity due to mitochondrial haplotypes (each, a set of genetic determinants located on a mitochondrial genome). These mitochondrial genetic variants arise from mutations in the DNA carried by mitochondria, the maternally inherited (from ova) cellular organelles that drive oxidative metabolism. These organelles have their own genomes that encode key proteins and RNA molecules found in mitochondria. Although a mitochondrial genome is quite small (16,000 nucleotide base pairs in length) compared to the nuclear genome (3 billion base pairs in length), each somatic cell contains many mitochondria; thus mitochondrial DNA can be readily prepared and haplotypes of such DNA in different individuals can be determined. The species-wide distribution of mitochondrial haplotypes in present-day human populations compared to the distribution in chimpanzees reveals that humans show less DNA diversity than chimpanzees. In fact, the diversity that humans show is no larger than that shown by a chimpanzee subspecies. While populations in different subspecies of

chimpanzees share only 28% of the total species diversity, human populations in different ethnic groups (European Norwegians and South African hottentots) and even within the same group share 85–90% of the total species diversity. The human genetic diversity due to mutations in nuclear DNA (the DNA carried by the chromosomes of cells) also shows this same pattern in which most of the variation within the human species as a whole is contained within single populations. These findings argue that all living humans are biologically not only members of the same species but also the same subspecies and even the same race, the human race. A further inference from the data on mitochondrial genetic diversity is that the birth of the human race took place in Africa about 200,000 years ago.

Recently, small pieces of mitochondrial DNA were recovered from bones of an undated Neandertal specimen. On comparing mitochondrial haplotypes from present-day humans to the Neandertal haplotype and to chimpanzee haplotypes, and assuming a molecular clock with a reference date of 6 Ma for the age of the last common ancestor of humans and chimpanzees, the estimated date for divergence of the Neandertal haplotype from the lineage to the last common ancestor of the present-day human haplotypes is about 800,000 years ago, as compared to about 200,000 years ago for the age of the last common ancestor of present-day human haplotypes. These age estimates as compared to those for the last common ancestor of haplotypes from different chimpanzee subspecies suggest that Neandertals belonged to the same species as humans do, but a different subspecies. The results so far gathered on nuclear genetic diversity in present-day human populations leave open the possibility that as the founder members of the human subspecies, *Homo sapiens sapiens*, multiplied and spread from one continent to another, the gene pool of these migrants could have been enriched by admixture of some genes from the preexisting archaic human populations. See MITOCHONDRIA.

Morris Goodman

Bibliography. M. Goodman, Epilogue: A personal account of the origins of a new paradigm, *Mol. Phylogenet. Evol.*, 5:269–285, 1996; M. Goodman et al., Toward a phylogenetic classification of primates based on DNA evidence complemented by fossil evidence, *Mol. Phylogenet. Evol.*, 9:585–598, 1998; M. Krings et al., Neandertal DNA sequences and the origin of modern humans, *Cell*, 90:19–30, 1997; E. H. McConkey and M. Goodman, A human genome evolution project is needed, *Trends Genet.* 13:350–351, 1997; M. Ruvolo, Genetic diversity in hominoid primates, *Annu. Rev. Anthropol.*, 26:515–540, 1997.

Molecular beams

Utilization of well-directed streams of atoms or molecules in vacuum. This is a cornerstone technique in the investigation of molecular structure and interactions. Molecular beams are usually formed at sufficiently low particle density for the interaction

of one beam molecule with another to be negligible. This ensemble of truly isolated molecules is available for the spectroscopic study of molecular energy levels using photon probes from the radio-frequency to optical portions of the electromagnetic spectrum. Some of the best-determined fundamental knowledge of physics comes from spectroscopic molecular-beam experiments. Beyond this, beams can be applied as probes of the multifaceted nature of gases, plasmas, surfaces, and even the structure of solids. An application intermediate in complexity is the study of molecular interactions determining the properties of plasma and electric discharge devices, the nature of the upper atmosphere, and some aspects of the cooler astrophysical regions. See SCATTERING EXPERIMENTS (ATOMS AND MOLECULES).

Production and detection. One simple means of forming a beam is to permit gas from an enclosed chamber to escape through a small orifice into a second chamber maintained at high vacuum by means of large pumps. **Figure 1a** shows such effusion into a collimating chamber from an oven chamber, generally heated to control the vapor pressure of the gas. The molecules coming from the orifice are distributed in angle according to a cosine law, illustrated by the circle downstream of the orifice which represents the envelope of relative beam flux vectors. A useful number of molecules passes forward along the horizontal axis of the apparatus. A well-collimated beam is then formed by requiring that those molecules entering the test chamber where an experiment is to be performed pass not only through the orifice but also through a second small hole separating the collimating and test chambers. A property of beams formed this way is that the velocities of the individual molecules have a large thermodynamic spread in values centered on a mean value of order 10^3 m/s (3.3×10^3 ft/s) determined by the oven temperature.

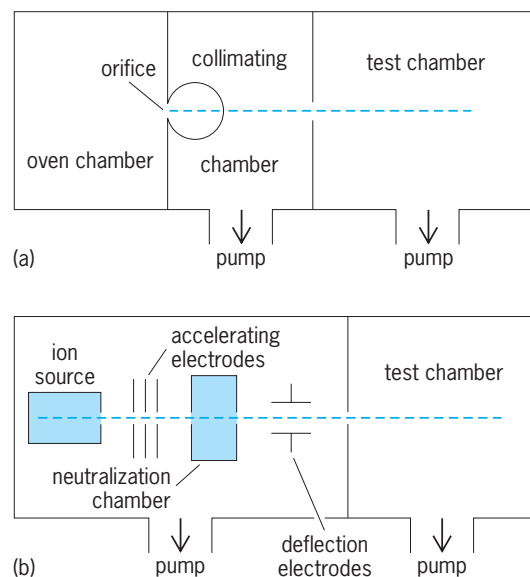


Fig. 1. Schematic diagrams of systems for producing molecular beams. (a) Conventional oven-beam system. (b) Charge-exchange beam system.

Charge-exchange system. If higher velocities are desired, then a charge-exchange beam system can be used. In this scheme, ions are produced by some ionizing process such as electron impact on atoms within a gas discharge. Since the ions are electrically charged, they can be accelerated to the desired velocity and focused into a beam using electric or magnetic fields. The last step in neutrally charged beam formation is to pass the ions through a neutralizing gas where electrons from the gas molecules are transferred to the beam ions in charge-exchange molecular collisions. If the acceleration voltage is relatively high, then the ion-beam velocity will be 10^5 m/s (3.3×10^5 ft/s) or greater upon entering the neutralizer. In many cases, such energetic charge-exchange collisions produce beam atoms or molecules in internally excited energy levels rather than just in the ground-state level of lowest energy occupied by atoms in low-temperature gases. On the other hand, useful beams of internally excited molecules can be formed from ground-state molecules which have been excited by energetic electron- or ion-beam impact or by photon absorption from a laser or other light beam. See EXCITED STATE; GROUND STATE; ION SOURCES.

Secondary electron detection. The faster types of neutral molecular-beam particles are easy to detect by secondary electron ejection from a solid. A collision of a beam molecule with a surface is sufficiently violent for one or more electrons to be ejected. For intense beams, the rate of electron production is so high that the electrons can be collected and measured electronically as an electric current. At lower rates, the effect of each electron can be multiplied by means of an alternating sequence of electron acceleration and surface ejection steps, to produce a burst of 10^6 or more electrons from a single beam molecule. This pulse of current is adequate for electronic pulse counting, leading to a very sensitive overall beam-detection technique useful at beam intensities as low as five molecules per minute.

Other detection techniques. For molecular beams effusing from an oven, a variety of detection techniques have been devised. Alkali atoms have such small ionization potentials that their valence electrons can be transferred to a heated metal surface having a large work function, such as tungsten. The resultant ions can then be detected by current measurement or particle multiplication techniques, depending again upon the beam intensity. A second special technique useful for some beams of reactive or excited molecules is electron ejection from a surface, powered by the internal energy of these special beam molecules. A universal detector for any kind of slow molecule employs initial conversion into ions by electron- or light-beam impact; the ions are then accelerated into a fast beam for easy detection.

Special beams. Occasionally a beam containing atoms or molecules in a specific quantum-mechanical state is needed. Energy-resonant transitions between the ground state and the specific excited state can be utilized, often induced by single-frequency laser radiation. Beams of slow molecules

with magnetic (or electric) dipole moments can be selected according to the direction of orientation of their moments; spatial separation into component beams is achieved through orientation-dependent interaction between the dipole moment and an externally applied strong nonuniform magnetic (or electric) field. Some precision spectroscopic molecular-beam experiments use such beams, and form the basis for atomic-clock precision time standards. See ATOMIC CLOCK.

Molecular-beam spectroscopy. Much of molecular spectroscopy involves the absorption or emission of light by molecules in a gas sample. The frequency of the light photon is proportional to the separation of molecular energy levels involved in the spectroscopic transition. However, the molecule density in typical gas samples is so high that the energy levels are slightly altered by collisions between molecules, with the transition frequency no longer characteristic of the free molecule. Using low-density molecular beams with their sensitive detection techniques can reduce this collision alteration problem, with the result that atomic properties can be measured to accuracies of parts per million or better. If the simplest atoms or molecules are employed, the basic electromagnetic interactions holding the component electrons and nuclei together can be precisely studied. This is of great importance to fundamental physics, since theoretical understanding of electromagnetic interactions through quantum electrodynamics represents the most successful application of quantum field theory to elementary particle physics problems. See QUANTUM ELECTRODYNAMICS; QUANTUM FIELD THEORY.

Properties measured. Among the basic quantities of physics measured by molecular beams are the fine-structure constant $\alpha = e^2/\hbar c = 1/137.0361$ (e is the electron charge in electrostatic units, \hbar is Planck's constant divided by 2π , and c is the speed of light), obtained from microwave spectroscopy studies of the fine and hyperfine energy-level splittings in one- and two-electron atoms; the value 1.5210326×10^{-3} of the magnetic moment of the proton in Bohr magnetons, obtained with the hydrogen maser; the purely quantum-electrodynamic shifts of atomic energy levels called Lamb shifts; the nuclear magnetic dipole moments of several hundred isotopes; the equality in magnitude of the unit electron and nuclear charges to an accuracy of better than parts in 10^{18} ; and the absence of intrinsic electric dipole moments for the electron and proton, which is a test of parity and time reversal as symmetry properties obeyed by the electromagnetic interaction. The isolated, unperturbed nature of atoms in a beam is needed for the device used as the length standard, based upon the 605.7-nanometer wavelength of krypton atoms being reproducible to parts in 10^9 . The time standard is the cesium-beam atomic clock operating on the ground-state hyperfine splitting microwave transition with a reproducibility over hours of parts in 10^{12} .

Apparatus. The molecular-beam magnetic resonance apparatus that is used in a number of the

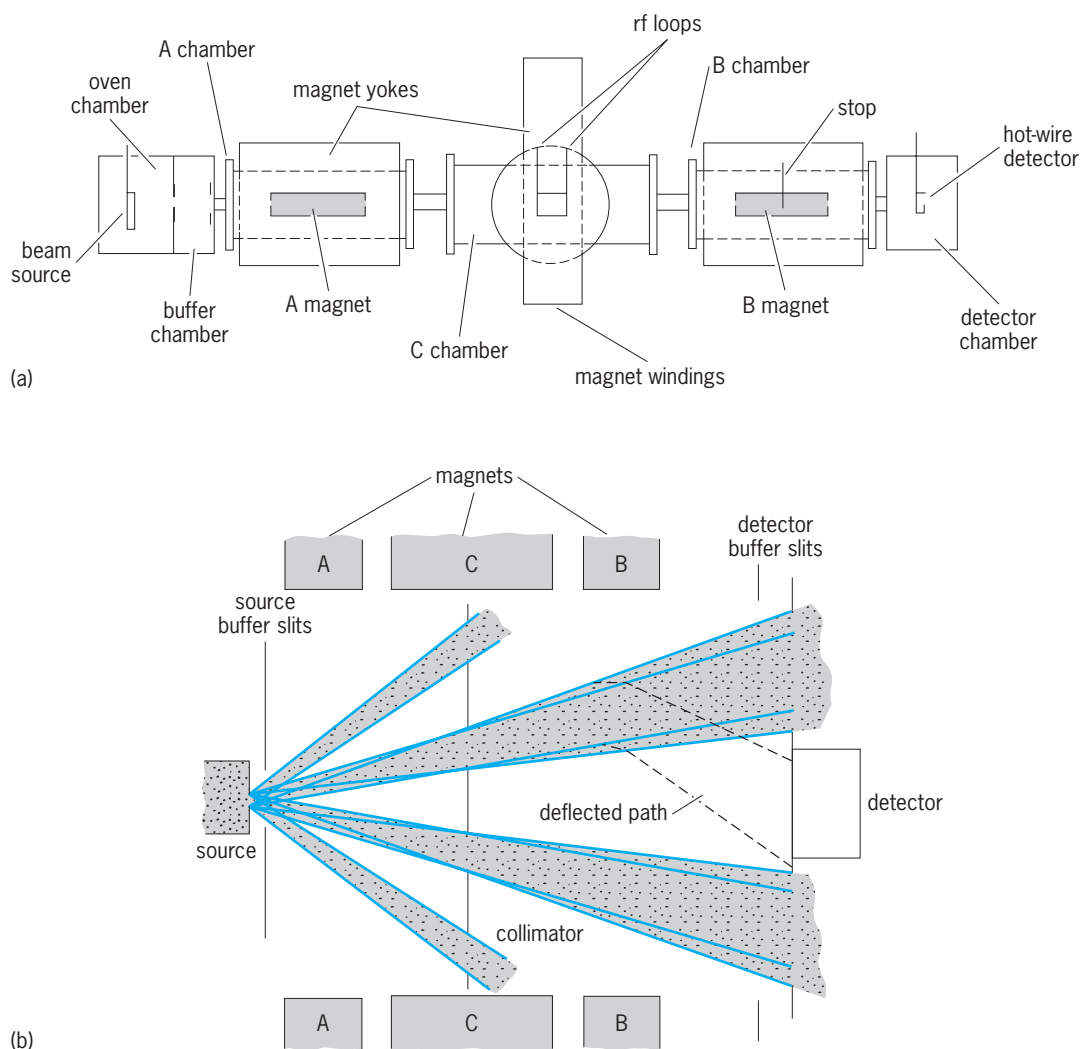


Fig. 2. Molecular-beam magnetic resonance spectroscopy experiment. (a) Apparatus used in the experiment. (b) Spatially resolved multiple-beam configuration showing typical molecular trajectories.

spectroscopy experiments is shown in **Fig. 2**, along with a diagram showing typical molecular trajectories. On the left end of **Fig. 2a** is an oven beam source, here of alkali atoms; on the right end is a hot-wire surface ionization detector of these atoms. Between the ends are three separate regions where external fields are applied to the beam. The A region is a state selector, and the B region a state analyzer; that is, some property of the beam atoms is well defined in the A region, and then this situation is checked in the B region. In the example of **Fig. 2** the selected property is the direction of the atoms' magnetic moment. The center of the apparatus is the transition region marked C, where a resonant radio-frequency transition can be induced between atomic energy levels associated with different directions for the magnetic moment. This alteration of the atoms entering the analyzing region results in a change in atom trajectories, with a corresponding change in the intensity of the beam striking the hot-wire detector. For example, in **Fig. 2b**, molecules undergoing a transition within the C region follow the deflected path and are detected. When the frequency of the radio waves in the C region is not resonant, such a change does not occur. Determining the frequency for maximum ef-

fect results in a measure of the atomic energy level splitting, which for the example of **Fig. 2** is directly related by theory to the atom's nuclear magnetic moment.

The power of molecular-beam techniques derives, in large measure, from the ability, in principle, of the A and B regions to contain any kind of state-selective device utilizing combinations of static or time-varying electromagnetic fields. Also, the interaction region C can involve resonant interactions with light or other photons or even nonresonant-state destructive interactions such as collisions with atoms in an introduced gas. If the A and B regions do not control beam atom trajectories, but instead some other property such as the state of internal energy, then correspondingly the nature of the beam detector might need to be sensitive to the controlled property.

Use of fast beams. The fast molecular beams produced by charge exchange increasingly have been employed in spectroscopy experiments and offer some advantages in addition to easy detection. Among these is the fractionally well-defined and controllable molecule velocity determined by the ion acceleration voltage, a feature useful in the

study of time-dependent quantum-mechanical interference effects on transition rates. The ability to transport rapidly decaying excited atoms through an apparatus is also enhanced using fast beams. A highly accurate atomic fine-structure measurement employed a fast hydrogen atom beam, as did an experiment on multiphoton microwave transitions between highly excited atomic states.

Scattering experiments. The molecular-beam study of interactions occurring during individual collisions of two species of electrons, ions, atoms, or molecules can be divided into two categories, depending upon whether the targets for the particles in one beam are those in a gas or those in another beam. In addition, the nature of the collisional interaction depends strongly on the velocity of impact of the two particles; if this is very low, all that can happen is that the two particles elastically scatter off each other, with no change in the internal nature of either particle being energetically possible. At higher collision velocities, inelastic processes can occur in which a part of the energy of collision is converted into a change in internal energy of one or both particles. Particle rearrangement can also occur, such as the electron transfer used previously in charge-exchange atomic-beam sources, or proton transfer in chemical reactions between molecules. Collisions involving fast electrons or fast bare nuclei are primarily probes of the structure of the target molecule. The slower collision processes are more concerned with the composite molecular system transiently formed during the collision time. An example is the negative ion state formed when an electron joins a neutral atom or molecule. Such compound states cannot live indefinitely since the total energy of the system is positive; in a rearrangement collision, however, the compound state breaks up into molecules different from the originally colliding ones.

Quantities measured. The simplest quantity to measure in a molecular-beam collision experiment is the probability for a particular elastic or inelastic event to result from a single collision of one beam particle

with one target particle. This probability is usually expressed in terms of an apparent size of the target particle for the process, called a total cross section. This quantity is a sum over contributions from all possible distances of closest approach of the two particles, since molecules are so microscopic that individual ones cannot be aimed at other individual ones in any controlled manner. One must work with collections of beam particles and target particles, and divide out their numbers to obtain a cross section. However, distances of approach are often uniquely correlated with definite angles of scattering of beam particles relative to the incident beam direction. Thus, the necessary averaging of the distances of approach contained in a total cross section can be avoided if one instead measured angular distributions of scattered beam particles. Different scattering angles are associated with beam particles probing the strength of the collisional interaction at different particle separations, and so angular distribution studies investigate the detailed nature of molecular interactions in a way similar to angular scattering experiments in nuclear and high-energy physics.

Electron-beam scattering. The apparatus for a comparatively sophisticated electron-beam angular-scattering study is indicated schematically in Fig. 3. The target for the electrons is a molecular beam such as previously discussed, and is shown here as a dot in the center of the experiment. The molecular beam is coming out of the paper. No collision-induced changes in the properties of that beam are detected, although this actually can be done in coincidence with observation of electrons scattered from the electron beam.

The electrons are produced by thermionic surface emission from a hot cathode, and have a maxwellian spread in energies of order 0.1 eV. They can be electrostatically accelerated if desired, and then energy-selected with a cylindrical electrostatic analyzer consisting of two 60° electric field plates on arcs of circles about a common origin. The resolution of such analyzers can be as good as 0.01 eV, a value sometimes needed to observe sharp resonance peaks in

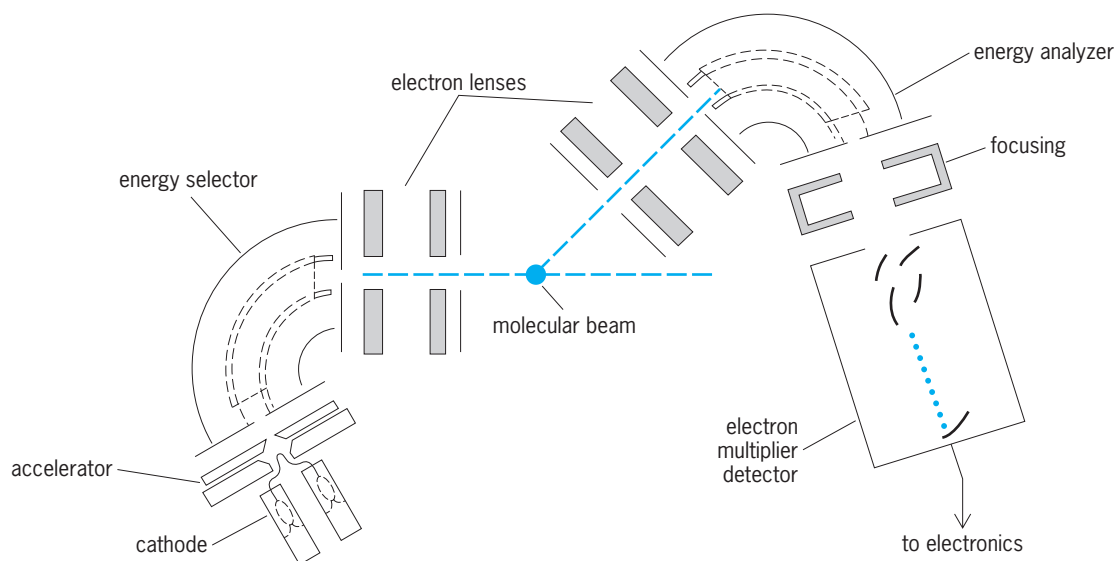


Fig. 3. Apparatus for electron-beam angular scattering study.

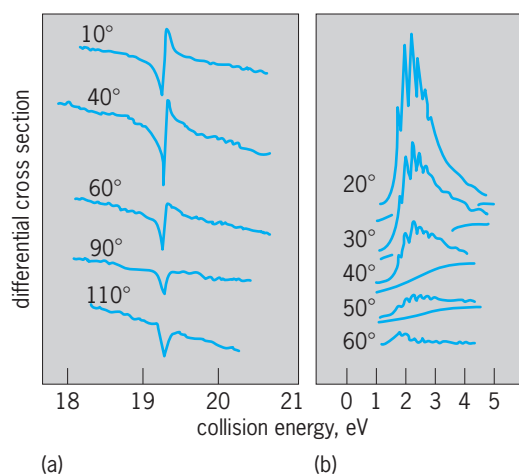


Fig. 4. Resonances observed at various scattering angles for elastic electron scattering by beams of (a) He and (b) N_2 molecules.

the energy dependence of cross sections that are associated with the excitation of long-lived compound negative-ion states. After focusing, the energy-selected electron beam is scattered by the molecular beam. Electrons scattered at various angles are selected by rotation of the remainder of the apparatus about the molecular-beam axis. Scattered electron focusing is followed by energy analysis in a second cylindrical analyzer and then by acceleration into an electron multiplier for detection by pulse counting. Thus, the experiment is divided into source, state-selector, interaction, state-analyzer, and beam-detector regions just as in the molecular-beam magnetic resonance experiment.

Figure 4 shows the collision energy dependence observed at various scattering angles for elastic electron scattering by beams of He and N_2 molecules. The change with angle in the shape of the He⁻ resonance near 19.3 eV makes possible a positive identification of the orbital angular momentum of the compound state. The resonances in the N_2 scattering curves are associated with the excitation of different amounts of internuclear vibration in the N_2^- ion. Searches for new states of molecules can be made by using electron scattering by ion beams. Detailed studies of resonances in slow electron scattering are used

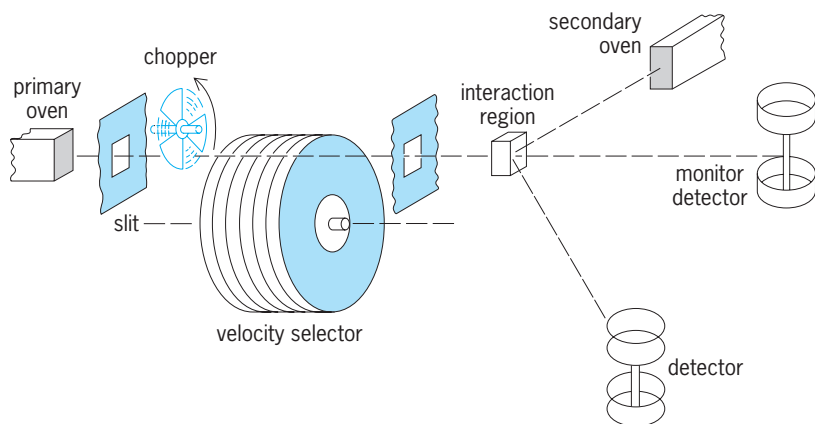


Fig. 5. A double-molecular-beam experiment on chemical reactions.

to investigate with considerable success the nature of energy exchange between the various bonds of molecules as large as benzene.

Double-molecular-beam scattering. Figure 5 schematically shows a typical double-molecular beam scattering apparatus used in studies of chemical reactions at the molecular level. Elastic scattering can also be studied to obtain data sensitive to the molecule-molecule interaction potential. The beams from the primary and secondary ovens collide in an interaction region. The intensity of the primary beam is monitored by an on-axis detector, and the scattered secondary-beam molecules or chemical reaction products are measured with a second detector. Shown in the primary beam line before the interaction region are two devices commonly used in such experiments: a beam chopper for phase-sensitive beam detection and a velocity selector. The latter is a set of disks rotating together, each with beam-transmission holes placed at different angles about the axis of rotation. Only those beam molecules having a certain velocity pass through all the holes as the disks rotate. The beam chopper breaks the beam into pulses with known timing, and detected scattered particles are electronically checked for the same timing pattern. Thus beam-beam scattering can be detected in the presence of a poor vacuum producing very large amounts of beam scattering in the background gas.

The results of such chemical molecular-beam experiments can tell much about the mechanisms underlying chemical reactions. For instance, the alkali halide product molecule formed in $K + Br_2 \rightarrow KBr + Br$ reactions is produced primarily in small-angle or forward scattering (called a stripping reaction), whereas in $K + CH_3I$ collisions this product is seen largely in backward scattering (a rebound reaction). Large-reaction total cross sections are correlated with stripping reactions, small ones with rebound reactions; this is observed even though the sums of reaction and elastic scattering cross sections are equal in the two cases. One concludes that the total effective sizes of the targets are the same in the two cases, and that the reactive process steals scattering probability away from the elastic process where no particle rearrangement occurs.

A detailed description of the reaction mechanism can be simplified into a sequence of events occurring as the collision progresses. The sequence involves the so-called harpoon model based upon a long-range sudden jumping of the valence electron on the primary alkali atom over to the electronegative halide molecule. The jump occurs at a molecular separation R_c , where the interaction potential energies for the composite systems $K + Br_2$ and $Kr^+ + Br_2^-$ happen to become almost equal, a point where little collision kinetic energy needs to be converted into a change in potential energy. The electron's jump is sufficiently violent for the Br_2^- to be vibrationally excited, and in addition the strong attraction of the two oppositely charged ions polarizes the Br_2^- , two factors leading to the second step in the model: a dissociation of Br_2^- into $Br^- + Br$. The reaction product KBr at last is formed by the attractive interaction of

the K^+ and Br^- ions while the neutral Br atom is still nearby.

The difference between the rebound and stripping cases now can be seen as associated respectively with small and large values of the electron jump distance R_c . When reached, small values of molecular separation correspond to strong molecular interaction and thus to large scattering angles. If R_c itself is small, then reaction occurs only for large scattering angles; at the smaller angles the molecules never become close enough for the electron to jump. The total reaction cross section πR_c^2 will be small in this case since R_c is small. On the other hand, if R_c is large, then the reaction occurs at small angles and the cross section is large as well.

Although the harpoon model is not always applicable, it does characterize the ideas pursued as a result of molecular-beam scattering experiments. Development of apparatus for the scattering of molecules polarized with their electric dipole orientations fixed by external fields has been undertaken, and this should permit further progress in understanding chemical reaction mechanisms, including those of interest to biology.

Laser excitation. The development of tunable, strong laser sources of single-frequency light beams added another dimension to molecular-beam experiments. With laser radiation resonantly tuned to excite a molecule from its normal ground state to one of its infinite number of vibrationally, rotationally, and electronically excited states, the number of possible studies and applications of excited molecular beams becomes enormous. Of basic importance is the fact that excited molecules can be either highly reactive or good carriers of stored potential energy. See MOLECULAR STRUCTURE AND SPECTRA; NUCLEAR STRUCTURE.

James E. Bayfield

Bibliography. N. F. Ramsey, *Molecular Beams*, 1956, reprint 1990; G. Scoles et al. (eds.), *Atomic and Molecular Beam Methods*, vol. 1, 1988, vol. 2, 1992.

Molecular biology

The study of structural and functional properties of biological systems, pursued within the context of understanding the roles of the various molecules in living cells and the relationship between them. Historically, molecular biology evolved as a biological discipline with roots derived from biophysics, genetics, and biochemistry. Since its inception, a prime focus of the field has been the molecular basis of genetics, and with the demonstration in the mid-1940s that deoxyribonucleic acid (DNA) is the genetic material, emphasis has been on structure, organization, and regulation of genes. Initially, molecular biologists restricted their studies to bacterial and viral systems, largely because of the systems' genetic and biochemical simplicity. However, a series of conceptual and technological developments occurred rapidly during the late 1970s that permitted molecular biologists to approach a broad spectrum of plant and animal cells with experimental techniques. One of the

major factors that rendered higher plant and animal cells amenable to molecular analysis has been the development and applications of genetic engineering. The implementation of recombinant DNA technology allowed the isolation and selective modification of specific genes, thereby reducing both their structural and functional complexity and facilitating the study of gene expression in higher cells.

Systems and Approaches

Although molecular biology initially developed as an independent discipline, the concepts and techniques used by molecular biologists have been rapidly and effectively employed to resolve numerous cellular, biological, and biochemical problems—becoming routine at both the basic and applied levels. The result is that molecular biology has emerged as a high-resolution approach to solving questions previously addressed in more traditional biochemical terms.

Intact cells and cell-free systems. While the ultimate objective of molecular biology is to understand biological function and its regulation at the molecular level, the complexity of intact higher organisms has in most cases precluded direct analysis. It has therefore been necessary to pursue the characterization and functional properties of biologically important molecules in a series of simplified systems. Since the 1960s many such investigations have been carried out using intact cells. *Escherichia coli* has been extensively examined because of its limited number of cellular functions and the corresponding restricted amount of genetic information encoded in the bacterial chromosome. Simple eukaryotic cells, such as protozoa and yeast, offer similar advantages and have also been studied. For these same reasons, bacteriophage and animal viruses have provided molecular biologists with the ability to study the structural and functional properties of molecules in intact cells.

The development of techniques for propagation of higher eukaryotic cells, including human cells, in vessels has permitted assessment of molecular function in living cells under conditions far simpler than those that exist in intact organisms. Although analysis of molecular events and regulatory mechanisms is complicated by the presence of the cell's complete set of genes and by a multitude of complex cellular functions, the ability to work with a single cell type under defined conditions has proved to be extremely advantageous. Cell culture is a widely utilized approach that has become important with advances in the understanding of specific cell growth requirements and of factors necessary for various differentiated functions. See TISSUE CULTURE.

Cell-free systems, in which components of cells are isolated and the properties of specific molecules and biochemical reactions are investigated in test tubes, have advanced the understanding of complex cellular events. Systems have been developed that will carry out a number of biologically essential molecular processes such as DNA synthesis, RNA synthesis, and protein synthesis. Most important, these systems allow the specific molecules

participating in the processes to be identified and characterized, frequently by addition, subtraction, or modification of the potentially relevant component. The ability to carry out biologically important reactions in cell-free systems reduces the number of variables and simplifies interpretation of experimental results. Yet, one must accept conclusions from experiments on such systems with the proviso that conditions in a test tube are somewhat artificial.

Gene mapping. The recognition of DNA as the genetic material coupled with the discovery that genes reside in chromosomes resulted in an intensive effort to map genes to specific chromosomes. Initially genes were assigned to chromosomes on the basis of correlations between modifications in cellular function, particularly biochemical defects, and the addition, loss, or modification of specific chromosomes. These early efforts at chromosomal location of genetic traits were largely restricted to lower animals and plants since they have relatively few chromosomes. Chromosome analysis was histochemical, that is, it was done by staining metaphase chromosomes spread on microscope slides and examining the preparations by light microscopy. Despite the limited resolution of this approach, chromosomal assignments of a number of genes were made. In fact, in the absence of knowledge of the biochemical or genetic basis for several human disorders such as Down syndrome, a series of chromosomal aberrations permitted identification of chromosomes encoding genes related to these diseases. Another approach which, early on, permitted assignment of genes to chromosomes was introduction of mutations by radiation or chemicals followed by a correlation between aberrations in metaphase chromosome morphology and biochemical or cellular abnormalities. *See* CHROMOSOME ABERRATION; MUTATION.

A major breakthrough in mapping genes was the development of somatic cell genetics. This is an approach in which, for example, human and hamster cells are fused, resulting in a hybrid cell initially containing the complement of human and hamster chromosomes. As the cells grow and divide in culture, the hamster chromosomes are retained while there is a progressive loss of human chromosomes. By correlating the loss of human biological or biochemical traits with the loss of specific human chromosomes, a number of human genes have been successfully mapped. *See* SOMATIC CELL GENETICS.

The development of methods for isolating genes and for determining the genetic sequences of the DNA in which the genes are encoded, led to rapid advances in gene mapping at several levels of resolution. Localization of specific genes to chromosomes is routinely carried out with cloned genes as probes. For example, isolated human genes are radioactively labeled and used to recognize and interact with an identical copy of the human gene in DNA isolated from human-hamster hybrid cells. Such analysis of a series of human-hamster hybrids, each hybrid containing limited numbers of various human chromosomes, offers an effective method for chromosome assignment. Further information about the segment

of a chromosome in which a specific gene resides can be obtained by directly determining the DNA sequences of both the gene itself and the surrounding region.

Chromosome localization of specific genes has numerous applications at both the basic and clinical levels. At the basic level, knowledge of the positions of various genes provides insight into potentially functional relationships. At the clinical level, chromosome aberrations are now routinely used in prenatal diagnosis of an extensive series of human genetic disorders, and several chromosomal modifications have been linked to specific types of cancer. Knowledge of genetic defects at the molecular level has permitted the development of diagnostic procedures that in some instances, such as sickle cell anemia, are based on a single nucleotide change in the DNA. *See* GENETIC MAPPING.

Recombinant DNA. Recombinant DNA technology has provided molecular biology with an extremely powerful tool. Though introduced only in the mid-1970s, genetic engineering has had a major impact on science, medicine, technology, and society in general, providing numerous opportunities for improvement of the quality of human existence. In broad terms, applications of recombinant DNA technology can be divided into four areas—biomedical, basic biological, agricultural, and industrial. Biomedical applications include the elucidation of the cellular and molecular bases of a broad spectrum of diseases, as well as both diagnostic and therapeutic applications in clinical medicine.

In a strictly formal sense, the term recombinant DNA designates the joining or recombination of DNA segments. However, in practice, recombinant DNA has been applied to a series of molecular manipulations whereby segments of DNA are rearranged, added, deleted, or introduced into the genomes of other cells. The processes associated with modifying DNA molecules are extremely precise, permitting addition, subtraction, or alterations of specific genes or of particular regions of genes.

The importance of recombinant DNA can be readily appreciated because of the pivotal role of genetic sequences in defining the structural and functional properties of cells, tissues, and organs. Selective expression of particular genes provides the unique properties characteristic of specialized cells, for example, hemoglobin production by erythropoietic cells and myosin production by muscle cells. A progressive expression of certain genes coupled with selective repression of others is necessary for the successful orchestration of development and differentiation. Prerequisite for maintaining the physiological integrity of cells, tissues, and organs is the capacity to modulate gene expression in response to changes in cellular requirements: for example, activation of genes associated with cell proliferation in conjunction with tissue regeneration, or modifications in the expression of genes encoding metabolic enzymes in association with fluctuations in nutrient levels.

Equally important are the findings that numerous

diseases, including cancer and disorders that involve inborn errors of metabolism, are related to a variety of genetic alterations, including rearrangements or mutations in specific DNA sequences. It therefore follows that by having the capacity to manipulate the organization of genes, the capacity to correct genetic defects is not an unrealistic expectation. And while the latter goal requires technological developments and ethical justifications, the available repertoire of genetic manipulations has enabled scientists to detect several clinically important genetic defects in the fetus (such as sickle-cell anemia) and to produce biologically important molecules in cell-free systems on a scale which allows them to be used for clinical treatment (for example, insulin, growth hormone, and interferon).

Techniques. Historically, the ability to manipulate or “engineer” genetic sequences is based on several developments, as discussed in the following sections.

1. *Methods for breaking and rejoining DNA.* The precise breaking and rejoining of DNA has been made possible by the discovery of restriction endonucleases, enzymes that have the ability to recognize specific DNA sequences and to cleave the double helix precisely at these sites. Also important is the ability to join fragments of DNA together with the enzyme DNA ligase, and the techniques to determine the nucleotide sequence of genes and thereby confirm the identity and location of structural and regulatory sequences.

2. *Carriers for genetic sequences.* Bacterial plasmids, that is, circular double-stranded DNA molecules that replicate extrachromosomally, have been modified so that they can serve as efficient carriers for segments of DNA, complete genes, regions of genes, or sequences contained within several different genes. Bacteriophage (viruses which infect and replicate in bacteria) and animal viruses, such as SV40 (a simian virus), retroviruses (RNA tumor viruses), and bovine papilloma virus, have also been successfully utilized as DNA carriers. In the vernacular of a genetic engineer, these carriers are referred to as cloning vectors. Host cells in which vectors containing cloned genes can replicate range from bacteria to numerous other cells, including normal, transformed, and malignant human cells.

3. *Introduction of recombinant DNA molecules.* Genetic sequences in the form of isolated DNA fragments, or chromosomes, or of DNA molecules cloned in plasmid vectors can be introduced into host cells by a procedure referred to as transfection or DNA-mediated gene transfer—a technique that renders the cell membrane permeable by a brief treatment with calcium phosphate, thereby facilitating DNA uptake. Genes cloned in viruses can also be introduced by infection of host cells.

4. *Selection of cells containing cloned sequences.* Several approaches have been developed for identification of bacterial, plant, or animal cells containing cloned genetic sequences. Bacterial cells containing plasmids with cloned genes can be detected by selective resistance or sensitivity to antibiotics. In addition, the presence of introduced genes in bacterial,

plant, or animal cells can be assayed by a procedure known as nucleic acid hybridization, in which a radioactive probe consisting of a region of a cloned gene is permitted to anneal with complementary sequences in the host cells. The recognition of the complementary sequence in the host cell is reflected by formation of a sequence-specific hybrid.

5. *Amplification.* Amplification of genetic sequences cloned in bacterial plasmids is efficiently achieved by treatment of host cells with antibiotics which suppress replication of the bacterial chromosome, yet do not interfere with replication of the plasmid with its cloned gene. Sequences cloned in bacterial or animal viruses are often amplified by virtue of the ability of the virus to replicate preferentially. *See* GENE AMPLIFICATION.

6. *Expression.* Expression of cloned human genes can be mediated by regulatory sequences derived from the natural gene, from exogenous genes, or by host cell sequences. Selection of regulatory sequences affixed to the cloned gene is based on the host cell or the regulatory signal to which expression is desired. For example, production of large quantities of a protein or RNA from a cloned human gene can often be achieved in a bacterial cell if the human gene is under the control of bacterial regulators and regulatory sequences. If a gene is introduced into a human cell to compensate for a genetic defect, however, responsiveness to normal physiological signals is an absolute necessity.

Applications. Two clinically important genes, human insulin and human growth hormone, have been cloned and introduced into bacteria under conditions where biologically active hormones can be produced. It would be difficult to underestimate the clinical relevance of cloned human insulin and growth hormone genes. Availability of biosynthetic human insulin will permit management of those patients with diabetes who cannot be successfully treated with bovine or porcine insulin and will ensure a continuous supply of insulin for everyone with diabetes. The unlimited availability of biosynthetic human growth hormone from cloned genes will effectively eliminate the shortage of this hormone and will make it possible for all children with growth hormone deficiency to be treated. *See* ADENOHYPOPHYSIS HORMONE; INSULIN.

Progress has been made in applications of recombinant DNA technology to the resolution of agricultural problems, especially for the improvement of both crops and livestock. For example, there are possibilities for nitrogen fixation by crops which presently deplete the soil of nutrients, for enhancement of photosynthetic capabilities, and for improvement of the nutritional levels of storage protein in grains. *See* BREEDING (ANIMAL); BREEDING (PLANT); GENETIC ENGINEERING.

Biophysical analysis. Understanding of the structural properties of molecules and the interaction between molecules that constitute biologically important complexes has been facilitated by biophysical analysis. For example, developments in the resolution offered by techniques such as electron

microscopy, x-ray diffraction, and neutron scattering have provided valuable insight into the structure of chromatin, the protein-DNA complex which constitutes the genome of eukaryotic cells. These techniques have also provided clues about modifications in chromatin structure that accompany functional changes. One possible application of biophysical analysis is the diagnosis of human disorders by adaptation of nuclear magnetic resonance for tissue and whole body evaluation of soft tissue tumors, blood flow, and cardiac function. *See* ELECTRON MICROSCOPE; NUCLEAR MAGNETIC RESONANCE (NMR); X-RAY DIFFRACTION.

Flow of Molecular Information

Information for all cellular activities is encoded in DNA; selective elaboration of this information is prerequisite to meeting both structural and biochemical requirements of the cell. In this regard, there are three major areas of investigations by molecular biologists: (1) the composition, structure, and organization of chromatin, the protein-DNA molecular complex in which genetic information is encoded and packaged; (2) the molecular events associated with the expression of genetically encoded information so that specific cellular biochemical requirements can be met; and (3) the molecular signals that trigger the expression of specific genes and the types of communication and feedback operative to monitor and mediate gene control.

Gene composition and organization. In both prokaryotic and eukaryotic cells, genetic information is contained in DNA molecules, but there are differences in the composition and organization of genes which are directly related to variations in expression and control. In all cases, cellular proteins are encoded in DNA sequences utilizing a triplet code of four deoxynucleotide bases (adenine, guanine, thymine, and cytosine), the building blocks of DNA. The DNA sequences also provide for major elements of secondary and possibly higher-order structural features of genes and serve as recognition sites for molecules that participate in gene structure and function. Typically, the central component of the gene is a series of nucleotides encoding a protein, and these protein coding sequences are flanked by stretches of nucleotides which influence expression. *See* GENETIC CODE.

Deoxyribonucleic acid is a long polymer with two polynucleotide chains coiled around a central axis in the configuration of a double helix. The backbone or constant component of each nucleotide chain consists of deoxyribose sugar moieties linked covalently by 3',5'-phosphodiester bonds. The variable component of DNA which renders the molecule capable of encoding genetic information is its sequence of four nitrogenous bases: two purines (adenine and guanine) and two pyrimidines (cytosine and thymine). The bases are covalently linked to the deoxyribose moiety of the backbone by *N*-glycosidic bonds. The bases of the two nucleotide chains are linked to each other by weak hydrogen bonds, with adenine always paired with thymine and guanine with cytosine. The two polynucleotide chains are antiparallel (the 3' end

of one strand faces the 5' end of the other) and oriented so that the purines and pyrimidines reside within the helix and the sugar-phosphate backbone is on the outside. The viability of the double-helical model for DNA has been due to its compatibility with two key molecular and biochemical properties of the molecule: the ability to replicate, and the ability to elaborate genetic information by serving as a template for the synthesis of complementary messenger ribonucleic acid (RNA) molecules which then act as template for the synthesis of genetically encoded proteins. *See* DEOXYRIBONUCLEIC ACID (DNA); NUCLEIC ACID.

The representation, organization, and packaging of genes is directly related to their expression; hence, much ongoing effort has been directed toward studying these properties of both prokaryotic and eukaryotic genes. In the bacterium *E. coli* there is a single chromosome containing a circular DNA molecule approximately 0.06 in. (1.5 mm) in length when unfolded, and with about 4×10^6 nucleotides encoding a limited number of genes (approximately 4000), most represented only once. In contrast, human cells contain 23 pairs of chromosomes, each with a linear DNA molecule. Several yards of DNA are found in the nucleus of every human somatic cell, encoding many thousands of genes, some represented only once per set of chromosomes and others represented by up to several hundred copies. *See* CHROMOSOME; GENE.

Protein-DNA complexes. Both prokaryotic and eukaryotic chromosomes exist as protein-DNA complexes. Although the representation of proteins in prokaryotic chromosomes is limited, several bacterial proteins have been extensively characterized and shown to interact with specific regions of genes, thereby regulating expression. Lac repressor is an example of a bacterial protein which interacts with a specific series of nucleotides in the gene encoding β -galactosidase and influences the extent to which β -galactoside messenger RNA (mRNA) can be synthesized.

In contrast, the chromosomes of eukaryotic cells are protein-DNA complexes with a 2:1 protein-to-DNA ratio. A large body of experimental results indicates that the chromosomal proteins of eukaryotic cells are responsible for packaging the huge DNA molecule within the confines of the cell nucleus, which is only several micrometers in diameter. Modifications in the interactions of chromosomal proteins with eukaryotic DNA have also been shown to occur in conjunction with changes in gene activity and have therefore been implicated in the control of gene expression. Accessibility of specific regions of genes to nucleases, enzymes which degrade DNA unprotected by chromosomal proteins, has provided an effective probe for establishing such structural-functional relationships. *See* DEVELOPMENTAL GENETICS.

There are five classes of basic chromosomal proteins known as histones which are enriched in arginine, lysine, and histidine residues. These histone proteins are present in an amount equivalent to DNA and have been shown to interact with one another and with DNA to form the primary unit of

eukaryotic chromatin structure, the nucleosome. The other chromosomal proteins of eukaryotic cells are designated nonhistone proteins and constitute a complex and heterogeneous array of molecules which interact with DNA and appear to participate in structural, enzymatic, and regulatory activities.

DNA rearrangements. It has become evident that genes are not necessarily static. Alterations in the representation and organization of DNA nucleotide sequences can occur in both bacteria and eukaryotic cells. For example, viruses can introduce DNA sequences into chromosomes, and a series of transpositions or movements of DNA sequences from one location to another has been documented in bacterial, plant, and animal cells. Flexibility in gene structure and organization appears to be related to biological function. There are several biological situations in which rearrangements of DNA sequences occur: (1) in conjunction with production of immunoglobulins; (2) with the development and onset of both solid tumors and human leukemias; and (3) with changes in the phenotypic properties of corn.

Extrachromosomal genes. A complete evaluation of a cell's genes and gene functions must take into account the presence of extrachromosomal DNA. Bacterial cells can carry extrachromosomal genes, such as those that confer antibiotic resistance, in the form of plasmids, which are covalently closed circular DNA molecules. Stable sets of extrachromosomal genes in eukaryotic cells are those in the mitochondria, and in chloroplasts, which encode several of the proteins required for the functional integrity of those organelles. Extrachromosomal DNA can also accumulate in eukaryotic cells in conjunction with the development of drug resistance, as has been shown for the amplification of dihydrofolate reductase genes as cells acquire tolerance to the folate antagonist methotrexate. *See* DRUG RESISTANCE; MITOCHONDRIA.

Gene expression. Expression of genetic traits both in bacteria and in eukaryotic cells requires the selective elaboration and processing of genetically encoded information. This occurs by transcription of DNA into RNA molecules that are complementary to the gene; processing of these RNA transcripts into mRNAs; and association of the mRNAs with ribosomes where the nucleotide sequences serve as a template for synthesis of a protein. Three types of RNA are required for the process: mRNA which carries the blueprints for proteins from DNA to ribosomes; ribosomal RNAs (rRNA) which serve as major components of the protein synthesis apparatus; and transfer RNAs (tRNA) which carry amino acids, the primary units of proteins, to the ribosome-associated mRNAs where they are assembled into proteins. *See* RIBONUCLEIC ACID (RNA); RIBOSOMES.

In prokaryotic cells, gene expression is rapid and considerably less complex than in eukaryotic cells. Prokaryotic mRNAs are generally encoded in contiguous DNA nucleotides, and with minimal modifications they function as templates for protein synthesis. In most situations, prokaryotic mRNAs are short-lived and some are polycistronic, that is, sev-

eral proteins are encoded in a single mRNA molecule regulated by a single set of control sequences. *See* BACTERIAL GENETICS.

Eukaryotic mRNAs in general are not encoded by contiguous DNA nucleotides; rather, a precursor RNA which contains encoded segments of the mRNA (exons) separated by intervening sequences (introns) is transcribed from the DNA. The intervening sequences are excised and the protein coding sequences are joined together, or "spliced." Further aspects of eukaryotic mRNA processing that occur within the nucleus involve enzymatic modifications of the 3' terminus [addition of a series of adenosine monophosphate (AMP) residues], the 5' terminus (addition of a methylguanine-containing moiety referred to as a CAP structure), and various bases within the mRNA molecule (for example, methylation). During this processing the mRNA becomes complexed with proteins, forming a ribonucleoprotein particle which is exported to the cytoplasm where it can associate with ribosomes. These complicated eukaryotic mRNA processing steps require from 20 to 45 min; present understanding of the specific molecular mechanisms involved is minimal. The diversity of eukaryotic mRNAs is reflected by the proteins they template, the manner in which they are processed, and their lifetimes which range from several minutes to extremely long periods of time and vary as a function of the biological situation. *See* EXON; INTRON.

Protein synthesis in both prokaryotic and eukaryotic cells is a series of critically orchestrated processes in which the information genetically encoded in the mRNA is utilized for the step-by-step covalent assembly of amino acids into a protein molecule. The process occurs entirely in the cytoplasm of eukaryotic cells after completion of mRNA processing, while in prokaryotic cells protein synthesis can be initiated during completion of mRNA synthesis. Protein synthesis requires availability of ribosomes, tRNAs, and amino acids, along with a series of enzymes, factors, and energy sources. In many instances, after completion of protein synthesis, assembly of protein subunits or enzymatic modifications such as addition of acetate, phosphate, methyl, carbohydrate, or polyadenosine diphosphate-ribose moieties are required to yield a biologically functional protein molecule. *See* AMINO ACIDS; PROTEIN.

Regulation of gene expression. The mechanisms operative in the selective expression of specific genes and those responsible for the inactivity of others are major issues in molecular biology. Attention has been directed primarily toward identification of nucleotide sequences that are involved in regulating gene transcription and toward characterization of molecules that influence the availability of genes for transcription and those that affect the processing of RNA transcripts. In prokaryotic cells, gene expression is largely mediated at the transcriptional level, that is, the extent to which many bacterial proteins are synthesized is directly related to the extent that the gene is transcribed into mRNAs. Yet, posttranscriptional control is also an element in prokaryotic gene regulation, since the extent to which some

bacterial proteins are synthesized is governed by the ability of mRNAs to serve as templates for protein synthesis or by the efficiency of specific biochemical steps which participate in the protein synthetic process.

In eukaryotic cells, regulation of gene expression is both transcriptionally and posttranscriptionally mediated. In those situations where expression is controlled at the transcriptional level, there appear to be similarities in the mechanisms operative in prokaryotic and eukaryotic cells. However, because of the numerous steps associated with the nuclear and cytoplasmic processing of eukaryotic mRNAs there are many potential targets where posttranscriptional control can occur. Data suggest that in normal biological situations processing of mRNA transcripts within the cell nucleus and the ability of cytoplasmic mRNAs to associate with ribosomes and serve as functional templates for protein synthesis can influence the expression of specific genes. Also, the molecular basis of several human diseases probably resides in defects at various posttranscriptional levels of gene expression. For example, in at least one of the β -thalassemias, a human blood disorder in which there is a reduced representation of β -globin, an impairment exists in the splicing of the β -globin mRNA precursor molecule. See MOLECULAR PATHOLOGY.

An area of primary relevance to understanding regulation of gene expression, and one where knowledge is extremely limited, is the identification and characterization of cellular and molecular signals which monitor the requirements for specific proteins and which consequently activate the molecular mechanisms required to modulate the transcription, processing, or turnover of the respective mRNAs. Considerable attention has been directed toward the involvement of molecules such as steroid hormones, polypeptide hormones, and cyclic nucleotides as mediators of gene expression. The concept of autogenous control has also been considered as a mechanism that in some situations can provide the molecular basis for regulating expression of a particular gene by using the gene product to destabilize its mRNA or to block further transcription of the gene. See GENETICS.

Gary S. Stein; Janet L. Stein

Bibliography. F. M. Ausubel, *Current Protocols in Molecular Biology*, 2 vols., 1988; A. Darbre, *Introduction to Practical Molecular Biology*, 1988; E. D. DeRobertis and E. M. DeRobertis, Jr., *The Cell and Molecular Biology*, 8th ed., 1987; A. M. Lesk (ed.), *Computational Molecular Biology: Sources and Methods for Sequence Analysis*, 1989; J. D. Watson et al., *Molecular Biology of the Gene*, vol. 1, 4th ed., 1987.

Molecular chaperone

A specialized cellular protein that binds nonnative forms of other proteins and assists them to reach a functional conformation, in most cases through the expenditure of adenosine triphosphate (ATP). Orig-

inally identified by their increased abundance after heat shock, chaperone proteins in general bind to exposed hydrophobic surfaces of nonnative proteins, preventing them from forming intermolecular interactions that lead to irreversible multimolecular aggregation. Thus, the role of chaperone proteins under conditions of stress, such as heat shock, is to protect proteins by binding to incipiently misfolded conformations, preventing aggregation; then, following return of normal conditions, they allow refolding to occur, associated with protein release. See ADENOSINE TRIPHOSPHATE (ATP); PROTEIN.

Protein folding. Chaperones also play essential roles in folding under normal conditions (Fig. 1). Proteins in general contain in their primary amino acid sequences all the information necessary for proper folding. Under cellular conditions, however, "off-pathway" misfolding becomes a major competing reaction, particularly for multidomain proteins, potentially lodging them in kinetic traps (local energetic minima that are separated by energetic barriers from the global energetic minimum typical of the native state). Chaperones provide kinetic assistance to the folding process, binding such trapped, misfolded states through exposed hydrophobic surfaces and expending energy of ATP to place the protein back on a productive folding path. Thus, chaperones are not true folding catalysts; that is, they do not accelerate on-pathway folding. Rather, as kinetic assistants, they serve to smooth out the energy landscape, thus improving the overall rate and extent of productive folding in vivo.

One of the most extensively investigated roles of chaperones is to assist in the proper folding of newly translated proteins in the cytosol. This involves both cotranslational interactions with translating polypeptide chains and posttranslational interactions. Cotranslational interactions are formed with heat shock protein (Hsp) 70 in the eukaryotic cytosol and with a component that has prolyl isomerase activity, known as trigger factor, in the prokaryotic cytosol. These interactions appear to prevent premature folding of nascent polypeptide chains. In the case of Hsp70, this is likely to be mediated by localized binding of the chaperone to hydrophobic segments of extended nascent chain (Fig. 2). Such binding may, therefore, permit the polypeptide to complete translation before folding begins, allowing, for example, the amino terminus and carboxyl terminus to interact if required for proper folding.

Posttranslational chaperone interactions can also involve Hsp70 proteins or partner DnaJ proteins, stabilizing the polypeptide against premature folding or aggregation (Fig. 3). Another major class of chaperones, the chaperonin ring complexes, also provide posttranslational assistance for many proteins. In contrast to localized binding to extended segments as in the case of Hsp70, these megadalton oligomeric complexes bind collapsed, globular, nonnative conformations in the central cavity of a ring, affording multivalent hydrophobic interactions between the protein and the subunits of the surrounding chaperonin ring (Fig. 4). In the eukaryotic cytosol, for example, newly translated actin and

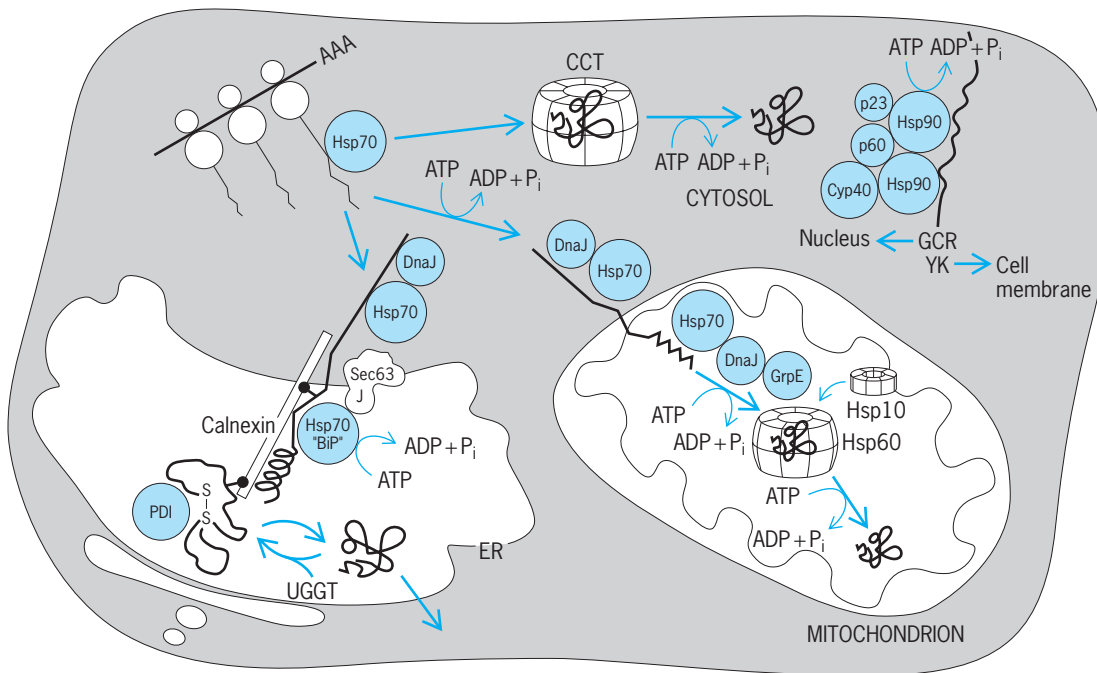


Fig. 1. Chaperone interactions with nonnative proteins in the eukaryotic cell. The interactions are with newly translated proteins in the cytosol and with proteins translocating into the endoplasmic reticulum (ER) and mitochondria, as well as between the Hsp90 complex and proteins involved in signal transduction. The nuclear compartment is not shown. Proteins are generally imported from the cytosol into the nucleus in an already-folded, native state, but chaperones such as Hsp70 proteins and CCT appear to play roles in this compartment under heat stress and in tubulin folding and assembly, respectively. Heavy lines denote nonnative protein; ball-and-stick on ER protein designates monoglucosylated N-linked glycan; PDI, protein disulfide isomerase; UGGT, UDP glucosyltransferase, which participates in the calnexin-calreticulin cycle reglycosylating nonnative N-linked glycan-containing proteins; GCR, glucocorticoid receptor; YK, tyrosine kinase; Cyp40, cyclophilin 40; other members of the Hsp90 complex are designated by molecular size.

tubulin are bound by the chaperonin CCT (containing tailless complex polypeptide 1), which is required for assisting folding into their native form. Similarly, in the bacterial cytoplasm, 5-10% of newly translated proteins interact with the essential GroEL chaperonin following translation. For all of these chaperone interactions, the subsequent binding of ATP to the chaperone allows release of the bound nonnative protein and an attempt at proper folding. In the case of the chaperonins, such release and folding occurs inside the sequestered cavity of a ring, with encapsulation provided for GroEL by association of the lidlike cochaperonin ring, GroES, in the presence of ATP. See CYTOPLASM.

Considering that there are often multiple chaperone family members present in the same cellular compartment (for example, both Hsp70 and GroEL reside in the bacterial cytoplasm), multiple chaperones can have significant affinity for the same nonnative protein conformation, setting up a kinetic competition. Thus, a network of chaperone interactions is present in the various cellular compartments, with nonnative states variously partitioning between chaperones that can assist their folding, as well as other components (such as proteases) that can remove irreparably damaged proteins that would otherwise clog the chaperone machinery.

Other functions. In addition to roles in protection against stress and in de novo cytosolic folding, molecular chaperones participate in the following:

1. Maintaining precursor forms of endoplasmic reticulum and mitochondrial proteins in unfolded,

| CHAPERONE | TOPOLOGY OF BINDING | ACTION |
|-------------------------------|----------------------|---|
| Hsp100 | side top | ATP-dependent disaggregation and unfolding for degradation |
| Hsp90 | Multiprotein complex | Conformational maturation of steroid hormone receptors and signal transducing kinases |
| Hsp70 (Dnak) | side | ATP-dependent stabilization of hydrophobic regions in extended polypeptide segments |
| GroEL (Hsp60) | side top | ATP-dependent facilitation of folding to the native state |
| Small Hsps (Hsp25, and so on) | side | Stabilization against aggregation during heat shock |
| Calnexin-calreticulin | ? | Folding of glucosylated proteins in the endoplasmic reticulum in cooperation with glucosyltransferase |

Fig. 2. Topology of polypeptide binding and action of chaperone families. Polypeptides are shown as wavy lines, and the thickened segments indicate sites that become directly associated with the chaperone, typically hydrophobic in character. (Adapted from B. Bukau and A. L. Horwich, *Cell*, 92:351-366, 1998)

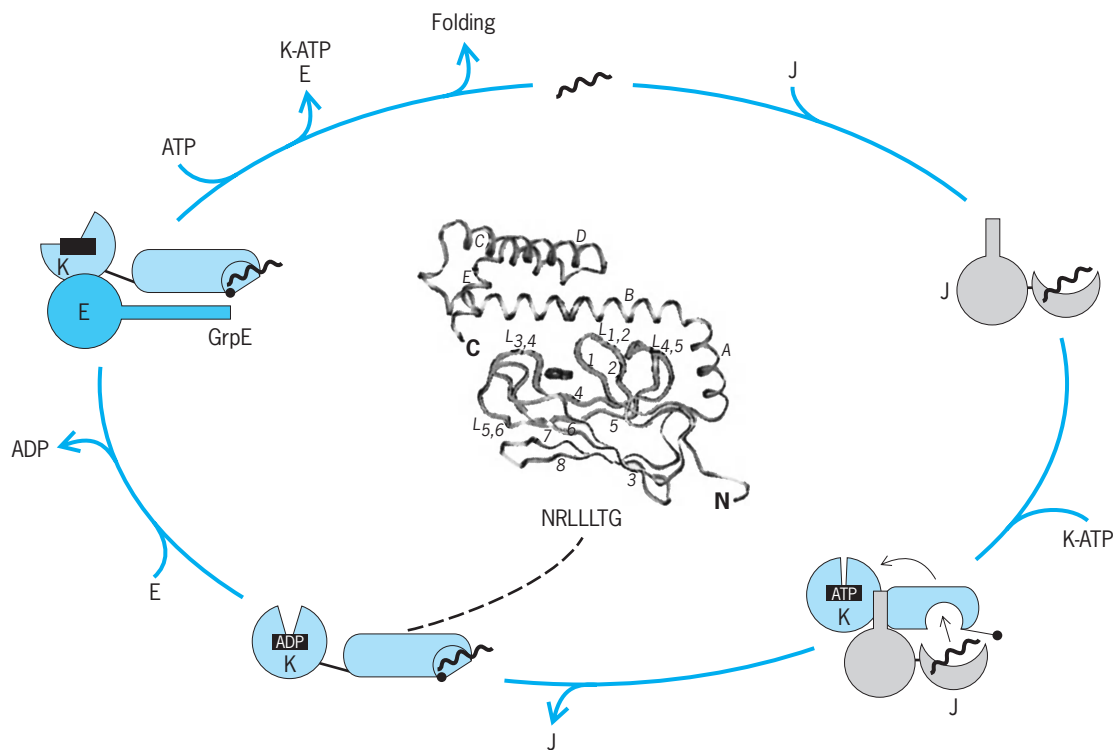


Fig. 3. ATP-directed cycles of nonnative protein binding and folding by DnaK (Hsp70)/DnaJ proteins. DnaJ protein associates with nonnative substrate protein and transfers it to the COOH-terminal peptide binding domain of DnaK. At the same time, association of DnaJ with DnaK stimulates ATP hydrolysis in the NH₂-terminal nucleotide binding domain of DnaK, which allosterically adjusts the COOH-terminal peptide binding domain to “lock in” the transferred substrate protein. (In the structural model in the center of the diagram, hydrophobic peptide, NRLLLTG, is bound to the peptide binding domain of DnaK; the extended peptide is buried in a hydrophobic pocket between loops of a β -sheet; the overlying α -helix may function as a lid.) The protein, stabilized by binding to DnaK, is then released by exchange of nucleotide, with ATP replacing ADP, mediated in the case of bacterial DnaK by an exchange factor known as GrpE, which acts to “pry open” the nucleotide binding site. ATP entry into the nucleotide binding site acts to reduce the affinity of the peptide binding domain, and the bound substrate protein is released to attempt to fold. The homologous eukaryotic Hsp70 proteins carry out similar actions, although the nature of the DnaJ and presence or absence of a GrpE component vary depending on the cellular compartment. (Model of the reaction cycle is from B. Bukau and A. L. Horwich, *Cell*, 92:351–366, 1998; the NRLLLTG structure is adapted from X. Zhu et al., *Science*, 272:1606–1614, 1996).

extended states in the eukaryotic cytosol, enabling the precursors to be recognized and translocated into these organelles.

2. Enabling translocation of proteins into the endoplasmic reticulum and mitochondria, by action at the trans (inside) aspect of the organellar membrane, mediated by distinct Hsp70 family members inside the organelles.

3. Assisting folding of proteins imported into mitochondria, involving the mitochondrial Hsp70/DnaJ/GrpE system, prolyl isomerases, and the Hsp60/Hsp10 chaperonin system.

4. Assisting folding of some proteins imported into the endoplasmic reticulum, involving another Hsp70 (BiP), thiol oxidoreductases, and the lectins calnexin and calreticulin, in cooperation with a glucosyltransferase (UGGT).

5. Binding near-native forms of signal transduction kinases and nuclear receptors in the cytosol, a role of complexes containing Hsp90 and partner proteins that include prolyl isomerases.

6. Functioning as ATP-driven enzymes to recognize appropriately tagged (for example, ubiquitinated) proteins, unfold them, and then bind

and translocate them for degradation into coaxially associated proteolytic cylinders such as the proteasome, mediated by hexameric Hsp100 ring assemblies and the 19S proteasome cap. See CELL (BIOLOGY); ENDOPLASMIC RETICULUM; MITOCHONDRIA; PROTEIN DEGRADATION.

Mechanics. The mechanics of ATP-driven action of chaperone proteins are best understood for the Hsp70 and chaperonin ring systems (Figs. 3 and 4). For both systems, it is the energy of ATP binding, not hydrolysis, that drives release of bound polypeptide and initiation of folding. For example, in the case of the GroEL chaperonin, binding of ATP and GroES is associated with large rigid-body movements of the apical and intermediate domains of the bound ring, including a 60° elevation and 90° clockwise turn of the apical domains, all occurring in less than a second. These conformational changes remove the hydrophobic binding surface from the central cavity and release the polypeptide into a cavity whose wall character has become hydrophilic, favoring its folding to the native state. ATP hydrolysis then pushes the chaperonin system forward on its pathway, relaxing the high-affinity interaction between GroEL

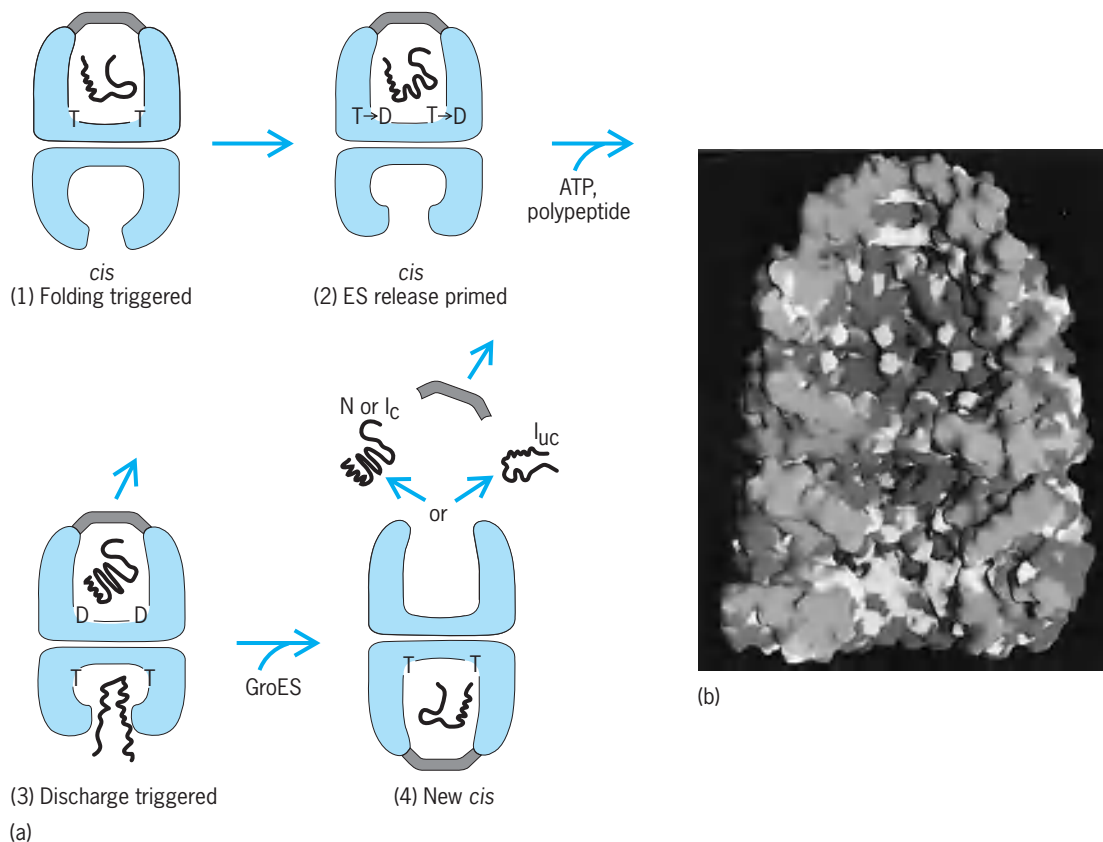


Fig. 4. GroEL/GroES system. (a) Nonnative substrate protein becomes bound to the open ring of a GroEL-GroES-ADP asymmetric complex (panels 2 and 3, bottom ring). In the presence of ATP, which also binds to the open ring, the GroES, folded polypeptide, and ADP ligands are rapidly discharged from the opposite ring (panels 3 and 4, top ring). GroES then binds to the polypeptide-ATP-bound ring (panel 4, bottom ring), producing a large conformational change of the ring, which removes the hydrophobic binding sites from contact with polypeptide and discharges it into the encapsulated, now hydrophilic, cavity, where folding commences. Such folding proceeds in this *cis* ATP complex (panel 4, bottom ring, and panel 1, complex inverted), which has no affinity for polypeptide or GroES on its open *trans* ring. ATP hydrolysis occurs after this longest phase of the chaperonin cycle, producing a *cis* ADP complex (panel 2), which has reduced affinity for GroES on its *cis* ring, and can bind polypeptide, ATP, and GroES on its open *trans* ring. While folding continues seamlessly in the *cis* ring of this ADP complex, the lifetime of this complex is short, ended by binding of polypeptide and ATP in the *trans* ring. Thus the chaperonin cycles back and forth in the presence of nonnative polypeptide and ATP, employing one ringful of 7 ATPs per folding cycle, using these ATPs to simultaneously nucleate a new folding-active *cis* complex while dissociating the old one. (b) Cutaway view of a crystallographic model of the asymmetric *cis* ADP-GroEL-GroES complex, with the hydrophobic cavity surface of the open *trans* ring, the site of capturing a nonnative protein, shown in white, and the hydrophilic character of the cavity of the folding-active *cis* ring, produced following the large-scale domain movements attendant to binding GroES, shown in dark gray. Dimensions of the complex, 184 Å height, 140 Å diameter. (Reaction cycle model taken from A. L. Horwich et al., *Proc. Nat. Acad. Sci. USA*, 96:11033–11040, 1999; GroEL-GroES-ADP model adapted from Z. H. Xu et al., *Nature*, 388:741–750, 1997)

and GroES and ultimately enabling the timed release of GroES and of the encapsulated, now-folded polypeptide. Such an orchestrated utilization of energy to carry out molecular movements and associated work, here the assistance of protein folding by alternation of the character of the cavity walls between hydrophobic and hydrophilic, qualifies such chaperones as molecular machines. Further study will delineate the mechanisms of the other chaperones, as well as define the precise effects of these components on protein conformation. See CELL (BIOLOGY).

Arthur Horwich; Wayne Fenton

Bibliography. R. A. Bradshaw et al. (eds.), *Molecular Biology of Intracellular Protein Sorting and Organelle Assembly*, 1988; A. Horwich (ed.), *Advances in Protein Chemistry*, vol. 59, 2001; G. E. Schultz and R. H. Schirmer, *Principles of Protein Structure*,

1988; W. Voelter et al. (eds.), *Chemistry of Peptides and Proteins*, vol. 3, 1985.

Molecular cloud

A large and relatively dense cloud of cold gas and dust in interstellar space from which new stars are born. Molecular clouds consist primarily of molecular hydrogen (H_2) gas, have temperatures in the range 10–100 K, and contain 10^{31} – 10^{36} kg of mass (for comparison the mass of the Sun is 2×10^{30} kg). Molecular clouds are among the most massive gravitationally bound objects in the Milky Way Galaxy.

Molecules. Molecular hydrogen is not directly observable under most conditions in molecular clouds. Therefore, almost all current knowledge about the

properties of molecular clouds has been deduced from observations of trace constituents, mostly simple molecules such as carbon monoxide (CO), which have strong emission lines in the centimeter-, millimeter-, and submillimeter-wavelength portions of the electromagnetic spectrum. The first molecules were detected as optical absorption lines in the spectra of bright stars lying behind molecular clouds. During the 1960s, centimeter-wavelength radio emission lines of hydroxyl (OH), water (H₂O), and ammonia (NH₃) were detected from the vicinity of bright optical nebulae such as the Orion Nebula (Messier 42). In 1970, the most important tracer of molecular gas in space, carbon monoxide, was discovered at a wavelength of 2.6 mm. *See* RADIO ASTRONOMY.

Properties. Molecular clouds are the principal sites of ongoing star formation. Therefore, they tend to be associated with young stars and star-forming regions. The nearest star-forming clouds are found in the constellations Ophiuchus, Taurus, and Perseus, at distances of 125, 140, and 300 parsecs (1 parsec = 3×10^{13} km or 2×10^{13} mi), where the nearest regions of active low- and intermediate-mass star formation are found. These cloud complexes have masses ranging from several thousand to perhaps over 10,000 times the mass of the Sun. Individual cloud cores can have masses as small as the Sun. However, most of the molecular gas in the Milky Way Galaxy is concentrated into giant clouds with masses more than 100,000 times the mass of the Sun. More than 10,000 giant molecular clouds have been found so far. The nearest giant molecular clouds are located at a distance of 460 parsecs toward the constellation Orion (*see* colorplate) where, over the last 10^7 years, they gave birth to tens of thousands of stars, including several dozen relatively rare high-mass stars. Some of the youngest massive stars light up the Great Nebula in Orion. *See* ORION NEBULA; PROTOSTAR.

Most molecular clouds have temperatures of only 10 K. Molecular clouds are orders of magnitude more dense than the general interstellar medium, with gas densities ranging from about 10 molecules per cubic centimeter on large scales to over 10^6 molecules per cubic centimeter in cloud cores. The sizes of individual clouds range from less than 0.1 parsec for small clouds and dense cores to over 100 parsecs for giant molecular clouds. In addition to star-forming molecular clouds concentrated toward the plane of the Milky Way, there are many smaller and lower-density molecular clouds visible most clearly far away from the galactic plane. These so-called high-latitude molecular clouds tend to be nearby, with the closest ones only about 50 parsecs from the Sun. While star-forming molecular clouds are bound by their own gravity, and are therefore long-lived objects with lifetimes ranging from millions to tens of millions of years, high-latitude clouds do not appear to be gravitationally bound. They may rapidly form and dissolve in the turbulent interstellar medium. Thus, these transient objects may survive as molecule-bearing clouds for less than 100,000 years.

Structure. Molecular clouds have a very complex internal structure consisting of clumps and filaments of dense gas surrounded by interclump gas of much

lower density. Individual clumps usually have supersonic internal motions with a velocity of several kilometers per second. The powerful outflows produced by young stars during the first 100,000 years of their existence may be a major source of these chaotic motions. Magnetic fields which thread molecular clouds may play a role in the longevity of turbulent motions and may support clouds against gravitational collapse. Clumps within a cloud frequently exhibit even larger relative motions produced by the gravitational potential of the entire cloud. Clouds near the galactic center have very large internal motion, frequently with internal velocities of more than 30 km/s (20 mi/s).

Composition. Carbon monoxide is the second most abundant molecule, after hydrogen. There is 1 carbon monoxide molecule for about every 10,000 hydrogen molecules. About 100 different chemical species have been so far identified within molecular clouds, indicating that there is a rich chemistry taking place. Over 1' of the mass of molecular clouds is in the form of interstellar dust grains that absorb starlight, making molecular clouds opaque at visible and ultraviolet wavelengths. Therefore, most nearby clouds can be seen in silhouette against the background of stars. *See* INTERSTELLAR EXTINCTION.

Molecular hydrogen is formed primarily from atomic hydrogen by chemical reactions taking place on the surfaces of dust grains. However, most other chemical species are produced by gas-phase chemical reactions involving the interactions of neutral species with ions. Energetic cosmic-ray particles moving through the molecular gas ionize a small fraction (about 1 part in 10^7) of the molecules, resulting in the formation of the highly reactive H₃⁺ ion, which drives the ion-neutral reactions leading to the formation of other observed chemical species. The small abundance of ions (and electrons) in molecular clouds is sufficient to couple cosmic magnetic fields to the gas. The magnetic field, together with the gravitational force, probably regulates the rate at which gas can undergo gravitational collapse to form stars. *See* COSMOCHEMISTRY; MAGNETOHYDRODYNAMICS.

Distribution. Since the discovery of carbon monoxide, surveys of its emission lines have demonstrated that molecular clouds are widespread in the plane of the Milky Way Galaxy. The majority of clouds lie in the broad Molecular Ring encircling the galactic center with an inner radius of about 3 kiloparsecs and an ill-defined outer radius extending to beyond 20 kpc. The number of clouds is greatest near the inner boundary of the Molecular Ring with a gradually diminishing concentration toward the outer Galaxy. While molecular clouds are widely distributed around the ring within 6 kpc of the galactic center, they tend to be highly concentrated in spiral arms in the outer Milky Way. There are relatively few clouds between the inner edge of the Molecular Ring and the immediate vicinity of the galactic center. However, the inner 0.5-kpc region of the Milky Way contains the greatest concentration of clouds in the entire Galaxy. This relatively small region near the galactic center contains nearly 10' of the molecular

gas in the Milky Way. It is thought that the stars in this part of the Galaxy are concentrated into an elongated cigar-shaped ridge called a galactic bar. Apparently, the rotation of this bar about the galactic center has swept up most of the interstellar gas lying inside the inner edge of the Molecular Ring, concentrating it within 0.5 kpc of the center. Nearly half of all of the mass of gas contained in the Milky Way's interstellar medium which fills the space between the stars is contained in molecular clouds.

Formation, evolution, and destruction. Molecular clouds are believed to survive for several times 10^7 years. The ultraviolet radiation produced by massive stars born from the clouds dissociates molecules, fully ionizes atoms and molecules, and heats the remaining gas. The resulting pressure gradients accelerate and disperse the molecular cloud, leading to its destruction.

The formation of molecular clouds is poorly understood at present. Compression of low-density atomic gas by the passage of a large-scale shock wave or a spiral arm of the Milky Way Galaxy, the tendency of atomic gas to cool by radiation, and the self-gravity of the gas may all play a role in the formation of molecular clouds.

Clouds in galaxies. Most spiral and irregular galaxies that form stars contain molecular clouds. On the other hand, elliptical galaxies contain very little molecular gas. Certain peculiar galaxies that are extremely luminous at infrared wavelengths contain very large amounts of molecular gas and are believed to be forming stars at a rate of many hundreds of solar masses per year. There is evidence that these starburst galaxies are produced by the merging of two gas-rich spiral galaxies. See STARBURST GALAXY.

In isolated spiral galaxies such as the Milky Way Galaxy, the total mass of molecular gas in the interstellar medium (at present about 2×10^9 times the mass of the Sun) probably decreases with time as gas is consumed by star formation. In the absence of substantial infall of fresh gas into the galaxy from the intergalactic medium, the mass of gas in the interstellar medium is halved approximately every few billion years. See GALAXY, EXTERNAL; INTERSTELLAR MATTER.

John Bally

Bibliography. F. Combes, Distribution of CO in the Milky Way, *Annu. Rev. Astron. Astrophys.*, 29:195-237, 1991; N. J. Evans II, Physical conditions in regions of star formation, *Annu. Rev. Astron. Astrophys.*, 37:311-362, 1999; R. A. James and T. J. Millar, Molecular clouds, *Proceedings of a Conference at the Department of Astronomy, University of Manchester, March 26-30, 1990*, Cambridge University Press, 1991; W. B. Latter et al. (eds.), CO: Twenty-Five Years of Millimeter-Wave Spectroscopy, *IAU Symp.*, no. 170, Kluwer Academic, 1997; V. Mannings, A. P. Boss, and S. S. Russell (eds.), *Protostars and Planets IV*, University of Arizona Press, Tucson, 2000; G. Winnewisser and G. C. Pelz, The physics and chemistry of interstellar molecular clouds, *Proceedings of the 2d Cologne-Zermatt Symposium, Zermatt, Switzerland, September 21-24, 1993*, XV, 393, Springer-Verlag, Berlin, 1995.

Molecular electronics

The use of molecules as active electronic devices. As contemporary integrated-circuit technology approaches its limits, molecular electronic devices are being pursued not just as a potential successor to electronic systems but also as a method for introducing new functionality or interfaces.

Quantum transport. Electronic transport in nanoscale systems deviates significantly from that in macroscopic systems due to quantization effects. These quantization phenomena are the quantization of charge due to the discreteness of charge on the electron, the quantization of energy due to quantum-mechanical size effects, and the quantization of conductance due to the quantum-mechanical transmission through a system. See QUANTUM MECHANICS.

Charge quantization manifests itself as the impediment or inability of current to flow through a circuit due to the electrostatic repulsion of electrons. When electronic devices are small enough, the addition of even a single electron is proportionally larger than the energy supplied by the bias voltage. This effect, called Coulomb blockade, causes the electrons to shuttle through the device one at a time, analogous to a turnstile.

Energy quantization is a basic tenet of quantum mechanics, the most obvious being the electronic shell structure of atoms. Synthetic structures also can show such quantization, although the energy separation is usually very weak (small) because the effect scales inversely with the length scale. See ELECTRON CONFIGURATION; ENERGY LEVEL (QUANTUM MECHANICS).

Conductance quantization is the most subtle effect, only becoming relevant when the length scale of the device is typically less than a few nanometers, which is the length scale over which the electron retains quantum-mechanical coherence. However, the effect is powerful because it predicts that the conductance, G , of a device is related solely to the quantum-mechanical transmission coefficient, T , by the Landauer formula,

$$G = \frac{2e^2}{h} \sum_{n=1}^N T_n$$

where e is the electronic charge, h is Planck's constant, and the sum is over the N channels through which electrons can transmit. The transmission coefficient (≤ 1) is the dominant quantity that can be designed or varied over many orders of magnitude.

Depending on details of the system, any combination of these effects can be observed in nanometer-scale devices. Molecular-scale devices are especially susceptible to these effects, which are increasingly important with decreasing length scale.

Molecular transport. The most simple case to consider is a single molecule bridging two electrical contacts. A pioneering advance in the field was the realization that functional end groups can be substituted onto the molecule of choice, allowing the molecule to be placed into a junction by chemical

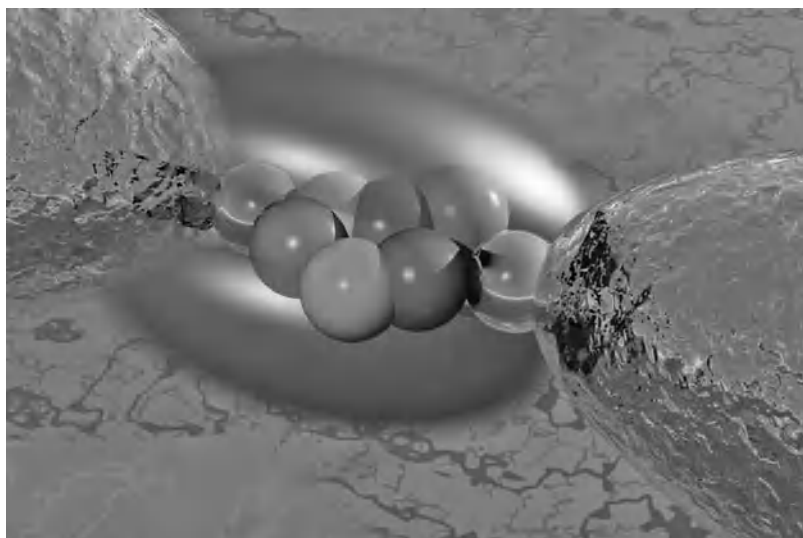


Fig. 1. Schematic of a break-junction apparatus for measuring the electrical properties of a molecular junction. (Courtesy of Mark A. Reed)

reaction. The most common example is the bonding of sulfur (thiol) end groups to coinage metals such as gold. An additional advance was the stable creation of nanometer-scale electrode structures bridged by a single molecule, which has been demonstrated by break junctions (Fig. 1). Scanning tunneling microscopy serves as a critical diagnostic and spectroscopic tool, although in a different regime as direct physical contact to the counter electrode (tip) is replaced by a vacuum gap. See SCANNING TUNNELING MICROSCOPE.

Molecular devices. A combination of experimental results for both single and multiple molecules in parallel (for larger devices, self-assembled molecular layers such as that in Fig. 2 are used) shows that molec-

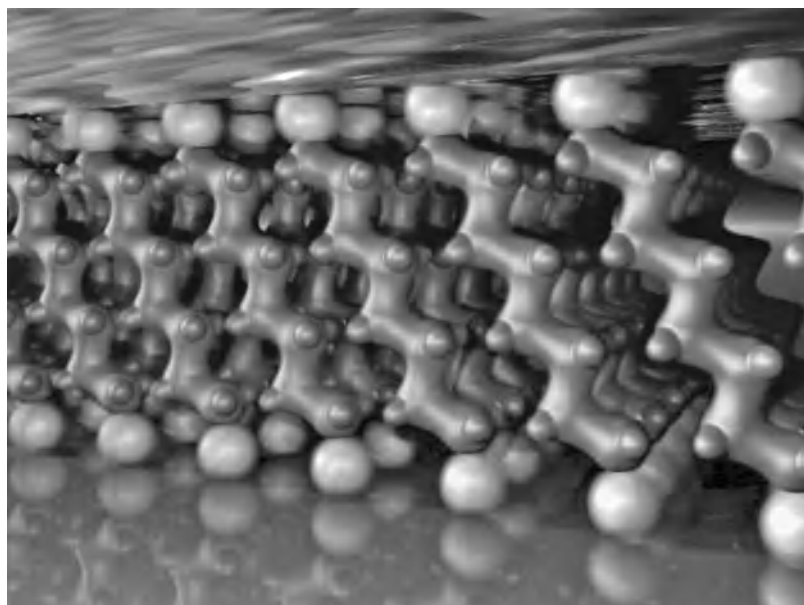


Fig. 2. Illustration of self-assembled monolayer. (Courtesy of Mark A. Reed)

ular devices can be tuned through the entire range of quantum transport regimes. In the low-transmission-coefficient regime, the conductance is shown to fit tunneling models well, with expected dependencies on molecular structure and length. If charge centers are placed in the molecule, the Coulomb blockade regime can be achieved; and this can act as the equivalent of an electronic memory. The energy quantization regime in this case reduces to the probing of molecular orbitals, and this potential tentatively has been demonstrated. See MOLECULAR ORBITAL THEORY; TUNNELING IN SOLIDS.

Molecular devices have two distinct differences from solid-state electronic devices. First, molecules will nearly always have strong vibronic coupling, which could have a dramatic effect on molecular transport regime or, in some cases, even switch the regime. These vibronic modes have been successfully demonstrated as a diagnostic tool, providing for the first time a “fingerprint” of the molecule in the junction, which had been problematic due to the lack of characterization capability. Second, molecules can be made exquisitely selective to a given reaction. Although this capability has yet to be fully explored and the best device structure has yet to be demonstrated, experiments have been demonstrated in which changing the charge state of the molecule affects the conductance of a nearby conventional electronic device. See MOLECULAR STRUCTURE AND SPECTRA.

Applications. Since the mid-1980s, the tremendous growth of molecular electronics is a reflection of the fabrication capability due to surface functionalization as well as novel device approaches that can reach to the ~ 1 nanometer length scale with repeatability. This progress bodes well not only to the pure intellectual understanding of molecular-scale electronics but also to its eventual system applications. The advantages of molecular device integration into complex systems are an active area of research, as contemporary integrated circuits struggle with power dissipation, materials properties, and interconnect problems. Molecular components offer the possibility of self-assembled processing and different functionality (as well as alternatives to charge-based computation), and represent the ultimate endpoint of any electronic system. Another application issue is precision, mainly of the interface to the active molecules. At the current stage of structural uncertainty, one sees fluctuations in expected device behavior due to this uncertainty. Advances in the field, as well as most other electronics approaches on this length scale, await the development of a truly planar (or 3D) atomic-scale technology with a precision better than 0.1 nm as well as structural characterization tools at this size scale. Mark A. Reed

Bibliography. G. Cuniberti, G. Fagas, and K. Richter (eds.), *Introducing Molecular Electronics*, Springer, Berlin, 2005; S. T. Pantelides et al. (eds.), *Molecular Electronics*, MRS Symp. Proc., vol. 582, Warrendale, PA, 2001; M. A. Reed and T. Lee (eds.), *Molecular Nanoelectronics*, American Scientific Publishers, 2003.

Molecular isomerism

The property of compounds (isomers) which have the same molecular formula but different physical and chemical properties. The difference in properties is caused by a difference in molecular structure (that is, molecular architecture). A typical example is dimethyl ether, CH_3OCH_3 , a chemically quite inert

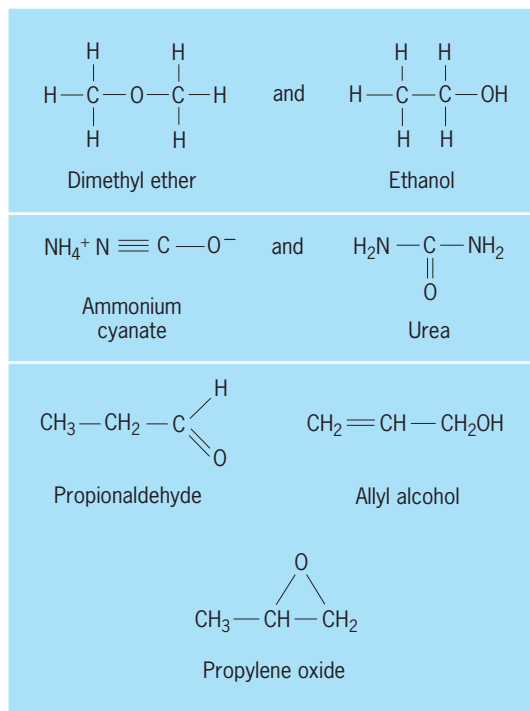


Fig. 1. Functional isomers.

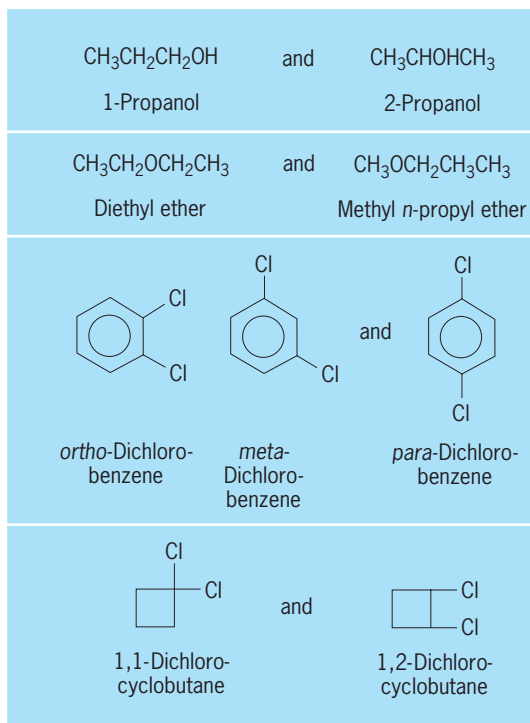


Fig. 2. Positional isomers.

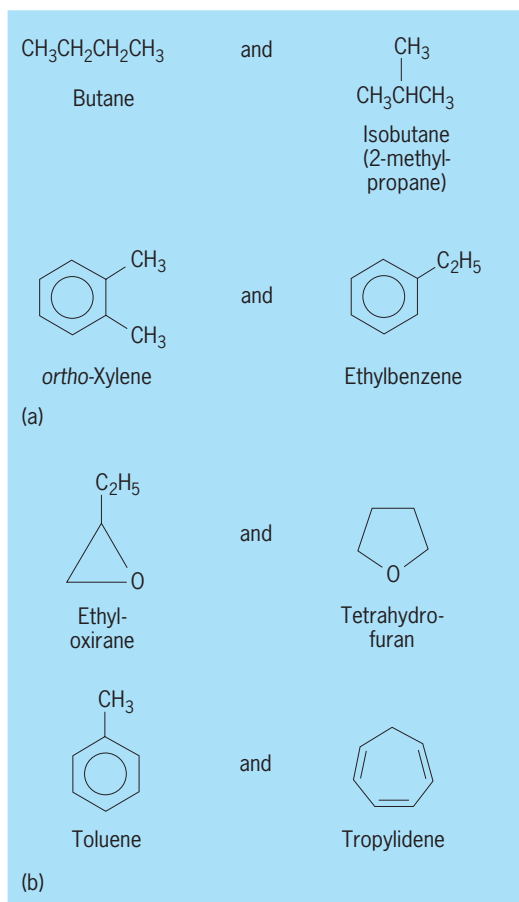


Fig. 3. Skeletal isomers. (a) Chain isomers. (b) Ring isomers.

gas which condenses at -24°C (-11°F), and ethyl alcohol, $\text{CH}_3\text{CH}_2\text{OH}$, a liquid of substantial chemical reactivity which boils at 78°C (172°F); both compounds have the molecular formula $\text{C}_2\text{H}_6\text{O}$.

Classification. Isomers may be classified as constitutional isomers or stereoisomers. Constitutional isomers differ in constitution or connectedness, relating to the question as to which atoms are linked to which others and how. Dimethyl ether and ethanol (Fig. 1) are constitutional isomers. In dimethyl ether each carbon is connected to three hydrogen atoms and the one oxygen atom; the two carbon atoms are thus equivalent. In ethyl alcohol (ethanol) one carbon is linked to three hydrogen atoms and the other carbon; the second carbon is linked to the first carbon, two hydrogens, and the oxygen atom which, in turn, is linked to the sixth hydrogen atom; the two carbon atoms are not equivalent. (Constitutional isomers can often be distinguished, and their constitution recognized, by carbon-13 nuclear magnetic resonance.) Stereoisomers, in contrast, have the same constitution but differ in the three-dimensional array of the atoms in space, called configuration. (In some cases, the difference in three-dimensional arrangement may, however, be due to rotation about single bonds, in which case it is spoken of as a difference in conformation.) See NUCLEAR MAGNETIC RESONANCE (NMR).

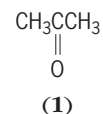
Constitutional isomers. Constitutional isomers have been subdivided into functional isomers, positional isomers, and chain isomers.

Functional isomers. Functional isomers (Fig. 1) differ in functional group, that is, the group (or groups) most material in determining chemical behavior. The ammonium cyanate–urea pair ($\text{CH}_4\text{N}_2\text{O}$) shown in Fig. 1 plays an important role in the history of chemistry inasmuch as the conversion of ammonium cyanate to urea by heating, effected by F. Wöhler in 1828, is considered the first example of an organic compound (urea) having been produced in the laboratory from a mineral one (ammonium cyanate). The third example, shown in Fig. 1, that of propionaldehyde (propanal), allyl alcohol (2-propen-1-ol), and propylene oxide (methyloxirane), illustrates the fact that functional isomers do not necessarily come in pairs. The three compounds all correspond to the molecular formula $\text{C}_3\text{H}_6\text{O}$, but the first one has an aldehyde function, the second combines a double bond with an alcohol function, and the third one has an epoxide function. Indeed, the number of possible isomers corresponding to a given molecular formula is generally remarkably large. Thus, even for the relatively simple composition $\text{C}_{10}\text{H}_{22}$, a saturated hydrocarbon with 10 carbon atoms, there are 75 isomers, and for a hydrocarbon with twice as many carbon atoms, $\text{C}_{20}\text{H}_{42}$, the number is well over 300,000.

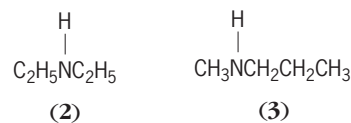
Positional and chain isomers. Positional isomers (Fig. 2) have the same functional group but differ in its position along a chain or in a ring. Closely related are chain isomers which also have the same functional group or groups but differ in the shape of the carbon chain (Fig. 3a); quite similar are ring isomers (Fig. 3b) which differ in the size of one or more rings. Ring and chain isomers together are sometimes called skeletal isomers.

Properties. It should be emphasized that these subclassifications of constitutional isomers are made for

the convenience of the chemist rather than because of any fundamental importance. All constitutional isomers differ in physical and chemical properties, such as melting and boiling points, density, refractive index, and free energy, as well as in all kinds of spectral properties, such as ultraviolet, infrared and nuclear magnetic resonance spectra, and, to a lesser extent, mass spectra. If crystalline, isomers can generally be assigned their proper structure by x-ray diffraction analysis. The above differences tend to be greatest for functional isomers and more subtle for positional and skeletal isomers. However, the last statement is not universally true; thus, there is a considerable difference between the fairly reactive ethyl oxirane and the fairly inert ring isomer tetrahydrofuran shown in Fig. 3b. Indeed, it is often not entirely clear when isomers should be called functional and when they should be called positional or chain isomers. Thus acetone (propanone; 1), and propi-



onaldehyde (propanal), $\text{CH}_3\text{CH}_2\text{CH}=\text{O}$, may be considered positionally isometric carbonyl compounds if one chooses not to distinguish between aldehyde and ketone functions. Yet these functions are substantially different (for example, aldehydes are readily oxidized to acids, while ketones are not), so an alternative might be to consider the ketone and the aldehyde as functional isomers. A similar situation occurs with amines: *t*-butylamine, $(\text{CH}_3)_3\text{CNH}_2$, and *n*-butylamine, $\text{CH}_3\text{CH}_2\text{CH}_2\text{CH}_2\text{NH}_2$, are clearly chain isomers, whereas diethylamine (2), and methylpropylamine (3), should be classified as positional



isomers. When one compares *n*-butylamine with diethylamine, however, the situation is less clear-cut. One could argue that these are positional isomers also, but in view of the functional differences (though slight) between primary and secondary amines, they are probably better classified as functional isomers. Such differences in classification are obviously quite tenuous.

Stereoisomers. Compounds which have not only the same molecular formula but also the same constitution (connectivity of the atoms) but which differ in the disposition of the atoms in space are called stereoisomers. Stereoisomers, in turn, are subdivided into two types: those that are mirror images of each other, called enantiomers, and those which are not mirror images, called diastereomers or diastereoisomers.

Enantiomers. These isomers are unique in that they always come in pairs (Fig. 4). Either a molecule is superposable with its mirror image, in which case it does not have an enantiomer, or it is not superposable with its mirror image, in which case it has one

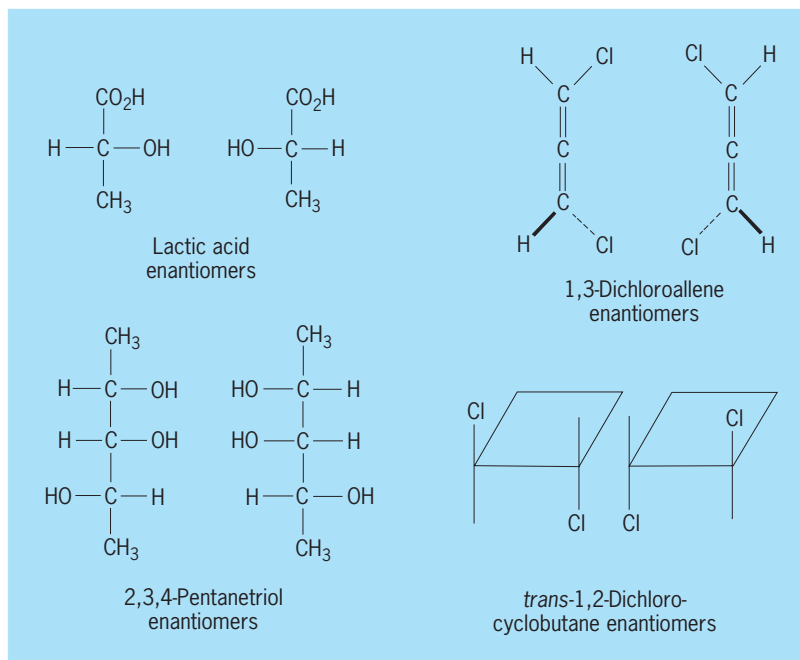


Fig. 4. Enantiomers, mirror-image isomers.

and only one enantiomer (since an object can have only one mirror image). Molecules which are not superposable with their mirror images are called chiral; those which are so superposable are called achiral. Enantiomers are much more alike than are other sets of isomers (constitutional isomers or diastereomers); thus they have the same melting point, boiling point, free energy, spectral properties, x-ray diffraction pattern, and so on. This is because their internal relationships are the same; for example, the distances between corresponding atoms are the same, much as the distances between corresponding fingers are the same in right or left hand. However, enantiomers differ in their behavior toward chiral reagents, much as a right hand and a left hand differ in their relation to a right glove, and they are also different in their behavior toward chiral physical agents, such as circularly polarized light. Thus they differ in circular dichroism and in the direction of rotation which they impart to plane polarized light (optical rotation, optical rotatory dispersion). Such differential physical properties have been called chiroptic properties.

Diastereomers. These isomers have the same constitution but different spatial arrangement and are not mirror images (Fig. 5). They resemble constitutional isomers in that there may be more than two isomers in a set and that their physical, energetic, and spectral properties are generally quite distinct. The example of *cis*- and *trans*-1,2-dibromoethene illustrates *cis*-*trans* isomerism in olefins. (It is recommended that the term geometrical isomers which was formerly used for this type of isomers be abandoned, just as optical isomers should no longer be used as a synonym for enantiomers.) 1,3-Dichlorocyclobutane and 1,2-dimethylcyclopropane illustrate diastereoisomerism (also of the *cis*-*trans* type) in cyclanes. The 1,3-dichlorocyclobutane exists in two achiral diastereomeric forms, whereas the 1,2-dimethylcyclopropane has a pair of (chiral) enantiomers (*trans*) which are diastereomeric with the (achiral) *cis* isomer (called *meso* because it is an achiral diastereomer in a set also containing chiral species). Pentane-2,3,4-triol illustrates a case with one chiral and two achiral (*meso*) diastereomers. The enantiomeric pair of the chiral 2,3,4-pentanetriols is shown in Fig. 4.

A set of stereoisomers containing n chiral centers will normally contain 2^n members. Since each member of the set has an enantiomer, there will be 2^{n-1} enantiomer pairs which, in relation to each other, are diastereomeric. However, when there is degeneracy, that is, when two or more of the chiral centers are equivalent (as in the 2,3,4-pentanetriols where the chiral centers at C-2 and C-4 are alike), there will be fewer isomers than the formula predicts. Thus there are only four stereoisomeric 2,3,4-pentanetriols instead of the eight (2^3) predicted by the formula (see Figs. 4 and 5). A general method for counting stereoisomers has been developed. See STEREOCHEMISTRY.

Limitations. Many isomers are quite stable and if they can be interconverted at all, the barrier between them is quite high (Fig. 6). For example, the barrier between *cis*- and *trans*-2-butene, $\text{CH}_3\text{CH}=\text{CHCH}_3$,

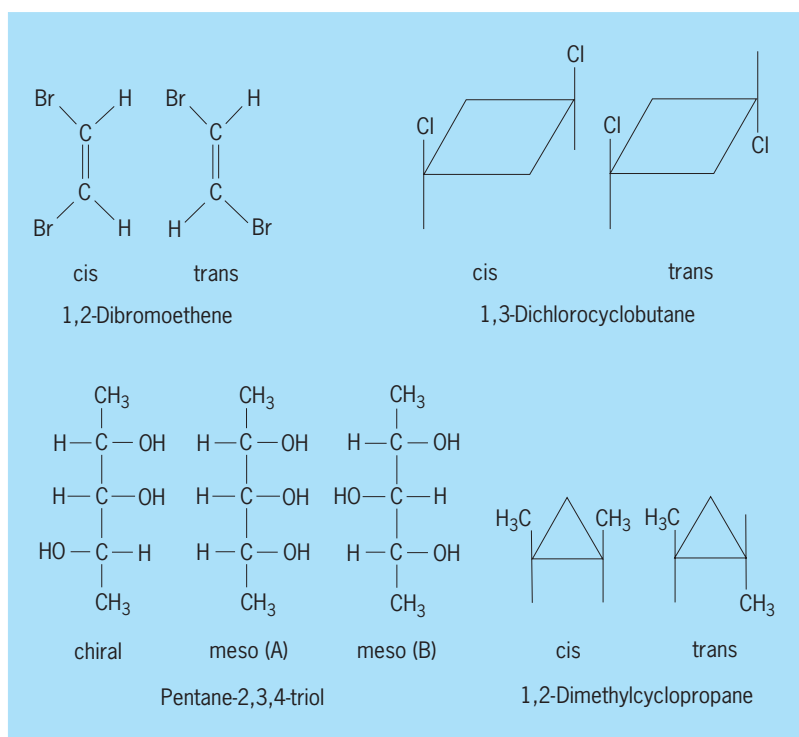


Fig. 5. Diastereomers, isomers having different spatial arrangements of the same groups of atoms.

is of the order of 65 kcal/mol (272 kilojoules/mol), and such isomers (the 1,2-dibromoethenes in Fig. 5 are another example) are spontaneously interconverted only at very high temperatures. In other cases, however, the barriers are intrinsically quite low or are readily lowered by deliberate or adventitious catalysis, frequently by acids or bases. An example relates to the keto and enol forms of ethyl acetoacetate (Fig. 7a). While these isomers can be separated by fractional distillation in clean quartz vessels, they are quickly interconverted in the presence of traces of acids or bases. Such easily interconvertible isomers which differ only in the position of an atom or group (in the case of ethyl acetoacetate, a hydrogen atom which is attached to carbon in the keto form and to oxygen in the enol form) are called tautomers, and the phenomenon is called tautomerism. A closely related type of isomerism in which only bonds shift and atoms remain in place (except for changes in bond distances) is represented by the cyclooctatriene-bicyclooctadiene interconversion (Fig. 7b) which takes place readily at room temperature. This kind of rapid isomerization is referred to as valence bond isomerism or valence tautomerism. Two other examples of rapidly interconverting isomers are shown in Fig. 8. The first is chlorocyclohexane which exists in an equatorial and an axial conformation rapidly interconverted by reversal (flipping) of the cyclohexane chair. The barrier to this interconversion is about 10 kcal/mol (42 kJ/mol), which is so low that the two conformational isomers (sometimes called conformers) can be isolated only at temperatures as low as -150°C (-238°F). However, the two isomers can be seen separately in nuclear magnetic resonance (NMR) spectra

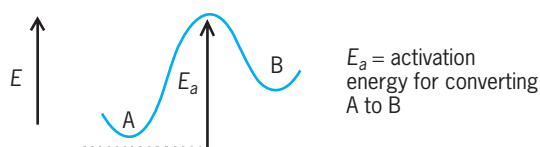


Fig. 6. Energy barriers between isomers A and B.

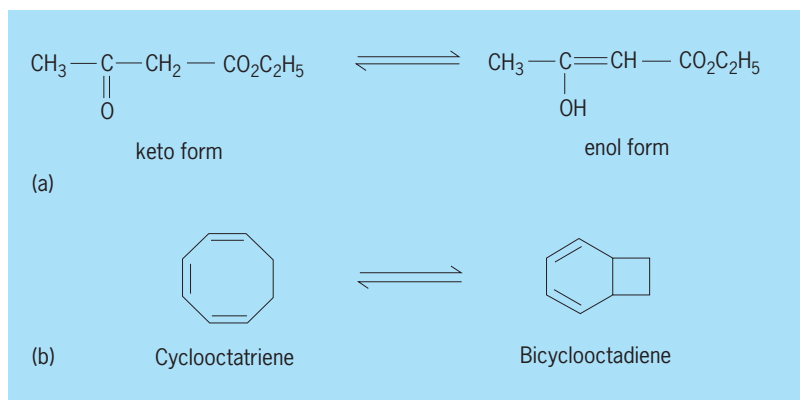


Fig. 7. Easily interconverted isomers. (a) Ethyl acetoacetate. (b) Cyclooctatriene and bicyclooctadiene.

below about -65°C or -85°F ; (the exact temperature depends on the frequency of the instrument and the nucleus observed). The axial and equatorial conformations of chlorocyclohexane are thus diastereomeric. In 1,2-dichloroethane (Fig. 8), three stereoisomeric conformational minima (two enantiomeric gauche conformations and the achiral anticonformation that is diastereomeric to the two others) can be discerned, but the barrier to rotation is so low (about 3 kcal/mol or 13 kJ/mol) that isolation is out of the question, and even many physical techniques (such as electron diffraction, NMR spectroscopy, and dipole moment measurement) yield only weighted average values of the physical properties of the three conformations. However, the gauche isomers and antiisomers can be seen distinctly in the infrared spectrum of the substance and are thus clearly different species.

Ultimately there is the question of what will hap-

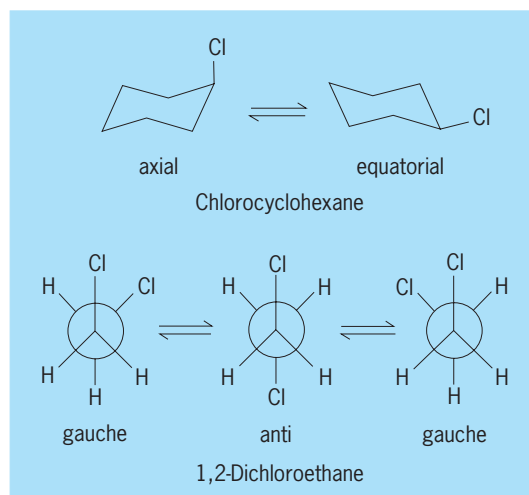


Fig. 8. Conformational isomers.

pen if the barrier to interconversion becomes even lower. A point must come where it is no longer possible to speak of two distinct isomeric molecules but where only a single molecule is deemed to exist. This happens when there is no longer an operational way of demonstrating the existence of two separate energy minima. It has been suggested that there are not two isomeric molecules when the barrier (Fig. 6) is lower than the product of the gas constant R on the absolute temperature T (about 0.6 kcal/mol or 2.5 kJ/mol at room temperature), since under those circumstances the molecule traverses the barrier in a single molecular vibration, but this limit is clearly somewhat arbitrary. In any case, it is clear that the differentiation between two distinct isomeric molecules and two energy states of a single molecule is not a sharp one in those instances where the energy barrier between isomers is very low. See CONFORMATIONAL ANALYSIS; OPTICAL ACTIVITY; TAUTOMERISM.

Ernest L. Eliel

Bibliography. D. Barton and W. D. Ollis (eds.), *Comprehensive Organic Chemistry*, vol. 1, 1979; E. L. Eliel and S. H. Wilen, *Stereochemistry of Organic Compounds*, 1993; Z. Slanina, *Contemporary Theory of Chemical Isomerism*, 1986.

Molecular machine

A molecular device is an assemblage of a discrete number of molecular components (that is, a supramolecular structure) designed to achieve a specific function. Each molecular component performs a single act, while the entire supramolecular structure performs a more complex function, which results from the cooperation of the various molecular components. Molecular devices operate via electronic or nuclear rearrangements. Like any device, they need energy to operate and signals to communicate with the operator. The extension of the concept of a device, so common on a macromolecular level, to the molecular level is of interest not only for basic research but also for the growth of nanoscience and nanotechnology. See NANOTECHNOLOGY.

A molecular machine is a particular type of molecular device in which the component parts can display changes in their relative positions as a result of some external stimulus. Such molecular motions usually result in changes of some chemical or physical property of the supramolecular system, resulting in a "readout" signal that can be used to monitor the operation of the machine (Fig. 1). The reversibility of the movement, that is, the possibility to restore the initial situation by means of an opposite stimulus, is an essential feature of a molecular machine. Although there are a number of chemical compounds whose structure or shape can be modified by an external stimulus (for example, photoinduced cis-trans isomerization processes), the term "molecular machines" is used only for systems showing large-amplitude movements of molecular components.

Natural molecular machines. The concept of machines at the molecular level is not new. The human

body can be viewed as a very complex ensemble of molecular-level machines that power motions, repair damage, and orchestrate an inner world of sense, emotion, and thought. Among the most studied natural molecular machines are those based on proteins such as myosin and kinesin, whose motions are driven by adenosine triphosphate (ATP) hydrolysis. One of the most interesting molecular machines of the human body is ATP synthase, a molecular-level rotary motor. In this machine, a proton flow through a membrane spins a wheellike molecular structure and the attached rodlike species. This changes the structure of catalytic sites, allowing uptake of adenosine diphosphate (ADP) and inorganic phosphate, their reaction to give ATP, and then the release of the synthesized ATP. See ADENOSINE TRIPHOSPHATE (ATP).

Artificial molecular machines. An artificial molecular machine performs mechanical movements analogous to those observed in artificial macroscopic machines (for example, tweezers, piston/cylinder, and rotating rings). The problem of the construction of artificial molecular-level machines was first posed by Richard P. Feynman, Nobel Laureate in Physics, in his address “There is Plenty of Room at the Bottom” to the American Physical Society in 1959: “What are the possibilities of small but movable machines? . . . An internal combustion engine of molecular size is impossible. Other chemical reactions, liberating energy when cold, can be used instead. . . . Lubrication might not be necessary; bearings could run dry; they would not run hot because heat escapes from such a small device very rapidly. . . .” Although clever examples of a few artificial molecular machines (for example, a phototweezer for metal ions) were reported in the 1980s, substantial growth in this research field has occurred only recently, after the development of supramolecular chemistry. Supramolecular species such as pseudorotaxanes, rotaxanes, and catenanes are particularly suitable structures for the design of artificial molecular machines.

Analogously to what happens for macroscopic machines, the energy to make molecular machines work (that is, the stimulus causing the motion of the molecular components of the supramolecular structure) can be supplied as light, electrical energy, or chemical energy. In most cases, the machinelike movement involves two different, well-defined and stable states, and is accompanied by on/off switching of some chemical or physical signal [absorption and emission spectra, nuclear magnetic resonance (NMR), redox potential, or hydronium ion (H_3O^+) concentration]. For this reason, molecular machines can also be regarded as bistable devices for information processing (Fig. 1). See ACID AND BASE.

Pseudorotaxanes, rotaxanes, and catenanes. Rotaxanes consist of a linear, dumbbell-shaped molecular component encircled by a macrocyclic component (Fig. 2a). Bulky groups (called stoppers) attached to both ends of the dumbbell-shaped molecule avoid the possibility of dethreading of the macrocyclic component. In pseudorotaxanes (Fig. 2b) there are no stoppers at the ends of the thread. Cate-

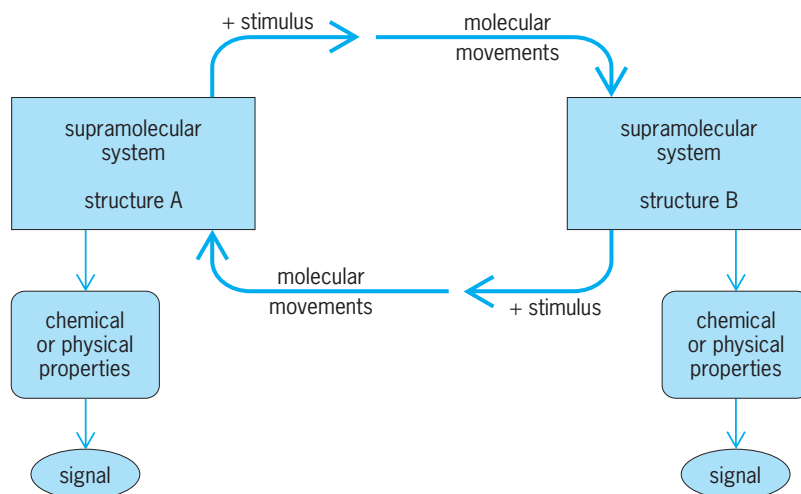


Fig. 1. Working scheme of a molecular machine.

nananes consist of mechanically interlocked macrocyclic rings (Fig. 2c). The rationale and efficient synthetic approaches for the preparation of complex supramolecular systems such as pseudorotaxanes, rotaxanes, and catenanes have been devised only recently. Such strategies usually rely on some kind of interaction (electron donor-acceptor, hydrogen bonding) which allows threading of the wirelike component into the macrocyclic one (pseudorotaxane structure), followed by blocking or cyclization reactions that lead to rotaxane and catenane structures, respectively. See CATENANES; MACROCYCLIC COMPOUND.

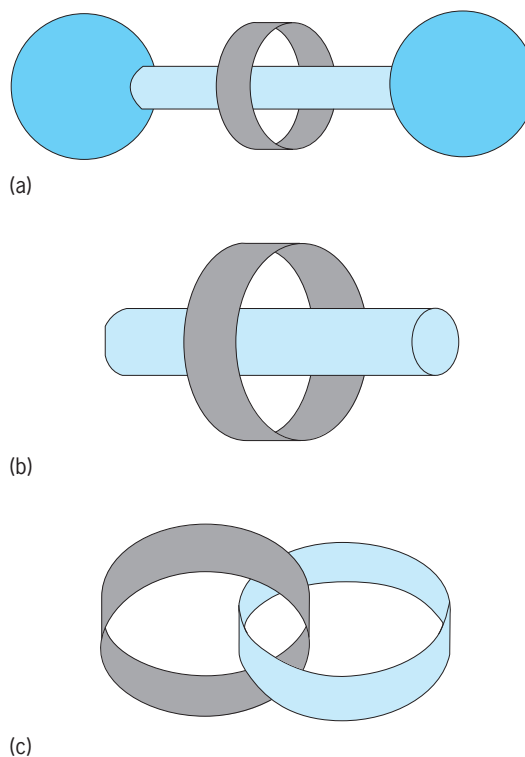


Fig. 2. Schematic representations of (a) rotaxanes, (b) pseudorotaxanes, and (c) catenanes.

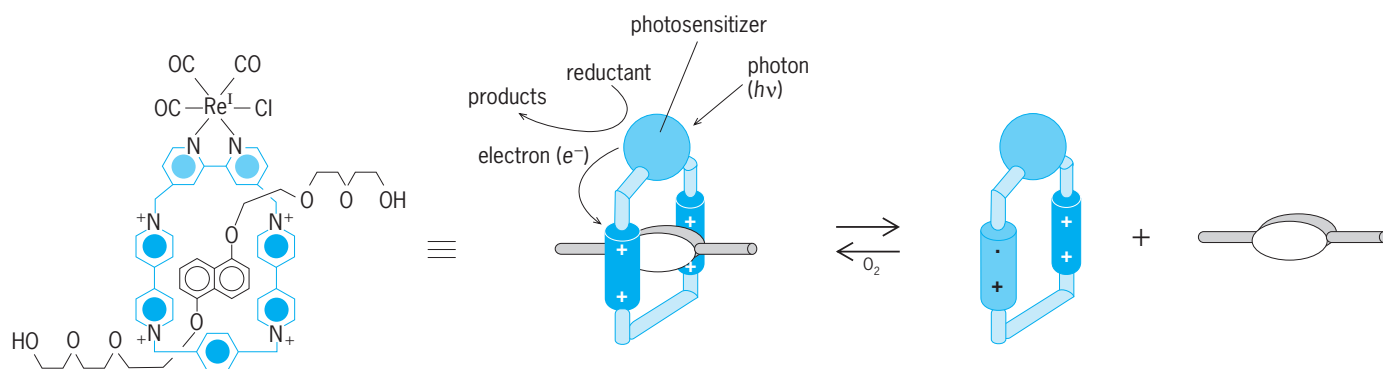


Fig. 3. Light-induced dethreading and oxygen-induced rethreading of a pseudorotaxane.

In the late 1990s several molecular machines based on pseudorotaxanes, rotaxanes, and catenanes were designed and investigated. In all cases the mechanical movements performed by the systems are very simple, so that the term “machine” is perhaps not fully appropriate. However, the extension of the concept of machine to the molecular level is a very stimulating exercise that helps the development of chemistry and underlines new aspects of the bottom-up approach to nanotechnology.

The construction of supramolecular species performing as molecular machines is not an easy task since it requires careful selection of the component units, the design of an appropriate supramolecular structure, and complex synthetic work.

The stimulus causing the motion in molecular machines can be light, electrical energy, or chemical energy. Examples of molecular machines are a pseudorotaxane stimulated by light, a rotaxane stimulated by chemical energy, and a catenane stimulated by electrical energy. The structural changes described below are always accompanied by strong changes in the absorption spectrum, fluorescence, NMR signals, and electrochemical behavior.

Light-fueled “piston/cylinder” molecular machine. In the pseudorotaxane shown in Fig. 3, the driving force for self-threading is the interaction between the bipyridinium electron-acceptor units of the cyclic component and the dioxynaphthalene electron-donor unit of the wirelike component. In order to cause dethreading, the electron donor-acceptor interaction must be destroyed. This can be done by reducing the electron acceptor or oxidizing the electron donor. A rhenium (Re) bipyridine complex, which is an excellent electron-transfer photosensitizer, is incorporated into the cyclic component. Light excitation of the photosensitizer is followed by electron transfer from the excited photosensitizer to one of the bipyridinium units of the cycle. The back electron-transfer reaction is prevented by scavenging the oxidized rhenium complex with a suitable sacrificial electron donor. Since a bipyridinium unit is reduced, the donor-acceptor interaction responsible for the stability of the pseudorotaxane structure is weakened so that dethreading occurs. If oxygen is allowed to enter the irradiated solution, the reduced bipyridinium unit is back-oxidized and rethreading takes place. In conclusion,

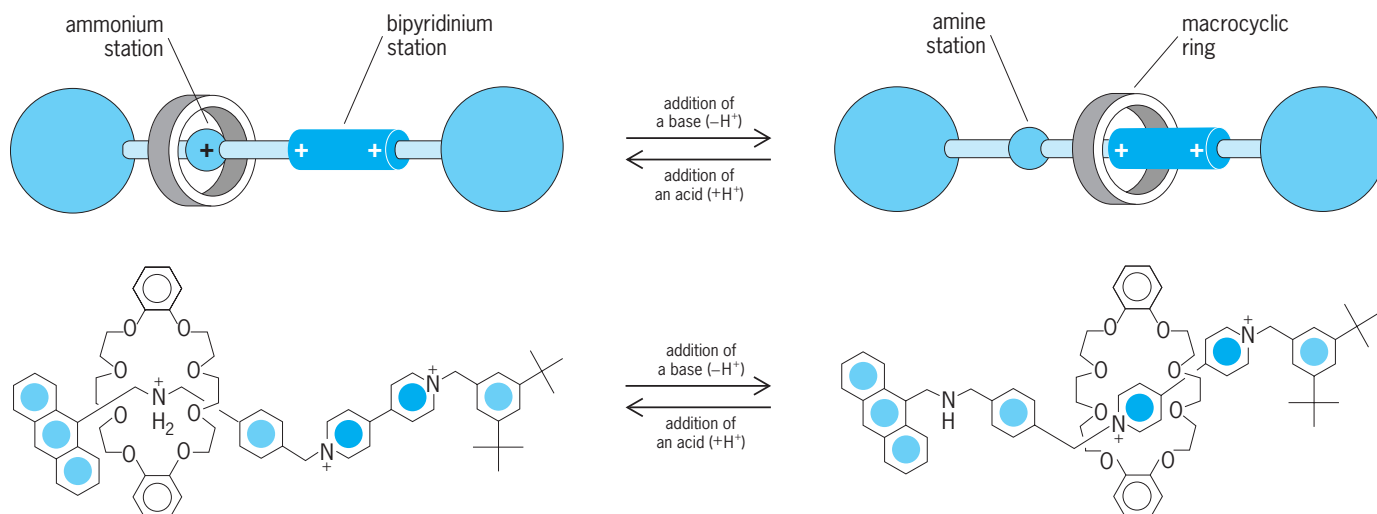


Fig. 4. Acid-base controllable molecular shuttle.

in this system a piston/cylinder-type movement can be caused by a light-fueled motor (namely, the electron-transfer photosensitizer). See PHOTO-CHEMISTRY.

Chemically driven molecular shuttle. Rotaxanes (Fig. 4) have been designed with the purpose of switching the ring component between two different positions (stations) along a wire-type component. The latter component contains two different recognition sites, namely, an ammonium center and an electron-acceptor bipyridinium unit. The ring component is a crown ether which incorporates two dioxybenzene units. Such a ring exhibits a strong affinity, based on hydrogen bonding, for the ammonium center and a lower affinity, based on an electron donor-acceptor interaction, for the bipyridinium unit. At the beginning, the rotaxane exists only as the translational isomer where the macrocyclic component resides on the ammonium recognition site. Upon addition of a base, the ammonium center is deprotonated and the hydrogen bonding interaction with the ring is destroyed, with consequent displacement of the ring on the bipyridinium station, where stabilization is obtained by donor-acceptor interaction. Addition of acid regenerates the ammonium center, and the cyclic component moves back to its original position. See ACID AND BASE; NANO-CHEMISTRY.

Electrochemically driven rotation. The catenane shown in Fig. 5 consists of a symmetric macrocycle containing two electron-acceptor bipyridinium units, and an asymmetric macrocycle containing two different electron-donor units (namely, a tetrathiafulvalene and a dioxynaphthalene unit). Initially, the catenane exists only as the translation isomer, with the better electron donor unit, namely, tetrathiafulvalene, inside the ring containing the two electron-acceptor units. Selective oxidation of the tetrathiafulvalene unit suppresses its electron-donor capacity and therefore destabilizes the original translational isomer with respect to the one in which the best electron donor, which is now the dioxynaphthalene unit, resides inside the electron-acceptor cyclophane. As a consequence, a circumrotation of the asymmetric ring with respect to the symmetric one takes place, with formation of the other translational isomer. Reduction of the oxidized tetrathiafulvalene unit regenerates the original structure.

Information processing. The interest in molecular machines arises not only from their mechanical movements but also from their switching aspects. Computers are based on sets of components constructed by the top-down approach. This approach, however, is now close to its intrinsic limitations. A necessary condition for further miniaturization to increase the power of information processing and computation is the bottom-up construction of molecular-level components capable of performing the functions needed (chemical computer). The molecular machines described above operate according to a binary logic and therefore can be used for switching processes at the molecular level. It has already been shown that suitable designed machinelike sys-

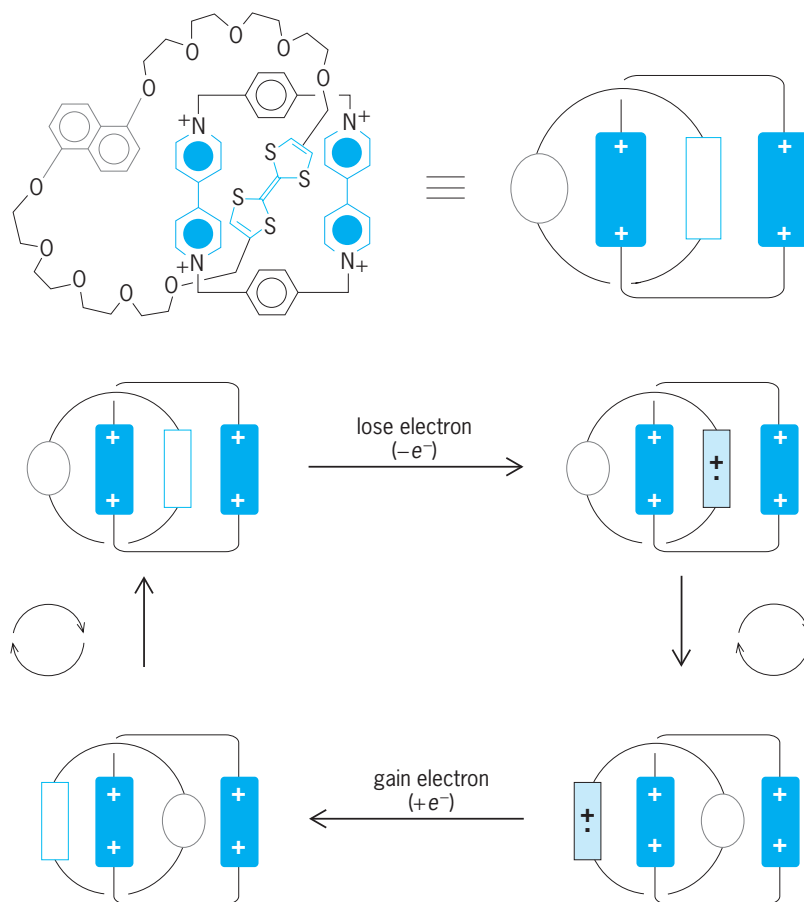


Fig. 5. Electrochemically driven rotation of a ring in a catenane.

tems can be employed to perform complex functions such as multipole switching, plug/socket connection of molecular wires, and XOR logic operation. See LOGIC.

V. Balzani

Bibliography. V. Balzani, M. Gómez-Lopez, and J. F. Stoddart, *Molecular machines*, *Acc. Chem. Res.*, 31:405–414, 1998; A. Credi et al., *Logic operations at the molecular level: An XOR gate based on a molecular machine*, *J. Amer. Chem. Soc.*, 119:2679–2681, 1997; J.-P. Sauvage, *Transition metal-containing rotaxanes and catenanes in motion: Toward molecular machines and motors*, *Acc. Chem. Res.*, 31:611–619, 1998.

Molecular mechanics

An empirical computational method that provides structural, energetic, and property information about molecules. One simple way to consider molecules is as a collection of balls (atoms) held together by springs (bonds)—the basis of many molecular model kits (Fig. 1). Molecular mechanics offers a way to model the behavior of matter mathematically in this manner. Mechanical spring-based theory begins with a fundamental assumption that matter consists of atoms and that the potential energy of a collection of atoms can be defined for every set of positions. The collection of atoms is treated as a mechanical system moving within this

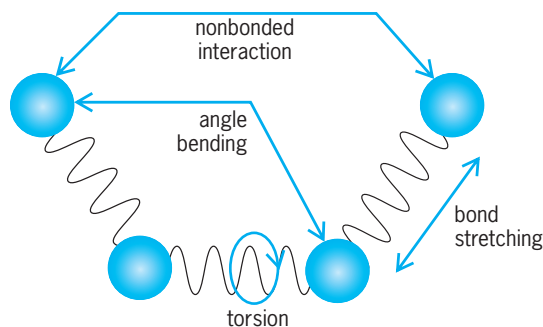


Fig. 1. Ball-and-spring model. Based on the history of physical and chemical analysis of structure and properties, it is logical to think of molecules as mechanical units made up of simple elements like balls (atoms), rods or sticks (bonds), and flexible joints (bond angles and torsion angles).

potential energy, just as a clockwork's motions are determined by its spring potentials. Molecular mechanics methods are a natural outgrowth of concepts of bonding between atoms in molecules and van der Waals forces between nonbonded atoms. In 1930, T. L. Hill proposed the first such potential energy function made up of a simple representation of this type, including van der Waals interactions together with stretching and bending deformation functions. The idea was to minimize this energy function, leading to information about the structure and steric energy in congested molecules. Subsequent studies established more elaborate formulations of the potential-energy functional form, and were used to understand increasingly more detail about molecular systems, such as the equilibrium structure, energetics, thermodynamic properties, and vibrational spectra. Further work in the 1940s verified the theory for more complex reactions. See CHEMICAL BONDING; ENERGY; VAN DER WAALS EQUATION.

Unlike quantum-mechanical approaches, electrons are not explicitly included in molecular mechanics calculations. This is possible due to the Born-Oppenheimer approximation, which states that the electronic and nuclear motions can be uncoupled from one another and considered separately. Molecular mechanics assumes that the electrons in a molecule find their optimum distribution, and approaches chemical problems from the standpoint of the nuclear structure. A molecule from this perspective is considered to be a collection of masses that are interacting with each other via nearly ideal harmonic forces—hence, the analogy to the ball-and-spring model (Fig. 1). Potential energy functions are used to describe the interactions between nuclei. With judicious parameterization, the electronic system is implicitly taken into account. The resulting energy function of structure, $\text{Energy} = f(\text{nuclear positions})$, is presumed to have a minimum corresponding to the most stable equilibrium geometry. Any deviation of the model from the ideal molecular geometry will correspond to an increase in energy. See NUCLEAR STRUCTURE; QUANTUM MECHANICS.

Mathematical formulation. Molecular mechanics methods treat molecules as collections of particles held together by simple forces. The various types of

forces are described in terms of simple individual potential functions, each summed over all respective atoms involved in that force. The total of all such contributions then constitutes the overall potential energy, or steric energy, of the molecular system. The resulting potential energy landscape of a molecule is also referred to as an empirical force field, since the derivative of the potential energy determines the forces acting on the atoms when in motion. In fact, the basic principles of molecular mechanics came out of the field of spectroscopy. Spectroscopists frequently use the term force field to mean a similar (but different) set of equations designed to reproduce or predict vibrational spectra. See MOLECULAR STRUCTURE AND SPECTRA; SPECTROSCOPY.

In its most simplistic representation, Equation (1)

$$E_T = E_b + E_\varphi + E_t + E_{nb} + E_{bb} + E_{el} + \dots \quad (1)$$

expresses the molecular mechanics force field, where E_b is the energy of bond distortion, E_φ that of angular distortion, E_t torsional angle motion, E_{nb} energy associated with atoms that are not bonded together (nonbonded), E_{bb} energy associated with explicit bonds that involve hydrogen, and E_{el} the electrostatic energy, that is, the charge interaction between atoms that are not bonded together. Each of these terms is a mathematical relationship corresponding to a different motion or interaction of the atoms in the molecule, and typically each involves parameters that are obtained from experiment. Other terms may be included in this expression, as indicated by “+...,” depending on the nature of the force field and the types of molecules and associated properties one would like to be able to use the model for. See ATOMIC STRUCTURE AND SPECTRA; INTERMOLECULAR FORCES.

In the simplest force fields, a handful of terms and perhaps a dozen or so parameters are necessary to describe a particular class of molecules, for example, the class of hydrocarbons. The actual mathematical form for the most basic terms in a force field are shown in Eq. 2, along with the various parameters.

$$E_{\text{total}} = \frac{1}{2} \sum_{\text{stretching}} k_b (b - b_0)^2 + \frac{1}{2} \sum_{\text{bending}} k_\theta (\theta - \theta_0)^2 + \sum_{\text{torsion}} k_\tau (1 + [\cos(n\tau - \delta)]) + \sum_i \sum_j \left[\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^6} + \frac{q_i q_j}{\epsilon r_{ij}} \right] + \dots \quad (2)$$

Figure 2 illustrates these basic terms in a manner that indicates the motion or interaction implied. The first term in Eq. (2), the stretching energy, is based on Hooke's law, which states that the energy change that results from movement away from some reference state can be well represented by a parabolic function—the so-called harmonic model of the energy. In this way, the stretching-potential function estimates the increase in energy associated with either stretching or compressing a bond length, l , away

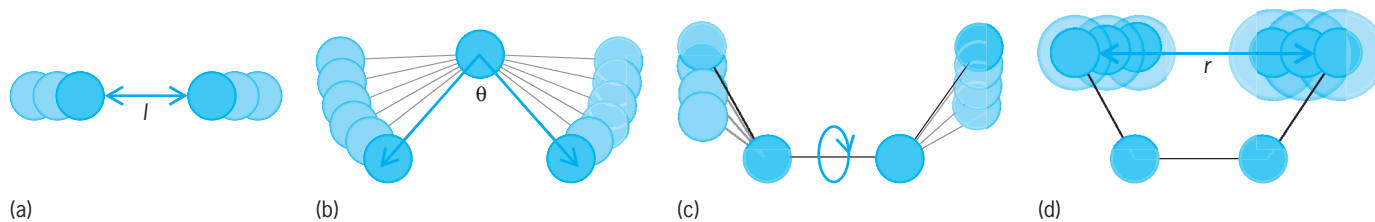


Fig. 2. Ball-and-spring representations of (a) bond stretching, (b) angle bending, (c) torsion, and (d) nonbonded interaction.

from resting position, l_0 (Fig. 2a). The k_s parameter in this term corresponds to the stiffness of the spring (bond), or the amount of force that would be required to restore the bond to the resting state if distorted. Both l_0 and k_s parameters are assigned from experimental data for each pair of bonded atoms and each possible atom type. The expression compares the current length of a bond (l) to its resting length (l_0), squares that difference, and then multiplies by $1/2$ of the restoring force constant, k_s . Then, the summation sign indicates that each bond stretch deformation energy should be summed together for all bonds in the molecule. See BOND ANGLE AND DISTANCE; HOOKE'S LAW.

Similarly, the potential function for angle bending is expressed as a Hooke's law potential corresponding to an angle compression and extension away from the reference state with angle θ_0 (Fig. 2b). The force constant, k_θ , represents the restoring force for the angle bend. The dihedral angle corresponds to an intramolecular rotation involving four bonded atoms, and is represented by a periodic function. The torsional energy involves several parameters, k_τ , and governs the amplitude of the motion; n is the periodicity of the motion, τ is the degree of the torsional motion, and δ is the phase (Fig. 2c). The interaction energy for atoms that are not bonded together (Fig. 2d) includes two types of interactions that to some extent provide information regarding the electrons in the molecule, and these are summed over all possible atom-atom pairs in the molecule. The first type describes the van der Waals attractions and repulsions [Eq. (2)] in the molecule, that is, how the nuclei are attracted and repelled by one another. In Fig. 2d, a Lennard-Jones potential function is used to best model this interaction. The second type of interaction is called the Coulomb energy and provides an energy associated with the charge-charge interaction on two different atoms. In this expression, the atomic charges, q_i and q_j , are parameters, ϵ is the macroscopic dielectric constant shielding the interaction due to the environment, and r_{ij} is the distance between the interacting pair of atoms.

The relative order of magnitude of the terms in Eq. (2) is roughly 100:10:1 for bond distortion:angle distortion:nonbonded interactions. The energy associated with deforming a bond is always much larger (60–200 kcal/mol) than that for angle deformation (6–20 kcal/mol), which is again larger than torsional angle deformation (1–5 kcal/mol). The nonbonded energy is expected to contribute overall the least amount of energy, around 2–3 kcal/mol. In general,

even a simple force field can provide a reasonable approximation to molecular structures and relative energetic differences provided the parameterization is optimal. In any case, absolute values obtained from such an expression (no matter how many terms) will likely be inferior to actual values measured experimentally.

The description of a molecule in terms of harmonic interactions is only a first approximation. As insights into molecular behavior are gained, and as more and better experimental and theoretical data become available and used, increasingly more sophisticated equations are used to reproduce molecular behavior. As such, there are many variations in the functional forms of force fields, typically falling into general classification types. To better reproduce the available experimental data, further optimization of the parameters may be necessary through modification of the equations that make up the force field. Because the resulting molecular mechanics force field is fit with parameters from experiment and then subsequently is tested against a large class of molecules for reproducibility, the individual parameters become highly interdependent, and the individual energy terms ultimately lose their meaning as individual motion components. Rather, the equation as a whole becomes a representation of the potential energy governing the structure and property of molecules within a particular class—a phenomenon known as transferability of the force field.

Application. Knowledge of the structure and properties of a series of molecules or understanding of very large macromolecular structures, such as enzymes and proteins, are problems particularly well suited for study using molecular mechanics. In fact, the mechanical molecular model formulation was developed out of a need to describe molecular structure and properties at a time when relatively little computer power was available, requiring the resulting method to be as practical as possible. The application to molecular systems with thousands of atoms, even with the advent of computational power, still mandates a more practical method, such as offered by molecular mechanics force-field methods. A variety of mathematical and computational approaches have been used over the years. This continues to be an area of active research.

The resulting equations for the total energy [Eq. (2)] are usually combined with algorithms for locating molecular structures, where the energy is at a minimum value over all possible conformations of

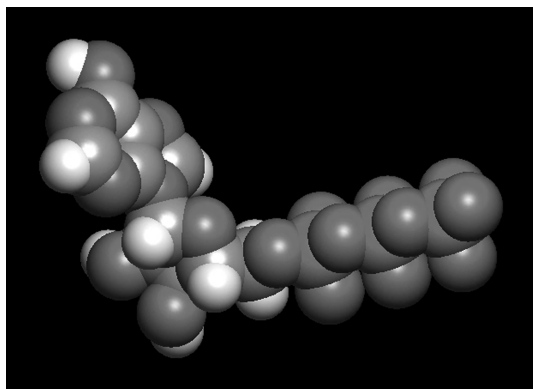


Fig. 3. Small organic molecule.

the molecule. Because this means that many possible conformations of the molecule must be considered, these are iterative procedures. One important consideration is the computational time involved in carrying out a calculation of a very large molecule using the force-field equation. The highest computational cost has to do with the terms in the equation that must sum over all atoms that are nonbonded. In studies involving thousands of atoms, any term involving all atom-atom interactions results in a very large number of evaluations that can ultimately add enormous overhead to the overall computation and, if not programmed efficiently, can bottleneck large computational investigations.

The computer algorithms that carry out the evaluations of the force field and find the minimum energy range from the most simplistic mathematical method (such as steepest descents or conjugate gradients) to sophisticated methods involving higher-order derivatives of the potential energy function (for example, the Newton-Raphson predictor-corrector method) and even grid-based methods. Often the computational procedure invokes a combination of methods to carry out the most practical and accurate method for determining the structure and properties of a molecular system. Many computer programs have been constructed for solving the force-field equations for specialized classes of molecules, and are available to the general research community. See COMPUTATIONAL CHEMISTRY.

Promises and perils. Many important and fundamental questions in science can be better understood through investigations using mathematical models. The general range of application of molecular mechanics includes biological or materials-related macromolecules having thousands of atoms, or even a large number of small molecules (Fig. 3) within a specialized class, such as organic hydrocarbons, organometallic compounds, or inorganic molecules. Computations involving molecular mechanics are mostly restricted to studies involving no bond-making or bond-breaking processes. This is particularly true if the force-field equation involves functions that well represent molecules only around their resting position, and not when they are stretched to a point of significant change in structure and reactivity.

In many investigations, molecular mechanics is used to study molecules in isolation, that is, studies performed in the gas (vacuum) phase. In biological investigations, it may be necessary to consider the effect that the environment has on the structure and how it might react. The environment is generally treated in one of two ways: an explicit or an implicit representation of the environment. In the explicit model the solvent molecules are actually present and taken into account in the evaluation of the total energy, whereas in the implicit approach the solvent is treated as a continuum governed by a function of the dielectric constant. Likewise, it could be important to consider the salt content of the system or other materials-related effects.

Because molecular mechanics-based methods tend to be computationally very efficient, application is practical for conformational analysis studies to predict the most likely conformations for highly flexible molecules (Fig. 4). In structural biology, this is quite important in the protein-folding problem, the goal of which is to predict the three-dimensional structure of a protein from the linear sequence of amino acids. Another wide use of molecular mechanics is in studies involving a small molecule binding into an active site, called docking studies. Estimating correct three-dimensional atomic structures of complexes between proteins and ligands is an important component of the drug-design process in the pharmaceutical industry. Basic aspects of ligand-protein interactions, categorized under the general term molecular recognition, are concerned with the specificity as well as the stability of the ligand to bind into a protein. See MOLECULAR RECOGNITION.



Fig. 4. Large flexible biological molecule.

Structural studies performed using the molecular mechanics model takes into account only the potential energy and not the possibility of any kinetic energy, or motion of the structure. In fact, all molecules contain potential energy and kinetic energy, and the atoms of a molecule are constantly undergoing small movement about their resting position. Molecular dynamics is also based on the empirical force field and basic assumptions described above for molecular mechanics, but allows the study of the movement of atoms over time, thereby providing a more realistic understanding of molecular structure and reactivity. In general, the molecular mechanics force-field method is largely applied for basic understanding of energetic and conformational aspects of molecules using strategies such as structure refinement, ligand docking simulations, molecular dynamics simulation, and/or Monte Carlo statistical simulations (similar to molecular dynamics, but based on statistics), all of which rely on the force field that was discussed. An outcome of the relatively long history of all such studies is the ability to construct quantitative structure–activity relationships for general classes of molecules, providing information to support our intuition about the behavior of molecules in terms of their general structure and property features. See CHEMICAL DYNAMICS; CHEMOMETRICS; MONTE CARLO METHOD.

Kim K. Baldridge; Celine Amoreira

Bibliography. U. Burkert and N. L. Allinger, *Molecular Mechanics*, ACS Monogr. 177, 1982; D. Frenkel and B. Smit, *Understanding Molecular Simulation: From Algorithms to Applications*, 2002; T. L. Hill, Steric effects, *J. Chem. Phys.*, 14:465, 1946; A. R. Leach, *Molecular Modeling: Principles and Applications*, 1996; A. Hinchliffe, *Molecular Modeling for Beginners*, 2003; K. B. Lipkowitz and D. B. Boyd (eds.), *Reviews in Computational Chemistry*, vol. 2, 1991; T. Schlick, *Molecular Modeling and Simulation: An Interdisciplinary Guide*, 2002.

Molecular orbital theory

A quantum-mechanical model concerned with the description of the discrete energy levels associated with electrons in molecules. One useful way to generate such levels is to assume that the molecular orbital wave function (ψ_j) may be written as a simple weighted sum of the constituent atomic orbitals (χ_i) [Eq. (1)]; this is called the linear combination of

$$\psi_j = \sum c_{ij} \chi_i \quad (1)$$

atomic orbitals approximation. The c_{ij} coefficients may be determined numerically by substitution of Eq. (1) into the Schrödinger equation and application of the variational theorem. The theorem states that an approximate wave function will always be an upper bound to the true energy; thus minimization of the energy of the system given by the wave function of Eq. (1) will provide the best values of c_{ij} . Once the wave function is known, its associ-

ated energy may be calculated. The energies of the occupied orbitals in molecules may be probed by using photoelectron spectroscopy, which gives a good check on the accuracy of the theory. There are some simple concepts that contribute to a qualitative understanding of these molecular orbital energy levels and hence an insight into chemical bonding in molecules. They may be illustrated with reference to the hydrogen molecule. See CHEMICAL BONDING; ELECTRON SPECTROSCOPY; QUANTUM MECHANICS; SCHRÖDINGER'S WAVE EQUATION.

Hydrogen molecule. First, the basis orbitals (χ_i) used in the expansion of Eq. (1) can usefully be restricted to include the valence orbitals only. For molecular hydrogen (H_2) the 1s orbitals on the two hydrogen atoms are then the only two orbitals to be included. Second, since hydrogen atoms are chemically identical, any observable characteristic whose value might be computed with Eq. (1) must be the same for both atoms. This leads to the requirement that $c_{1j}^2 = c_{2j}^2$, where the labels 1,2 refer to hydrogen atoms 1 and 2. As a consequence $c_{1j} = \pm c_{2j}$.

When the signs of the two coefficients are the same, the two hydrogen orbitals are mixed in phase; when they are different, the two hydrogen orbitals are mixed out of phase. **Figure 1** shows how the wave function ψ_j and its square (giving the electron probability distribution) may be constructed for those two cases. When the atomic orbitals are mixed in phase, then electron density is built up between the two hydrogen nuclei and the potential energy of the nuclei and electrons is lowered. In fact, a reduction of kinetic energy also occurs. An electron lying in the molecular orbital corresponding to Fig. 1a is then of lower energy than an electron associated with an isolated hydrogen 1s orbital. It is called a

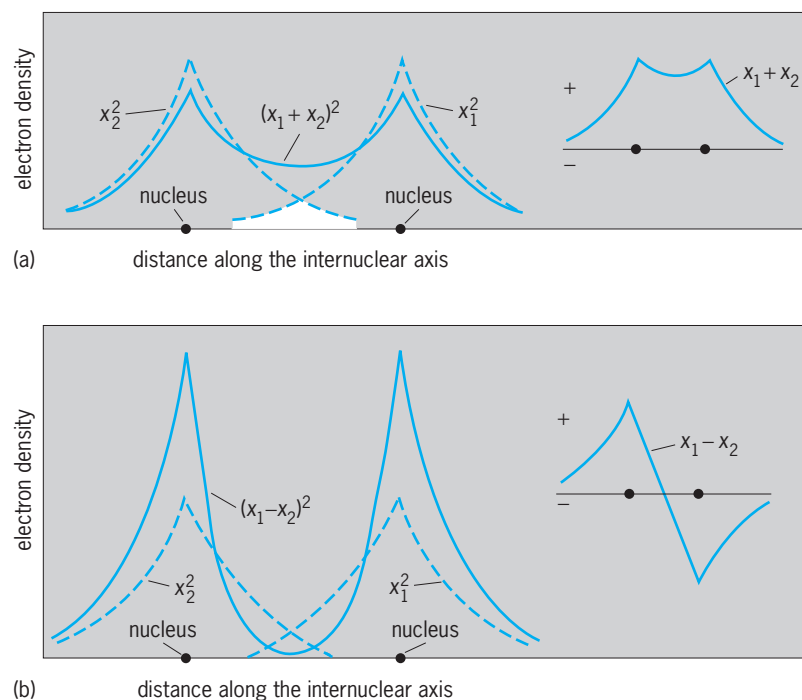


Fig. 1. In-phase and out-of-phase addition of atomic hydrogen wave functions (x_1, x_2) to give (a) bonding and (b) antibonding molecular orbitals.

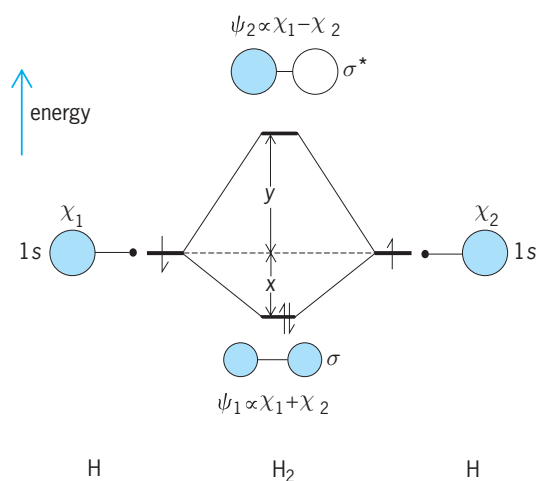


Fig. 2. Molecular orbital diagram of H_2 .

bonding orbital. The increase in electron density between the two nuclei is the electronic “glue” holding the nuclei together. When the atomic orbitals are mixed out of phase, the opposite behavior occurs (Fig. 1*b*). Electron density is removed from the region between the two nuclei, resulting in an increase of both potential and kinetic energy of the electrons. An electron lying in such a molecular orbital would experience an energetic destabilization relative to an electron associated with an isolated hydrogen $1s$ orbital.

Such a molecular orbital is called an antibonding orbital. **Figure 2** shows how this information may be collected together on a molecular orbital diagram. The shading convention of the orbitals has been adopted to indicate the in- and out-of-phase mixing of the basis orbitals. Just as the energy levels of atoms are filled in an Aufbau process, so the orbitals of the molecule may be analogously filled up with electrons, each level accommodating two electrons of opposite spin. In H_2 there are two electrons to be accounted for. They lie in the bonding orbital, and the stabilization energy relative to two isolated hydrogen atoms (the bond energy) is $2x$. Antibonding orbitals are invariably destabilized more than their bonding counterparts are stabilized. This is shown in Fig. 2 by making $y > x$. With four electrons to be accommodated in this collection of orbitals (this would correspond to the hypothetical case of the He_2 molecule), one electron pair resides in the bonding orbital and one pair in the antibonding orbital. Since $y > x$, this molecule is less stable relative to two isolated helium atoms, and as a result the molecule does not exist as a stable entity. He_2^+ , however, with only three electrons is known.

The size of the interaction energy associated with two atomic orbitals (x in Fig. 2) is controlled by the extent of their spatial overlap. This overlap integral is proportional to the shaded region of Fig. 1*a*, and its magnitude clearly depends upon the internuclear separation. The equilibrium bond length in the hydrogen molecule (and indeed in all molecules) is then a balance between the attractive forces associated

with bonding orbital formation and the electrostatic repulsion between the nuclei. Such a molecular parameter is amenable to numerical calculation.

The description of the bonding in the H_2 molecule using this model is one where two electrons occupy a bonding orbital and give rise to a simple two-center-two-electron bond traditionally written as $H-H$. Since the electron density associated with the bonding orbital is cylindrically and symmetrically located about the $H-H$ axis, this bond is called a σ bond. In Fig. 2 the bonding orbital is labeled with a σ and the corresponding antibonding orbital with a σ^* .

First-row diatomics. Ideas similar to those above are readily extended to diatomic molecules from the first row of the periodic table, such as N_2 and O_2 , where the valence orbitals to be considered are the one $2s$ and the three $2p$ orbitals of the atoms. The $2s$ orbitals lie deeper in energy than the triply degenerate $2p$ orbitals. As shown in **Fig. 3**, the atomic $2s$ orbitals form bonding and antibonding orbitals (labeled σ and σ^*) just as in the case of elemental hydrogen described above, but the behavior of the $2p$ orbitals is a little different. Here there are three possible types of interaction between the p orbitals on one center and those on the other. The end-on overlap of two p orbitals gives rise to a σ interaction (**Fig. 4*a***), and the sideways overlap of two p orbitals gives rise to a π interaction (**Fig. 4*b***). The interaction in Fig. 4*c* can be ignored since the overlap between the two orbitals in this orientation can be seen to be identically zero. The result shown in Fig. 3 is a σ -bonding orbital and a σ -antibonding orbital (labeled

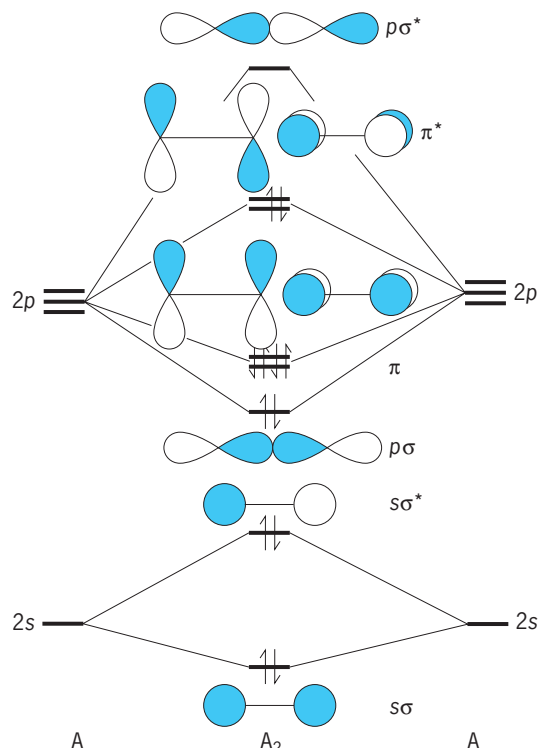


Fig. 3. Molecular orbital diagram for a first-row diatomic molecule (A_2). The electronic configuration corresponds to that of O_2 .

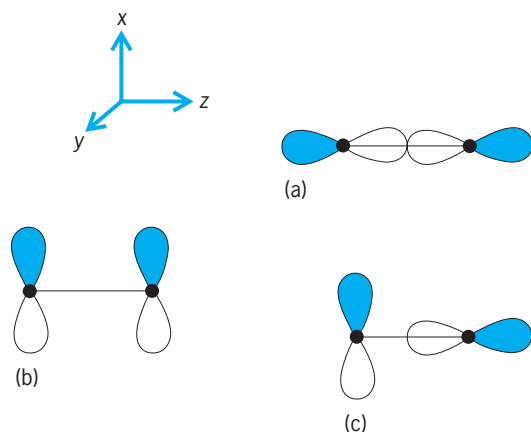


Fig. 4. Possible orientations of the p orbitals on adjacent atomic centers. (a) End-on overlap. (b) Sideways overlap. (c) Zero overlap.

$p\sigma$ and $p\sigma^*$), and a pair of π -bonding and a pair of π -antibonding orbitals (labeled π and π^*). The larger interaction energy associated with $p\sigma$ compared to $p\pi$ is due to the larger σ overlap compared to π overlap in Fig. 4.

Filling these orbitals with electrons allows comment on the stability of the resulting diatomics. The molecule Li_2 ($s\sigma$)² is known and, like H_2 , may be written as $\text{Li}-\text{Li}$ to emphasize the single, two-center, two-electron bond between the nuclei. The molecule Be_2 which would have the configuration $(s\sigma)^2(s\sigma^*)^2$ is unknown since, just as in He_2 , $s\sigma^*$ is destabilized more than $s\sigma$ is stabilized relative to an atomic $2s$ level. If the molecular orbital bond order is written as expression (2), then the bond order in Li_2 is one but the bond order in Be_2 is zero.

$$\text{Molecular bond order} = \left(\frac{\text{number of bonding electron pairs}}{\text{electron pairs}} \right) - \left(\frac{\text{number of antibonding electron pairs}}{\text{electron pairs}} \right) \quad (2)$$

By filling up the molecular orbital levels derived from the $2p$ orbitals, the bond order associated with the other diatomics may be generated: $\text{B}_2(1)$, $\text{C}_2(2)$, $\text{N}_2(3)$, $\text{O}_2(2)$, $\text{F}_2(1)$, and $\text{Ne}_2(0)$. All of these species are known except Ne_2 , which is predicted, like He_2 and Be_2 , to have a zero bond order and therefore not to exist as a stable molecule. The molecular orbital bond orders for the three best-known diatomics are consistent with their traditional formulation as $\text{N}\equiv\text{N}$, $\text{O}=\text{O}$, and $\text{F}-\text{F}$. N_2 , for example, would be described as having one σ and two π bonds. Figure 3 shows the electron occupancy for O_2 . With the configuration $(s\sigma)^2(s\sigma^*)^2(p\sigma)^2-(\pi)^4(\pi^*)^2$, there are four bonding pairs of electrons and two antibonding pairs giving rise to a net bond order of two. The pair of π^* orbitals is only doubly occupied, whereas there is space for four electrons. Hund's rules (which for the electronic ground state maximize the number of electrons with parallel spins) identify the lowest-energy arrangement as the one where each of the degenerate π^* components is singly occupied, the spins of the two electrons being parallel. Unpaired electrons

give rise to paramagnetic behavior, and gaseous oxygen is indeed paramagnetic.

Multicenter bonding. Not all molecules can be described as being held together by simple two-center, two-electron bonds. Multicenter delocalized bonding, for example, is the best way of describing the π bonding which occurs in organic acyclic and cyclic polyenes. Figure 5 shows the molecular orbitals that result from interaction of the $p\pi$ orbitals that lie perpendicular to the molecular plane in benzene and cyclobutadiene. For benzene there are three bonding orbitals, one lying deeper than the other degenerate pair and three analogous antibonding orbitals. In cyclobutadiene there is a bonding and antibonding orbital in addition to a degenerate pair of nonbonding orbitals. A nonbonding molecular orbital has the

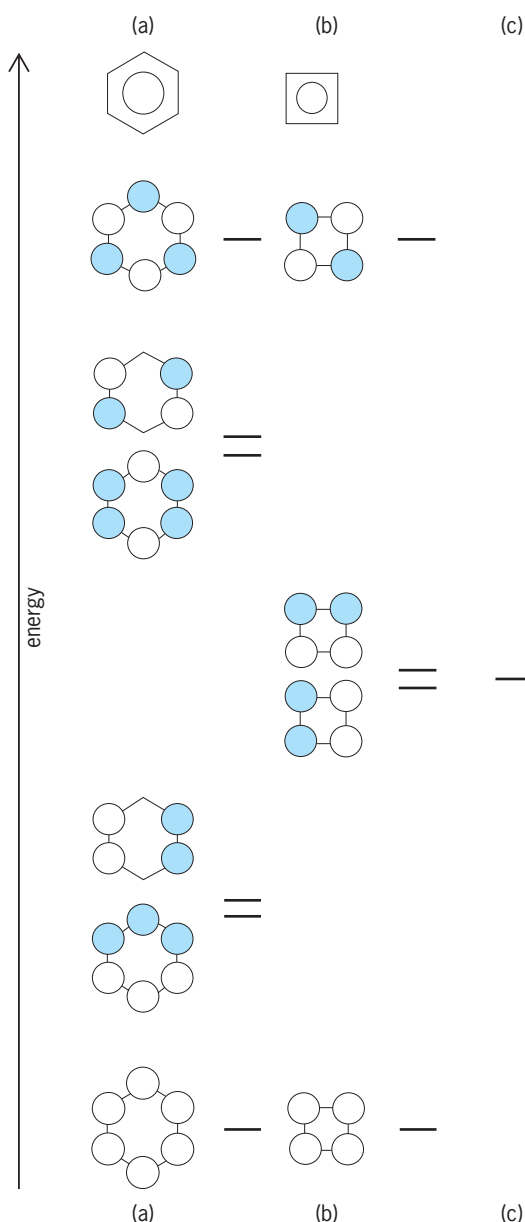


Fig. 5. Molecular π orbitals for (a) benzene and (b) cyclobutadiene viewed from above so that the relative phases only of the upper half of the $p\pi$ orbitals are shown. (c) The energy of an isolated $p\pi$ orbital.

same energy as the isolated atomic orbitals from which it is constructed. In these examples of neutral hydrocarbons the number of molecular orbitals is equal to the number of atomic orbitals used in synthesizing the diagram (Fig. 5). Each carbon atom contributes one electron to this π -orbital manifold. Thus in benzene the lower three orbitals are doubly occupied. As far as the π manifold of benzene is concerned, there are then three bonding pairs but six close contacts between carbon atoms. As a result, the π -bond order is one-half, and the π bonding in this molecule is referred to as being delocalized over the carbon framework. Similar ideas employing delocalized bonding can be applied to three-dimensional molecules such as the boranes, carboranes, and transition-metal cluster compounds. See BORANE; CARBORANE; METAL CLUSTER COMPOUND.

There is a simple rule that has its basis in the π -orbital structure of these hydrocarbons. If a molecule possesses $4n + 2\pi$ electrons ($n = 0, 1, 2, \dots$), then it will be stable as a planar aromatic molecule. If the π -electron count is otherwise, then the planar symmetric molecule will be antiaromatic and unstable. This is called Hückel's rule. Benzene, with six electrons, satisfies the rule ($n = 1$), but cyclobutadiene does not. The latter molecule distorts away from

the square geometry, probably to a rhombus, and is kinetically unstable. C_8H_8 is another species which does not satisfy the rule, and this molecule is nonplanar. $C_8H_8^{2-}$ and $C_4H_4^{2-}$ do satisfy Hückel's rule and are found as planar, symmetric units bound to transition metals. The species S_4^{2+plus} , isoelectronic with $C_4H_4^{2-}$, is also a stable, square molecule. Hückel's rule has its foundation in the requirement that all bonding and nonbonding orbitals of the π network be full of electrons for structural stability. See RESONANCE (MOLECULAR STRUCTURE).

Transition-metal complexes. Similar orbital ideas applied to transition-metal complexes have led to an understanding concerning their structures and reactivity. Molecular orbital models have virtually replaced the older concepts of crystal field and valence bond theory in the area. **Figure 6** shows a molecular orbital diagram for an octahedral ML_6 complex, where L is a ligand that possesses an accessible orbital with which to bond in a σ fashion to the metal. (Examples are hydride, water, and halide.) This orbital diagram is different in several ways from those previously described. The ionization potential for an electron located in a ligand orbital is larger than for a metal-located electron. This is shown by drawing the metal atomic levels higher in energy than the ligand levels. The six ligand orbitals split into three sets when the ligands are coordinated to the metal. (The reasons for this are described in the next section.) The five atomic metal d orbitals split into two sets in ML_6 , one doubly degenerate and one triply degenerate. The set labeled t_{2g} is nonbonding between metal and ligands, whereas the set labeled e_g is metal-ligand antibonding. The energy separation between the two is conventionally labeled δ . Electronic transitions between these orbital sets occur in the visible part of the spectrum for many first-row transition-metal complexes (scandium through copper) and are thus the source of their color. The number of electrons that occupy these orbitals is simply given by counting the number of $d + s$ electrons of the corresponding gaseous ion. Thus Cr(II) is a d^4 system while Zn(II) is a d^{10} system. In the latter, the e_g and t_{2g} orbitals are completely filled. In organometallic compounds where δ is large, a rule similar to Hückel's rule is found to apply. If all bonding and nonbonding orbitals are occupied, then kinetic and structural stability is conferred upon the system. This occurs when a total of 18 electrons (or nine electron pairs) are associated with the metal. In Fig. 6 this would occur for six pairs of electrons in ligand-located orbitals plus three pairs of electrons in the nonbonding t_{2g} set. This 18-electron rule has proven to be very useful in understanding organometallic chemistry. See ORGANOMETALLIC COMPOUND.

Symmetry considerations. Symmetry is very useful in understanding many aspects of molecular orbital theory. In qualitative terms, the more symmetric a molecule, the larger the number of elements of symmetry it possesses. **Figure 7** shows the symmetry elements—the mirror planes (σ_v), rotation axes (C_n , for a rotation of $360^\circ/n$), and center of symmetry (i)—for the A_2 diatomic molecule shown in Fig. 3.

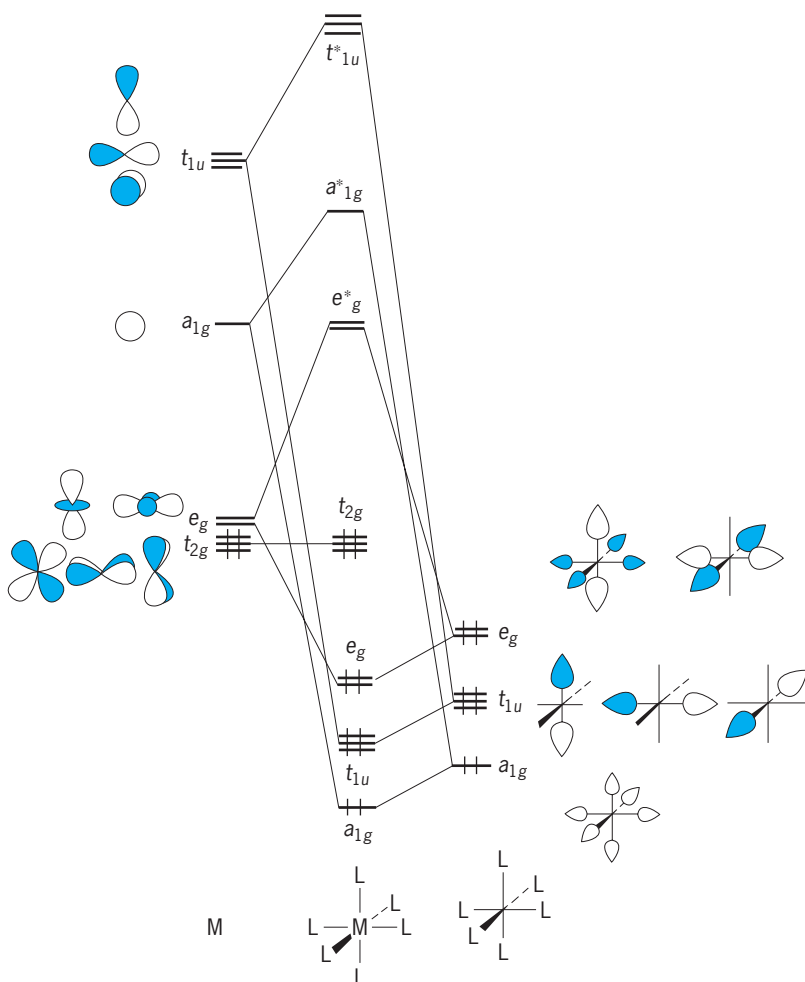


Fig. 6. Molecular orbital diagram for an ML_6 transition-metal complex. The labels a_{1g} , e_g , t_{1u} , and t_{2g} describe the symmetry properties of the orbital levels.

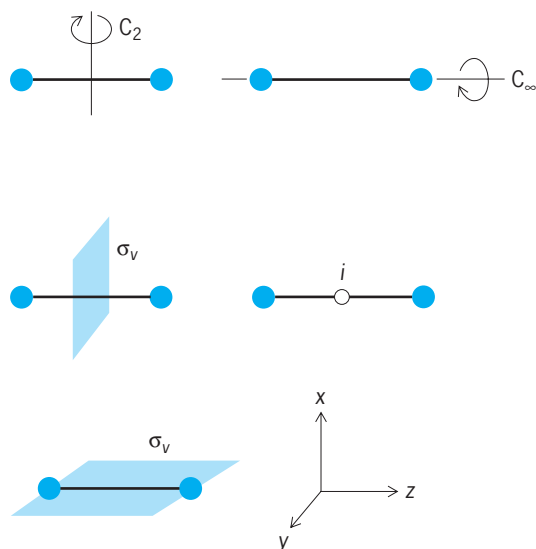


Fig. 7. Symmetry elements for the H_2 or A_2 diatomic molecule.

Although only one mirror plane containing the internuclear axis is shown, there are an infinite number of them, just as there are an infinite number of C_2 axes perpendicular to the internuclear axis.

The mathematical theory that allows manipulation of such observations is called group theory. It can be seen that the directions x and y in Fig. 7 are symmetry-equivalent in such a molecule, but that the z direction is unrelated to either x or y . A more formal way of putting this is to note that, although there are symmetry operations possessed by the molecule which will interconvert x and y , there is no such operation which will interconvert z with either x or y . A direct result of this observation is that the p_x and p_y orbitals will always be found in degenerate orbital situations, but that p_z will be energetically separate. In the linear molecule the symmetry label π is used to refer to such a double degeneracy. The nondegenerate p_z orbital is described by the symmetry label σ . The symmetry description of the three valence p orbitals on an A atom in the A_2 diatomic is then $\sigma + \pi$. As the number of symmetry elements associated with a molecule increases, the size of possible degeneracies also increases. In octahedral and tetrahedral molecules, for example, double and triple degeneracies are possible, and in icosahedral molecules fivefold degeneracies may occur. In the octahedral molecule shown in Fig. 6, the x , y , and z directions are symmetry-equivalent and are described by the symmetry label t_{1u} representing a triple degeneracy. (Greek letters are used for the symmetry labels of linear molecules, for example, σ or π and Roman letters, for example, a_{2g} or t_{2u} for nonlinear molecules.)

Here a useful subscript is used to describe a function's behavior with respect to inversion. The label g applies to functions that are symmetric with respect to such an operation (s and d orbitals on the central atom, for example) and the label u for antisymmetric functions (p and f orbitals on the central atom, for example). The five d orbitals become of $e_g + t_{2g}$ sym-

metry in the octahedral molecule, just as the three p orbitals become of $\sigma + \pi$ symmetry in the diatomic molecule shown in Fig. 3.

There is one very powerful result which is extremely useful in constructing orbital diagrams. Orbitals of different symmetry have zero overlap and therefore do not interact with each other. For example, in Fig. 4b the two p_x orbitals have a nonzero interaction since they are both of π symmetry, but the overlap integral for the combination of Fig. 4c is zero since one orbital is of π and the other of σ symmetry. The molecular orbital diagrams shown in Figs. 2, 3, and 6 may then be generated simply by matching orbitals of the same symmetry at the left- and right-hand sides of the picture and constructing bonding and antibonding orbital combinations from each pair. In Fig. 6 the central metal atom orbitals of t_{2g} symmetry find no symmetry match with the orbitals of the ligands and so remain nonbonding. Symmetry arguments such as these are behind many of the basic orbital concepts in molecules. See GROUP THEORY; SELECTION RULES (PHYSICS); SYMMETRY LAWS (PHYSICS).

Orbital symmetry. The Woodward-Hoffmann rules, which collect together within a theoretical framework many important organic reactions, are based on the symmetry control of orbital interaction. Figure 8 shows orbital correlation diagrams linking the orbitals of a 1,4 substituted butadiene and the product of ring closure, the cyclobutene. Two different modes of closure may be envisaged, the conrotatory and disrotatory motion of the groups attached at the 1 and 4 positions. In the former a twofold rotation axis is preserved, and in the latter a mirror plane is preserved during the reaction. In Fig. 8a all the levels may then be classified as being either symmetric (S) or antisymmetric (A) with respect to the twofold rotation axis (C_2). Correspondingly in Fig. 8b the levels may be classified analogously according to their behavior with respect to the mirror plane operation. Orbitals of the same symmetry will interact with each other and will not cross along the reaction coordinate, but orbitals of different symmetry may cross.

The orbital correlation diagrams of Fig. 8 may thus be generated by connecting pairwise the orbitals at the left- and right-hand sides. Notice how in Fig. 8b the highest occupied molecular orbital (HOMO) of the reactant correlates with the lowest unoccupied molecular orbital (LUMO) of the product. For this reaction a high activation energy is expected as a result. Contrast the process in Fig. 8a which will experience no such energetic penalty. The thermal reaction of Fig. 8b is then orbitally forbidden, whereas that of Fig. 8a is orbitally allowed. On photochemical excitation of an electron from HOMO to LUMO of the reactant, the opposite is now true. The disrotatory motion is photochemically allowed, but the conrotatory motion is a photochemically forbidden process. This generalization has had a profound effect on organic chemistry. See ORGANIC REACTION MECHANISM; STEREOCHEMISTRY; WOODWARD-HOFFMANN RULE.

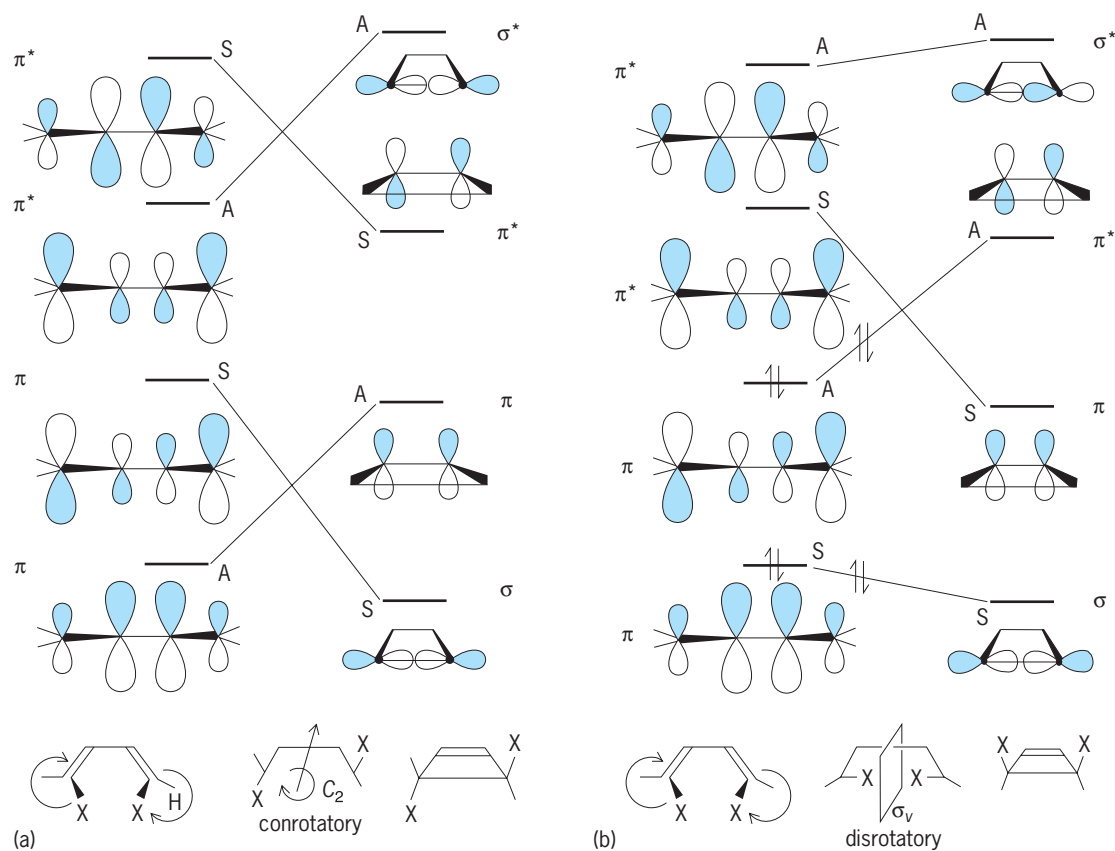


Fig. 8. Orbital correlation diagram for (a) the conrotatory and (b) the disrotatory ring closing of 1,4-butadienes.

Frontier orbitals. The use of valence orbitals alone is sufficient to build up the molecular orbitals and a qualitative picture of chemical bonding. This occurs because the core orbitals (for example, the 1s orbital on the oxygen atom) are very tightly bound and have a negligible overlap with the orbitals of a neighboring atom. When two molecules, or fragments, are brought together, it is often found that their mutual interaction is qualitatively well described by considering, in turn, a rather small number of these valence-orbital-based molecular orbitals. These so-called frontier orbitals are the orbitals of the two molecules which are spatially arranged so that the overlap integral between them is significant. Other interactions are either zero by symmetry, or small because their geometrical arrangement precludes good overlap.

Many aspects of chemical reactivity and structure may be understood by considering the details of their interaction. It often turns out that the energetically most important interactions occur between the HOMO of one molecule and the LUMO of another, giving rise to a stabilizing interaction. **Figure 9** shows how consideration of the frontier orbitals of these molecules allows understanding of the observation that maleic anhydride reacts easily with butadiene but with difficulty with ethylene. The HOMO-LUMO overlap in the former case (Fig. 9a) is nonzero and gives rise to a stabilizing interaction. In the latter

such overlap is exactly zero since the positive overlap at the left-hand side is canceled by the negative overlap at the right-hand side and the interaction of the reactants is zero (Fig. 9a). In the area of organic chemistry, frontier orbital theory allows explanation of many aspects of the Woodward-Hoffmann rules. In the areas of molecular geometry such ideas are very useful in understanding the relative orientation of different parts of a molecule by maximizing such interactions.

Status. In the discussion above, the emphasis has been on the qualitative aspects of molecular orbital theory. Indeed, many of the advances in chemistry that have been led by theoretical considerations have employed arguments such as these, based on symmetry, electronegativity, and overlap. However, the numerical aspects of molecular orbital theory should not be underestimated. Quantitative calculations designed to estimate bond energies and map out the details of reaction pathways are becoming increasingly possible as a result of improvements in the size and speed of computers. The bond dissociation energies of small molecules that contain a few atoms from the second row of the periodic table can be calculated to chemical accuracy (± 1 kcal/mole). However, it is not yet possible to predict the outcome of most experiments by calculations of this type. Even as that time approaches, it will still be necessary to have a qualitative understanding of the numbers

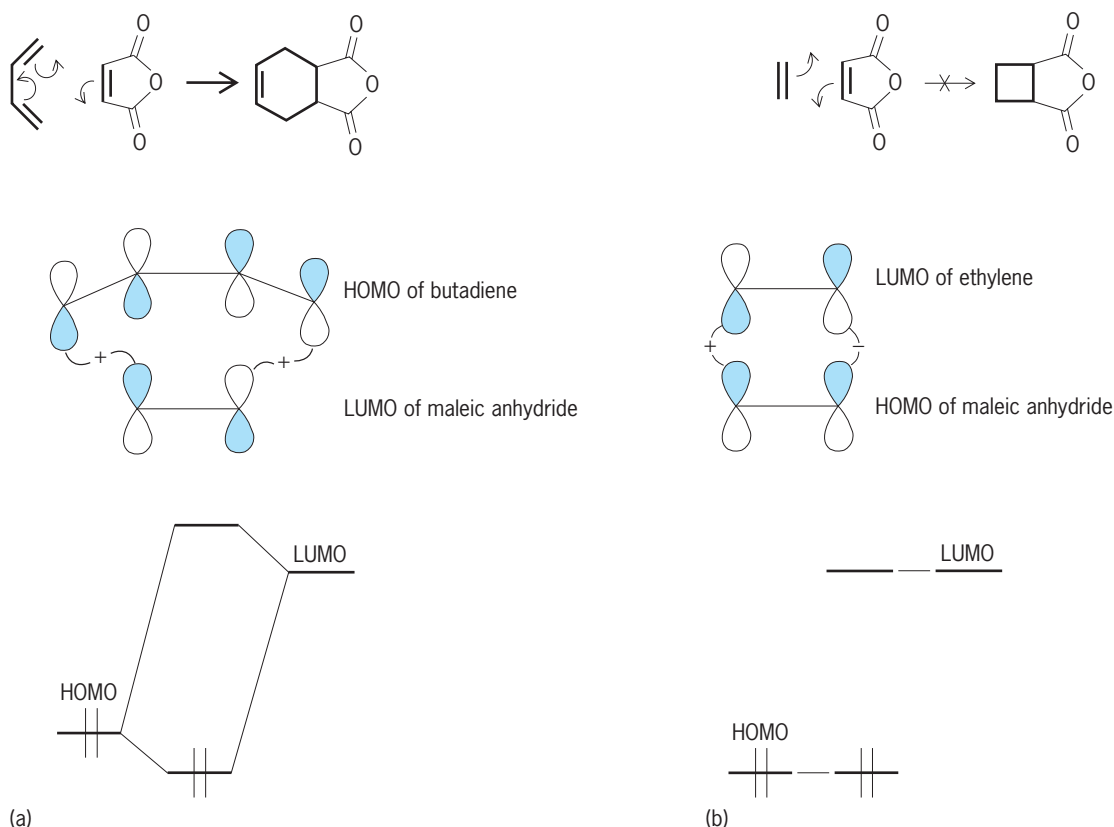


Fig. 9. HOMO-LUMO interactions between maleic anhydro and (a) butadiene and (b) ethylene.

generated by the computer. See BAND THEORY OF SOLIDS; COMPUTATIONAL CHEMISTRY; MOLECULAR STRUCTURE AND SPECTRA; QUANTUM CHEMISTRY.

Jeremy K. Burdett

Bibliography. T. A. Albright, J. K. Burdett, and M. H. Whangbo, *Orbital Interactions in Chemistry*, 1985; O. K. Anderson, V. Kumar, and A. Mookerjee, *Methods of Electronic Structure Calculations*, 1994; F. A. Cotton, *Chemical Applications of Group Theory*, 3d ed., 1990; I. Fleming, *Frontier Orbitals and Organic Chemical Reactions*, 1976; H. Rauk, *Orbital Interaction Theory of Organic Chemistry*, 2d ed., 2000; A. Streitwieser, *Molecular Orbital Theory for Organic Chemists*, 1961; R. B. Woodward and R. Hoffmann, *The Conservation of Orbital Symmetry*, 1970.

Molecular pathology

A discipline that deals with the origins and mechanisms of diseases at their most fundamental level, that of macromolecules such as deoxyribonucleic acid (DNA) and protein, in order to provide precise diagnoses and discover possible avenues for treatment. It is interdisciplinary, including infectious disease, oncology, inherited genetic disease, and legal issues such as parentage determination or forensic identity testing. While a variety of biophysical and biochemical techniques can be applied to study the

molecular basis of disease, antibodies and nucleic acid probes are two of the principal approaches. See ANTIBODY; NUCLEIC ACID; ONCOLOGY.

Antibodies. Monoclonal antibodies recognize and noncovalently bind specific molecular shapes called epitopes on macromolecules. When these antibodies are either tagged to permit their detection or immobilized on a chromatographic column to purify their specific target molecule, they serve as powerful tools for analyzing pathologic processes. A monoclonal antibody conjugated to an enzyme generating a colored reaction product is the basis of the enzyme-linked immunosorbent assay (ELISA), which is widely applied in many diagnostic tests. See MONOCLONAL ANTIBODIES.

Autoimmune diseases result when the body produces antibodies against its own molecular components. Autoantibodies from the sera of affected individuals are often used as highly specific reagents for understanding the nature of these diseases and the role that the affected molecules or organelles normally perform in the cell. For example, autoantibodies directed against the skeletal-muscle acetylcholine receptor result in myasthenia gravis; individuals with scleroderma, a collagen vascular disease associated with excess collagen deposition in the skin and other organs, may have autoantibodies directed against components of the centriole. These autoantibodies can be detected in tissues by using fluorescently tagged antibodies directed against human

immunoglobulins. See AUTOIMMUNITY; FLUORESCENCE MICROSCOPE; MYASTHENIA GRAVIS.

In situ hybridization. In an analogous manner to immunohistochemistry, traditional histopathology (that is, the microscopic study of diseased tissues) can be enhanced by using in situ hybridization. With this technique, an infectious agent such as a virus or a specific messenger ribonucleic acid (mRNA) can be localized within a specific cell or tissue. The final appearance of the tissue section resembles an immunohistochemical result, although these distinct molecular probes have different targets; antibodies are generally used to detect proteins, while nucleic acid probes detect DNA or RNA.

The ability to localize specific molecular targets with DNA probes can be applied in cytogenetics by using fluorescent in situ hybridization (FISH). This technique can precisely map genes to a specific region of a chromosome in an appropriately prepared karyotype, or can enumerate chromosomes, or can detect chromosomal deletions, translocations, or gene amplifications in the interphase nucleus of cancer cells.

Nucleic acid probes. Diseases often result from germline or somatic mutations in the individual's DNA, such as are seen in sickle cell disease or cancer, respectively. These abnormalities can be detected by using two basic techniques of molecular genetics: the Southern blot and the polymerase chain reaction (PCR).

Southern blot. For the Southern blot, high-molecular-weight DNA isolated from a specimen (most commonly peripheral-blood white cells) is digested by using an appropriate restriction endonuclease. The resulting fragments are then separated by molecular weight by means of gel electrophoresis. This DNA is transferred in its single-stranded form from the gel to a nylon membrane. This membrane is incubated with a solution containing a specific, labeled probe, also in single-stranded form. Probe-target hybrids formed by annealing of their complementary sequences can be detected by autoradiography or a colorimetric reaction.

The Southern blot technique can detect DNA polymorphisms, mutations (including deletions, expansions, amplifications, or translocations), or the presence of viral, bacterial, or specific sex chromosomes (such as the Y chromosome, indicating male sex). By using a Southern blot, fragile X syndrome, the most common cause of inherited mental retardation, can be diagnosed by the presence of a larger band on the autoradiograph.

Polymerase chain reaction. The polymerase chain reaction has revolutionized molecular genetics and provides a powerful tool for diagnosis. Short oligonucleotide primers flank the specific gene region or RNA sequence to be amplified and are combined with the target specimen and free nucleotides, which are synthesized into new DNA by *Taq* DNA polymerase. An automated thermal cycler repeatedly alters the temperature to denature the target DNA, to allow the primers to reanneal to the target, and then

to synthesize the product. This amplified polymerase chain reaction product is typically detectable as a band in a gel. Simultaneous amplification of several genes or gene regions in the same reaction is known as multiplex polymerase chain reaction, which permits more rapid testing.

Cystic fibrosis provides an example of the clinical utility of polymerase chain reaction diagnosis. It is an autosomal recessive disease resulting from a defective gene located on the long arm of chromosome 7. When this gene is mutated, the cell's ability to adjust concentrations of chloride ion is altered, resulting in sweat with elevated chloride levels (the basis for the traditional sweat chloride determination), and in thickened, relatively dehydrated secretions in the body. These secretions obstruct various glands such as the pancreas, ultimately resulting in atrophy and fibrosis, while the lungs become susceptible to repeated infections and subsequent tissue injury. The most common mutation causing cystic fibrosis is a deletion at a position that normally encodes a phenylalanine. After polymerase chain reaction amplification, this deletion is detected as a smaller band when examined in an electrophoretic gel.

Advantages. Understanding diseases at the genetic level has several advantages. The central dogma of molecular biology—that biological information flows from DNA to RNA to protein—permits an impressive understanding of a disease based on the DNA sequence of the involved gene. Even when the gene's protein product is not expressed, definitive diagnoses can be made by using DNA-based techniques. For example, fetal DNA from an amniocentesis permits the prenatal diagnosis of sickle cell disease even though the defective beta chain of hemoglobin is not sufficiently expressed within the fetus to be detected with standard hemoglobin electrophoresis. Diseases that are similar clinically (phenotypically) can, in fact, be due to different mutations (genotypes) within a single gene or due to mutations in different genes, often related in an enzyme complex or as a portion of a group of structural proteins. See MUTATION.

Another advantage of molecular diagnosis is the ability to detect phenotypically normal carriers of genetic diseases in order to provide information for appropriate genetic counseling and prenatal diagnosis. See HUMAN GENETICS; PATHOLOGY. Mark A. Lovell

Bibliography. N. Lemoine, J. Neoptolemos, and T. Cooke, *Cancer: A Molecular Approach*, 1994; E. Rubin and J. L. Farber, *Pathology*, 3d ed., 1999.

Molecular physics

The study of the physical properties of molecules. Molecules possess a far richer variety of physical and chemical properties than do isolated atoms. This is attributable primarily to the greater complexity of molecular structure, as compared to that of the constituent atoms. Molecules also possess

additional energy modes because they can vibrate; that is, the constituent nuclei oscillate about their equilibrium positions and rotate when unhindered. These modes give rise to additional spectroscopic properties, as compared to those of an atom; molecular spectroscopy in the optical, infrared, and microwave regions is one of the physical chemist's most powerful means of identifying and understanding molecular structure. Molecular spectroscopy has also given rise to the rapidly growing field of molecular astronomy.

Molecular physics is primarily concerned with the study of properties of isolated molecules, as contrasted to the more general study of molecular reactions, which is the domain of physical chemistry. Such properties, in addition to the broad field of spectroscopy, include electron affinities (for the formation of molecular negative ions); polarizabilities (the "distortability" of the molecule along its various symmetry axes by external electric fields); magnetic and electric multipole moments, attributable to the distributions of the electric charge; currents and spins of the molecule; and the (nonreactive) interactions of molecules with other molecules, atoms, and ions.

One of the most important areas of molecular physics is the application of computational techniques to the calculation of molecular wave functions from basic principles and, consequently, of molecular charge distributions and of all the physical properties that are noted above.

Intermolecular forces are responsible for such varied phenomena as condensation, surface tension, gaseous diffusion, and, most importantly, for the formation of crystalline and noncrystalline solids. Thus, the field is basic to achieving an understanding of the physical world. It would be difficult to list all the articles in the present volumes which could be classed under the domain of molecular physics. For a representative sampling see COSMO-CHEMISTRY; INFRARED SPECTROSCOPY; INTERMOLECULAR FORCES; MICROWAVE SPECTROSCOPY; MOLECULAR BEAMS; MOLECULAR STRUCTURE AND SPECTRA; SPECTROSCOPY.

Benjamin Bederson

Molecular recognition

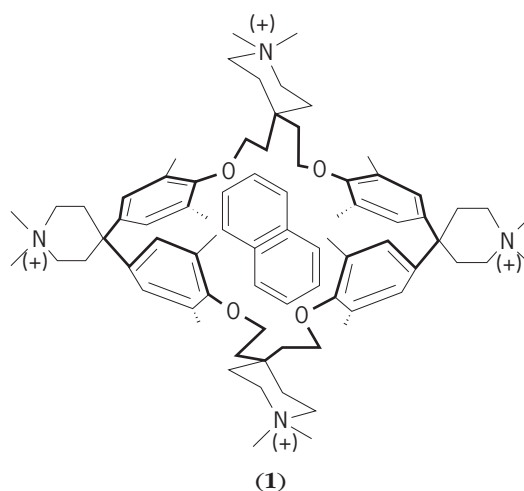
The ability of biological and chemical systems to distinguish between molecules and regulate behavior accordingly. How molecules fit together is fundamental in disciplines such as biochemistry, medicinal chemistry, materials science, and separation science. A good deal of effort has been expended in trying to evaluate the underlying intermolecular forces. The weak forces that act over short distances (hydrogen bonds, van der Waals interactions, and aryl stacking) provide most of the selectivity observed in biological chemistry and permit molecular recognition. The recognition event initiates behavior such as replication in nucleic acids, immune response in antibodies, signal transduction in receptors, and regulation

in enzymes. Most studies of recognition in organic chemistry have been inspired by these biological phenomena. It has been the task of bioorganic chemistry to develop systems capable of such complex behavior with molecules that are comprehensible and manageable in size, that is, with model systems. See CHEMORECEPTION; ENZYME; HYDROGEN BOND; INTERMOLECULAR FORCES; NUCLEIC ACID; SYNAPTIC TRANSMISSION.

Macrocycles. The advantage of cyclic structures lies in their ability to restrict conformation or flexibility. A rigid matrix of binding sites, that is, preorganized sites, is usually associated with high selectivity in binding. A flexible matrix tends to accept several binding partners. Although sacrificing selectivity, this has the advantage of transmitting conformational information and is relevant to biological signaling events.

Crown ethers. Early work in the area focused on macrocyclic (crown) ethers. Such structures can bind and transport ions and imitate biological processes involving macrolides. Large ring structures that are lined with oxygen present an inner surface which is complementary to the spherical outer surface of positively charged ions. See MACROCYCLIC COMPOUND.

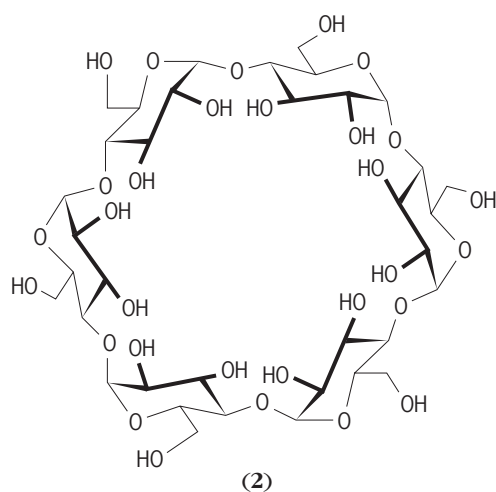
Cyclophane-type structures. Cyclophane-type structures offer considerable rigidity because of the aromatic nuclei. The further incorporation of charged groups on the periphery imparts sufficient water solubility for complexing in aqueous systems. Binding forces between host and guest are largely hydrophobic, with some contribution of aryl-stacking interactions involving the inner surfaces of the cyclophane and the outer π surfaces of the guest species. A typical system is a cyclophanenaphthalene complex (1), in which a naphthalene



guest is bound by a water-soluble cyclophane derivative. See COORDINATION COMPLEXES.

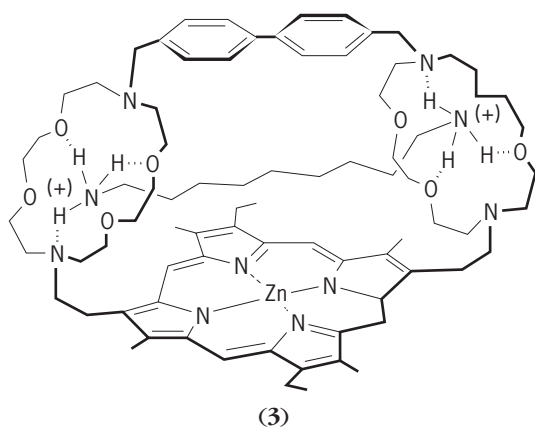
Cyclodextrins. Naturally occurring cycloamylose derivatives (cyclodextrins), such as cyclohexa-

amylose (α -cyclodextrin; **2**), likewise present a



hydrophobic inner surface and a hydroxylic outer surface for complexation of neutral structures in water. Aromatics, adamantanes, and even ferrocene derivatives fit snugly within. Such complexes can place electrophilic sites of the guest near the hydroxyl groups of the host and result in acyl transfer reactions relevant to enzymatic processes. See AROMATIC HYDROCARBON.

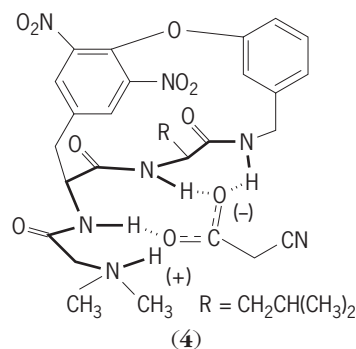
Hybrid structures. Molecules with large cavities can be assembled in modular fashion by using macrocyclic subunits. In one such supramolecular assembly (**3**), a large aromatic porphyrin surface pro-



vides the spacer between two crown ether binding sites. A number of rigid spacing elements other than the aryls shown have been used in this way to permit binding to diammonium ions of appropriate size. In addition, macrocycles composed of both carbohydrate and diaryl spacers have been prepared.

Other cyclic structures. The lining of macrocyclic interiors with acidic sites has permitted the complexation of anions. A synthetic receptor capable of finding certain carboxylates has been prepared. These

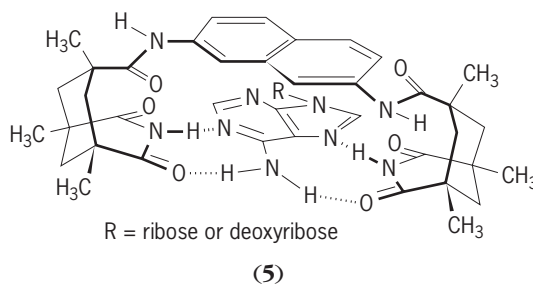
structures, such as (**4**), imitate the vancomycin class



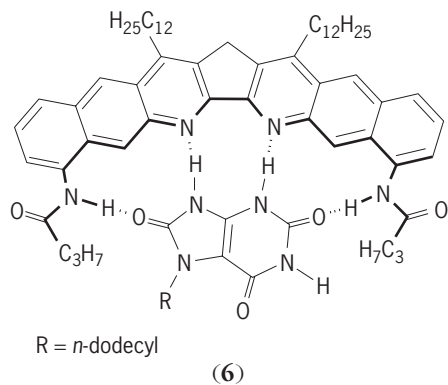
of antibiotics, which bind to peptide acids involved in bacterial cell-wall biosynthesis. A number of modified cyclic structures bind heterocycles such as imidazoles.

Concave structures. Because the encircling of larger, more complex molecules with macrocycles poses structural problems, other molecular shapes have been explored. Cleft molecules offer advantages in this regard. The principle underlying these systems involves the shape of the small organic target molecules: convex in surface and bearing functional groups that diverge from their centers. Accordingly, designing a trap for such targets requires molecules of a concave surface in which functional groups converge. This complementarity is also a feature of the immune system: the "hot spots" of an antigen tend to be convex, whereas the binding sites of the antibody are concave.

Systems featuring a cleft have been developed to bind adenine derivatives and other heterocyclic systems through chelation, as shown in (**5**). The

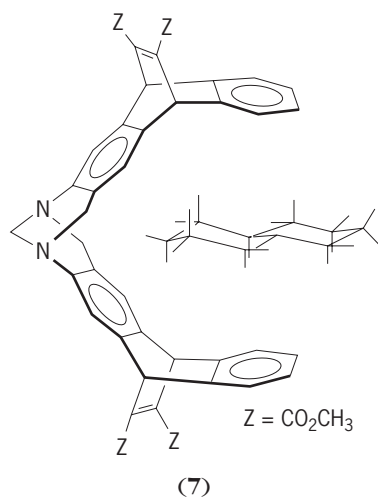


concave surface can provide both Watson-Crick and Hoogsteen base pairing, whereas aromatic stacking interactions help fix the adenine to the synthetic receptor. The high lipophilicity of these structures permits the transport of adenosine across liquid membranes between two aqueous phases. A sickle-shaped system takes advantage of the rigid aromatic structures bearing heterocyclic nitrogen. These provide convergent acidic and basic binding sites closely tailored to uric acid derivatives, which they complex [structure (**6**)] and solubilize in organic media.

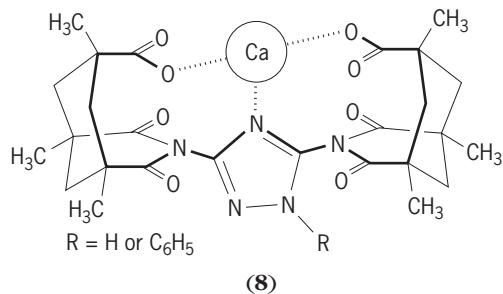


Related molecules bind urea. See CHELATION; UREA; URIC ACID.

Cleft structures also offer hydrophobic and van der Waals forces. In one system, decalin is sandwiched [structure (7)] between the π systems of a rigid syn-



thetic receptor. At the other extreme, highly polar microenvironments have been engineered into concave surfaces. An example is provided by the metal chelate structure (8). The convergence of the more



basic electron pairs of the unusual dicarboxylic acid provides high selectivity (a molecular vise) for divalent ions.

Experimental methods. The complexation of the synthetic receptors with the target molecules is most frequently assessed by observation of solubilization. For example, the extractability of metal ions from aqueous solutions into organic solvents is evidence

Association constants for complexes of the acridine-derived diacid*

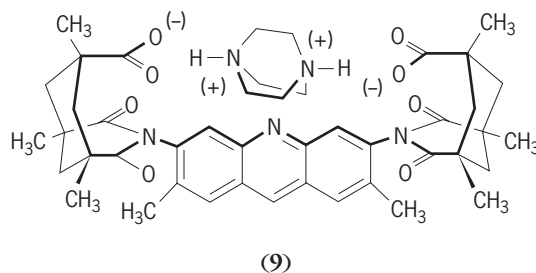
| Base | K_a (M^{-1}) | pK_a (BH ⁺) |
|----------------------|--|---------------------------|
| Triethylene diamine* | 1.6×10^5 | 8.2, 4.2 |
| Pyrazine | 1.4×10^3 | 0.65 |
| Quinoxaline | 23×10^3 | 0.56 |
| Phenazine | 2.2×10^3 | 1.2 |
| Pyrimidine | 0.7×10^3 | 1.3 |
| Quinazoline | 1.6×10^3 | 1.9 |
| Imidazole† | $K_1 = 1.0 \times 10^6$ $K_2 = 5.5 \times 10^4$ | 6.9 |
| Benzimidazole† | $K_1 = 1.5 \times 10^4$ $K_2 = 7.5 \times 10^3$ | 5.5 |
| Purine | $\sim 8 \times 10^3$ | 2.4 |
| Pyridine† | $K_1 = 1.2 \times 10^2$ $K_2 < 1$ | 5.2 |

* In chloroform-*d* (CDCl₃) solution, at 77°F (25°C).

† K_1 and K_2 are association constants for one and two molecules, respectively.

of an unusually favorable relationship between host and guest. Similar observations can be made with host species and amino acids, barbiturates, nucleosides, and uric acid derivatives. It is also possible to perform solid-liquid extractions, because the complexes tend to be more soluble than the individual components. Another way of establishing unusual affinities is through transport studies. Active or passive transport of amino acids, metal ions, and nucleosides can be observed by using liquid membranes that consist of a chloroform layer between two aqueous solutions. See AMINO ACIDS; BARBITURATES.

Quantitative determinations are available through titrations. Association constants can be determined between host and guest by spectroscopic techniques. Frequently, nuclear magnetic resonance (NMR) is convenient because it determines response and concentration simultaneously. A special advantage of nuclear magnetic resonance deals with the use of Overhauser effects and two-dimensional spectra. These permit the assignment of contact points between host and guest that map the geometry of the complexes. The table shows typical data for titration of molecular clefts featuring convergent acidic sites with appropriate guest species that contain divergent basic sites in chloroform-*d* (CDCl₃) solution; the molecular chelation of triethylene diamine (DABCO) within a cleft derived from acridine is shown in (9). Such data reveal features



of size and shape complementarity, acid-base effects, and stereoelectronic details within the complexes. See CONFORMATIONAL ANALYSIS; NUCLEAR

MAGNETIC RESONANCE (NMR); SPECTROSCOPY; STEREOCHEMISTRY; TITRATION.

Future applications. Apart from the abstract questions concerning articulation of molecules, some practical applications in the pharmaceutical industry may be envisioned. Many of the target structures are biologically active, and the use of synthetic sequestering agents for metabolic substrates can represent a new approach to biochemical methods and drug delivery. This represents a departure from most current strategies, in which small convex molecules (penicillin, mitomycin, beta blockers) are directed at the folds of a biological macromolecule. In addition, development of sensing devices at the molecular level and the synthesis of new composites are expected. Another issue to be addressed will be how the molecular fit alters the chemical behavior of the interacting partners.

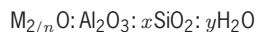
Julius Rebek, Jr.

Bibliography. J. L. Atwood (ed.), *Inclusion Phenomena and Molecular Recognition*, 1990; A. D. Buckingham, A. C. Legon, and S. M. Roberts (eds.), *Principles of Molecular Recognition*, 1993; M. Delaage (ed.), *Molecular Recognition Mechanisms*, 1991; E. R. Weber (ed.), *Supramolecular Chemistry, I: Directed Synthesis and Molecular Recognition*, 1993.

Molecular sieve

Any one of the crystalline metal aluminosilicates belonging to a class of minerals known as zeolites. These minerals are found widely scattered in nature in relatively small quantities. Synthetic forms of the naturally occurring minerals, as well as many species having no known natural counterpart, have been prepared by a hydrothermal process. An important characteristic of the zeolites is their ability to undergo dehydration with little or no change in crystal structure. The dehydrated crystals are honeycombed with regularly spaced cavities interlaced by channels of molecular dimensions which offer a very high surface area for the adsorption of foreign molecules.

Structure. The basic formula for all crystalline zeolites can be represented as where M represents a



metal ion and n its valence. In general, a particular crystalline zeolite will have values for x and y that fall into a definite range. For example, in two synthetic varieties of molecular sieve, designated type A and type X, the values of x are typically about 2.0 and 2.5, respectively. When the crystal is fully dehydrated, the value of y for both types is zero. The crystal structure consists basically of a three-dimensional framework of SiO_4 and AlO_4 tetrahedrons (Fig. 1). The tetrahedrons are cross-linked by the sharing of oxygen atoms, so that the ratio of oxygen atoms to the total of silicon and aluminum atoms is equal to 2. The electrovalence of the tetrahedrons containing aluminum is balanced by the inclusion of cations in the crystal. One cation may be exchanged for an-

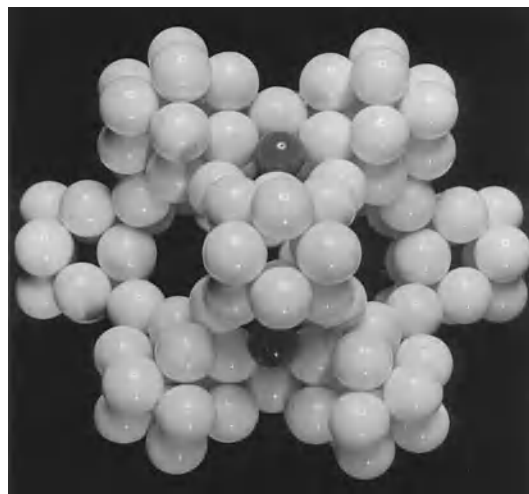


Fig. 1. Molecular sieve type-A crystal model. Dark spheres represent the included cations, and light spheres the SiO_4 or AlO_4 tetrahedrons.

other by the usual ion-exchange techniques. The size of the cation and its position in the lattice determine the effective diameter of the pore in a given crystal species.

The crystal habit of molecular sieve type X is similar to that of diamond, in which the carbon atoms are replaced by silica-alumina polyhedrons. With alkali metal ions present in the structure, effective pore diameter is between 0.9 and 1.1 nanometers; with the alkaline earth cations present, effective diameter is between 0.8 and 0.9 nm.

Properties. The properties of molecular sieves as adsorbents which distinguish them from nonzeolitic adsorbents are (1) the relatively strong coulomb fields generated by the adsorption surface and (2) the uniform pore size; the pore size is controlled, in a given crystal species, by the associated cation. The strong surface forces are reflected in the peculiar shape of the adsorption isotherm, the character of the adsorption isobar, and the relatively high heats of adsorption. The isotherm, found by plotting the capacity for a given adsorbate against pressure or concentration at constant temperature, is of the so-called Langmuir type (Fig. 2). The shape of the isotherm is approximately rectangular, rising steeply at low partial pressures or concentrations and leveling off when maximum load is attained. The isobar, found by plotting capacity against temperature at constant pressure, shows that molecular sieves have an unusually high capacity at elevated temperatures. The relatively high exothermic heats characteristic of adsorption on molecular sieves necessitate somewhat higher heat requirements to effect desorption than are necessary with other adsorbents.

The basic characteristics of molecular sieves are utilized commercially in several production and research applications. Their absorption properties make them useful for drying, purification, and separations of gases and liquids. Conversely, molecular sieves can be preloaded with chemical agents, which are thereby isolated from the reactive system

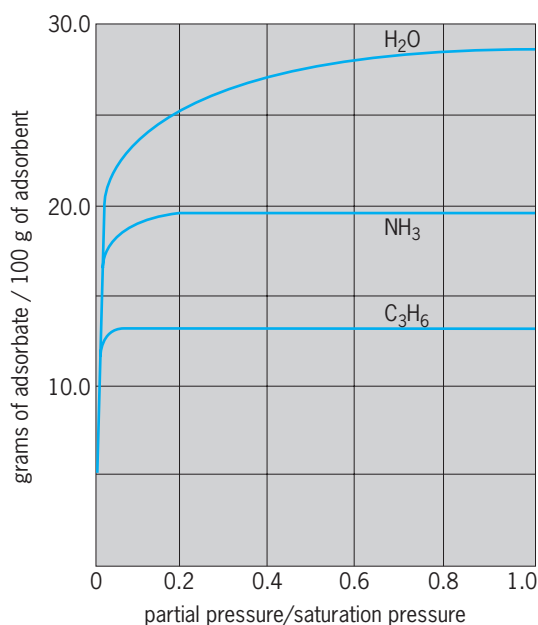


Fig. 2. Isotherms for type-5A molecular sieves at 25°C (77°F).

in which they are dispersed until released from the adsorbent either thermally or by displacement by a more strongly adsorbed compound. The presence in the crystal lattice of an associated exchangeable metal ion provides the basis for their use as a cation exchange medium. Their chemical composition and crystal structure make them novel catalysts and catalyst supports.

As is predictable from the water-adsorption isotherm (Fig. 2), molecular sieves are capable of drying gases and liquids to extremely low residual water concentrations. The isotherm also shows that, even at low initial water concentrations in a gas or liquid, the relative desiccant capacity is high.

By virtue of the uniform pore size of a given molecular-sieve crystal type, molecules having a minimum projected cross section larger than the effective diameter of the zeolite pore are excluded from the internal surface. Molecules having a minimum projected cross section smaller than the effective pore diameter are adsorbed internally. This phenomenon is utilized in separating molecules of fluid mixtures on the basis of their size or shape. For example, the 0.5-nm molecular sieve type has an effective pore size such that straight-chain hydrocarbons are adsorbed and thus effectively separated from branched-chain and cyclic hydrocarbons, which are excluded from the pore system in the selective adsorption process.

In general, when two molecules of similar volatility are sufficiently small to enter the pore system, separation is based on the degree of unsaturation or on the polarity of the molecules. The more unsaturated or more polar molecule is more strongly adsorbed. See ADSORPTION; GAS CHROMATOGRAPHY; ION EXCHANGE; ZEOLITE.

R. L. Mays

Bibliography. D. W. Breck, *Molecular Sieves: Structure, Chemistry and Use*, 1974; E. M. Flanigen and

L. B. Sand (eds.), *Molecular Sieve Zeolites II*, 1971; W. H. Flank and T. E. White (eds.), *Perspectives in Molecular Sieve Science*, 1988; J. R. Katzer (ed.), *Molecular Sieves II*, 1977; W. M. Meier and J. B. Uytterhoeven (eds.), *Molecular Sieves*, 1973.

Molecular simulation

Molecular simulation and computational quantum chemistry are rapidly evolving tools which are having an increasingly significant impact on the chemical, pharmaceutical, materials, and related industries. These tools provide a mechanism for predicting entirely computationally many useful functional properties of systems of interest in these industries. Included are thermodynamic properties (such as equations of state, phase equilibria, and critical constants), thermochemical properties (such as energies of formation and reaction, and reaction pathways), spectroscopic properties (such as dipole moments and vibrational and other spectra), mechanical properties (such as stress-strain relationships and elastic moduli), transport properties (such as viscosity, diffusion, and thermal conductivity), and morphological information (such as location and shape of binding sites on a biomolecule and crystal structure). This list continues to grow as algorithmic and computer hardware advances make it possible to access additional properties.

Definitions. The two main molecular simulation techniques are molecular dynamics and Monte Carlo simulation, both of which are rooted in classical statistical mechanics. Given mathematical models for the internal structure of each molecule (the intramolecular potential which describes the energy of each conformation of the molecule) and the interaction between molecules (the intermolecular potential which describes the energy associated with molecules being in a particular conformation relative to each other), classical statistical mechanics provides a formalism for predicting properties of a macroscopic collection of such molecules based on statistically averaging over the possible microscopic states of the system as it evolves under the rules of classical mechanics. Thus, the building blocks are molecules, the dynamics are described by classical mechanics, and the key concept is statistical averaging. In molecular dynamics, the microscopic states of the system are generated by solving the classical equations of motion as a function of time (typically over a period limited to tens of nanoseconds). Thus, one can observe the relaxation of a system to equilibrium (provided the time for the relaxation falls within the time accessible to molecular dynamics simulation), and so molecular dynamics permits the calculation of transport properties which at the macroscopic scale describe the relaxation of a system in response to inhomogeneities. In Monte Carlo simulation, equilibrium configurations of systems are generated stochastically according to the probabilities rigorously known from classical statistical mechanics. Thus, Monte Carlo simulation

generates equilibrium states directly (which has many advantages, including bypassing configurations which are not characteristic of equilibrium but which may be difficult to escape dynamically) and so can be used to study equilibrium configurations of systems which may be expensive or impossible to access via molecular dynamics. The drawback of Monte Carlo simulation is that it cannot yield the kind of dynamical response information that leads directly to transport properties. Many hybrid methods exist that combine the best features of molecular dynamics and Monte Carlo simulation. Today, intra- and intermolecular forces are determined by a combination of computational quantum chemistry methods and fitting to available experimental data, and the development of these functions (also known as force fields) is a very active subfield of molecular simulation. *See* CHEMICAL DYNAMICS; COMPUTATIONAL CHEMISTRY; MONTE CARLO METHOD; SIMULATION; STOCHASTIC PROCESS.

In computational quantum chemistry, the Schrödinger equation for the electronic degrees of freedom of a collection of atoms (nuclei and electrons) is solved numerically to determine the minimum energy configuration of the atoms. The strength of computational quantum chemistry is that it makes few, if any, assumptions (unlike molecular dynamics and Monte Carlo simulation, in which a model must be assumed for the intra- and intermolecular interactions). The problem is that the computational cost is very high, limiting the methods to relatively small system sizes on current computers. The success of computational quantum chemistry methods resulted in one of the pioneers in the field, John Pople, receiving the 1998 Nobel Prize in Chemistry. One of the developments that has significantly sped up computational quantum chemistry calculations is density functional theory, which solves equations for electron density rather than positions of individual electrons. The primary developer of density functional theory, Walter Kohn, shared the 1998 Nobel prize with Pople. In the course of finding the minimum energy configuration of a collection of molecules, computational quantum chemistry yields information about intra- and intermolecular interactions, and so represents a route to these functions needed by molecular dynamics and Monte Carlo simulation. In fact, by combining molecular dynamics with the density functional theory form of computational quantum chemistry (so that the interactions needed by molecular dynamics are provided by density functional theory calculations) one can perform so-called *ab initio* molecular dynamics, the best-known version being Car-Parrinello. The Car-Parrinello and similar hybrid methods are presently restricted to fairly small systems simulated over periods of 1–10 picoseconds, yet they offer a glimpse into the future of molecular simulation and computational quantum chemistry when improvements in hardware and algorithms will make *ab initio* hybrid methods the primary tool for predicting the properties of systems about which little or nothing is known

experimentally. *See* COMPUTATIONAL CHEMISTRY; QUANTUM CHEMISTRY.

The use of molecular simulation and computational quantum chemistry has become widespread, well beyond the research specialists, by the development of commercial versions of research codes, often with graphical interfaces that facilitate use of the methods. Companies which market such codes include Molecular Simulations Inc. (molecular simulation and computational quantum chemistry), Gaussian Inc. (computational quantum chemistry), Oxford Molecular (molecular simulation and computational quantum chemistry), and Schrödinger Inc. (molecular simulation/computational quantum chemistry hybrid). This list is not exhaustive. In addition, many noncommercial ventures (located at national laboratories or universities) make available molecular simulation and computational quantum chemistry codes. Examples include the NWChem package from Pacific Northwest National Laboratory (molecular simulation and computational quantum chemistry) and the DL-POLY code (molecular simulation) available from Daresbury Laboratory in the United Kingdom.

Capabilities and impact. Computational quantum chemistry and molecular simulation methods can be used to predict properties that once were only accessible experimentally, resulting in several significant applications in basic and industrial research. These applications include providing estimates of properties for systems for which little or no experimental data are available, which is especially useful in the early stages of chemical process design; yielding insight into the molecular basis for the behavior of particular systems, which is very useful in developing engineering correlations, design rules, or quantitative structure-property relations; and providing guidance for experimental studies by identifying the interesting systems or properties to be measured.

In the fields of chemical engineering and materials science, academic and industrial interest in molecular simulation and computational quantum chemistry has grown exponentially over the past decade. Two developments are making molecular simulation and computational quantum chemistry and practical tools for application to systems of industrial interest. One is the continuing exponential increase in computing power, often referred to as Moore's law, which states that the speed of computers is increasing by a factor of two every 18 months (equivalently, an order of magnitude every 5 years). The practical consequence of Moore's law is that today's desktop work stations and laptops have the computational power of a supercomputer of a decade ago, thus making it possible for an ordinary user to perform calculations that once would have been prohibitively expensive. The second development is the proliferation of efficient algorithms for calculating properties of practical interest. In molecular simulation, for example, the Gibbs ensemble Monte Carlo algorithm, introduced just over a decade ago, made it possible to predict the phase envelope of a fluid system directly from knowledge of the intermolecular forces,

thus opening up the possibility of calculating many properties relevant to the engineering design separations processes (such as distillation). Algorithm advances (which can result in order-of-magnitude or more increases in efficiency), combined with the steady exponential increase in computational power, will continue to dramatically broaden the range of applicability of computational quantum chemistry and molecular simulation methods in the coming years. The expectations for molecular simulation and computational quantum chemistry are very high, and they have been identified as key technologies for the chemical industry to achieve its vision of completely automated product and process design in the year 2020.

Examples. A sampling of companies, and examples of the industrial problems which they have tackled using molecular simulation and computational quantum chemistry techniques, is provided in a National Science Foundation report (1997). Applications are diverse and include molecular design of zeolites for air separation, design and optimization of catalysts for homogeneous and heterogeneous catalysis, evaluation of the feasibility of methane adsorption, molecular modeling of polysaccharide rheology, molecular design of detergent enzymes, and molecular design of photocopying materials. The 2000 report was expected to provide additional comparative information on the ways that European, Japanese, and American companies are applying these tools.

One area which is increasingly benefiting from molecular simulation methods is rheology. Nonequilibrium molecular dynamics methods have become very powerful tools for direct simulation of systems subject to flow fields, such as the shear flow field present in lubrication and the extensional flow fields present in polymer processing. The theory for deriving and validating methods for performing nonequilibrium molecular dynamics simulations has been perfected over the past 15 years. Recently, this technique has been used to predict the viscosity index of several typical lubricant components. The viscosity index is a key performance indicator for lubricants, measuring the degree to which the viscosity diminishes with increasing temperature. Nonequilibrium molecular dynamics methods are also being used to understand how nanoscale confinement, such as in the case of hard-disk drive lubrication, affects the rheological properties of lubricants (Fig. 1). These studies also shed light on the lubrication of microelectromechanical systems.

Rational catalyst design is one example of the use of computational quantum chemistry and ab initio molecular dynamics. Density functional methods make it possible to study catalytic processes qualitatively, and in some cases quantitatively, with increasingly sophisticated questions becoming answerable. For example, Figure 2, based on ab initio molecular dynamics simulations, shows a snapshot from a simulation study of the effect of water on the dissociation of acetic acid over Pd(III). The protons remain solvated in the water layer (as H_5O_2^+ species),

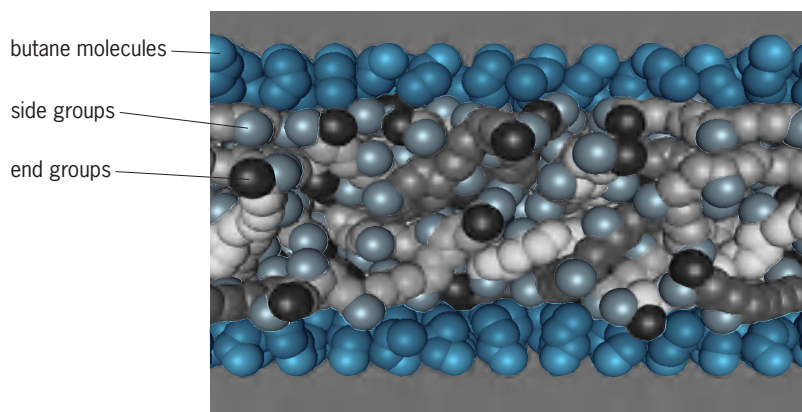


Fig. 1. A lubricant component, squalane (a 30-carbon alkane with a 24-carbon backbone and 6 methyl side groups), is confined to a nanoscale gap created by a surface to which butane molecules have been attached. The visualization depicts the carbons only. Molecules are shaded differently to distinguish them. The goal of such simulations is to understand the changes in lubrication properties induced by confinement to such narrow gaps.

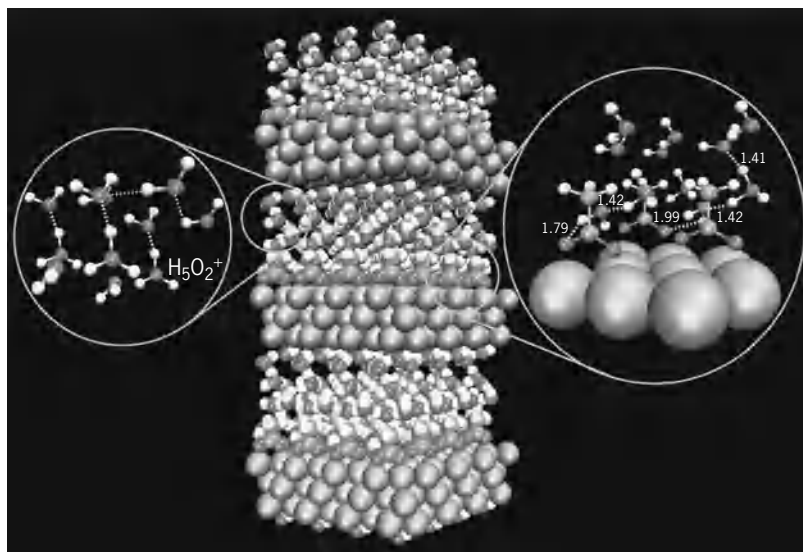


Fig. 2. Snapshot from an ab initio molecular dynamics simulation probing the effect of water on the dissociation of acetic acid over Pd(III). The change in binding energy of the acetate anion on Pd(III) in the liquid phase is measured in the simulation to be -142 kJ/mol, while the same quantity in the vapor phase (that is, in the absence of the aqueous solvent) is -212 kJ/mol. (Simulation performed by, and image courtesy of, Matt Neurock, University of Virginia)

while acetate anions are bound to the surface. The solvent lowers the interaction energy significantly. The water solvent interacts with the naked negative charge of the anion, thus decreasing the anion's affinity for the surface.

The rapidly evolving capabilities of molecular simulation and computational quantum chemistry, and hybrids of these methods, suggest that they are beginning to fulfill their promise as equal and complementary partners with experiments in chemical and related industries.

Peter J. Cummings

Bibliography. M. P. Allen and D. J. Tildesley, *Computer Simulation of Liquids*, Oxford University Press, Oxford, 1987; R. Car and M. Parrinello, Unified approach for molecular dynamics and density-functional theory, *Phys. Rev. Lett.*, 55:2471, 1985;

P. T. Cummings et al., *Report: NSF Workshop on Future Directions for Molecular Modeling and Simulation: Fundamentals and Applications*, National Science Foundation, Arlington, VA, November 3–4, 1997; D. J. Evans and G. P. Morriss, *Statistical Mechanics of Nonequilibrium Liquids*, Academic Press, New York, 1990; A. Z. Panagiotopoulos, Direct determination of phase coexistence properties of fluids by Monte Carlo simulation in a new ensemble, *Mol. Phys.*, 61:813, 1987; A. Z. Panagiotopoulos et al., Phase equilibria by simulation in the Gibbs ensemble: Alternative derivation, generation, and application to mixture and membrane equilibria, *Mol. Phys.*, 63:527, 1988; *Technology Vision 2020: The U.S. Chemical Industry*, a joint report, 1996; S. Sarman, D. J. Evans, and P. T. Cummings, Recent developments in non-equilibrium molecular dynamics, *Phys. Rep.*, 305:1–92, 1998.

Molecular structure and spectra

Until the advent of quantum theory, ideas about the structure of molecules evolved gradually from analysis and interpretation of the facts of chemistry. Chemists developed the concept of molecules as built from atoms in definite proportions, and identified and constructed (synthesized) a great variety of molecules. Later, when the structure of atoms as built from nuclei and electrons began to be understood with the help of quantum theory, a beginning was made in explaining why atoms can combine in definite ways to form molecules; also, infrared spectra began to be used to obtain information about the dimensions and the nuclear motions (vibrations) in molecules. However, a fundamental understanding of chemical binding and molecular structure became possible only by application of the present form of quantum theory, called quantum mechanics. This theory makes it possible to obtain from the spectra of molecules a great deal of information about the nature of molecules in their normal as well as excited states, and about dissociation energies and other characteristics of molecules. For an important aspect of molecular structure which is treated separately see CHEMICAL BONDING.

Molecular sizes. The size of a molecule varies approximately in proportion to the numbers and sizes of the atoms in the molecule. Simplest are diatomic molecules. These may be thought of as built of two spherical atoms of radii r and r' , flattened where they are joined. The equilibrium value R_e of the distance R between their nuclei is then smaller than the sum of the atomic radii (Fig. 1). However, the nuclei of atoms in two different molecules cannot normally approach more closely than a distance $r + r'$; r and r' are called the van der Waals radii of the atoms. The smallest molecule is hydrogen (H_2), with two electrons whose negative charges equal the positive charges of the two nuclei. Here r is about 0.12 nanometer giving $r + r' = 0.24$ nm, but R_e is only 0.074 nm. In HCl $r = 0.12$ nm for H and 0.18 nm for Cl, but R_e is only 0.127 nm.

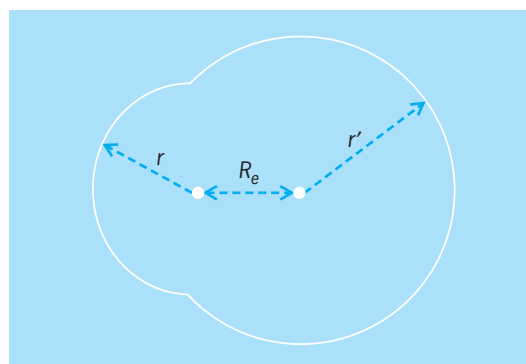


Fig. 1. Diatomic molecule with nuclei at distance R_e apart, built from atoms of radii r and r' .

To describe a polyatomic molecule, one must specify not merely its size but also its shape or configuration. For example, carbon dioxide (CO_2) is a linear symmetrical molecule, the $O-C-O$ angle being 180° . The $H-O-H$ angle in the nonlinear water (H_2O) molecule is 105° . Many molecules which are essential for life contain thousands or even millions of atoms. Proteins are often coiled or twisted and cross-linked in curious ways which are important for their biological functioning.

Dipole moments. Most molecules have an electric dipole moment. In atoms, the electron cloud surrounds the nucleus so symmetrically that its electrical center coincides with the nucleus, giving zero dipole moment; in a molecule, however, these coincidences are disturbed, and a dipole moment usually results.

Thus, when the atoms of HCl come together, there is some shifting of the H-atom electron toward the Cl. A complete shift would give H^+Cl^- , which would constitute an electric dipole of magnitude eR_e , where e is the electronic charge. But in fact the dipole moment is only $0.17eR_e$. This is because the actual electronic shift is only fractional. See ELECTRONEGATIVITY.

Although in molecules such as H_2 , N_2 , and CO_2 partial shifts of electronic charge from the original atoms do take place, these necessarily occur so symmetrically that no dipole moment results. Many larger molecules also have zero dipole moments by virtue of high symmetry. Examples are methane (CH_4), uranium hexafluoride (UF_6), and benzene (C_6H_6).

In general, the dipole moment of a neutral molecule is defined as the vector sum of quantities $+QS$ for the nuclei and $-es$ for the electrons. Here Q is the charge on any nucleus and S its vector distance from any fixed point in the molecule; s is the average vector distance of any electron from the same point. To calculate a dipole moment with these definitions, quantum mechanics must be used.

However, a study of what is known experimentally about molecular dipole moments has led to useful semiempirical generalizations. Bond moments and group moments have been obtained for various types of chemical bonds and of chemical groups or radicals. By adding these vectorially, the actual dipole moment of a large molecule can often be reproduced

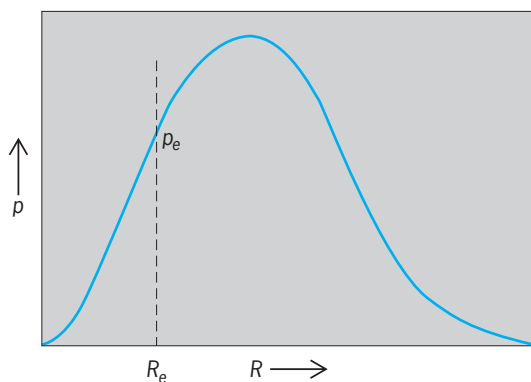


Fig. 2. Electric dipole moment p of typical diatomic molecule as function of internuclear distance R ; p_e is the dipole moment at the equilibrium distance R_e .

fairly accurately. In CH_4 or CO_2 , one can assume a moment for each $\text{C}-\text{H}$ or $\text{C}=\text{O}$ bond, even though these cancel out vectorially to give a zero resultant. In the linear molecule OCS , the unequal moments of the $\text{C}=\text{O}$ and $\text{C}=\text{S}$ bonds give a nonzero resultant. Because of the zero moment of CH_4 , the CH bond moment and the CH_3 group moment must be equal and opposite. In CH_3Cl , the total moment can be thought of as the vector sum of the H_3C group moment and the $\text{C}-\text{Cl}$ bond moment. See PERMITTIVITY.

When molecules vibrate, their dipole moments usually vary. Figure 2 shows how the dipole moment \mathbf{p} in a diatomic molecule may vary with R ; the quantity previously discussed is \mathbf{p}_e , the value of \mathbf{p} at R_e . When \mathbf{p}_e is zero because of symmetry, it remains zero for symmetrical vibrations but, in polyatomic molecules, varies during unsymmetrical vibrations.

Molecules may possess magnetic as well as electric dipole moments. See DIPOLE MOMENT; MICROWAVE SPECTROSCOPY.

Molecular polarizability. In the preceding consideration of dipole moments, the discussion has been in terms of atoms and molecules free from external forces. An electric field pulls the electrons of an atom or molecule in one direction and pushes the nuclei in the opposite direction. This action creates a small induced dipole moment, whose magnitude per unit strength of the field is called the polarizability.

Molecular polarizabilities can be expressed as sums of atomic polarizabilities, plus corrections depending on the types of bonds present. Polarizabilities increase rather rapidly in such series as F , Cl , Br , I , and also from HF to HI , or F_2 to I_2 .

Molecular polarizabilities can also be expressed approximately as sums of bond polarizabilities. These polarizabilities are anisotropic, being greater along bonds than perpendicular to bonds.

Molecular energy levels. The stationary states of motion of nuclei and electrons in a molecule, or of electrons in an atom, are restricted by quantum mechanics to special forms with definite energies. (Nonstationary states, which vary in the course of time, are constructed by mixing stationary states of different energies.) The state of lowest energy is

called the ground state; all others are excited states. In analogy to water levels in a river, the energies of the stationary states are called energy levels. Excited states exist only momentarily, following an electrical or other stimulus. See ENERGY LEVEL (QUANTUM MECHANICS); NONRELATIVISTIC QUANTUM THEORY; QUANTUM CHEMISTRY; QUANTUM MECHANICS.

Energy levels are either discrete or continuous. The levels of a self-contained atom or molecule are restricted to special, sharply defined values (discrete levels). When an atom or molecule is ionized, that is, when one of its electrons has enough energy to escape completely, the energy can take on any value exceeding the minimum escape energy. This range of energies is called a continuous level or an ionization continuum. Molecules also have dissociation continua, which are discussed below.

Excitation of an atom consists of a change in the state of motion of its electrons. Electronic excitation of molecules can also occur, but alternatively or additionally, molecules can be excited to discrete states of vibration and rotation.

A diatomic molecule can vibrate similar to the motion of two masses separated by a distance R connected by a spring. In a diatomic vibration, R varies periodically above and below R_e , the equilibrium internuclear distance. The possible vibration energies E_v (in joules) are given by Eq. (1),

$$E_v/h = \omega_e(v + 1/2) - \omega_e x_e(v + 1/2)^2 + \dots \quad (1)$$

where ω_e is just the small-amplitude vibration frequency (in hertz, with dimensions s^{-1}), and h is the Planck constant (6.626×10^{-34} J·s); $\omega_e x_e$ is a small quantity (relative to ω_e) that is nearly always positive. The vibrational quantum number v can take whole number values 0, 1, 2, and so forth. The $+\dots$ in Eq. (1) indicates small correction terms. The zero-point vibration energy $1/2 h \omega_e - 1/4 h \omega_e x_e$ present in the ground vibration state ($v = 0$) is a characteristic manifestation of quantum phenomena.

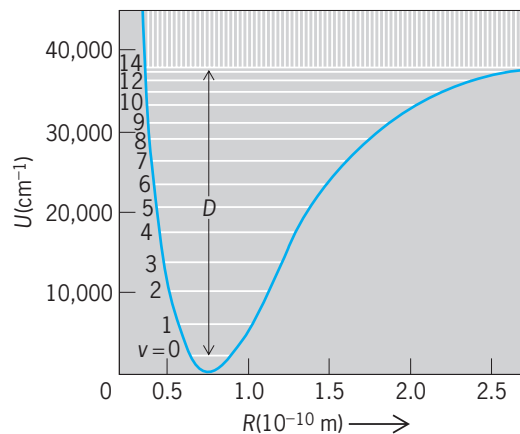


Fig. 3. $U(R)$ curve of ground electronic state of H_2 with vibrational levels and dissociation continuum. D indicates the dissociation energy. Maximum v here is 14. (After G. Herzberg, *Molecular Spectra and Molecular Structure*, vol. 1, 2d ed., Van Nostrand, 1950)

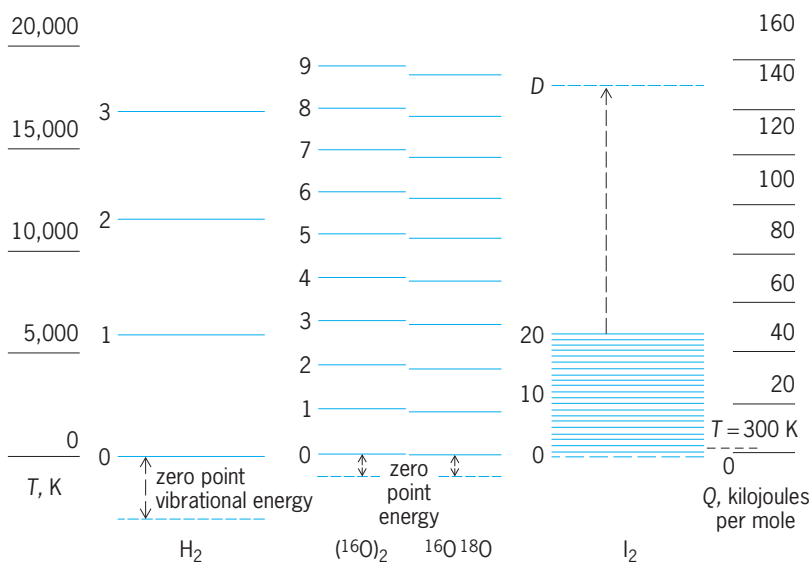


Fig. 4. Lowest vibrational levels of H_2 , O_2 , and I_2 , numbered by vibrational quantum number v . Vibration level spacings decrease with increasing v . Where spacing reaches zero, the molecule dissociates; dissociation level D is indicated for I_2 . Energies are given by the scale at right. The scale at left shows the average energy of vibration at various temperatures. $^\circ\text{F} = (\text{K} \times 1.8) - 459.67$.

Units can be confusing in spectroscopy because energies, vibrational frequencies, and other quantities are often expressed in wavenumbers ($\tilde{\nu}$), traditionally defined as the number of wavelengths per centimeter (that is, $E = h\nu = hc\lambda^{-1} = 100hc\tilde{\nu}$). In physics the symbol ω is generally an angular frequency (units of radians s^{-1}) given by $2\pi\nu$, but in spectroscopy ω_e is a frequency or wavenumber given typically in megahertz (hertz/ 10^6) or cm^{-1} units. It is also customary to treat $\omega_e x_e$ as a single symbol, rather than as product of ω_e and x_e .

The value of ω_e (in hertz) depends on the masses m_1 and m_2 of the atoms and the force constant k , as

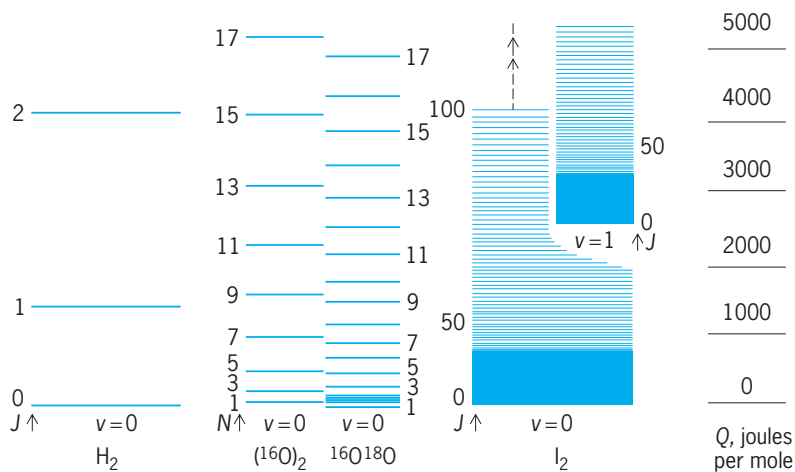


Fig. 5. Lowest rotational levels of H_2 , O_2 , and I_2 . For H_2 and I_2 , J is the rotational quantum number, according to Eq. (5) in the text. O_2 is in a Hund's case b triplet state, and the rotational levels are designated by N , where the total angular momentum $J = N + 1$, N , and $N - 1$. This narrow spin tripling is indicated for the $N = 1$ level of $^{16}\text{O}^{18}\text{O}$ only. Energies are given by the scale shown at right.

shown in Eq. (2), with the reduced mass, μ , defined

$$\omega_e = \frac{1}{2\pi} \sqrt{k/\mu} \quad (2)$$

as $m_1 m_2 / (m_1 + m_2)$.

The vibrational motion of a diatomic molecule is governed by the potential energy curve, which shows how the energy of attraction $U(R)$ of the atoms varies with R (Fig. 3). The quantities R_e , k , and the dissociation energy D are the most important properties of a potential curve; k is d^2U/dR^2 taken at R_e . The $U(R)$ curve and vibrational levels for the ground electronic state of H_2 are shown in Fig. 3. Similar curves, but with other R_e , k , and D values, exist for other electronic states and other molecules. Molecules have also repulsive electronic states, whose $U(R)$ curves rise steadily with decreasing R . These are often important for spectroscopy and in atomic collisions. For stable (attractive) $U(R)$ curves, the vibrational levels decrease in spacing as v increases, until finally, as the spacing approaches zero, a maximum v is reached just before dissociation; in Fig. 3 this is 14. After a small gap, a dissociation continuum of energy levels then sets in. Here the atoms have enough mutual kinetic energy to fly apart. For repulsive states, there is only a dissociation continuum, with no vibrational levels. Figure 4 illustrates how strongly vibration level spacings can vary: both k and $1/\mu$, and therefore ω_e , decrease from H_2 to O_2 to I_2 . Figure 4 likewise illustrates the effect of mass in isotopic molecules called isotopologues.

The total internal energy of a molecule is a sum of several terms: the largest is the electronic energy, E_{el} , followed by the vibrational energy, E_v . The total energy can be written as Eq. (3). Both the electronic

$$E = E_{el} + E_v + (E_r + E_{fs} + E_{hfs} + E_{ext}) \quad (3)$$

energy E_{el} and vibration energy E_v can be discrete or continuous. The quantities E_r , E_{fs} , and E_{hfs} denote rotational, fine-structure, and hyperfine-structure energies, respectively. The last two appear as small or minute splittings of the rotation levels. Fine-structure splittings are caused by any unpaired electrons, and hyperfine splittings result from the presence of nuclear spins. The spacings ΔE of adjacent discrete levels of each type are usually in the order given in Eq. (4).

$$\Delta E_{el} \gg \Delta E_v \gg \Delta E_r \gg \Delta E_{fs} \gg \Delta E_{hfs} \quad (4)$$

The fine structures of rotational levels differ strongly for different types of electronic states. The simplest diatomic electronic states are called $^1\Sigma$ states, and include $^1\Sigma^+$ and $^1\Sigma^-$ types for heteropolar and $^1\Sigma_g^+$, $^1\Sigma_u^+$, $^1\Sigma_g^-$, and $^1\Sigma_u^-$ for homopolar molecules. Most even-electron diatomic and linear polyatomic molecule ground states are $^1\Sigma^+$ states ($^1\Sigma_g^+$ if homopolar). The rotational levels of $^1\Sigma$ states have no fine structure; hyperfine structure, because of interaction with nuclear spins, is usually on too small a scale to detect by optical or infrared spectroscopy, but is often seen by rotational

spectroscopy in the microwave region. The E_{ext} term in Eq. (3) refers to additional fine structure which appears on subjecting molecules to external magnetic fields (Zeeman effect) or electric fields (Stark effect). See FINE STRUCTURE (SPECTRAL LINES); HYPERFINE STRUCTURE; STARK EFFECT; ZEEMAN EFFECT.

The rotational levels of any $^1\Sigma$ state are given by Eq. (5). The quantity B_v is related to the moment of inertia

$$E_r/b = B_v J(J+1) - D_v [J(J+1)]^2 + \dots \quad (5)$$

of a diatomic molecule, and to v , by Eq. (6). The small positive

$$B_v = \frac{h}{8\pi^2} \overline{(1/I)}_v = B_e - \alpha_e(v + 1/2) + \dots \\ = B_0 - \alpha_e v + \dots \quad (6)$$

quantity D_v in Eq. (5) is the centrifugal distortion constant that accounts for the stretching of the bonds by centrifugal forces as the molecule rotates. The rotational quantum number J can have any whole number value from 0 up, and corresponds to an angular momentum of magnitude $\sqrt{J(J+1)}(h/2\pi)$. The averaging of $1/I$ in Eq. (6) normally results in a slow decrease of B with increasing v (α_e is usually a small positive quantity). The quantity B_e refers to a hypothetical nonvibrating molecule ($R = R_e$). The B , D , and α_e in constants in Eqs. (5) and (6) are given in frequency units (hertz), and the conversion to the traditional wavenumbers ($\tilde{\nu}$ in cm^{-1}) is again accomplished with the equations: $E = h\nu = hc\lambda^{-1} = 100hc\tilde{\nu}$.

Figure 5 illustrates how enormously rotational level spacings can vary because of differences in m and R_e (both are much greater for I_2 than H_2). The effect of mass for isotopic molecules is illustrated for O_2 . Comparison with Fig. 4 illustrates the relation $\Delta E_v \gg \Delta E_r$, mentioned earlier.

Polyatomic molecules have much more complicated patterns of vibrational and (usually) of rotational energy levels than diatomic molecules. The number of normal modes (independent motions) of vibration for a molecule with n atoms is $3n - 6$ for nonlinear molecules, and $3n - 5$ for linear molecules. Each normal mode is a cooperative vibration of some or all the atoms moving with the same frequency, characteristic of the mode. Sometimes two or even three modes are so related in form that their frequencies are identical. These are called degenerate vibrations.

Figure 6 depicts the normal modes of H_2O and CO_2 . They are labeled by symbols which also denote their frequencies. The arrows indicate the directions of motion of the atoms during one phase of vibration. The CO_2 frequency ν_2 is twofold degenerate: there are two independent modes corresponding to motion in either of two planes at right angles. The other two CO_2 modes, and all three H_2O modes, are nondegenerate.

Molecular spectra. Radiation emitted by a molecule initially in an upper state level E' as it makes a transition to a lower state level E'' obeys

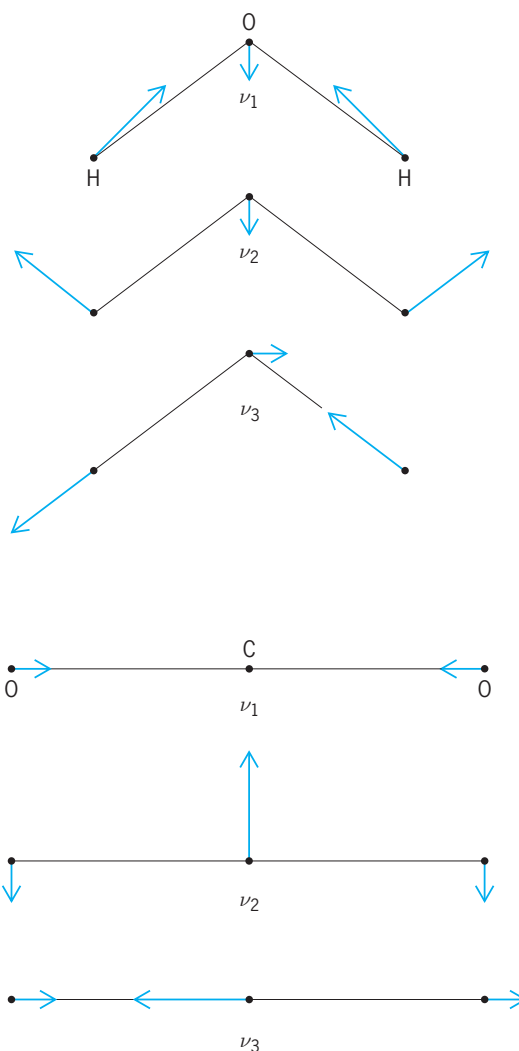


Fig. 6. Normal vibration modes of H_2O and CO_2 . Synchronized displacements of atoms occur in proportion to lengths of the arrows. Diagram corresponds to snapshot taken at one phase of vibration.

the Einstein-Bohr equation (7). Molecular emission

$$h\nu = E' - E'' \quad (7)$$

spectra accompany jumps in energy from higher to lower levels; absorption spectra accompany transitions from lower to higher levels. Both E' and E'' can be either discrete or continuous levels. If both are discrete, they give a spectrum of discrete frequencies; otherwise, they give a continuous spectrum. Discrete spectra are the main type considered here. Discrete frequencies are usually called "lines" because of their appearance when recorded by an optical spectrograph using photographic plates. The wealth and precision of spectroscopic observations have been increased by orders of magnitude by the advent of laser sources and Fourier transform spectrometers. See LASER SPECTROSCOPY.

Besides its frequency, the intensity and width of a spectral line are important. Intensities vary over wide ranges. In the extreme case of nearly zero intensity for a spectroscopic transition, the transition is called

forbidden. Only a small minority of all pairs of levels yield allowed transitions. These are governed by selection rules derivable from quantum theory. See SELECTION RULES (PHYSICS).

Under disturbing influences, however, some lines that violate these rules are seen weakly. Further, the usual selection rules are electric dipole rules, and additional transitions become very weakly allowed if magnetic dipole, electric quadrupole, and other selection rules are also considered. The following discussion is confined to spectra which obey the electric dipole rules.

Molecular spectra can be classified as fine-structure or low-frequency spectra, rotation spectra, vibration-rotation spectra, and electronic spectra. Low-frequency spectra are discussed elsewhere. See ELECTRON PARAMAGNETIC RESONANCE (EPR) SPECTROSCOPY; MAGNETIC RESONANCE; MICROWAVE SPECTROSCOPY; MOLECULAR BEAMS; SPECTROSCOPY.

Pure rotation spectra. Transitions between energy levels differing only in rotational state give rise to pure rotation spectra. For diatomic molecules in $^1\Sigma$ states, Eq. (5), the relation is given by Eq. (8). The transitions

$$\nu = (E'_r - E''_r)/h = B_v[J'(J' + 1) - J''(J'' + 1)] + \dots \quad (8)$$

obey the selection rule $\Delta J = 1$ (ΔJ means $J' - J''$). Setting $J' = J'' + 1$, Eq. (9) is obtained. Equation (9)

$$\nu = 2B_v(J'' + 1) + \dots \quad (9)$$

represents a sequence of lines spaced almost equidistantly ($2B_v, 4B_v, 6B_v, \dots$), and lying in the far infrared or (for small B or low J'') the microwave region. Their intensities are proportional to $|\mathbf{P}_e|^2$, where \mathbf{P}_e is the electric moment at R_e (Fig. 2); hence homopolar molecules (H_2, N_2 , and so on) show no pure rotation spectra. The intensities are proportional also to the lower-state (v'', J'') level population and to ν (for absorption) or ν^4 (for emission).

Pure rotation spectra of linear polyatomic molecules are like those of diatomic molecules. Polyatomic molecules having $\mathbf{P}_e = 0$, whatever their shape (examples are $\text{CO}_2, \text{CH}_4, \text{C}_6\text{H}_6$), have no pure rotation spectra. In other cases, the spectra can be obtained using $h\nu = E'_r - E''_r$ with appropriate E_r expressions and selection rules.

Vibration-rotation bands. Spectra involving only vibrational and rotational state changes lie mainly in the infrared. For a $^1\Sigma$ diatomic state, using Eqs. (1), (5), and (7), Eq. (10) is obtained, with ν_0 defined in Eq. (11). In Eq. (10) B' and B'' mean B_v for v' and v'' ,

$$\nu = \nu_0(v', v'') + B'J'(J' + 1) - B''J''(J'' + 1) + \dots \quad (10)$$

$$\nu_0(v', v'') = (\omega_e - \omega_e x_e)(v' - v'') - \omega_e x_e(v'^2 - v''^2) + \dots \quad (11)$$

respectively. Each band consists of two sets of rotational lines, one on each side of ν_0 , which is called

the band origin. Each line corresponds to a particular rotational transition conforming to $\Delta J = \pm 1$. The two series (branches) have frequencies defined in Eq. (12) for the R or positive branch ($J' = J'' + 1$), and in Eq. (13) for the P or negative branch ($J' = J'' - 1$). Both can be represented by a single equation, Eq. (14), by letting $m = J'' + 1$ for the R and

$$\nu = \nu_0 + (B' + B'')(J'' + 1) + (B' - B'')(J'' + 1)^2 + \dots \quad (12)$$

$$\nu = \nu_0 - (B' + B'')J'' + (B' - B'')J''^2 + \dots \quad (13)$$

$$\nu = \nu_0 + (B' + B'')m + (B' - B'')m^2 + \dots \quad (14)$$

$m = -J''$ for the P branch. Neglecting the term in m^2 , Eq. (14) represents a series of equidistant lines with one missing ($m = 0$) at ν_0 . **Figure 7** shows how the line positions are related to the upper (v') and lower (v'') sets of rotational levels. (For transitions involving at least one non- $^1\Sigma$ state for diatomic and linear molecules as well as for all other polyatomic molecules, Q branches, corresponding to $\Delta J = 0$, can also exist.)

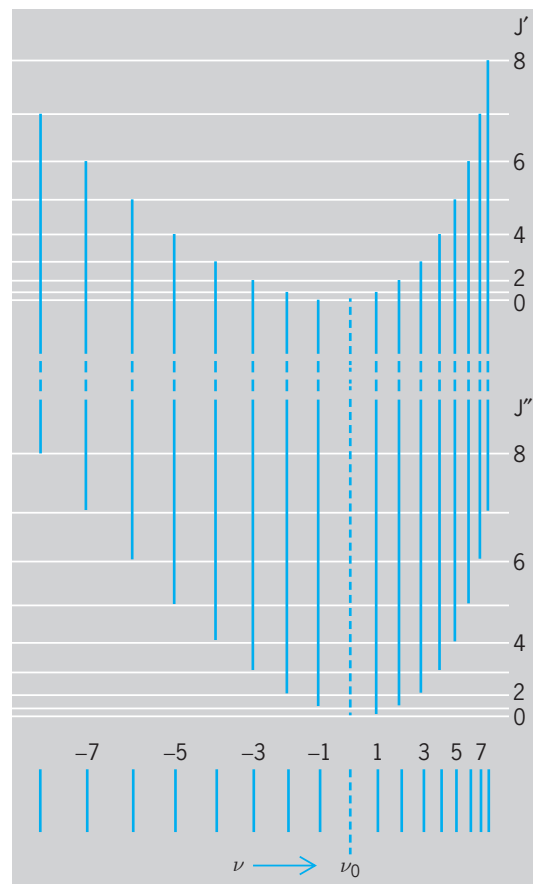


Fig. 7. Relation of band lines (lower part) [see Eqs. (8) and (9)] to rotational levels [see Eq. (6)] for a vibration-rotation band or an electronic band. In the former case, the upper and lower sets of rotational levels belong to two vibrational levels of a $^1\Sigma$ electronic state. In the latter case, they belong to two different $^1\Sigma$ states. Positive m values, R branch; negative m values, P branch.

Since $B' - B''$ is a small negative quantity [see Eq. (6), noting that $\nu' > \nu''$], the m^2 term makes the P line spacing increase and the R line spacing decrease slowly as m increases. This is shown, exaggerated, in Fig. 7. At some large m value, the R branch turns back on itself, but usually the lines have become weak before this value is reached.

The relative intensities of band lines depend primarily on the initial rotational distribution of molecules. More precisely, Eq. (15) holds. Here n in

$$\text{Intensity} = C(\nu', \nu'') \nu^n (J' + J'' + 1) e^{-E/(kT)} \quad (15)$$

ν^n is 4 for an emission band and 1 for an absorption band, while E is the upper state rotational energy level from which emission occurs or the lower state level for the case of absorption. **Figure 8** shows diagrammatically how the values of B and T affect the appearance of a typical absorption band ($B' = B''$ has been assumed for simplicity in Fig. 8). **Figure 9** shows the appearance of an actual DCI band. The weaker $D^{37}\text{Cl}$ lines are at slightly lower frequencies than the $D^{35}\text{Cl}$ lines, mainly because ω_e is smaller [see Eqs. (2) and (11)].

The factor $C(\nu', \nu'')$ is largest by far for fundamental bands (1, 0), and falls rapidly with increasing $\Delta\nu$ in the overtone bands or harmonics ($\Delta\nu$ is $\nu' - \nu''$). For fundamental bands, C depends on the slope of the $P(R)$ curve (Fig. 2), being approximately proportional to $(dP/dR)^2$ taken at R_e . For overtone bands, C depends on the detailed shapes of both the $P(R)$ and $U(R)$ curves. Fundamental or overtone bands arising from $\nu'' > 0$ are called hot bands.

Vibration absorption bands of liquids and solutions are widely used in chemical analysis. For many purposes, it is sufficient to know empirically the spectrum of each molecule which may be present. Also, groups of atoms which recur in many molecules often have nearly constant frequencies, of use for identification and in determining molecular structure. See INFRARED SPECTROSCOPY.

Electronic band spectra. These are the most general type of molecular spectra. The characteristic feature is a change of electronic state. From Eqs. (3) and (7), neglecting fine structure, Eqs. (16) and (17) are

$$h\nu = (E'_{el} - E''_{el}) + (E'_v - E''_v) + (E'_r - E''_r) \quad (16)$$

$$\nu = \nu_{el} + \nu_v + \nu_r = \nu_0 + \nu_r \quad (17)$$

obtained. Diatomic electronic spectra are often observed in emission, while the electronic spectra of polyatomic molecules are usually absorption spectra. Depending mainly on the magnitude of ν_{el} , electronic spectra occur in the infrared, visible, ultraviolet, or vacuum ultraviolet.

For any one electronic transition, the spectrum consists typically of many vibrational bands, each labeled by (ν', ν'') . These lie in general at frequencies both above and below ν_{el} , since ν_v can be positive or negative. They constitute a band system. Each band consists of numerous rotational lines arranged in two or more branches and lying on both sides of a central position $\nu_0(\nu', \nu'')$, called the band origin.

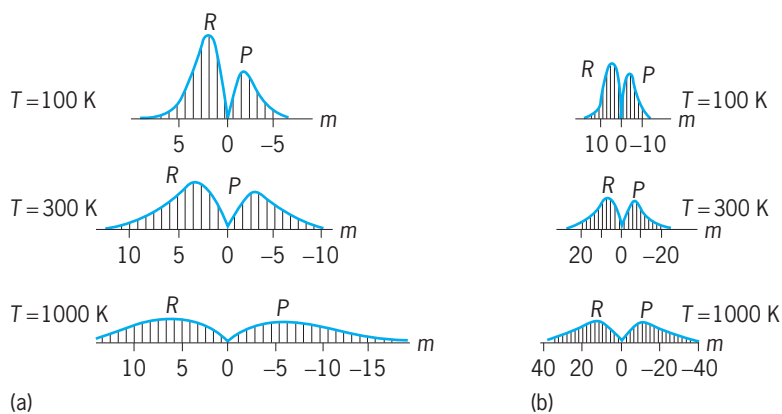


Fig. 8. Intensity distribution at several temperatures for a diatomic absorption band. Line positions are based on Eq. (9) assuming $B' = B''$ for simplicity; frequency increases toward the left (opposite to Fig. 7). (a) and (b) correspond respectively to B values of HCl ($B = 10.44 \text{ cm}^{-1}$) and of 2 cm^{-1} (approximately the value for CO, for which $B = 1.93 \text{ cm}^{-1}$). 100 K = -280°F ; 300 K = 80°F ; 1000 K = 1340°F . (After G. Herzberg, *Molecular Spectra and Molecular Structure*, vol. 1, 2d ed., Van Nostrand, 1950)

For diatomic molecules, the band origin ν_0 depends on a single ν' and ν'' and, using Eq. (1) for each electronic state, is given by Eq. (18). Equation (10)

$$\begin{aligned} \nu_0(\nu', \nu'') = & \nu_{el} + [\omega'_e(\nu' + 1/2) - \omega_e x'_e(\nu' + 1/2)^2 + \dots] \\ & - [\omega''_e(\nu'' + 1/2) - \omega_e x''_e(\nu'' + 1/2)^2 + \dots] \quad (18) \end{aligned}$$

also applies to electronic spectra but the band origin, $\nu_0(\nu', \nu'')$ is given by Eq. (18). The relative intensities of the bands depend on (1) the initial distribution of molecules among vibrational levels, and (2) the relative transition probabilities from any initial to various final vibrational levels.

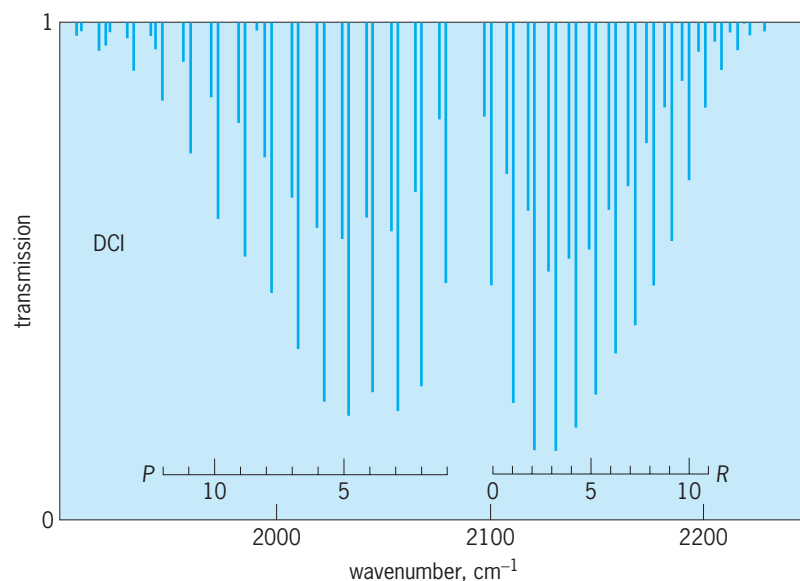


Fig. 9. Fundamental (1,0) vibration-rotation band of DCI in absorption showing the intensity distribution of the lines. The wavenumber scale increases to the right (in contrast to Fig. 8), so the R branch is to the right of the band origin, P branch to the left. The stronger lines are $D^{35}\text{Cl}$; the weaker companions, at lower wavenumbers, are $D^{37}\text{Cl}$. (After P. F. Bernath, *Spectra of Atoms and Molecules*, 2d ed., Oxford University Press, 2005)

The simplest example is the absorption spectrum of a cool gas of low molecular weight, for which all molecules initially have $v'' = 0$. The spectrum then consists of one " v' progression," a single series of bands with various values of v' ; the frequencies are given approximately by $\nu = \nu_0(0, 0) + (\omega'_e - \omega_e X'_e)v' - \omega_e X'_e v'^2$, in which $\nu_0(0, 0)$ is the frequency of the $(0, 0)$ band center. For a hot or a heavy gas, additional weaker v' progressions with $v'' > 0$ also appear.

In emission spectra, the initial population usually ranges over a number of v' values, from each of which transitions occur to a number of v'' values, so that the system contains many bands on both sides of $\nu_0(0, 0)$. In the special case of fluorescence spectra, the molecule is excited to various v' values by absorbing light; it then emits light belonging to the same (or sometimes another) electronic transition. From each v' , it can descend not only to the original v'' but also to various other, mainly larger, values. Hence fluorescence bands lie mainly at lower frequencies than the absorption bands used to excite them. See FLUORESCENCE.

Relative transition probabilities are governed by the Franck-Condon principle. This takes note of the very great rapidity of electronic motions as compared with those of the far more massive nuclei, and concludes that during the extremely brief time for an electronic transition, the nuclei tend to remain unchanged in their positions and momenta. It is applicable to both polyatomic and diatomic spectra. Consider first a diatomic molecule starting from the $v'' = 0$ level of a ground state $U(R)$ curve like the lower curves in Fig. 10. A vertical line drawn from the bottom point A ($v'' = 0$ if zero-point vibration is neglected) to point B on any one of the upper curves corresponds to an electronic absorption transition in which the nuclei have not moved. See FRANCK-CONDON PRINCIPLE.

In the case of Fig. 10a, point B corresponds to

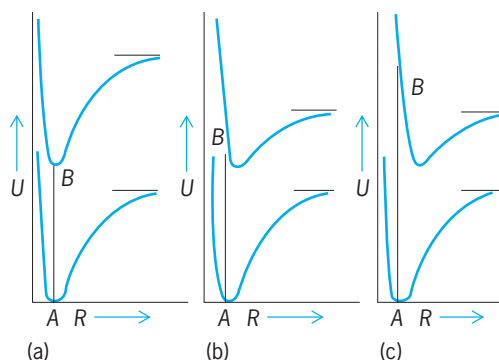


Fig. 10. Diatomic $U(R)$ curves for three cases to explain the vibrational intensity distribution according to the Franck-Condon principle. The asymptote of each curve for large R corresponds to dissociation into atoms, with one or both atoms excited in the case of the upper curves. Starting in each case from the bottom of the lower curve (essentially $v'' = 0$), the most probable transition in absorption is (a) to $v' = 0$, (b) to $v' = 3$ or 4, and (c) to the dissociation continuum, as shown by vertical lines. (After G. Herzberg, *Molecular Spectra and Molecular Structure*, vol. 1, 2d ed., Van Nostrand, 1950)

$v' = 0$, and the conclusion is that this is the most probable v' for $v'' = 0$. In the case of Fig. 10b, point B corresponds to an excited molecule at the inner turning point of a vibration with a v' of possibly 3 or 4, in a typical case. One then concludes (with J. Franck) that the strongest absorption bands for $v'' = 0$ have $v' = 3$ and 4. To obtain more exact information, a quantum-mechanical calculation (first carried out by E. U. Condon) is necessary.

In the case of Fig. 10c, point B corresponds to an energy level in the dissociation continuum above the asymptote of the upper $U(R)$ curve. According to the Franck-Condon principle, the absorption spectrum will have maximum intensity in a continuous range of frequencies, with $h\nu$ about equal to the energy difference AB . The quantum-mechanical calculation shows that the actual spectrum will extend with appreciable intensity over a range of both higher and lower frequencies than this, including, on the lower-frequency side, a number of high- v'' bands. Actual examples of such spectra (a long v'' progression followed by a strong continuum) are the far-ultraviolet Schumann-Runge bands of oxygen and the visible bands of iodine. By measuring the frequency at which the continuum begins, one obtains an exact value of the dissociation energy of these molecules. In so doing, any excitation energy of the atomic dissociation products to which the upper $U(R)$ curve leads is subtracted.

The Franck-Condon method is useful in understanding intensity distributions and structure in emission as well as absorption band systems. For diatomic spectra, various patterns of intensity as functions of v'' and v' occur, depending largely on the R_e values of the two $U(R)$ curves and, of course, also on the initial distribution among v'' levels. Sometimes the upper-state $U(R)$ curve is stable (has a minimum) but the lower state is repulsive. Continuous emission spectra then occur, with the atoms flying apart on reaching the lower state. The H_2 molecule shows such a spectrum, as do rare gas molecules such as He_2 and Kr_2 , which are stable only in excited or ionized states.

Molecular electronic states. Before discussing the structures of electronic bands, one must consider the nature of molecular electronic states. Each electronic state has orbital and spin characteristics. The spin quantum number S has a whole-number value if the number of electrons is even, a half-integral value if it is odd. Electronic states with $S = 0$ are called singlet states, all others multiplet states. The orbital characteristics differ sharply for linear (including diatomic) and nonlinear molecules.

For linear molecules only, there is a quantum number Λ such that $\pm \Lambda h/2\pi$ is the component of angular momentum around the line of nuclear centers. Linear-molecule electronic states can be discussed under three headings: (1) singlet states; (2) multiplet states with strong spin coupling (Hund's case *a*); and (3) multiplet states with weak spin coupling (Hund's case *b*). Strictly speaking, actual multiplet states are intermediate between cases *a* and *b*, or between these and certain other cases called *c* and

d. The discussion to follow is largely restricted to singlet electronic states.

Singlet states with $\Lambda = 0$ include $^1\Sigma^+$ and $^1\Sigma^-$ states: states with $\Lambda = 1, 2, \dots$ are called $^1\Pi, ^1\Delta,$ and so on. In linear molecules with a center of symmetry (H_2, CO_2 and so on), one must further distinguish even and odd (g and u) states: $^1\Sigma_g^+, ^1\Sigma_u^+, ^1\Sigma_g^-, ^1\Sigma_u^-$, $^1\Pi_u, ^1\Pi_g, ^1\Delta_g, ^1\Delta_u$, etc. The rotational levels of singlet states obey formula (19). Here J is restricted to integral values equal to or greater than Λ .

$$E_r/b = B_v[J(J+1) - \Lambda^2] \dots \quad (19)$$

For $\Lambda > 0$, each rotational level is a narrow doublet (Λ -doubling). Corresponding fine structure [see E_{fs} in Eq. (3)] can usually be detected in electronic bands, but (for ground states only) it can be much more accurately studied in low-frequency spectra. Hyperfine structure [see E_{hfs} in Eq. (3)] is usually on too small a scale to be detected in electronic band lines, but has been found in a few cases. Hyperfine structure is best studied in low-frequency spectra in the microwave region.

Electronic band structure. The simplest electronic bands occur for transitions between singlet electronic states. The possible types of electronic transitions are limited by the selection rule $\Delta\Lambda = 0, \pm 1$. The structure of $^1\Sigma - ^1\Sigma$ bands is essentially the same as for the $^1\Sigma$ -state vibration-rotation bands described earlier. Equations (12) to (15) and Fig. 7, also Fig. 8, for the intensities in absorption are still applicable if Eq. (18) instead of Eq. (11) is used for the band origin ν_0 , and it is recognized that B' and B'' now belong to two different electronic states.

The quantity $B' - B''$ in Eq. (14), instead of always being a small negative quantity, may now be either positive or negative, and $(B' - B'')/(B' + B'')$ is often fairly large (although it can also be nearly zero). As a result, it is usual in electronic bands to find so-called heads. A head is a position of maximum or minimum frequency; by using Eq. (14) to obtain $dv/dm = 0$, one finds $m_{\text{head}} = (B' + B'')/[2(B' - B'')]$. Then, on inserting m_{head} into Eq. (14), one obtains $\nu_{\text{head}} = \nu_0 - (B' + B'')^2/[4(B' - B'')]$. [Since $(B' + B'')/[2(B' - B'')]$ is not usually a whole number, the actual m_{head} is the nearest whole number m to that calculated.] According to whether $B' - B''$ is negative or positive, the positive (R) or the negative (P) branch forms the head. Figure 7, if continued to somewhat larger m values, illustrates the formation of an R -branch head at a calculated m of 10.5; the actual head is formed by the two coincident lines $m = 10$ and 11.

Although homopolar molecules (H_2, N_2 , and so on) have no pure rotation or vibration-rotation spectra, they do have electronic spectra. For homonuclear homopolar molecules, the band lines show alternating intensities. The lines in each branch are alternately stronger and weaker as m increases, this effect being superposed on the otherwise smoothly varying intensity distribution. The alternation ratio depends on the nuclear spin I and has been, in several cases, the means of determining I . When $I = 0$, alternate lines are completely missing. Heteronu-

clear molecules, even if homopolar (for example, HD or $^{16}\text{O}^{18}\text{O}$) do not show alternating intensities.

Polyatomic electronic spectra. These differ from diatomic electronic spectra because several initial and final vibration quantum numbers are involved, and because the rotational structure (except for linear molecules) is usually much more complicated. However, the detailed structure of the electronic spectra of a number of simple molecules and radicals in the vapor state in emission and in absorption has been studied. Nevertheless, for the most part, the spectra of polyatomic molecules are examined as absorption spectra in solution. The rotational structure is then completely absent, but the vibrational structure can be seen.

The Franck-Condon principle is here a useful guide. One of its corollaries is that only totally symmetric vibrations (vibrations during which the equilibrium symmetry of the molecule is preserved) undergo quantum number changes. This greatly simplifies the vibrational structure, especially of absorption spectra where most molecules are initially mainly in the $\nu'' = 0$ state of all vibrations. One finds then mostly ν'' progressions of one or a very few totally symmetric vibrations, and combinations of these.

Rather often, polyatomic band systems do not even show obvious vibrational structure. This can happen for any of several reasons: The upper state may involve dissociation; in CH_3I , for example, the first ultraviolet absorption region yields $\text{CH}_3 + \text{I}$; there may be so many low-frequency, upper-state vibrations that the spectrum looks continuous; or there may be a combination of these and other reasons. Such continuous or pseudocontinuous band systems are often loosely referred to as bands. For complicated molecules, the spectra of several different electronic transitions often overlap strongly so that it is difficult even to separate one system from another. See ATOMIC STRUCTURE AND SPECTRA; ELECTRON SPIN; INTERMOLECULAR FORCES; MOLECULAR WEIGHT; NEUTRON SPECTROMETRY; RAMAN EFFECT; RESONANCE (MOLECULAR STRUCTURE); SCATTERING EXPERIMENTS (ATOMS AND MOLECULES); VALENCE.

Robert S. Mulliken; U. Fano; Peter Bernath

High-resolution spectra. Molecular spectra exhibit varying amounts of detail, depending upon the inherent structure of each band and upon the ability of the recording instrument to distinguish transitions of nearly equal wavelength. This ability, called the resolving power, has increased dramatically since about 1970. Dispersive spectrometers have been superseded by Fourier-transform interferometers and tunable lasers, improving resolution as much as a millionfold. The resulting insight into the details of molecular structure and dynamics has revolutionized the field. See INFRARED SPECTROSCOPY; LASER SPECTROSCOPY.

Figure 11 shows a thoroughly studied infrared vibration-rotation absorption band of sulfur hexafluoride (SF_6) at four stages of resolution. This transition is from the vibrational ground state ($\nu'' = 0$) to an excited state containing one quantum of the

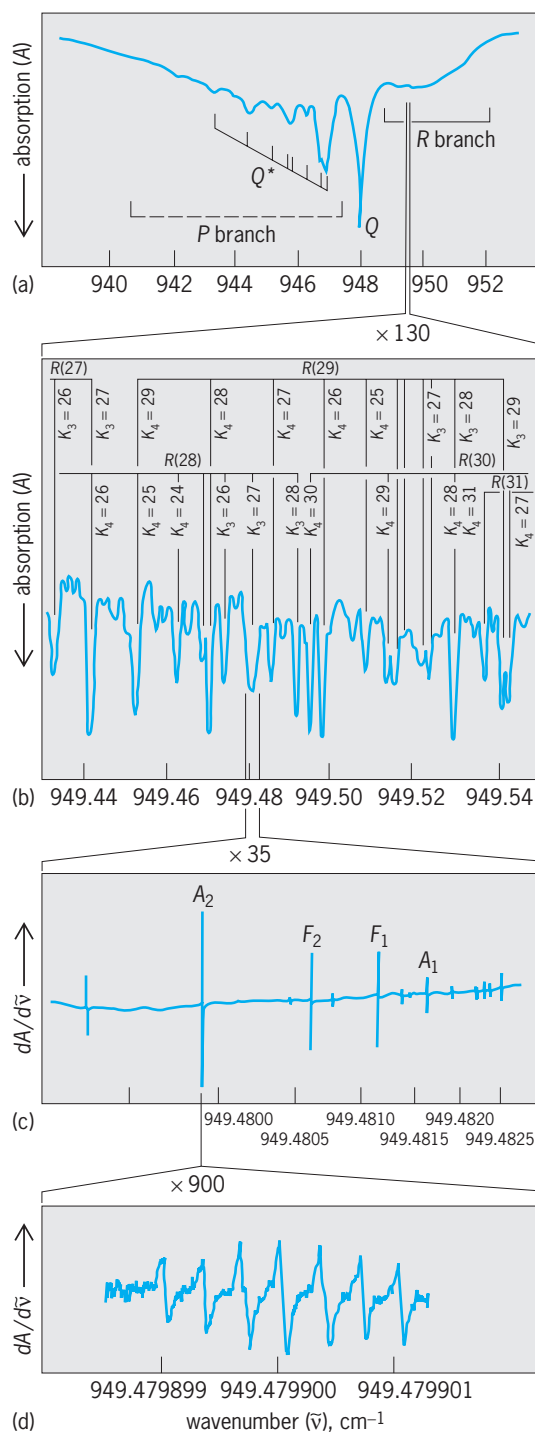


Fig. 11. Assignment of individual transitions in the spectrum of the S-F stretching fundamental ν_3 of sulfur hexafluoride (SF_6) as resolving power increases. (a) Fourier transform spectrum; resolution of 0.06 cm^{-1} [2 GHz]. (b) Spectrum recorded with tunable semiconductor diode laser; effective resolution limited by the SF_6 Doppler linewidth of 0.0010 cm^{-1} [30 MHz]. (Parts a and b after R. S. McDowell et al., *Identification of the SF_6 transitions pumped by a CO_2 laser*, *Opt. Commun.*, 17:178–183, 1976) (c) Saturation spectrum (derivative trace) recorded with a free-running waveguide carbon dioxide (CO_2) laser at a resolution of about 10^{-6} cm^{-1} [30 kHz]. (d) Saturation spectrum from a frequency-controlled CO_2 laser spectrometer with resolution of about $3 \times 10^{-8} \text{ cm}^{-1}$ [1 kHz]. (Parts c and d after J. Bordé and C. J. Bordé, *Superfine and hyperfine structures in the ν_3 band of $^{32}\text{SF}_6$* , *Chem. Phys.*, 71:417–441, 1982)

infrared-active stretching fundamental ν_3 ($\nu'' = \nu_3 = 1$). In Fig. 11a the unresolved band envelope appears with P, Q, and R branches, corresponding to series of transitions having $\Delta J = -1, 0$, and $+1$, respectively. The P branch is partially obscured by hot-band Q branches (labeled Q^*), prominent here because at room temperature 70% of the sulfur hexafluoride molecules are in excited vibrational states ($\nu'' > 0$) even without any optical excitation.

Higher resolution (Fig. 11b) reveals rotational fine structure in a portion of the R branch. In $R(28)$, for example, the rotational quantum number changes from $J'' = 28$ to $J' = 29$ during the vibrational transition. Centrifugal forces will distort a rotating sulfur hexafluoride molecule from its equilibrium octahedral configuration to a lower symmetry, and so rotations about different molecular axes will have slightly different energies. The states resulting from this dynamic symmetry breaking are identified in Fig. 11b by semiclassical quantum numbers K_3 and K_4 that specify the angular momentum about the axes with threefold or fourfold symmetry. These $R(J'')$ manifolds are wider than their mutual separation, and their overlap produces the dense structure observed. Weaker unidentified absorptions arise from hot bands.

Saturation spectroscopy can probe below the Doppler limit to reveal further details. Since the angular momentum can be oriented in either direction along any of the four threefold or three fourfold symmetry axes of sulfur hexafluoride each K_3 or K_4 cluster has eightfold or sixfold degeneracy, respectively, which can be partly removed by quantum tunneling. Figure 11c shows this superfine splitting of the $R(28)$ ($K_3 = 27$) cluster into four transitions that are labeled according to the symmetry species of the octahedral rotational group. The $R(28)(K_3 = 27)$ cluster has the structure $A_2 + F_2 + F_1 + A_1$, and since the F species are triplets, this accounts for the required eight levels. These four species have statistical weights $10 + 6 + 6 + 2$ due to nuclear spin degeneracy, which determines their relative intensities.

Finally Fig. 11d shows the $R(28)(K_3 = 27)$ A_2 transition at the highest resolution, revealing magnetic hyperfine structure arising from the interaction of the nuclear spin with the magnetic field of the rotating molecule. The tenfold A_2 spin degeneracy is resolved into a septet and a nearly coincident overlapping triplet.

Many thousands of lines in this one band of sulfur hexafluoride have been assigned to specific transitions, yielding very accurate molecular constants. For example, the moment of inertia implies a S—F bond length of 156.0588 picometers in the vibrational ground state (which differs very slightly from the equilibrium value R_e). The richness and complexity of such spectra pose a rigorous challenge to theoretical analysis, but they reveal subtleties of molecular energy levels and structure that were inaccessible before the development of high-resolution techniques.

Robin S. McDowell

Lasers and synchrotron sources. Modern sources of radiation have changed the field of molecular

spectroscopy enormously. Ultrafast pulsed lasers can offer time resolutions of a few femtoseconds (10^{-15} s) and very high peak powers of the order of terawatts (10^{12} W). Laser linewidths of a few hundred hertz for continuous-wave semiconductor lasers operating at 445 terahertz (674 nm) are readily achieved so that the spectral resolution is limited by Doppler or pressure broadening of the molecular lines of the gas sample. Special laser techniques such as saturation spectroscopy (as described above for SF_6) are required to take advantage of these exceedingly narrow laser linewidths. See LASER.

Synchrotron radiation also has many advantages for molecular spectroscopy. This radiation is emitted by a relativistic electron beam (that is, the electrons are traveling at close to the speed of light) that is accelerated in an electron storage ring. The required acceleration can be achieved with dipole magnets (bending magnets), which change the direction of motion of the electron beam, or with special "insertion" devices, which change the direction of motion of the electron beam many times in a short distance. An insertion device (wiggler or undulator) uses a periodic array of magnets (electromagnets or more commonly permanent magnets), with the direction of magnetic field alternating from magnet to magnet (say, field up and field down) to deflect the electrons back and forth through small angles as they travel through the magnetic array. The effect is to produce radiation that is a sum (a coherent sum in the case of an undulator, and an incoherent sum in the case of a wiggler) of the radiation from each deflection, resulting in high brightness (for an undulator) and high flux (for a wiggler). See SYNCHROTRON RADIATION.

Synchrotron radiation has long been used at short wavelengths, for example, in the photoionization of molecules and surfaces and in x-ray crystallography. More recently synchrotron radiation has been used also for infrared and far-infrared spectroscopy. Synchrotron radiation is typically used to replace the standard sources (the glowbar in the infrared and the mercury arc lamp in the far infrared) used for absorption spectroscopy with a Fourier transform spectrometer. The far-infrared radiation emitted from the edge of a bending magnet that turns the electron beam is many orders of magnitude brighter than that from a traditional thermal or arc source. In this context, brightness means power (watts) emitted per unit frequency interval, per unit emitting area, and per unit solid angle. The high brightness of the radiation can be used to improve the signal-to-noise ratio or the sensitivity of a typical infrared absorption measurement. Alternatively the high brightness can be used for very fast time resolution or for imaging.

In part the high brightness is due to the very small cross-sectional area (less than a millimeter) of the electron beam emitting the radiation. Synchrotron radiation (like a laser) can therefore be focused to a very small diffraction-limited spot (similar in size to the wavelength of the radiation) without excessive loss of radiation. This small spot can be used

for hyperspectral imaging of biological or heterogeneous industrial samples. Hyperspectral imaging means that each pixel of a two-dimensional slice of a sample has an associated spectrum. Hyperspectral imaging creates a three-dimensional data cube. Two of the dimensions are the intensity values at the x and y coordinates of the pixels of the image of the sample. The z direction is then the spectral axis, and each pixel has a spectrum along this third dimension. A typical sample might be a slice of biological tissue and the measurements made with a microscope. The sample is moved in a raster pattern and a spectrum recorded at each position to generate the hyperspectral image. Lasers or synchrotron sources are not required for hyperspectral imaging, but the time required to generate a data cube is generally prohibitively long with conventional light sources.

Peter Bernath

Bibliography. C. N. Banwell and E. M. McCash, *Fundamentals of Molecular Spectroscopy*, 4th ed., 1994; P. F. Bernath, *Spectra of Atoms and Molecules*, 2d ed., 2005; W. Demtröder, *Laser Spectroscopy*, 3d ed., 2003; J. D. Graybeal, *Molecular Spectroscopy*, 1993; W. G. Richards and P. R. Scott, *Structure and Spectra of Molecules*, 1985; W. S. Struve, *Fundamentals of Molecular Spectroscopy*, 1989; S. Svanberg, *Atomic and Molecular Spectroscopy: Basic Aspects and Practical Applications*, 4th ed., 2004; A. P. Thorne, U. Litzén, and S. Johansson, *Spectrophysics: Principles and Applications*, 1999.

Molecular weight

The sum of the atomic weights of all atoms making up a molecule. Actually, what is meant by molecular weight is molecular mass. The use of this expression is historical, however, and will be maintained. The atomic weight is the mass, in atomic mass units, of an atom. It is approximately equal to the total number of nucleons, protons and neutrons, composing the nucleus. Since 1961 the official definition of the atomic mass unit (amu) has been that it is $1/12$ the mass of the carbon-12 isotope, which is assigned the value 12.000 exactly. See ATOMIC MASS; ATOMIC MASS UNIT; RELATIVE ATOMIC MASS; RELATIVE MOLECULAR MASS.

The microscopic atomic weight is connected to the fundamental macroscopic mass unit, the kilogram, through the definition of the mole (mol). The kilogram is defined as the mass of a platinum-iridium bar kept at the International Bureau of Weights and Measures in Paris, France. A mole is an amount of substance containing Avogadro's number, N_A , approximately 6.022×10^{23} , of molecules or atoms. Molecule, in this definition, is understood to be the smallest unit making up the characteristic compound. Originally, the mole was interpreted as that number of particles whose total mass in grams was numerically equivalent to the atomic or molecular weight in atomic mass units, referred to as gram-atomic or gram-molecular weight. This is how the

above value for N_A was calculated. As the ability to make measurements of the absolute masses of single atoms and molecules has improved, however, modern metrology is tending to alter its approach and define Avogadro's number as an exact quantity, thereby changing slightly the definition of the atomic mass unit and removing the need to define atomic weight with respect to a particular isotopic species. The latest and most accurate value for Avogadro's number is $6.0221415(10) \times 10^{23} \text{ mol}^{-1}$. See AVOGADRO'S NUMBER; MOLE (CHEMISTRY).

Mass spectrometry. The best values for microscopic atomic and molecular masses are derived from measurements utilizing mass spectrometry, an experimental technique employed both for mass detection and for absolute mass determinations based on deflection of charged particles in electric or magnetic fields. Of the two methods, magnetic deflection yields the most accurate results and is the one most often used. A magnetic deflection apparatus which is used for mass determination is called a mass spectrometer. While particular types may vary as to the details of construction, the principle of operation for all is based on the fact that a charged particle moving in a magnetic field will follow a curved path, the curvature of which depends on the mass and charge of the species.

For a spectrometer employing deflection in a constant magnetic field with induction \vec{B} , the curvature can be easily calculated. Let \vec{v}_0 be the initial velocity with which the particle enters the magnetic field region, m the mass of the particle, and q the charge. The magnetic force does not alter the magnitude of the particle's velocity; the sole action is to cause the particle to execute a circular motion in the field. If \vec{v}_0 is perpendicular to \vec{B} , so that \vec{v}_0 is equal to \vec{v}_\perp , the component of the velocity perpendicular to \vec{B} , the plane of the orbit will be perpendicular to the region in which \vec{B} is constant. The radius of curvature of the circle r is then given by Eq. (1). Thus, the particle's mass is given by Eq. (2).

$$r = \frac{mv_\perp}{qB} \quad (1)$$

$$m = \frac{qrB}{v_\perp} \quad (2)$$

Mass spectroscopic measurements are made by first ionizing the atom or small molecule by using bombardment with high-energy electrons or absorption of ultraviolet light, then accelerating it across an electric potential difference V to produce a initial velocity given by Eq. (3), and finally allowing it to

$$v = \left(\frac{2qV}{m} \right)^{1/2} \quad (3)$$

be deflected in the magnetic field and strike a detector. With q known, m is extracted by measuring the radius of curvature r of the circle of motion, at least in principle. In practice, r and \vec{B} will be equal to the charge on the electron, as most ionization will produce only the singly ionized species. As the mass

of the electron is so much smaller than the mass of the proton, approximately 1/1836, the loss of one electron will not appreciably alter the mass of the atom or molecule.

Mass spectrometry has made possible determinations of absolute masses of virtually all the elements as well as many small molecules. The drawback to this technique is that the species to be measured must be in the vapor phase in order to execute motion in the fields. Its primary advantage is that it, taken together with Avogadro's number, puts the atomic mass scale on an absolute level. See MASS SPECTROMETRY; MASS SPECTROSCOPE.

Importance of determination. As the masses of all the atomic species are now known so well, masses of molecules can be determined once the composition of the molecule has been ascertained. Alternatively, if the molecular weight of the molecule is known and enough additional information about composition is available, such as the basic atomic constituents, it is possible to begin to assemble structural information about the molecule. Thus, the determination of the molecular weight is one of the first steps in the analysis of an unknown species. Given the increasing emphasis on the study of biologically important molecules, particular attention has been focused on the determination of molecular weights of larger and larger units. There are a number of methods available, and the particular one chosen will depend on the size and physical state of the molecule. All processes are physical macroscopic measurements and determine the molecular weight directly. Connection to the absolute mass scale is straightforward by using Avogadro's number, although, for extremely large molecules, this connection is often unnecessary or impossible, as the accuracy of the measurements is not that good. The main function of molecular weight determination of large molecules is elucidation of structure.

Weighing of gases. Although molecular weight determination for gases has been largely taken over by the mass spectroscopic analysis, it is still possible to determine molecular weights by direct weighing of gases. This method has some historical value in that much of present-day knowledge about chemical reactions was originally derived on the bases of observations concerning gases. For any real gas, if the pressure is low enough, the relationship between the pressure P , volume V , temperature T , and number of moles n of that gas can be expressed through the ideal gas law in the form of Eq. (4). Here R is the uni-

$$R = \lim_{P \rightarrow 0} \frac{PV}{nT} \quad (4)$$

versal gas constant, equal to 8.31441 J/mol K. The number of moles of the gas is related to the molecular weight by Avogadro's number, as discussed above. Thus, by measuring the pressure and weight of a known volume of gas, it is possible to derive the molecular weight M of the gas from Eq. (5). Here

$$M = RT \lim_{P \rightarrow 0} \frac{\rho}{P} \quad (5)$$

ρ is the mass density, equal to the total mass of the gas as measured in the experiment divided by the volume V . See GAS.

Colligative properties of solutions. Molecular weight determination of materials which are solid or liquid at room temperature is best achieved by taking advantage of one of the colligative properties of solutions, boiling-point elevation, freezing-point lowering, or osmotic pressure, which depend on the number of particles in solution, not on the nature of the particle. The choice of which to use will depend on a number of properties of the substance, the most important of which will be the size. All require that the molecule be small enough to dissolve in the solution but large enough not to participate in the phase change or pass through a semipermeable membrane. Freezing-point lowering is an excellent method for determining molecular weights of smaller organic molecules, and osmometry, as the osmotic pressure determination is called, for determining molecular weights of larger organic molecules, particularly polymeric species. Boiling-point elevation is used less frequently. See POLYMER.

The basis of all the methods involving colligative properties of solutions is that the chemical potentials of all phases must be the same. (Chemical potential is the partial change in energy of a system as matter is transferred into or out of it. For two systems in contact at equilibrium, the chemical potentials for each must be equal.) See CHEMICAL EQUILIBRIUM; CHEMICAL THERMODYNAMICS.

Phase change methods. For the two methods involving a phase change, analysis requires that it be possible to form a solution of the molecule in some known solvent which is dilute enough so that the solution may be considered ideal but concentrated enough that a measurable raising of the boiling point or lowering of the freezing point will be produced. It must also be possible to make the assumption that there are no solute particles in either the vapor or the solid phase. The balance of chemical potentials then demands that the solvent remain in solution rather than freeze out or vaporize. Under these restricted conditions of a dilute, ideal solution, the extent ΔT of freezing-point lowering or boiling-point elevation is related to the concentration of solute, and alternatively its molecular weight M through Eq. (6). Here R is the universal gas constant, T_0

$$\Delta T = \frac{RT_0^2 M}{1000 \Delta H} c \quad (6)$$

the temperature at which freezing or boiling would occur in the pure solvent, ΔH the enthalpy of fusion or vaporization of the solvent (the amount of energy which must be supplied or removed at constant pressure to cause 1 mole to change phase), and c the molal concentration of the solute. (1 molal equals 1 mole of solute per 1 kg of solvent.) Because of the restrictions on the solutions which must be formed and for which the above expression is valid, combined with the accuracy with which temperature differ-

ences can be measured, these two methods are best employed for molecules with molecular weights less than 10,000.

Osmometry. The basis for the application of the osmotic pressure method (osmometry) is that the chemical potentials of a solution and a pure solvent separated from each other by a semipermeable membrane must be the same. The membrane must be coarse enough to allow solvent molecules to pass through, but fine enough to prohibit passage of solute. In an attempt to equalize the chemical potentials on either side of the membrane, solvent will flow from the side containing only solvent to the solution. The resulting change in chemical potential causes an increase in pressure on the side containing the solution, which will continue until the pressure produced prohibits further flow of solvent, that is, the chemical potentials are equal. This pressure Π is a real physical pressure which can be measured and is related to the concentration in the solution through the van't Hoff equation (7). Here R is again

$$\Pi = RTc \quad (7)$$

the universal gas constant, T the temperature, and c the concentration in number of moles per given amount of solvent. Because of restrictions on passage through the membrane, osmometry can be used to determine molecular weights from about 1000 to about 30,000. For larger molecules other methods are necessary. See OSMOSIS; SOLUTION.

Sedimentation. Of all techniques available for determination of molecular weights of large molecules, sedimentation, or the settling of large molecules out of solution under the action of an external force, is perhaps the most common. This process may take a number of varied forms, but its successful utilization for weight determination depends on the fact that the solute molecules in a solution are never at rest but are free to move among the solvent molecules. The practical application of sedimentation involves producing a directed motion of these particles. This is done in a high-speed centrifuge. A particle of mass m when rotated in a circle experiences a force which is equal to $m\omega^2 r$, where ω is the angular velocity and r the distance of the particle from the center of rotation. In the absence of any other forces, the particle would tend to move away from the force center and to the edge of the container, where it would be collected as soon as it can move no further. However, as the molecules move away from the center of the rotation, a concentration gradient is set up. This causes a diffusion of the particles in the opposite direction to the gradient, and a tendency for the particles to remain distributed throughout the solvent. In addition, there is also a buoyant force on the particles produced by collisions with solvent molecules which tends to maintain the particles in solution. At equilibrium the external centrifugal force driving the sedimentation must balance the total forces produced by the buoyant effect and the concentration gradient. If V is the volume of the solute molecule, ρ_M its density, and ρ_0 the density of the solvent, this

balancing of forces leads to Eq. (8). Here D is the

$$N_A V \rho_M (1 - \rho_M^{-1} \rho_0) \omega^2 r = \frac{RT \dot{r}}{D} \quad (8)$$

diffusion coefficient, N_A Avogadro's number, R the universal gas constant, and T the temperature. The molecular weight M is related to the mass density by Eq. (9). At equilibrium the rate of sedimentation

$$M = N_A V \rho_M \quad (9)$$

(rate of deposit of particles across an area in the direction of the rotation axis) will equal the rate of diffusion across that same boundary. Combining this requirement with the force balance yields Eq. (10)

$$M = \frac{2RT}{(1 - \rho_M^{-1} \rho_0) \omega^2} \frac{\ln(c_2/c_1)}{r_2^2 - r_1^2} \quad (10)$$

for the molecular weight. Successful application of this method requires determination of the concentrations c_1 and c_2 at two values of the radius r_1 and r_2 . This is carried out by optical refractometry. Light when passing from one medium to another will be refracted so as to follow a different path from the original one. If the second medium happens to be a solution, the angle of refraction will depend on the concentration of the solution. A modern centrifuge has containers for the solution which are constructed of quartz to provide for this measurement while the solution is being centrifuged. See REFRACTION OF WAVES.

A variant on the sedimentation equilibrium method described above involves dissolving the molecule in a concentrated salt solution and centrifuging the resulting mixture. The buoyant force produced by the dense salt solution will force the molecular species to be concentrated in that region in which its density is equal to that of the concentrated solution, $\rho_M = \rho_0$. Continuous motion of the molecules, however, prevents the concentration from being a perfect line, but some distribution about the central point r_0 occurs. From the width σ of this distribution about the center and the density gradient $\Delta\rho/\Delta r$ of the salt solution, the molecular weight can be extracted according to Eq. (11).

$$M = \frac{2RT}{\rho_M^{-1} (\Delta\rho/\Delta r) \omega^2 r_0 \sigma^2} \quad (11)$$

See ULTRACENTRIFUGE.

Light scattering. Another measurement from which molecular weights can be obtained is based on the scattering of light from the molecule. A beam of light falling on a molecule will induce in the molecule a dipole moment which in its turn will radiate. The interference between the radiated beam and the incoming beam produces an angular dependence of the scattered radiation which depends on the molecular weight of the molecule. This occurs whether the molecule is free or in solution. While the theory for this effect is complicated and varies according to the size of the molecule, the general result for molecules whose size is considerably less

than that of the wavelength λ of the radiation (less than $\lambda/50$) is given by Eq. (12). Here $I(\theta)$ is the

$$\frac{I(\theta)}{I_0} = \text{constant} (1 + \cos^2 \theta) M c \quad (12)$$

intensity of radiation at angle θ , I_0 the intensity of the incoming beam, M the molecular weight, and c the concentration in grams per cubic centimeter of the molecule. If the molecules are much larger than $\lambda/50$ (about 9 nanometers for visible light), this relationship in this simple form is no longer valid, but the method is still viable with appropriate adjustments to the theory. In fact, it can be used in its extended version even for large aggregates. See SCATTERING OF ELECTROMAGNETIC RADIATION.

Sample purity. One requirement which all these methods demand is purity of the sample if the resulting weight is to be very accurate. Impurities result in a molecular weight which is not absolute for the species but is the average weight of the species plus the impurities. Depending on whether or not the influence of the impurities is due to their number (for low-molecular-weight impurities) or their weight (high-weight impurities), the average is either a number average M_n or a weight average M_w . The two average molecular weights are defined by Eqs. (13) and (14), where n_i is the number of particles of species i with molecular weight M_i .

$$M_n = \frac{\sum_i (n_i M_i)}{\sum_i n_i} \quad (13)$$

$$M_w = \frac{\sum_i (n_i M_i^2)}{\sum_i (n_i M_i)} \quad (14)$$

Measurement of auxiliary parameters. All the methods in use in biophysics require the knowledge of parameters other than the one actually being measured in the experiment. In most cases this will be the volume of the molecule, or ρ_M^{-1} . For these methods to give accurate results, this number must be obtained by other means. One of the best, but often one of the hardest to apply, is x-ray crystallography, the same technique used to determine Avogadro's number. See X-RAY CRYSTALLOGRAPHY.

Other determination methods. Other methods of molecular weight determination include end-group labeling, useful for proteins; viscosity, useful for relative molecular weight determination; gel electrophoresis, also useful for relative measurements; and electron microscopy, useful if the molecule is large enough to be well resolved in the microscope and again only for relative measurements. For a truly accurate molecular weight determination, a combination of these techniques may be necessary. See ELECTRON MICROSCOPE; ELECTROPHORESIS; VISCOSITY.

C. Denise Caldwell

Bibliography. P. W. Atkins and J. de Paula, *Atkins' Physical Chemistry*, 8th ed., 2006; A. R. Cooper (ed.), *Determination of Molecular Weight*, 1989; D. Freifelder, *Principles of Physical Chemistry with Applications to the Biological Sciences*, 2d ed., 1985; I. N. Levine, *Physical Chemistry*, 5th ed., 2001.

Molecule

The smallest unit of a compound, which consists of atoms bonded together in a unique arrangement. A diatomic molecule consists of two atoms linked together by chemical bonds. Examples include the oxygen and nitrogen of the air (O_2 and N_2 , respectively) and the neurotransmitter nitric oxide (NO). A polyatomic molecule consists of more than two linked atoms. Examples include water (H_2O), carbon dioxide (CO_2), ethanol (C_2H_5OH), and the vast molecule of deoxyribonucleic acid (DNA) with thousands of atoms.

The bonds that hold the atoms together are covalent. That is, they consist of pairs of electrons that are shared by the neighboring atoms. This type of bonding is in contrast to ionic bonding, in which there is complete transfer of one or more electrons from one atom to another and large numbers of the resulting ions (charged atoms) clump together, as in sodium chloride (NaCl). Some chemists refer to the smallest unit of an ionic compound (in the case of sodium chloride, a sodium ion, Na^+ , and a chloride ion, Cl^-) as constituting a “molecule” of the compound, but this usage is uncommon and the term “formula unit” is preferred. In this article, coverage is confined to covalently bonded molecules, of which there are millions of kinds. See CHEMICAL BONDING; IONIC CRYSTALS.

Organic and inorganic molecules. Many molecules are organic. That is, they contain at least one atom and commonly many atoms of carbon. Molecules that do not contain carbon, with a few exceptions, are inorganic. Thus ethanol consists of organic molecules, and water consists of inorganic molecules. The principal exception to this general classification is carbon dioxide, which is treated as an honorary inorganic compound. The term “organic” arose when it was thought erroneously that such compounds could be produced only by living organisms. Advances in chemistry over decades have blurred the distinction between organic and inorganic, and many compounds are now known which, though they contain carbon, are regarded as lying within the domain traditionally regarded as inorganic chemistry (the so-called organometallic compounds), and many organic compounds that are actually produced by organisms contain metal atoms (some enzymes and vitamins, such as vitamin B_{12} , which contains cobalt atoms). Nevertheless, the distinction between organic and inorganic is valuable as a broad-brush classification and is still widely used. This article concentrates on organic molecules. See ORGANOMETALLIC COMPOUND; VITAMIN B_{12} .

Multiplicity of bonds. The bonds between atoms may be single, double, or triple (in rare cases, quadruple), in which neighboring atoms share one, two, or three electron pairs. The OH bonds in H_2O are single (represented as $H-O-H$), the CO bonds in CO_2 are double (represented as $O=C=O$), and the NN bonds in N_2 are triple ($N\equiv N$). Broadly speaking, the greater the multiplicity of bonds, the more tightly

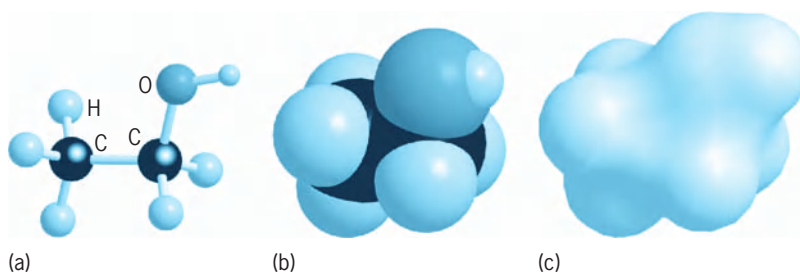


Fig. 1. Three representations of the ethanol molecule, C_2H_5OH : (a) ball-and-stick, (b) space-filling, (c) surface.

bonded the atoms. This tightness of bonding is manifest, for example, in the inertness of molecular nitrogen and its role as a diluent of the dangerously reactive molecular oxygen of the atmosphere. It also accounts for the difficulty of converting atmospheric nitrogen into ammonia and nitrate fertilizers for use in agriculture.

Shapes. Molecules have not only a characteristic atomic composition but also, because the atoms are linked together in a characteristic array, a characteristic shape. Thus, H_2O is an angular molecule in which the two OH bonds make an angle of about 104° to each other, and CO_2 is linear (all three atoms in a straight line). The four bonds that carbon commonly forms (as in methane, CH_4) are typically arranged in a tetrahedral array with bond angle close to 109° . The resulting shape of the ethanol molecule is seen in Fig. 1, which shows both the bonds as sticks joining balls representing the atoms (Fig. 1a), and a “space-filling” molecule (Fig. 1b), which gives a better notion of the overall bulk of the molecule by representing each atom by a sphere that corresponds to its actual size. Computer graphics are enormously useful for displaying molecular shapes, and can take a variety of forms. Figure 1c aims to display the general shape of the molecule without distinguishing the individual atoms.

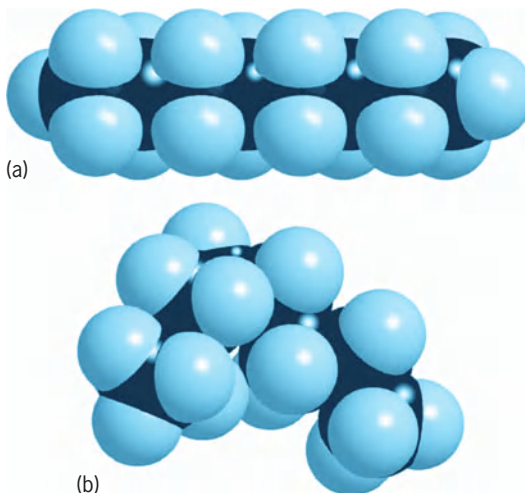


Fig. 2. Octane molecule (C_8H_{18}) in its (a) linear and (b) coiled conformations.

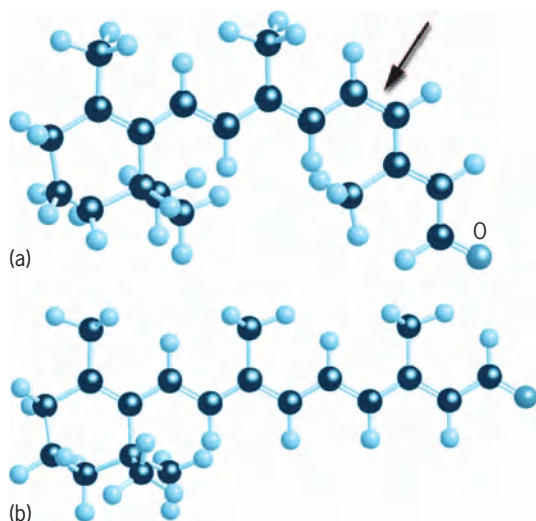


Fig. 3. Ball-and-stick models of (a) *cis*-retinal and (b) *trans*-retinal. The arrow indicates the double bond that is loosened by the absorption of light.

Flexibility and rigidity. The tetrahedral geometry of the bonds to carbon applies only when all the bonds are single. Moreover, neighboring groups of atoms are free to rotate around a single bond, which acts as a miniature axle. As a result, a polyatomic molecule consisting of many singly bonded atoms should not be thought of as a rigid framework, but as ceaselessly

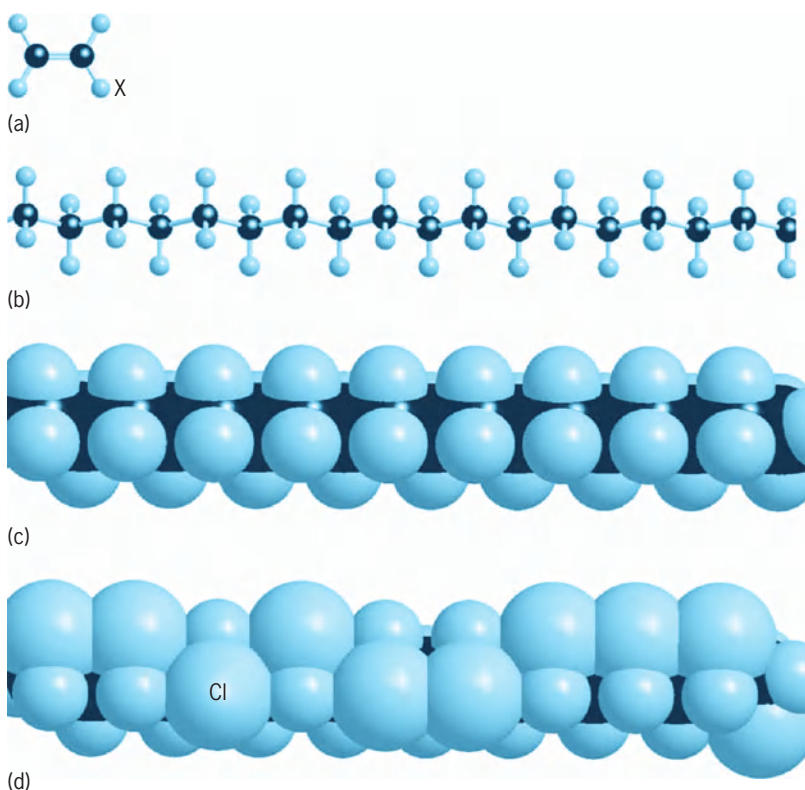


Fig. 4. Polymerization based on ethene (ethylene). (a) Ethene molecule; the atom marked X may be replaced by atoms of other elements. (b) Ball-and-stick and (c) space-filling models of a fragment of a polyethylene chain. (d) Space-filling model of a fragment of a polyvinyl chloride (PVC) chain.

writhing and twisting into different shapes or “conformations.” Figure 2 shows two of the conformations of the hydrocarbon molecule octane (C_8H_{18}), a component of gasoline. See CONFORMATIONAL ANALYSIS.

In contrast, the presence of a multiple bond confers structural rigidity because the neighboring groups of atoms joined by a double bond are not free to rotate relative to one another. This structural rigidity has a number of important consequences. For instance, beef fat consists of molecules (tristearin) in which three 18-carbon-atom-long chains are joined together. As all the bonds are single, these long chains are flexible and the entire molecule can roll up into a ball; these ball-like molecules can pack together closely, and tristearin is a solid. On the other hand, similar molecules produced in olives have a double bond in each of their chains. As a result, the chains are less flexible, the molecule cannot roll up into a ball, the molecules cannot pack together as closely, and the compound is an oil rather than a fat. See FAT AND OIL.

Role of double bonds between carbon atoms. Double bonds between carbon atoms play important roles in physiology and industry. The primary act of vision, for instance, can be traced to the properties of double bonds. The retinas of human eyes are populated by the molecule *cis*-retinal (Fig. 3a). As can be seen from the line structure, one feature of the molecule is the alternation of single and double bonds between neighboring carbon atoms. This alternation has two consequences: the pairs of electrons forming the bonds are quite loosely attracted to the atomic nuclei and can easily be moved around by incident light; and the molecule is rigid and adopts the shape shown in Fig. 3a. However, when light is incident on the molecule, one of the double bonds is weakened, the long side-chain becomes free to rotate, and the molecule snaps into the form known as *trans*-retinal (Fig. 3b). This molecule can no longer fit into the protein molecule that normally houses it, and it is expelled. The protein molecule responds by relaxing into a new shape, and as a result a signal is pulsed along the optic nerve into the brain, which interprets it as “vision.” Various biochemical reactions then take place to restore the molecule to its original shape, in which it is primed to act again. See PHOTORECEPTION.

Polymerization. The industrial importance of double bonds between carbon atoms lies in the reactivity they confer on organic molecules (in contrast to the inertness conferred on nitrogen). One aspect of this reactivity is the ability to combine small double-bonded organic molecules into long chains. That is, certain molecules act as monomers that may be polymerized to form the plastics characteristic of the modern world. The process is illustrated in Fig. 4, which shows a variety of polymers based on ethene (ethylene, $CH_2=CH_2$). The most primitive of these polymers, but far from being useless, is polyethylene itself, the polymer of ethylene, which consists of long chains of thousands of $-CH_2-CH_2-$ groups, often with branch

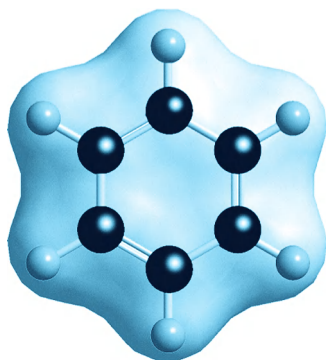


Fig. 5. Ball-and-stick model of benzene (C_6H_6) superimposed on a model of its surface.

lines where the polymerization has proceeded in a slightly different way. When the atom marked X in Fig. 4a is replaced by a chlorine atom to give the monomer $CHCl=CH_2$, the polymerization process results in chains of $-CHCl-CH_2-$ groups and the polymer polyvinyl chloride (PVC). When X is a benzene-like ring, the monomer is styrene and the resulting polymer is polystyrene. Chemists are able to modify the monomers in a wide variety of ways and to produce polymers with a correspondingly wide range of properties. Thus, if all the hydrogen atoms of the monomer are replaced by fluorine atoms, to give the monomer $CF_2=CF_2$, the resulting polymer is polytetrafluoroethylene (PTFE), used in nonstick coatings. See ETHYLENE; POLYFLUOROOLEFIN RESINS; POLYMER; POLYMERIZATION; POLYOLEFIN RESINS; POLYSTYRENE RESIN; POLYVINYL RESINS; STYRENE.

Aromatic compounds. The members of one very important class of molecules with double bonds have a greatly reduced reactivity (and thus are like organic counterparts of nitrogen). The parent of these so-called aromatic compounds is benzene, C_6H_6 . Like benzene itself, many of these compounds have an aroma, but many do not, and the term applies to any compound that contains a benzene-like group. A benzene molecule consists of a planar hexagonal array of carbon and hydrogen atoms (Fig. 5). Although the carbon atoms are joined by alternating single and double bonds, and so might be expected to be highly reactive, this arrangement actually confers stability (for quantum-mechanical reasons) and the molecule is not very reactive. As a result, the ring is a motif that is found in many organic compounds, including polystyrene. See AROMATIC HYDROCARBON; BENZENE.

This brief survey only skims the subject of molecules. Whatever we breathe, eat, drink, and wear, whatever medications we take, whatever cosmetics we use, almost whatever we touch and see, is made of molecules.

Peter Atkins

Bibliography. P. Atkins, *Atkins' Molecules*, Cambridge University Press, 2003; P. Atkins and L. Jones, *Chemical Principles*, W. H. Freeman, New York, 2007.

Mollusca

A major phylum of the animal kingdom comprising an extreme diversity of external body forms (oysters, clams, chitons, snails, slugs, squid, and octopuses among others), all based on a remarkably uniform basic plan of structure and function. The phylum name is derived from *mollis*, meaning soft, referring to the soft body within a hard calcareous shell, which is usually diagnostic. Soft-bodied mollusks make extensive use of ciliary and mucous mechanisms in feeding, locomotion, and reproduction. Most molluscan species are readily recognizable as such.

The Mollusca constitute a successful phylum; there are probably over 110,000 living species of mollusks, a number second only to that of the phylum Arthropoda, and more than double the number of vertebrate species. More than 99% of living molluscan species belong to two classes: Gastropoda (snails) and Bivalvia. Ecologically, these two classes can make up a dominant fraction of the animal biomass in many natural communities, both marine and fresh-water. Certain bivalve species are the most abundant marine benthic animals, and computations have suggested that one of these may encompass a larger "standing crop" biomass of animal tissue than any other single animal species on the planet.

Classification

The phylum Mollusca is divided into seven distinct extant classes, three of which (Gastropoda, Bivalvia, and Cephalopoda) are of major significance in terms both of species numbers and of ecological bioenergetics, and one extinct class. An outline of their classification follows.

- Class Monoplacophora (mainly fossil; but one living genus *Neopilina*)
- Class Aplacophora
 - Subclass Neomeniomorpha
 - Subclass Chaetodermomorpha
- Class Polyplacophora
- Class Scaphopoda
- Class Rostroconchia (fossil only)
- Class Gastropoda
 - Subclass Prosobranchia
 - Order: Archaeogastropoda
 - Mesogastropoda
 - Neogastropoda
 - Subclass Opisthobranchia
 - Order: Bullomorpha (or Cephalaspidea)
 - Aplysiomorpha (or Anaspidea)
 - Thecosomata
 - Gymnosomata
 - Pleurobranchomorpha (or Notaspidea)
 - Acochlidiaacea
 - Sacoglossa
 - Nudibranchia (or Acoela)
 - Subclass Pulmonata
 - Order: Systellommatophora
 - Basommatophora
 - Stylommatophora

- Class Bivalvia (or Pelecypoda)
 - Subclass Protobranchia
 - Subclass Lamellibranchia
 - Order: Taxodonta
 - Anisomyaria
 - Heterodonta
 - Schizodonta
 - Adapedonta
 - Anomalodesmata
 - Subclass Septibranchia
- Class Cephalopoda
 - Subclass Nautiloidea
 - Subclass Ammonoidea (fossil only)
 - Subclass Coleoidea
 - Order: Belemnoidea
 - Sepioidea
 - Teuthoidea
 - Vampyromorpha
 - Octopoda

The subclass Opisthobranchia is sometimes divided into two superorders: Tectibranchia (the first seven orders) and Nudibranchia. The planktonic opisthobranch orders Thecosomata and Gymnosomata are sometimes united in the order Pteropoda. There are several alternative ordinal arrangements for the many well-established superfamilies of lamellibranch bivalves. An attractive one on functional grounds (but unacceptable phyletically) is to divide the subclass into two: Filibranchia and Eulamellibranchia.

The use by paleontologists of subphyla, including *Cyrtosoma* (for Monoplacophora with Gastropoda and Cephalopoda) and *Diasoma* (for Rostroconchia with Bivalvia and Scaphopoda) in the classification of the phylum Mollusca has not been generally adopted by students of living forms (neontologists). There is no disagreement over the further division of the phylum into the eight clearly distinguishable classes listed above.

Functional Morphology

The unique basic plan of the Mollusca involves the different modes of growth and of functioning of the three distinct regions of the molluscan body (Fig. 1). These are the head-foot with some nerve concentrations, most of the sense organs, and all the locomotory organs; the visceral mass (or hump) containing organs of digestion, reproduction, and excretion; and the mantle (or pallium) hanging from the visceral mass and enfolding it and secreting the shell. In its development and growth, the head-foot shows a bilateral symmetry with an anterioposterior axis of growth. Over and around the visceral mass, however, the mantle-shell shows a biradial symmetry, and always grows by marginal increment around a dorsoventral axis (Fig. 1 and Fig. 2). It is of considerable functional importance that a space is left between the mantle-shell and the visceral mass forming a semi-internal cavity; this is the mantle cavity or pallial chamber within which the typical gills of the mollusk, the ctenidia, develop. This mantle cav-

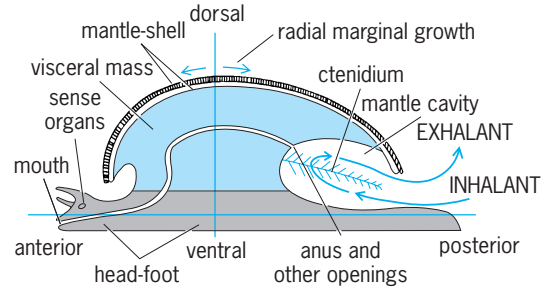
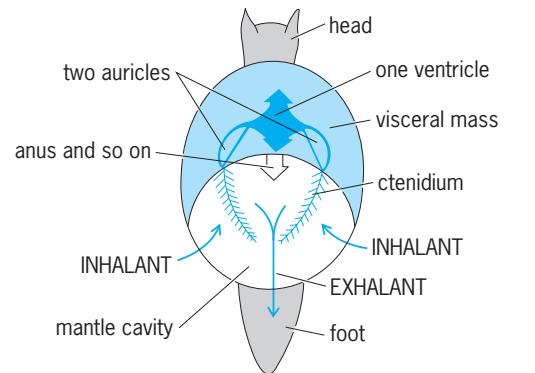


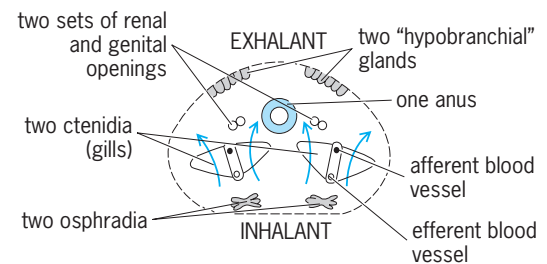
Fig. 1. Generalized model of a stem mollusk (or archetype) in side view. There are three distinct regions in the molluscan body: head-foot, visceral mass, and mantle-shell. Water circulation through the mantle cavity, gills (ctenidia), and pallial complex is from ventral inhalant to dorsal exhalant. (After W. D. Russell-Hunter, *A Life of Invertebrates*, Macmillan, 1979)

ity is almost diagnostic of the phylum; it is primarily a respiratory chamber housing the ctenidia, but with alimentary, excretory, and genital systems all discharging into it (Fig. 2). The basic functional plan is always recognizable, although it has undergone remarkable modifications of structure and of function in different groups of mollusks.

Mantle-shell and body form. In looking at any mollusk, it is important to realize that whatever the shape of the shell, it is always underlain by the mantle, a



(a)



(b)

Fig. 2. Mantle cavity and its associated structures in a molluscan archetype. (a) Dorsal view of the symmetrical posterior mantle cavity, with a pair of feather gills (aspidobranch condition) and two auricles in the heart (diotocardiac condition). (b) Cross section through the mantle cavity of the archetype showing all the elements of the pallial complex in a bilaterally symmetrical array. (After W. D. Russell-Hunter, *A Life of Invertebrates*, Macmillan, 1979)

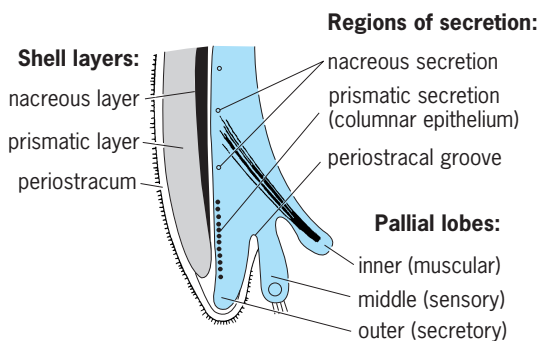


Fig. 3. Arrangement of the mantle and shell layers in the mollusks. The fleshy mantle margin is divided into three lobes, and the outer pallial lobe is largely responsible for secretion of the three layers of the shell. (After W. D. Russell-Hunter, *A Life of Invertebrates*, Macmillan, 1979)

fleshy fold of tissues which has secreted it. The detailed structure of the shell and of the mantle edge (with three functionally distinct lobes; **Fig. 3**) is also consistent throughout the Mollusca. The shell is made up of calcium carbonate crystals enclosed in a meshwork of tanned proteins. It is always in three layers: the outer periostracum, which is highly organic; the prismatic layer, which is the most massive part; and the innermost or nacreous layer, which has the least organic content (**Fig. 3**).

Each of the eight classes of the Mollusca has a characteristic body form and shell shape. Two classes are enormous (Gastropoda and Bivalvia), one of moderate extent (Cephalopoda), the others being minor by comparison. The Gastropoda constitute a diverse group with the shell usually in one piece. This shell may be coiled as in typical snails—that is, helicoid or turbanate—or it may form a flattened spiral, or a short cone as in the limpets, or it may be secondarily absent as in the slugs. Most gastropods are marine, but many are found in freshwaters and on land; in fact, they are the only successful nonmarine mollusks.

The Bivalvia are a more uniform group, with the shell in the form of two calcareous valves united by an elastic hinge ligament. Mussels, clams, and oysters are familiar bivalves. The group is mainly marine with a few genera in estuaries and in freshwaters. There can be no land bivalves since their basic functional organization is as filter feeders. The third major group, the Cephalopoda, includes the most active and most specialized mollusks. There is a chambered, coiled shell in *Nautilus* and in many fossil forms; this becomes an internal structure in cuttlefish and squids, and is usually entirely absent in octopods.

There are only about 40 genera of chitons, or “coat-of-mail shells” with eight plates, placed in the Polyplacophora. The other three living minor groups (class Rostroconchia is completely extinct) are marine, and there are only a few species in each. The Monoplacophora encompass several fossil families and the one living genus *Neopilina*, all with limpet-like shells. The Scaphopoda are the “elephant’s-tusk-

shells,” and the Aplacophora have a wormlike body whose mantle secretes discrete calcareous spicules but never a shell.

Ctenidia and pallial complex. In the more primitive living snails and bivalves, there are paired feather-like (aspidobranch) ctenidia (**Fig. 4**). Filaments or gill plates alternate on either side of an axis which contains both a dorsal afferent blood vessel (deoxygenated blood) from the body and a ventral efferent blood vessel (oxygenated blood) running into the heart. Water flow (created by the lateral cilia) is always from ventral to dorsal, in the direction opposite to blood flow (**Fig. 4**). This physiologically efficient counterflow system is functionally homologous throughout the mollusks (with few, adaptively explicable, exceptions), and is based on homologous structural arrangements of ciliated tracts, blood vessels, skeletal elements, and renopericardial connections. A diversity of gill patterns have evolved in the major molluscan groups (**Fig. 5**), paralleling the evolution of the mantle-shell patterns, but all clearly derived from the archetypic ctenidial arrangement (**Fig. 4**).

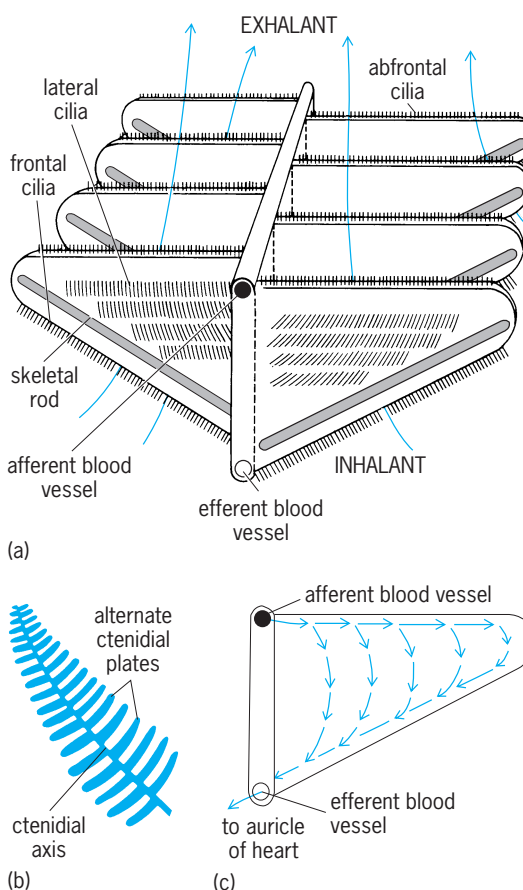


Fig. 4. Aspidobranch ctenidium (feather or plume gill) as found in primitive gastropods and other molluscan groups. (a) Stereogram showing the water current from ventral inhalant to dorsal exhalant between adjacent gill plates—a current created by the lateral cilia. (b) Entire gill viewed from above. (c) A single gill plate showing the counter-current flow of blood within the plate, which results in physiological efficiency of oxygen exchange. (After W. D. Russell-Hunter, *A Life of Invertebrates*, Macmillan, 1979)

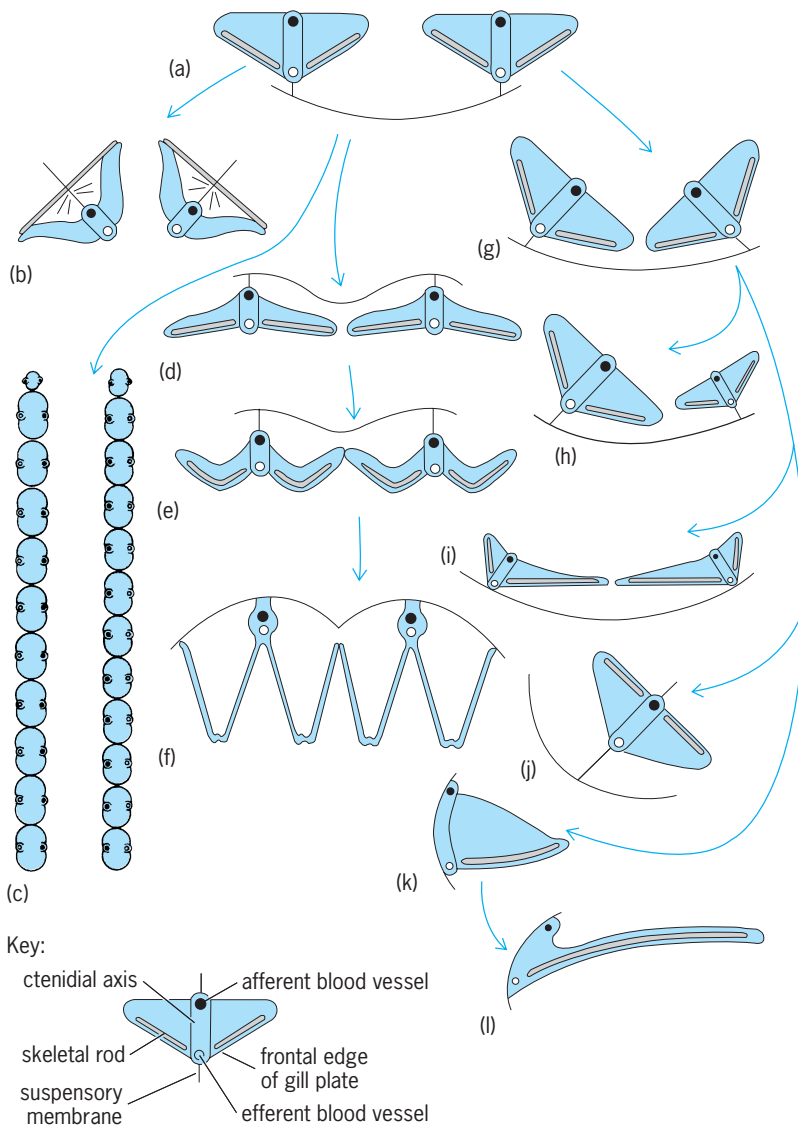


Fig. 5. Evolution of molluscan gills (ctenidia) in cross sections, viewed from the posterior. General (a) molluscan and (g) gastropod archetypes are illustrated, but the only other hypothetical pattern is (e) that shown as an intermediate bivalve condition. All the other gill patterns shown occur in presently living forms: (b) in modern cephalopods, (c) in chitons, (d) in protobranch and (f) lamellibranch bivalves, and (h-l) in a series of gastropods. (After W. D. Russell-Hunter, *A Life of Invertebrates*, Macmillan, 1979)

The more advanced gastropods show reduction from a pair of aspidobranch ctenidia to a single one, and from that to a one-sided pectinibranch ctenidium (or comb gill), and subsequently to no gill at all in the pulmonate snails. The bivalves show enlargement of gill leaflets to longer filaments and their subsequent folding into the true lamellibranch condition (Fig. 5), used in filter feeding. The gills in the cephalopods, while still structurally homologous, are modified with new skeletal elements to resist the stresses of water pumping by muscles (rather than by cilia). In the chitons, ctenidia are replicated, though not necessarily symmetrically (Fig. 5).

Besides the gills, the other organs of the mantle cavity (termed collectively the pallial complex; Fig. 2) again show morphological and functional consistency throughout the main groups of the Mol-

lusca. The ctenidia form a curtain functionally dividing the mantle cavity into an inhalant part (usually ventral) containing the osphradia (pallial sense organs which sample the incoming water), and an exhalant part (usually dorsal) containing hypobranchial glands and both the anus and the openings of the kidney and genital ducts (Fig. 2). Thus feces, gametes, and nitrogenous wastes are all discharged downstream from the gill curtain in the exhalant current. This arrangement of the pallial complex as a functionally integrated group of structures remains consistent in mollusks whether they have ctenidia asymmetrically reduced (higher Gastropoda), or enlarged into feeding lamellae (higher Bivalvia), or replicated in lateral series (Polyplacophora).

Cardiac and renogenital complex. Throughout the Mollusca, adult kidneys and genital ducts have developed from true mesodermal coelomoducts associated with the pericardial cavity (Fig. 6). The cardiac structures of mollusks are also closely linked to the pallial complex. If there is a symmetrical pair of ctenidia, there will be a symmetrical pair of auricles on either side of the muscular ventricle of the heart; if one ctenidium, one auricle; if four ctenidia, four auricles. There are auriculoventricular valves to prevent backflow, and the ventricle drives blood both anteriorly and posteriorly into the major arterial systems, again through nonreturn valves. In return, these serve the hemocoelic spaces with a relatively low-pressure circulation of relatively large volumes of blood. Note that body fluids in mollusks are almost all blood, just as body cavities are almost all hemocoel.

The respiratory pigment is usually hemocyanin in solution, so that neither circulatory efficiency nor blood oxygen-carrying capacity is high (compared to even lower vertebrates like fishes). However, mollusks are mostly sluggish animals with low metabolic (and hence respiratory) rates. The blood in the hemal meshwork of all mollusks has another functional importance, since it is used as a hydraulic skeleton to transmit forces generated by distant skeleton to transmit forces generated by distant muscle contraction. The characteristically extensible soft parts of mollusks, such as tentacles, the foot, and the siphons, can all be rapidly withdrawn by muscular contraction, but are only slowly extended again by blood pressure, by blood being shifted into them from another part of the molluscan body. Mechanically, this is a case of distant sets of antagonistic muscles transmitting forces through a hemocoelic, hydraulic skeleton of fixed volume.

As shown in Fig. 6, the basic arrangement of one pair of gonads and one pair of kidneys is modified in the bivalves and chitons principally by separation of genital from renal functions while retaining paired structures. On the other hand, in gastropod evolution the asymmetry developed in gills and pallial structures (and hence in cardiac organization) has allowed various levels of asymmetry (and hence functional separation) in the renogenital structures. As shown in Fig. 6, in the majority of advanced snails the left coelomoduct becomes the functional renal

organ, while the right one is the gonoduct for the solitary gonad.

Ciliary sorting and gut function. Uniquely molluscan is the use of cilia in “sorting surfaces,” which can segregate particles into different size categories and send them to be disposed of in different ways in several parts of the organism. In a simpler type of sorting surface, the epithelium is thrown into a series of ridges and grooves, the cilia in the grooves beating along them and the cilia on the crests of the ridges beating across them. Thus, fine particles impinging on the surface can be carried in the direction of the grooves, while larger particles are carried at right angles. Such sorting surfaces occur both externally on the feeding organs and internally in the gut of many mollusks. For example, on the labial palps of bivalves, they are used to separate the larger sand grains (which are rejected) from the smaller microorganisms which then pass to the mouth. Although such sorting is by size alone, it can result in a separation of valuable food from inedible particles. Some molluscan sorting surfaces can achieve segregation into four or more size categories. Further, many mollusks have the capacity to alter the thresholds of discrimination on these surfaces, since such organs as labial palps can be expanded to a varying extent by blood pressure.

The alimentary canal in mollusks is usually extensively ciliated, with many sorting surfaces. Such a gut is organized to deal slowly but continuously with a steady stream of finely divided plant material passing in from the feeding organs, whether these be filter mechanisms or the characteristic rasping tongue (the radula) of the grazing mollusks. In the majority, digestion is functionally mixed: both extracellular and intracellular breakdown is carried out. In all lamellibranchs, and in both grazing and filter-feeding snails, a peculiar secreted structure termed the crystalline style is used as a cilia-driven gastric stirring rod and enzyme store, allowing a slow, continuous release of amylase and glycogen-breaking enzymes. The digestive diverticula (often miscalled liver lobes) are masses of blind tubules, the cells of which are phagocytic, where final intracellular digestion takes place. In the cephalopods and a few carnivorous gastropods, digestion is simpler, largely extracellular, and cyclic.

Nervous system. The range in levels of complexity of molluscan nervous systems is comparable to that found in the phylum Chordata. The four-strand nervous system with one pair of tiny ganglia found in chitons is not dissimilar to the neural plan in turbellarian flatworms. In contrast, the nervous system and sense organs of a cephalopod like an octopus are equaled and exceeded only by those of some birds and mammals.

In the majority of mollusks the nervous system is in an intermediate condition. This state involves five groups of paired ganglia associated with the buccal structures, the other head organs, the foot, the visceral mass, and the mantle. These five pairs of “local brains” are linked by long connectives and commissures. In gastropods the process of torsion has re-

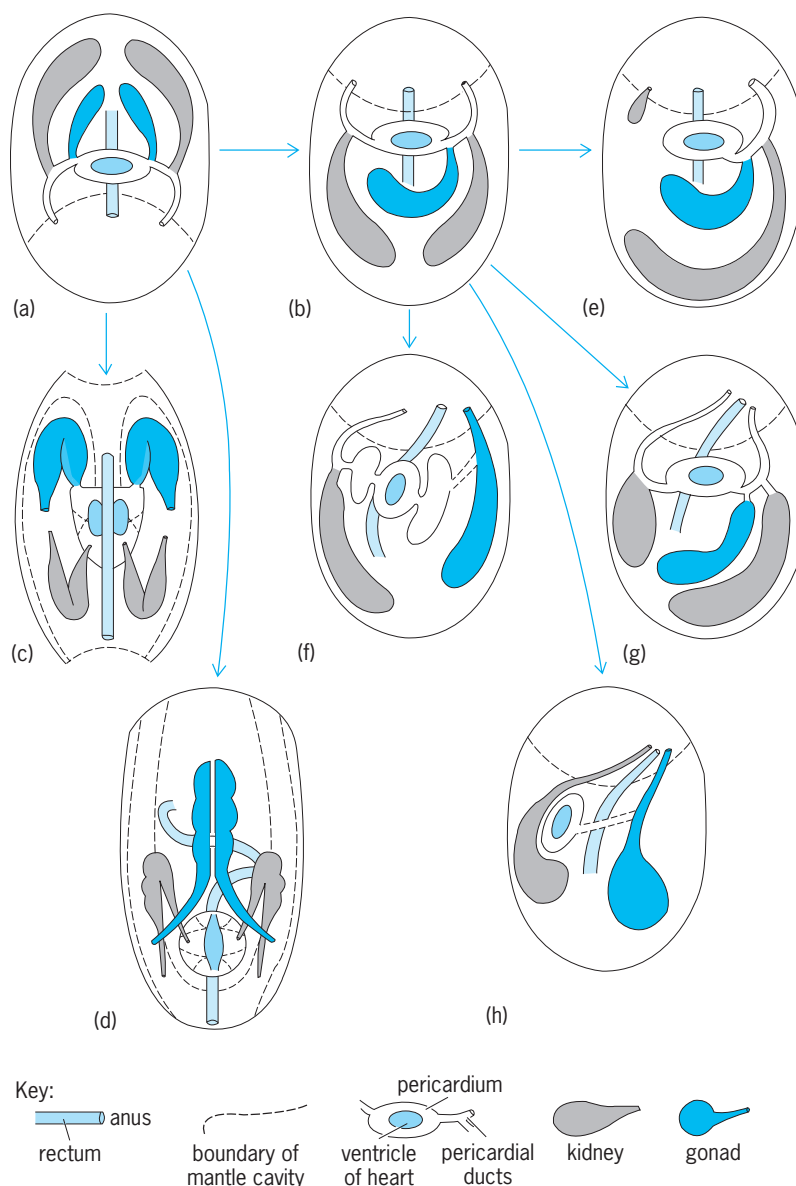


Fig. 6. Evolution of the pericardial cavity and its associated organs (heart, kidneys, gonads, rectum, and various ducts) in a series of representative mollusks. All are dorsal diagrams and, while (a) molluscan and (b) gastropod archetypes are illustrated first, all the other conditions shown occur as actual states of these organ systems in presently living forms: (c) in bivalves, (d), chitons, (e) fissurellids, (f) neritaceans, (g) trochids, and (h) pectinibranch prosobranchs (and higher gastropods). (After W. D. Russell-Hunter, *A Life of Invertebrates*, Macmillan, 1979)

sulted in twisting of these long trunk nerves (the so-called streptoneurous condition). This decentralized pattern of relative simplicity is probably related to the fact that, in mollusks other than cephalopods, the main effectors controlled by the nervous system are cilia and mucous glands. In fact, apart from the muscles which withdraw it into its shell, the typical mollusk is a slow-working animal with little fast nervous control or quick reflexes. In the brain of modern cephalopods, these paired molluscan ganglia have been fused into a massive structure, with over 300 million neurons and extensive “association” centers providing considerable mnemonic and learning capacities. See OCTOPUS.

Reproduction. In all primitive mollusks, the sexes are separate, and external fertilization follows the spawning of eggs and sperm into the sea. The zygote undergoes spiral cleavage and eventually becomes a trochophore larva, with a characteristic ring of locomotory cilia. This later develops into a typical wheel-shaped velum bearing long cilia, and this veliger larva then differentiates a mantle rudiment (secreting a shell) and a distinct headfoot. Molluscan larvae first become recognizable as gastropods, as bivalves, or as chitons when their mantle-shell rudiments first take on the growth patterns characteristic of each class. After this differentiation, the ciliated larva settles to the bottom and metamorphoses into a crawling spat (usually a miniature version of the adult mollusk).

In more advanced mollusks, eggs are larger (and fewer), fertilization may become internal (with complex courtship and copulatory procedures), and larval stages may be sequentially suppressed. A remarkably large number of mollusks (including many higher snails) are hermaphroditic. Although some are truly simultaneous hermaphrodites, many more show various kinds of consecutive sexuality. Most often the male phase occurs first, and these species are said to show protandric hermaphroditism.

Distributional Ecology

Mollusks are largely marine. The extensive use of ciliary and mucous mechanisms in feeding, locomotion, reproduction, and other functions demands a marine environment for the majority of molluscan stocks. Apart from a small number of bivalve genera living in brackish and fresh waters, all nonmarine mollusks are gastropods.

Despite the soft, hydraulically moved bodies and relatively permeable skins typical of all mollusks, some snails are relatively successful as land animals, although they are largely limited to more humid habitats. The primary physiological requirements for life on land concern water control, conversion to air breathing, and temperature regulation. Other nonmarine adaptations involve changes in reproduction (to larger eggs or viviparity) and in nitrogenous excretion (for water conservation). Some groups of littoral snails exhibit the serial acquisition of such terrestrial adaptations, as they occur in the upper seashore. It is a remarkable evolutionary fact that the successful land snails of the class Pulmonata have taken third place (admittedly well behind the arthropods and amniote vertebrates) in the race to exploit nonmarine habitats, both terrestrial and fresh water. Pulmonate land snails occur on suitable mountains at altitudes of nearly 19,700 ft (6000 m). The dominant snails of fresh waters are pulmonates of the higher limnic Basommatophora, representing an air-breathing stock of snails showing progressive readaptation to aquatic life.

In the sea, all classes of mollusks are found, and all habitats have mollusks. Protobranchiate bivalves are found at depths of over 29,500 ft (9000 m). Although ecologically cephalopod mollusks are limited to the sea, there are sound reasons for claiming modern cephalopods as the most highly organized in-

vertebrate animals. The functional efficiencies of jet propulsion and of massive brains in squid, cuttlefish, and octopuses have not been paralleled in their other physiological systems. The processes of excretion and reproduction, and the respiratory function of the blood, are still typically molluscan and could not be readily adapted for nonmarine life. It is clear that the pelagic cephalopods evolved along with the swimming vertebrates, with size and speed ratios being kept in pace by selection pressures of alternate predation and competition. In other ways, pressure to keep up with the competitive vertebrates must have helped evolve the speed, the size, and the brains of modern cephalopods from the molluscan body plan.

In addition to the extreme diversity of external body form exhibited by different mollusks, they show a remarkable diversity in their ecological distribution and life styles. However, the basic molluscan plan of structure and function always remains recognizable. See APLACOPHORA; BIVALVIA; CEPHALOPODA; GASTROPODA; LAMELLIBRANCHIA; MONOPLACOPHORA; POLYPLACOPHORA; SCAPHOPODA; SNAIL.

W. D. Russell-Hunter

Bibliography. V. Fretter and A. Graham, *British Prosobranch Molluscs*, 1962; J. E. Morton, *Molluscs*, 4th ed., 1967; S. P. Parker (ed.), *Synopsis and Classification of Living Organisms*, 2 vols., 1982; W. D. Russell-Hunter, *A Life of Invertebrates*, 1979; K. M. Wilbur and C. M. Yonge (eds.), *Physiology of Mollusca*, 1964 and 1967; C. M. Yonge and T. E. Thompson, *Living Marine Molluscs*, 1976.

Molybdenum

A chemical element, Mo, atomic number 42, and atomic weight 95.94, in the periodic table in the triad of transition elements that includes chromium (atomic number 24) and tungsten (atomic number 74). Research has revealed it to be one of the most versatile chemical elements, finding applications not only in metallurgy but also in paints, pigments, and dyes; ceramics; electroplating; industrial catalysts; industrial lubricants; and organometallic chemistry. Molybdenum is an essential trace element in soils and in agricultural fertilizers. Molybdenum atoms have been found to perform key functions in enzymes (oxidases and reductases), with particular

| | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
|----|----|----|----|----|----|----|----|----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| 1 | | | | | | | | | | | | | | | | | 18 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 1 | H | | | | | | | | | | | | | | | | | 2 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 3 | Li | Be | | | | | | | | | | | | | | 5 | B | 6 | C | 7 | N | 8 | O | 9 | F | 10 | Ne | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 11 | Na | Mg | 12 | | | | | | | | | | | 13 | Al | 14 | Si | 15 | P | 16 | S | 17 | Cl | 18 | Ar | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 19 | K | Ca | Sc | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 | 31 | 32 | 33 | 34 | 35 | 36 | 37 | 38 | 39 | 40 | 41 | 42 | 43 | 44 | 45 | 46 | 47 | 48 | 49 | 50 | 51 | 52 | 53 | 54 | 55 | 56 | 57 | 58 | 59 | 60 | 61 | 62 | 63 | 64 | 65 | 66 | 67 | 68 | 69 | 70 | 71 | 72 | 73 | 74 | 75 | 76 | 77 | 78 | 79 | 80 | 81 | 82 | 83 | 84 | 85 | 86 | 87 | 88 | 89 | 90 | 91 | 92 | 93 | 94 | 95 | 96 | 97 | 98 | 99 | 100 | 101 | 102 | 103 | 104 | 105 | 106 | 107 | 108 | 109 | 110 | 111 | 112 | 113 | 114 | 115 | 116 | 117 | 118 | 119 | 120 |
| 87 | Rb | Sr | Y | Zr | Nb | Mo | Tc | Ru | Rh | Pd | Ag | Cd | In | Sn | Sb | Te | I | Xe | 121 | 122 | 123 | 124 | 125 | 126 | 127 | 128 | 129 | 130 | 131 | 132 | 133 | 134 | 135 | 136 | 137 | 138 | 139 | 140 | 141 | 142 | 143 | 144 | 145 | 146 | 147 | 148 | 149 | 150 | 151 | 152 | 153 | 154 | 155 | 156 | 157 | 158 | 159 | 160 | 161 | 162 | 163 | 164 | 165 | 166 | 167 | 168 | 169 | 170 | 171 | 172 | 173 | 174 | 175 | 176 | 177 | 178 | 179 | 180 | 181 | 182 | 183 | 184 | 185 | 186 | 187 | 188 | 189 | 190 | 191 | 192 | 193 | 194 | 195 | 196 | 197 | 198 | 199 | 200 | 201 | 202 | | | |
| 87 | Ra | Lr | Rf | Db | Sg | Bh | Hs | Mt | 110 | 111 | 112 | 113 | 114 | 115 | 116 | 117 | 118 | 119 | 120 | 121 | 122 | 123 | 124 | 125 | 126 | 127 | 128 | 129 | 130 | 131 | 132 | 133 | 134 | 135 | 136 | 137 | 138 | 139 | 140 | 141 | 142 | 143 | 144 | 145 | 146 | 147 | 148 | 149 | 150 | 151 | 152 | 153 | 154 | 155 | 156 | 157 | 158 | 159 | 160 | 161 | 162 | 163 | 164 | 165 | 166 | 167 | 168 | 169 | 170 | 171 | 172 | 173 | 174 | 175 | 176 | 177 | 178 | 179 | 180 | 181 | 182 | 183 | 184 | 185 | 186 | 187 | 188 | 189 | 190 | 191 | 192 | 193 | 194 | 195 | 196 | 197 | 198 | 199 | 200 | 201 | 202 | | |

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

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

interest being directed toward its role in nitrogenase, which is employed by bacteria in legumes to convert inert nitrogen (N_2) of the air into biologically useful ammonia (NH_3). See NITROGEN FIXATION; PERIODIC TABLE.

Molybdenum is widely distributed in the Earth's crust at a concentration of 1.5 parts per million by weight in the lithosphere and about 10 parts per billion in the sea. It is found in at least 13 minerals, mainly as a sulfide [molybdenite (MoS_2)] or in the form of molybdates [for example, wulfenite ($PbMoO_4$) and magnesium molybdate ($MgMoO_4$)].

Although molybdenum is closer to chromium in atomic weight and atomic number, its chemical behavior is usually very similar to that of tungsten, which has nearly the same atomic radius. (This is due to the so-called lanthanide contraction in which atomic radii decrease for elements 57 to 71 found in the period between molybdenum and tungsten.) See CHROMIUM; LANTHANIDE CONTRACTION; TUNGSTEN.

Molybdenum atoms contain six valence electrons ($4d^55s^1$), which are employed with great versatility in forming compounds and complexes in which electronic configurations vary from d^0 (no d electrons in oxidation state +6) to d^8 (8 d electrons in oxidation state -2). The +6 state is preferred, but all states from -2 to +6 are known. States usually exhibit a variety of coordination numbers (4 to 9), and include polynuclear complexes and metal-metal bonds in metallic clusters with two to six metal atoms in their metallic cores. Molybdenum forms a very large number of compounds with oxygen. Low-valent molybdenum [for example, $Mo(CO)_6$ and Mo_2 , Mo_3 , and Mo_6 clusters] has a very rich organometallic chemistry, including clusters that are being studied as models for molybdenum metal surfaces that catalyze organic reactions employed in industrial syntheses and oil refining. The ability of molybdenum atoms to vary oxidation state, coordination number, and coordination geometry and to form metal-metal bonds in clusters accounts in part for the large number of industrial catalysts and biological enzymes in which Mo atoms are found at the active site for catalysis. See CHEMICAL BONDING; COORDINATION CHEMISTRY; ELECTRON CONFIGURATION.

Molybdenum is a high-melting silver-gray metal, strong even at high temperatures, hard, and resistant to corrosion (see table). It also exhibits high conductivity, a high modulus of elasticity, high thermal conductivity, and a low coefficient of expansion. Its major use is in alloy steels, for example, as tool steels ($\leq 10\%$ molybdenum), stainless steel, and armor plate. Up to 3% molybdenum is added to cast iron to increase strength. Up to 30% molybdenum may be added to iron-, cobalt-, and nickel-based alloys designed for severe heat- and corrosion-resistant applications. It may be used in filaments for light bulbs, and it has many applications in electronic circuitry. See ALLOY; IRON ALLOYS; STAINLESS STEEL.

Molybdenum trioxide, molybdates, sulfo-molybdates, and metallic molybdenum are found in thou-

| Physical properties of molybdenum metal | |
|---|---|
| Property | Value |
| Density | 10.22 g/cm ³ (5.911 oz/in. ³) |
| Heat of vaporization | 491 kJ/mol |
| Heat of fusion | 28 kJ/mol |
| Specific heat | 0.267 J/g °C |
| Thermal conductivity | 1.246 J/s/cm ² /cm °C (200 °C) 0.923 J/s/cm ² /cm °C (2200 °C) |
| Electrical conductivity | 34% International Copper Standard |
| Electrical resistivity | 5.2 microhm-cm, 20 °C 78.2 microhm-cm, 2525 °C |
| Magnetic susceptibility | 0.93×10^{-6} emu, 25 °C 1.11×10^{-6} emu, 1825 °C |
| Mean linear expansion coefficient | $6.65 \times 10^{-6}/°C$, 20–1600 °C |
| Modulus of elasticity | 0.324 N/m ² |
| Lattice parameter | 0.314767 nm (body-centered cube) |

sands of industrial catalysts used in oil refining, ammonia synthesis, and industrial syntheses of organic chemicals. Monomeric molybdenum (IV) in aqueous solution is a powerful catalyst for the reduction of inert oxo-anions such as perchlorate (ClO_4^-) or nitrate (NO_3^-) as well as other oxidized nonmetals such as azide ion (N_3^-) and dinitrogen (N_2). The trinuclear cation $MO_3O_4^{4+}$ is inert, unreactive, and noncatalytic. See CATALYSIS; HOMOGENEOUS CATALYSIS.

The molybdenum enzymes comprise two major categories. The first category contains the single, highly important enzyme nitrogenase, which is responsible for biological nitrogen fixation. The second category contains all other known molybdenum enzymes, which are crucial for the metabolism of bacteria, plants, and animals, including humans.

Edward I. Stiefel

Bibliography. F. A. Cotton and G. Wilkinson, *Advanced Inorganic Chemistry*, 6th ed., 1999; C. K. Gupta, *Extractive Metallurgy of Molybdenum*, 1993; A. Sigel and H. Sigel (eds.), *Metals Ions in Biological System*, vol. 39: *Molybdenum and Tungsten: Their Roles in Biological Processes*, 2002; T. G. Spiro, *Molybdenum Enzymes*, 1985; E. I. Stiefel, D. Coucouvanis, and W. E. Newton (eds.), *Molybdenum Enzymes, Cofactors, and Model Systems*, ACS Symp. Ser., no. 535, 1993.

Molybdenum alloys

Solid solutions of molybdenum and other metals. Molybdenum is classified as a refractory metal by virtue of its high melting point (2623 °C or 4750 °F), and many of its applications result from its strength at high temperatures. A number of other physical and mechanical properties make it attractive for use in a wide variety of applications. Molybdenum is used extensively as electrodes in electric-boost furnaces because it erodes very slowly and does not contaminate the glass bath. Its high-temperature strength allows it to support significant structural loads

imposed during operation of the furnaces. *See* MOLYBDENUM; REFRACTORY.

Molybdenum has a body-centered cubic crystal structure and displays the ductile-brittle transition behavior typical of such metals. This behavior is illustrated by bend tests on molybdenum sheet. In this test, a sheet is bent over a mandrel having a radius equal to the sheet thickness. The angle of bend at fracture is a measure of the sheet's ductility. The data from such a test indicate that the microstructure of the material has a strong influence on the ductile-brittle transition temperature. Higher temperatures are required to form recrystallized sheet through a given angle as compared to stress-relieved (recovery-annealed) sheet. Recrystallized microstructures are to be avoided because they are susceptible to brittle intergranular fracture. *See* BRITTLENESS; PLASTIC DEFORMATION OF METAL.

Manufacturing practices. Molybdenum alloys are manufactured and made into usable forms by primary consolidation, forming and machining, and joining.

Primary consolidation. Molybdenum and its alloys are consolidated commercially either by powder metallurgy techniques or by vacuum-arc casting. Both mechanical pressing and cold isostatic pressing are used to consolidate powder metallurgy billets, although most powder metallurgy mill products start as cold-isostatic-pressed billets. Hydrogen atmospheres are typically used to sinter powder metallurgy billets because they aid in the removal of oxygen from the material during sintering by reaction to form water vapor. Vacuum sintering is also employed in production of powder metallurgy billets by some manufacturers. In all cases, oxygen in the finished product must be minimized, because it is quite detrimental to the ductility of molybdenum. The vacuum-arc casting process produces ingots that are low in oxygen and need no further treatment to remove it.

Extrusion and forging can be used to work large powder metallurgy billets; vacuum-arc-cast ingots must first be processed by hot extrusion, because of the as-cast material's propensity for intergranular fracture under tensile stresses. Once large billets and ingots are reduced to convenient sizes, both powder metallurgy and vacuum-arc materials can be processed by conventional techniques such as hot rolling, cold rolling, or swaging. *See* FORGING; METAL CASTING; METAL ROLLING; POWDER METALLURGY; SINTERING.

Forming and machining. Forming or stamping of finished bars or sheets should take into account the ductile-brittle transition-temperature behavior of the material. This will require higher forming temperatures as the sheet gauge or bar diameter increases.

Molybdenum can be machined quite readily, though machinists must be prepared to employ more exacting practices in machining molybdenum and its alloys as compared to conventional materials such as steel, aluminum, and brass. Carbide cutting tools are recommended because of the abrasive nature of the chips formed during machining. Adequate lubrication and appropriate machining practices allow

production of intricate and exacting parts. Molybdenum and its alloys are readily machined by electrical-discharge-machining techniques, but the material's high melting point and high thermal conductivity require higher power levels than are encountered in machining conventional materials. *See* MACHINING; METAL, MECHANICAL PROPERTIES OF; METAL FORMING.

Joining. Molybdenum can be joined by both brazing and welding. The fusion and heat-affected zones of welded components are typically very low in ductility because of the high ductile-brittle transition temperature of recrystallized molybdenum. Welding techniques that create narrow heat-affected zones, such as electron-beam or laser welding, are usually more successfully employed than arc welding. Oxygen contamination of welds can result in intergranular oxygen segregation and embrittlement of the already low-ductility fusion zone, and therefore the use of low-oxygen inert-atmosphere glove boxes is necessary for best results when arc welding. (Glove boxes are sealed chambers that are fitted with gloves protruding into the chamber, providing a controlled atmosphere for the operator.) *See* BRAZING; WELDING AND CUTTING OF MATERIALS.

Alloys. Four main classes of commercial molybdenum-base alloys exist. One class relies on the formation of fine metal carbides that strengthen the material by dispersion hardening, and extend the resistance of the microstructure to recrystallization above that of pure molybdenum. A second class relies on solid-solution hardening to strengthen molybdenum. These two classes of materials are typically produced in both vacuum-arc-casting and powder metallurgy grades. The third class uses combinations of carbide formers and solution hardeners to provide improved high-temperature strength. This class of alloys is normally produced by powder metallurgy techniques, but some of the alloys are also amenable to vacuum-arc-casting processing. A final class of alloys, known as dispersion-strengthened alloys, relies on second-phase particles (usually an oxide of a ceramic material) introduced or produced during powder processing to provide resistance to recrystallization and to stabilize the recrystallized grain structure, enhancing high-temperature strength and improving low-temperature ductility. These latter materials by their very nature must be produced by powder metallurgy techniques. *See* HIGH-TEMPERATURE MATERIALS.

Carbide-strengthened alloys. The most common of the carbide-strengthened alloys is known as TZM, containing about 0.5% titanium, 0.08% zirconium, and 0.03% carbon. Other alloys in this class include TZC (1.2% titanium, 0.3% zirconium, 0.1% carbon), MHC (1.2% hafnium, 0.05% carbon), and ZHM (1.2% hafnium, 0.4% zirconium, 0.12% carbon). The high-temperature strength imparted by these alloys is their main reason for existence.

Both TZM and MHC have found application as metalworking tool materials. Their high-temperature strength and high thermal conductivity make them quite resistant to the collapse and thermal

cracking that are common failure mechanisms for tooling materials. A particularly demanding application is the isothermal forging process used to manufacture nickel-base superalloy gas-turbine engine components. In this process the dies and workpiece are both heated to the hot working temperature, and the forging is performed in a vacuum by using large hydraulic presses.

Solid solution alloys. Tungsten and rhenium are the two primary alloy additions in this class of alloys. A variety of compositions may be available as special orders, but the most common compositions are 30% tungsten (Mo-30W), 5% rhenium (Mo-5Re), 41% rhenium (Mo-41Re), and 47.5% rhenium (Mo-50Re). With the exception of the Mo-30W alloy which is available as a vacuum-arc-cast product, these alloys are normally produced by powder metallurgy. The tungsten-containing alloys find application as components in systems handling molten zinc, because of their resistance to this medium. They were developed as a lower-cost, lighter-weight alternative to pure tungsten and have served these applications well over the years. The 5% rhenium alloy is used primarily as thermocouple wire, while the 41% and 47.5% alloys are used in structural aerospace applications. These high-rhenium alloys trace their existence to the discovery of the rhenium effect, which produces a dramatic improvement in molybdenum's ductile-brittle transition temperature when alloyed with rhenium. This effect is also utilized by employing the Mo-50Re alloy as welding rod to improve the ductility of welds in molybdenum. A major stumbling block to the use of rhenium alloys is their high cost, which is related to the expense and price volatility of rhenium. See RHENIUM; TUNGSTEN.

Combined alloys. The beneficial effects of solid-solution hardening and dispersion hardening found in the carbide-strengthened alloys have been combined in the HWM-25 alloy (25% tungsten, 1% hafnium, 0.07% carbon). This alloy offers high-temperature strength greater than that of carbide-strengthened molybdenum, but it has not found wide commercial application because of the added cost of tungsten and the expense of processing the material.

Dispersion-strengthened alloys. The dispersion-strengthened alloys rely exclusively on powder metallurgy manufacturing techniques. This allows the production of fine stable dispersions of second phases that stabilize the wrought structure against recrystallization, resulting in a material having improved high-temperature creep strength as compared to pure molybdenum. Once recrystallization occurs, the dispersoids also stabilize the interlocked recrystallized grain structure. This latter effect produces significant improvements in the ductility of the recrystallized material.

The potassium- and silicon-doped alloys such as MH (150 ppm potassium, 300 ppm silicon) and KW (200 ppm potassium, 300 ppm silicon, 100 ppm aluminum) are the oldest of this category; they are analogs to the doped tungsten alloys in common use for tungsten lamp filament. They were first developed to satisfy the requirements of the lighting in-

dustry for creep-resistant molybdenum components. They do not possess particularly high strength at low temperatures, but they are quite resistant to recrystallization and they possess excellent creep resistance due to their stable, interlocked recrystallized grain structure. They are used in applications requiring high-temperature creep resistance such as nuclear fuel sintering boats. See CREEP (MATERIALS).

Composites. A family of copper-molybdenum-copper (CMC) laminates has been developed to serve the needs of the high-performance electronics industry. Laminating molybdenum with copper raises the coefficient of thermal expansion and improves the heat-transfer characteristics of the composite relative to pure molybdenum. This allows better matching between the composite, which serves as a packaging or mounting material, and the solid-state devices being employed. The coefficient of thermal expansion of the composite can be tailored to a specific application by selection of the copper/molybdenum ratio. Similar control can be exerted on thermal conductivity and elastic modulus, both properties of great interest to circuit designers. See ALLOY; CIRCUIT (ELECTRICITY); ELASTICITY; ELECTRICAL CONDUCTIVITY OF METALS; SOLID SOLUTION.

John A. Shields

Bibliography. J. Z. Briggs and R. Q. Barr, Arc-cast molybdenum-base TZM alloy: Properties and applications, *High Temperatures-High Pressures 1971*, vol. 3, pp. 363-409, 1971; Climax Molybdenum Co., *Molybdenum Metal*, 1960; J. B. Conway and P. N. Flagella, *Creep-Rupture Data for the Refractory Metals to High Temperatures*, 1971; J. A. Shields, Jr., Refractory Metals Forum: Molybdenum and Its Alloys, *Adv. Mater. Proc.*, 142(4):28-36, 1992.

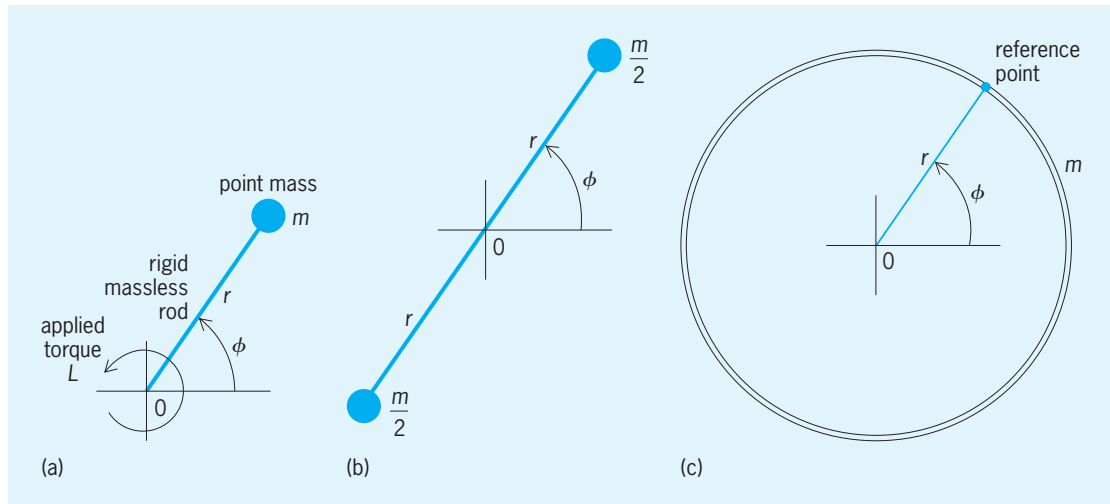
Moment of inertia

A measure of the "opposition to rotation" around a fixed axis that a rigid body presents to a torque applied about that axis. The concept is used in analyzing the dynamics of systems in which rotation occurs, for example, in attitude control of spacecraft. It extends naturally into quantum mechanics and the notion of quantization of rotational energy.

Two-dimensional systems. The simplest system to illustrate moment of inertia is the rigid rotor (*illus. a*) lying in a horizontal plane. This is a rigid massless rod of length r , pivoted at the origin of a coordinate system, at the end of which is a point mass m . The state of the system is specified by one quantity: the angle coordinate ϕ . The angular velocity ω is given by Eq. (1). It follows from Newton's first law of mo-

$$\omega \equiv \frac{d\phi}{dt} \quad (1)$$

tion that if no external forces are applied, ω remains constant. However, if a torque of magnitude L is applied to the system, centered on the origin, then from Newton's second law of motion it can be shown that



Three rotating systems for which the moments of inertia about the origin are all equal to mr^2 . (a) Rigid rotor. (b) Baton. (c) Hoop.

L is related to the angular acceleration by Eq. (2). For

$$L = mr^2 \frac{d\omega}{dt} \quad (2)$$

any given system, the factor preceding the derivative is defined as the moment of inertia, usually symbolized by the letter I . Thus, the angular acceleration is given by Eq. (3). Evidently, for the rigid rotor, the moment of inertia is given by Eq. (4).

$$\frac{d\omega}{dt} = \frac{L}{I} \quad (3)$$

$$I_{\text{rotor}} = mr^2 \quad (4)$$

See ACCELERATION; VELOCITY; TORQUE.

For a rigid system of N mass particles lying in the plane, the total moment of inertia is simply the sum of the individual moments of inertia, that is, the sum given by Eq. (5). Thus, the baton in illus. *b*, having

$$I_{\text{total}} = \sum_{i=1}^n m_i r_i^2 = \sum_{i=1}^n I_i \quad (5)$$

masses $m/2$ at each end of a rod that is twice as long, has exactly the same moment of inertia as the rigid rotor, as given in Eq. (6).

$$I_{\text{baton}} = \frac{m}{2} r^2 + \frac{m}{2} r^2 = mr^2 = I_{\text{rotor}} \quad (6)$$

Spreading the total mass m evenly around a circle of radius r (illus. *c*) gives a hoop. It can be shown, using calculus, that $I_{\text{hoop}} = mr^2$ as well. However, if the mass is spread uniformly to form a disk of radius r , the moment of inertia is less, as given by Eq. (7).

$$I_{\text{disk}} = \frac{1}{2} mr^2 = \frac{1}{2} I_{\text{rotor}} \quad (7)$$

The radius of gyration k of a body is defined as the radius at which all its mass could be concentrated to yield the same moment of inertia; that is, k is given

by Eq. (8). Thus, for the disk, the radius of gyration is

$$k \equiv \sqrt{\frac{I}{m}} \quad (8)$$

given by Eq. (9). Tables are available listing the values

$$k_{\text{disk}} = \frac{r}{\sqrt{2}} \quad (9)$$

of k for other shapes.

The angular momentum of a rotating system is given by Eq. (10), and its kinetic energy of rotation is given by Eq. (11).

$$J = I\omega \quad (10)$$

$$T_{\text{rotor}} = \frac{1}{2} I\omega^2 \quad (11)$$

See ANGULAR MOMENTUM; ENERGY.

Three-dimensional systems. The basic principles discussed above apply equally well to rigid bodies, or to systems of rigidly connected masses in three dimensions. The situation is more complex because not only can one choose any line in space as an axis of rotation, but the body itself may have any orientation with respect to that axis. However, analysis of rotation dynamics is greatly simplified for such systems in two ways. See RIGID-BODY DYNAMICS.

The first simplification is by the parallel-axis theorem. Assume that a rigid body of mass M is spun about an axis that passes through its center of mass and for which the moment of inertia is I_{COM} . Now consider a new axis, parallel to the first but shifted to a perpendicular distance R away. The parallel-axis theorem states that the new moment of inertia is given by Eq. (12).

$$I' = I_{\text{COM}} + MR^2 \quad (12)$$

See CENTER OF MASS.

The second simplification arises from the fact that no matter what shape the body may have, only six quantities are needed to allow full analysis. These are

the components of the matrix called the moment of inertia tensor \mathbf{I} , is given by Eq. (13), where the

$$\mathbf{I} = \begin{pmatrix} I_{xx} & I_{xy} & I_{xz} \\ I_{xy} & I_{yy} & I_{yz} \\ I_{xz} & I_{yz} & I_{zz} \end{pmatrix} \quad (13)$$

terms on the diagonals are the moments of inertia for rotation about the x , y , and z axes, respectively, and the off-diagonal quantities are the products of inertia.

To illustrate the practical meaning of the products of inertia, consider a rigid body rotating on an axle that passes through the center of mass and lies on, say, the x axis. Mechanical bearings at each end of the axle hold it in place. When the body spins, the fact that the products of inertia are not zero means that the bearings will be subjected to rotating sideways forces. In practice, this can lead to undesirable mechanical vibration and wear. For this reason it is not enough to balance automobile wheels statically such that the wheels' centers of mass lie on their axes of rotation: the wheels must also be dynamically balanced to avoid their vibration when the car is moving.

It can be shown that the matrix can always be transformed, through rotation of the assumed coordinate axes, such that the products of inertia are reduced to zero. The resulting axes are known as the principal axes, corresponding to which are the principal moments of inertia, I_a , I_b , and I_c , so the moment of inertia tensor is given by Eq. (14). By convention,

$$\mathbf{I}_{\text{principal}} = \begin{pmatrix} I_a & 0 & 0 \\ 0 & I_b & 0 \\ 0 & 0 & I_c \end{pmatrix} \quad (14)$$

the three principal moments of inertia are ordered in size as $I_a \geq I_b \geq I_c$.

The tennis-racket theorem applies to freely spinning bodies in free fall for which the principal moments of inertia are all different. It states that the body will spin stably about the axes corresponding to the largest principal moment (I_a) or the smallest (I_c); however, an initial spin about the axis corresponding to I_b will be unstable and quickly change to a spin state that is a combination of spins about the other two axes.

A useful visualization of the magnitude of the moment of inertia around various axes is the momental ellipsoid or inertial ellipsoid, given by Eq. (15). (The

$$x^2 I_x + y^2 I_y + z^2 I_z = 1 \quad (15)$$

principal axes are taken here to coincide with the x, y, z -coordinate axes). In the special case of a uniform sphere, any orthogonal set of axes with origin at the sphere's center are principal axes. The three principal moments are equal so that the momental ellipsoid is also a sphere.

Application to molecular dynamics. Among various spectroscopic techniques to deduce the structure of molecules is rotational spectroscopy. Molecules are generally small enough for quantum effects to be-

come evident in that the spectra show discrete rates of rotation, each separated by one quantum of energy at the far-infrared and microwave frequencies. For a diatomic molecule spinning as a rigid rotor confined to a plane, the time-independent part of Schrödinger's equation reduces to Eq. (16), where

$$\frac{d^2 \Phi(\phi)}{d\phi^2} = -\frac{2IE}{\hbar^2} \Phi(\phi) \quad (16)$$

Φ gives the variation of the molecule's wavefunction with ϕ , E is rotational energy, \hbar is Planck's constant divided by 2π , and I is the molecule's moment of inertia. The solutions that satisfy the necessary periodicity of Φ with ϕ yield quantized energies given by Eq. (17), where K is restricted to the integer values

$$E_K = \frac{K^2 \hbar^2}{2I} \quad (17)$$

$K = 0, \pm 1, \pm 2, \dots$ See MOLECULAR STRUCTURE AND SPECTRA; SCHRÖDINGER'S WAVE EQUATION.

Application to gravitational radiation. Einstein's general theory of relativity predicts the existence of gravitational waves. It is calculated that only very massive physical processes, such as stars collapsing into black holes, can generate gravitational waves powerful enough to be detectable with current technology. In principle, however, a spinning rod of any size will generate gravitational waves. According to theory, the rotational kinetic energy of a freely spinning rod is dissipated, by the gravitational waves it generates, at a rate given by Eq. (18), where G is the

$$\frac{dT_{\text{rotor}}}{dt} = \frac{32G\omega^6}{5c^5} I_{\text{rod}}^2 \quad (18)$$

gravitational constant and c is the speed of light. See GRAVITATION; GRAVITATIONAL RADIATION; RELATIVITY.

Andrej Tenne-Sens

Bibliography. J. Foster and J. D. Nightingale, *A Short Course in General Relativity*, 3d ed., Springer-Verlag, 2006; G. R. Fowles and G. L. Cassiday, *Analytical Mechanics*, 7th ed., Thomson Brooks/Cole, 2004; L. N. Hand and J. D. Finch, *Analytical Mechanics*, Cambridge University Press, 1998; G. K. Vemulapalli, *Physical Chemistry*, Prentice Hall, 1993.

Momentum

A foundational quantity in physics that is always conserved in isolated systems, defined in classical mechanics as the product of mass and velocity. An equivalent definition that is better suited to modern physics is that momentum is the rate of change of energy with respect to velocity: this allows extending the conservation laws to include entities of zero mass. Whenever energy is transported from one system to another, it is always associated with some form of momentum. See CLASSICAL MECHANICS.

The word momentum (Latin for motion) by itself usually means linear momentum, the vector quantity obtained by multiplying (or scaling) the particle's velocity vector by its mass. The linear

momentum of a system is the sum of the linear momenta of the particles composing it. By Newton's second law of motion, a particle's linear momentum increases with time in direct proportion to the applied force. *See* DYNAMICS; FORCE; LINEAR MOMENTUM; MOTION; NEWTON'S LAWS OF MOTION.

A particle's angular momentum is a vector quantity defined as its moment of (linear) momentum about a specified axis. The angular momentum of a system is the sum of the angular momenta (about the same axis) of the particles composing it. A system's angular momentum increases with time in direct proportion to the total applied moment of force, or torque, about the axis. *See* ANGULAR MOMENTUM; ROTATIONAL MOTION; TORQUE.

The concept of momentum is of fundamental importance because in isolated physical systems (that is, ones not acted on by some outside influence) both linear and angular momentum do not change with time: they are said to be conserved. (This statement also applies to total system energy.) The law of conservation of linear momentum comes from the fact that the laws of physics do not depend on where the system is located (the universe is homogeneous). The law of conservation of angular momentum comes from the fact that the laws of physics do not depend on the orientation of the system (the universe is isotropic). The two conservation laws of momentum along with the law of conservation of energy are the three foundational axioms of classical physics. *See* CONSERVATION LAWS (PHYSICS); CONSERVATION OF ENERGY; CONSERVATION OF MOMENTUM; KINEMATICS; MECHANICS; SYMMETRY LAWS (PHYSICS).

For continuous systems such as semirigid bodies, fluids, and electromagnetic radiation, momentum is often expressed in the form of momentum density, that is, momentum per unit volume.

When velocities approach the speed of light, observed mass increases and Newton's laws of motion must give way to Einstein's more accurate laws of relativity. Nevertheless, the same form of law of conservation of linear momentum remains valid if the relativistic increase of mass is taken into account. Relativity views space and time united into a single entity, called spacetime, and correspondingly unites momentum and energy into momentum-energy, or momenergy. Just as the spacetime interval between events is invariant (the same for all observers), the rest-mass component of a particle's momentum-energy is also. This invariance implies that particles with finite momentum but no mass, such as photons, can travel only at the speed of light. *See* PHOTON; RELATIVITY.

In systems where quantum-mechanical effects are significant, the exact outcome of an experiment cannot be forecast, only the probability of obtaining any one of all the possible outcomes. Nevertheless, each experiment will show that linear and angular momentum are conserved. *See* QUANTUM MECHANICS.

Generalized momentum. The conservation laws of classical mechanics are embodied in the powerful analysis methods of J. L. Lagrange and W. R. Hamilton, applicable to constrained dynamical systems whose

degrees of freedom are limited (or at least countable). *See* CONSTRAINT; DEGREE OF FREEDOM (MECHANICS).

The state of a constrained system can be uniquely specified by an appropriate choice of generalized coordinates represented by q_j , one for each degree of freedom. In Lagrangian analysis, the system's total kinetic energy T is written as a function of all the q_j and the generalized velocities $\dot{q}_j \equiv dq_j/dt$. Inserting T into Lagrange's equations (1) yields one differential

$$\frac{d}{dt} \left(\frac{\partial T}{\partial \dot{q}_j} \right) = \frac{\partial T}{\partial q_j} \quad (1)$$

equation (of second degree in time) for each degree of freedom, $j = 1, 2, 3, \dots$, and so forth.

The resulting set of simultaneous differential equations are then solved for the desired dynamic variables. The term $\partial T / \partial \dot{q}_j$ is called the generalized momentum; it is conserved in isolated systems. For systems that do not dissipate energy (through friction, say), T is replaced by the Lagrangian function $L \equiv T - V$, where V is the system's total potential energy. *See* LAGRANGE'S EQUATIONS.

As a simple example, consider a bead of mass m constrained to move on a fixed circle of radius R ; this system has only one degree of freedom. A natural (but not unique) choice of generalized coordinate is the angle θ , as seen from the circle's center, between the bead and a fixed reference direction. It can be shown that the bead's kinetic energy is $T = \frac{1}{2} m (R\dot{\theta})^2$. Differentiating T with respect to $\dot{\theta}$ gives the generalized momentum: $\partial T / \partial \dot{\theta} = mR^2\dot{\theta}$. For this particular case, the familiar formula for angular momentum is obtained.

It is sometimes convenient to convert the Lagrangian into the Hamiltonian function H , a function of the generalized coordinates and their generalized momenta, represented by p_j . In conservative systems, H equals the total energy $T + V$. Analysis proceeds by plugging H into Hamilton's equations (2), which generate two differential equations

$$\dot{q}_j = \frac{\partial H}{\partial p_j} \quad \dot{p}_j = -\frac{\partial H}{\partial q_j} \quad (2)$$

(of first degree in time) for each degree of freedom j .

This set of simultaneous differential equations is then solved for the desired results. Because of the symmetry of these equations, p_j and q_j are said to be canonically conjugate, and the p_j are called canonical momenta or conjugate momenta. *See* CANONICAL COORDINATES AND TRANSFORMATIONS; HAMILTON'S EQUATIONS OF MOTION.

In the equations of nonrelativistic quantum mechanics, the generalized coordinates are replaced formally by their corresponding operators. *See* NONRELATIVISTIC QUANTUM THEORY; SCHRÖDINGER'S WAVE EQUATION.

Momentum in unified theories. A major area of research is to reconcile the current contradictions between quantum theory and general relativity. Just as Newton's laws of motion are approximations to those of relativity, it may turn out that the laws of relativity are themselves approximations. For example, the hypothesis of loop quantum gravity, that

spacetime is quantized, carries the implication that the speed of light is not constant. All candidate theories, however, must be compatible with current conservation laws in the limit of everyday mass, time, and length scales. Conversely, any future experiments revealing situations in which momentum (or energy) is not conserved will illuminate the way to a self-consistent unified theory. *See* QUANTUM GRAVITATION.

Andrej Tenne-Sens

Bibliography. R. P. Feynman, R. B. Leighton, and M. L. Sands, *The Feynman Lectures on Physics: The Definitive and Extended Edition*, Addison-Wesley, Reading, MA, 2005; H. Goldstein, C. P. Poole, and J. L. Safko, *Classical Mechanics*, 3d ed., Pearson Education, San Francisco, 2002; R. Penrose, *The Road to Reality: A Complete Guide to the Laus of the Universe*, Jonathan Cape, London, 2004; L. Smolin, *Three Roads to Quantum Gravity*, Basic Books, New York, 2001; A. Sommerfeld, *Lectures on Theoretical Physics*, vol. 1: *Mechanics*, Academic Press, New York, 1964; E. F. Taylor and J. A. Wheeler, *Spacetime Physics: Introduction to Special Relativity*, 2d ed., W. H. Freeman, New York, 1992.

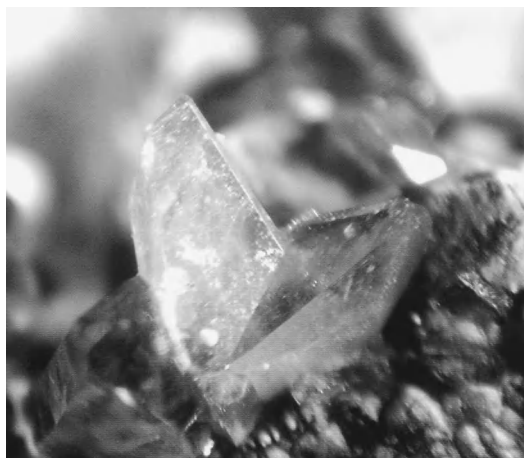
Monazite

A rare mineral that incorporates the light rare-earth elements (lanthanum, cerium, praseodymium, neodymium, promethium, samarium, europium, gadolinium) and also yttrium. Monazite has a general formula of $(La,Ce,Nd)PO_4$, but Pr, Sm, Eu, Gd, and Y substitute for La, Ce, and Nd in solid solution in minor amounts. The dominant rare-earth element in a particular monazite is denoted by the atomic suffix, such as monazite-(Ce) in which cerium exists in amounts greater than other rare-earth atoms. Monazite-(Ce), monazite-(La), and monazite-(Nd) are officially recognized by the International Mineralogical Association. *See* CERIUM; MINERAL; RARE-EARTH ELEMENTS; YTTRIUM.

Monazite is one of the main ore minerals for the rare-earth elements that are used in the manufacture of television and computer screens, fluorescent light bulbs, and highly efficient batteries, among other industrial applications.

Monazite occurs mainly as a minor mineral in granites, and is concentrated in sediments and other rock types following the erosion of the host granites. In the United States, monazite is found in economic amounts in metamorphic rocks and stream deposits in North Carolina. Most of the world's supply of monazite comes from beach sands eroded from granites in Australia, Brazil, India, and Malaysia. *See* GRANITE.

The atomic arrangement of monazite is formed of a packing arrangement of (PO_4) tetrahedra and distorted (REO_9) polyhedra, where RE = the rare-earth elements in the particular monazite mineral. The arrangement is formed of chains of alternating phosphate tetrahedra and RE polyhedra, parallel to the *c* axis. Monazite is similar in structure and chemistry to the tetragonal mineral xenotime, $Y(PO_4)$, that selectively incorporates the heavy rare-earth elements. *See* PHOSPHATE MINERALS.



Crystals of monazite.

Monazite crystallizes in the monoclinic crystal system (see **illus.**), crystal class $2/m$, space group $P2_1/n$, where $a \approx 6.8$, $b \approx 7.0$, $c \approx 6.5\text{\AA}$, and $\beta \approx 103.4^\circ$. The mineral is variably green, yellow, brown, or red-brown, and rarely occurs in crystals large enough to discern with the unaided eye. Mohs hardness is 5–5.5, and the specific gravity is 4.6–5.5, varying with substitution of different elements. *See* HARDNESS SCALES.

Monazite is often radioactive because of the substitution of the element thorium for the rare-earth elements. Although monazite typically incorporates relatively minor amounts of thorium, the rarity of this element makes monazite the chief thorium ore. Thorium is used for the generation of atomic energy, although in minor amounts. Its principal use is in the manufacture of mantles for incandescent gas lanterns. *See* RADIOACTIVE MINERALS.

Because of the ability of monazite to incorporate radioactive elements, such as thorium, extensive research has been undertaken to determine if the mineral can be synthesized with large amounts of radioactive elements in it. Much of this work has been pioneered by researchers at the Oak Ridge National Laboratory (Tennessee) in order to determine if synthetic monazite can be used to isolate radioactive waste from nuclear power plants and weapons production facilities. This important work has shown that synthetic monazite can incorporate significant amounts of radioactive waste, and is thus an effective method of encapsulating the waste in a solid form.

John M. Hughes; John Rakovan

Bibliography. L. A. Boatner and B. C. Sales, Monazite, in W. Lutze and R.C. Ewing (eds.), *Radioactive Waste Forms for the Future*, pp. 495–564, Elsevier Science, Amsterdam, 1988; L. L. Y. Chang, R. A. Howie, and J. Zussman, *Rock-Forming Minerals*, vol. 5B: *Non-silicates, Sulphates, Carbonates, Phosphates and Halides*, Longman Group, England, 1996; C. Klein and C. S. Hurlbut, Jr., *Manual of Mineralogy*, 21st ed., John Wiley, 1993; Y. Ni, J. M. Hughes, and A. N. Mariano, Crystal chemistry of the monazite and xenotime structures, *Amer. Mineral.*, 80:21–26, 1995.

Mongoose

Any of a group of 38 species of carnivorous mammals in the family Herpestidae. Mongooses inhabit a broad band along both sides of the Equator from western Africa to Southeast Asia. Although native to southern Asia, the East Indies, Africa, and Madagascar, one genus (*Herpestes*) was introduced into Spain and Portugal as well as to other parts of Europe and to many islands around the world including the West Indies, Fiji, and the Hawaiian Islands (see **table**).

Description. Mongooses are long slender carnivores with long faces, small rounded ears, short legs, and long tapering bushy tails. Most have long coarse, usually grizzled fur and lack body markings,

although a few have stripes (see **illustration**). Color ranges from dark gray through brown to yellowish or reddish. *Galidia elegans* from Madagascar has five to seven dark bands on its tail. Four or five toes are present on each foot, and the claws are nonretractile. Most species have a large anal sac containing at least two glandular openings. Adult mongooses have a head-body length of 24–60 cm (9.5–23 in.), a tail length of 19–44 cm (7.5–17 in.), and a weight ranging from 320 g to 5 kg (11 oz to 11 lb). The dwarf mongoose (*Helogale parvula*) is the smallest species, while the white-tailed mongoose (*Ichneumia albicauda*) is among the largest. The dental formula is I 3/3 C 1/1 Pm 3-4/3-4, M 2/2 × 2 for a total of 36–40 teeth.

Mongooses and their distribution

| Name | Range |
|---|---|
| Malagasy ring-tailed mongoose <i>Galidia elegans</i> | E., W., and N. Madagascar |
| Malagasy broad-striped mongooses <i>Galidictis fasciata</i> <i>G. grandidieri</i> | E. Madagascar S.W. Madagascar |
| Malagasy narrow-striped mongoose <i>Mungotictis decemlineata</i> | W. and S.W. Madagascar |
| Malagasy brown-tailed mongoose <i>Salanoia concolor</i> | N.E. Madagascar |
| Mongooses | |
| <i>Herpestes ichneumon</i> | Spain, Portugal, Morocco, Tunisia, Egypt, Libya, Africa |
| <i>H. javanicus</i> | Iraq, India, S. China, Malay Peninsula |
| <i>H. edwardsii</i> | Arabian Peninsula to Afghanistan, India, and Sri Lanka |
| <i>H. smithii</i> | India, Sri Lanka |
| <i>H. fuscus</i> | S.W. India, Sri Lanka |
| <i>H. vitticollis</i> | S.W. India, Sri Lanka |
| <i>H. urva</i> | Nepal, China, Taiwan, Malaysia |
| <i>H. semitorquatus</i> | Sumatra, Borneo |
| <i>H. brachyurus</i> | Malaysia, Sumatra, Borneo |
| <i>H. nasa</i> | Nigeria, Zaire |
| Gray and slender mongooses | |
| <i>Galerella pulverulenta</i> | South Africa |
| <i>G. nigrita</i> | Angola, Namibia |
| <i>G. sanguinea</i> | Africa |
| <i>G. ochracea</i> | Somalia |
| Banded and Gambian mongooses | |
| <i>Mungos gambianus</i> | Gambia, Nigeria |
| <i>M. mungo</i> | Gambia, Ethiopia, South Africa |
| Cusimanses | |
| <i>Crossarchus obscurus</i> | Sierra Leone to Ghana |
| <i>C. platycephalus</i> | Benin, Cameroon, Congo |
| <i>C. alexandri</i> | Zaire, Uganda, Central African Republic, Zambia |
| <i>C. ansorgei</i> | Zaire, Angola |
| Liberian mongoose <i>Liberiictis kuhni</i> | Liberia, Ivory Coast |
| Dwarf mongooses | |
| <i>Helogale parvula</i> | Ethiopia, Angola, South Africa |
| <i>H. hirtula</i> | Ethiopia, Somalia, Kenya |
| African tropical savannah mongoose <i>Dologale dybowskii</i> | Central African Republic, Zaire, Sudan, Uganda |
| Black-legged mongooses | |
| <i>Bdeogale crassicauda</i> | Yemen, Kenya, Mozambique |
| <i>B. nigripes</i> | Nigeria, Zaire, Angola |
| <i>B. jacksoni</i> | Uganda, Kenya |
| Meller's mongoose <i>Rhynchogale melleri</i> | Zaire, Tanzania, South Africa, Angola |
| White-tailed mongoose <i>Ichneumia albicauda</i> | Arabian Peninsula, Senegal, Egypt, Namibia, South Africa |
| Marsh mongoose (water mongoose) <i>Atilax paludinosus</i> | Senegal, Ethiopia, South Africa |
| Yellow mongoose <i>Cynictis penicillata</i> | Angola, Namibia, Botswana, Zimbabwe, South Africa |
| Gray meerkat (Soule's mongoose) <i>Paracynictis selous</i> | Angola, Zambia, Malawi, Namibia, Botswana, Zimbabwe, Mozambique, South Africa |
| Suricate (slender-tailed meerkat) <i>Suricata suricatta</i> | Angola, Namibia, Botswana, South Africa |



Banded mongoose, *Mungos mungo*. (Photo by Gerald and Buff Corsi; © 2004 California Academy of Sciences)

Mongoose are agile and active. Most species are terrestrial, although some are semiaquatic and arboreal. They seek shelter in hollow logs or trees, in rock crevices, or in holes in the ground. Some, such as *Cynictis* and *Suricata*, live in colonies in underground burrows. They may be either nocturnal or diurnal. Depending upon the species, they inhabit a wide range of habitats ranging from dense forests to open woodland, savanna, semidesert, and desert. The marsh mongoose (*Atilax paludinosus*) is a good swimmer and inhabits marshes, reed-grown streambeds, and tidal estuaries. Mongooses are omnivorous, feeding on insects, crabs, fish, frogs, snakes, birds and their eggs, small mammals, fruits, and other vegetable matter.

Breeding and development. Breeding may be seasonal or continue throughout the year with females having two or three litters annually. Litters of one to four are born after a gestation period ranging from 42 to 105 days. Studies of dwarf mongooses (*Helogale*) in Serengeti National Park revealed groups of four or five adults of each sex living together, with all of the adults cooperating to raise the offspring. Depending upon the species, sexual maturity may be reached as early as 9–10 months or by 2 years of age. A normal life-span in the wild would be about 6–10 years, although in captivity the dwarf mongoose (*Helogale parvula*) has lived 13 years, the marsh mongoose (*Atilax paludinosus*) for 19+ years, and the Malagasy ring-tailed mongoose (*Galidia elegans*) for 24 years 5 months.

Impact. Mongooses, particularly the genus *Herpestes*, have been introduced into many regions of the world. For example, the small Indian Mongoose (*Herpestes javanicus*) has been introduced into Costa Rica, Cuba, Antigua, Dominican Republic, Jamaica, Puerto Rico, West Indies, Fiji, Hawaii, Japan, and other areas. Originally introduced to help control populations of rodents and venomous snakes in sugarcane plantations, they quickly multiplied and

spread to become pests since there are no natural predators on many of these islands. They have destroyed not only the rats and snakes but also harmless reptiles, birds, and mammals. They have contributed to the extinction or endangerment of many desirable species of wildlife, as well as preying upon domestic poultry and consuming eggs. They have caused the extinction of seven species of amphibians and reptiles in Puerto Rico, and are one of the major factors in the extinction of a lizard, a snake, two birds, and a rodent in Jamaica. Mongooses are widespread in Hawaii and have had a significant impact on native species. Now the importation or possession of mongooses is forbidden by law in some countries, including the United States and Australia. In many areas, rats now live in trees, where they are safe from mongooses.

Threats. Mongooses may be preyed upon by larger carnivores and raptors. Humans hunt some species as pests. Habitat destruction has caused some species to be listed as vulnerable (*Galidictis fasciata*, *Mungotictis decemlineata*, *Salanoia concolor*, *Bdeogale jacksoni*) or endangered (*Galidictis grandidieri*, *Liberiictis kubni*, *Bdeogale crassicauda omnivora*) by the International Union for the Conservation of Nature and Natural Resources (IUCN). Some species can be easily domesticated. They are fairly intelligent and can be taught simple tricks, so they are often kept as pets to protect the home from vermin. See CARNIVORA; SCENT GLAND.

Donald W. Linzey

Bibliography. D. Macdonald (ed.), *The Encyclopedia of Mammals*, Andromeda Oxford, 2001; R. M. Nowak, *Walker's Mammals of the World*, 6th ed., Johns Hopkins University Press, 1999.

Monhysterida

An order of nematodes in which, generally, the stoma is funnel shaped and lightly cuticularized; however, in some families the stoma is spacious and heavily cuticularized, and is armed with protrusible teeth. The amphids vary from simple spirals to circular forms. Usually the second and third circllets of cephalic sensilla are combined, but in some taxa the third circllet of four distinct setae is separate. The normal pattern of distribution is often disrupted by numerous cervical setae. The near-cylindrical esophagus is sometimes swollen posteriorly. The cuticle may be smooth or may have annuli or ornamentation. When the annuli are distinct, the somatic setae may be long and in four to eight longitudinal rows. The female gonads are outstretched and either single or paired.

There are three monhysterid superfamilies, Linhomoeoidea, Monhysteroidea, and Siphonolaimoidea. The Linhomoeoidea and Siphonolaimoidea are primarily marine forms; feeding habits among the groups are unknown. The Monhysteroidea are free-living nematodes found in all environments, from marine waters to fresh waters and soil; feeding habits are unknown. See NEMATODA (NEMATODA).

Armand R. Maggenti

Monkey

An adaptive or evolutionary grade among the primates, represented by members of two of the three modern anthropoid superfamilies. The New World, platyrrhine monkeys (Ateloidea) and Old World, catarrhine forms (Cercopithecoidea) probably reached a monkey level of adaptation independently some time after their separation from a common ancestor, perhaps 45 million years ago (Ma; **Fig. 1**). The term monkey is not indicative of taxonomic or phylogenetic relationship: the closest relatives of the cercopithecoids are not the ateloid monkeys but the Old World apes and humans. *See* APES; FOSSIL APES; FOSSIL PRIMATES; PRIMATES.

The Ateloidea comprise two families, while the living Cercopithecoidea are considered today to comprise only one family, with two subfamilies. A modern classification of the Anthropoidea follows:

- Hyporder Anthropoidea
 - Infraorder Platyrrhini
 - Superfamily Ateloidea (New World or platyrrhine monkeys)
 - Family Atelidae

- Subfamily Atelinae (howler and spider monkeys)
- Subfamily Pitheciinae (saki, owl, and titi monkeys)
- Family Cebidae
 - Subfamily Cebinae (capuchin and squirrel monkeys)
 - Subfamily Callitrichinae (marmosets and tamarins)
- Family Branisellidae (extinct early ateloids)
- Infraorder Catarrhini (Old World anthropoids)
 - Parvorder Eucatarrhini (modern catarrhines)
 - Superfamily Hominoidea (gibbons, great apes, and humans)
 - Superfamily Cercopithecoidea
 - Family Cercopithecidae (Old World or catarrhine monkeys)
 - Subfamily Cercopithecinae (cheek-pouched monkeys: macaques, baboons, geladas, mangabeys, and guenons)
 - Subfamily Colobinae (leaf eaters: langurs and colobus)
 - Subfamily Victoriapithecinae (extinct early cercopithecids)
 - Parvorder Eocatarrhini (archaic catarrhines)
 - Family Propiopithecidae (early archaic catarrhines)
 - Family Pliopithecidae (later archaic catarrhines)
 - Infraorder Paracatarrhini (extinct early anthropoids)
 - Family Oligopithecidae (extinct archaic anthropoids)
 - Family Parapithecidae (extinct Egyptian monkeys)

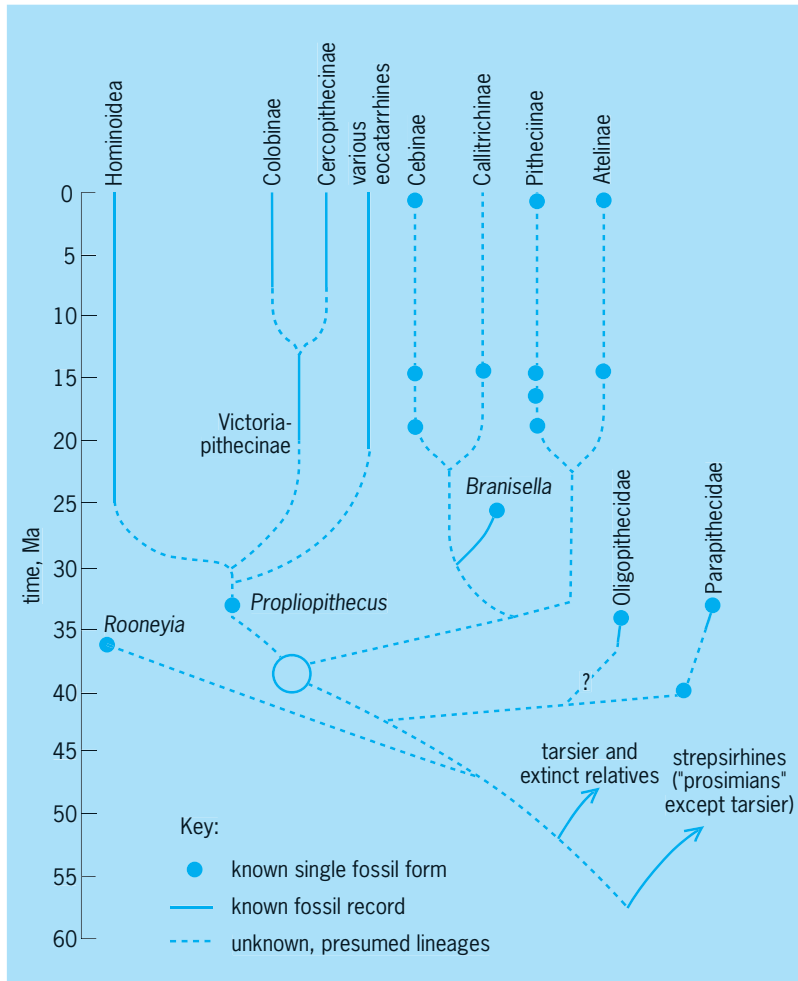


Fig. 1. Evolutionary relationships of monkeys among the primates.

Monkeys are hard to characterize as a group because of their great diversity, and because much of the discussion reflects a comparison with the apes. Both monkeys and apes contrast with the prosimian grade in that they are typically large, diurnal animals that live in social groups. Monkeys differ from apes in their possession of a tail, a smaller brain, quadrupedal pronograde (with long axis of body horizontal) posture, and a usually longer face (**Fig. 2**). They are generally smaller than apes, but large monkeys outweigh gibbons. Like almost all primates, monkeys are pentadactyl (**Fig. 3**), with nails rather than claws on the digits in most cases. They have pectoral mammary glands and well-developed vision. Monkeys are primarily vegetarian and inhabit forested tropical or subtropical regions of Africa, Asia, and South America. The differences between the New and Old World monkeys are summarized in the **table**.

Origins. The origin of the anthropoid or higher primates is still under debate in terms of the source group, area, and timing. The tarsier (*Tarsius*) is probably the living primate sharing the most recent common ancestry with anthropoids, and the tarsier-like extinct Omomyidae is accepted as a likely ancestral stock. The omomyids are known in all northern continents, especially between 55 and

| Major contrasts between New and Old World monkeys and special features of each | |
|--|---|
| New World species (Ateloidea) | Old World species (Cercopithecoidea) |
| Nose platyrrhine (nasal septum wide, nostrils opening to sides) | Nose catarrhine (septum narrow, nostrils opening downward) |
| Tail long, prehensile only in atelines and <i>Cebus</i> | Tail short to long, nonprehensile |
| 3 premolar teeth in each quadrant | 2 premolars in each quadrant |
| 24 deciduous, 36 permanent teeth (4 fewer in Callitrichinae), I 2/2 C 1/1 P 3/3 M 3/3/(M 2/2 in Callitrichinae) | 20 deciduous, 32 permanent teeth, I 2/2 C 1/1 P 2/2 M 3/3 |
| No ischial callosities | Ischial callosities present |
| Jaws and teeth lightly built in Cebidae; more robust in Atelidae, with deep lower jaw | Cheek pouches in Cercopithecinae; sacculated stomach in Colobinae |
| Opening on skull to internal ear circular but flat ("ringlike") | Opening extended into tube |
| Fingers and toes with curved nails (clawlike in Callitrichinae) | All nails tending to be flattened |
| Big toe opposable, thumb not fully so and sometimes reduced in Cebidae | Thumb and big toe opposable, thumb reduced in Colobinae |

35 Ma. At that time, South America and Africa were much closer, and the ancestors of New World monkeys may have traveled there from Africa on natural vegetation rafts, following paleocurrents across a 600-mi (1000-km) ocean gap. However, this has seemed unlikely to some researchers, because such rafted small primates would probably die of thirst or exposure and because known African fossils were not reasonable ancestors. Instead, it was hypothesized that anthropoid origins are from omomyids in North America or eastern Asia, regions that were connected during the Eocene (50–40 Ma at least) by a Bering land bridge. Early anthropoids might then have entered South America from Central America or the Caribbean by crossing narrow water gaps on rafts or by island hopping. New finds of protoanthropoid primates in northern Africa have reenergized the trans-Atlantic hypothesis, and both views are now strongly supported. Moreover, recent studies have suggested that the omomyids are not a close-knit group but may need to be broken into several distinct units; it is possible that the relatively late (36 Ma) Texas "protoanthropoid" *Rooneyia* might be the closest known relative of the common ancestor of all anthropoids listed above.

In China and Burma, several fossil sites 45–40 Ma old have yielded the remains of primates claimed to be protoanthropoids, but these are better interpreted as early strepsirhines or tarsier relatives. Others are known from North African localities dated about 55–40 Ma. Most of these, again, are either strepsirhines or tarsiiforms. Some authors have suggested that one or more of these newly recovered fossils represent a previously unknown group of early primates (unrelated to the omomyids) which gave rise directly to the anthropoids, while others consider them to be omomyid offshoots. Whichever hypothesis is supported, it is generally agreed that the most complete remains of early anthropoids are known from the Fayum area of northern Egypt.

In this region, three groups of fossil anthropoids have been found. The Parapithecidae (about 37–30 Ma) are monkeylike in adaptation and may be considered a third type of monkey that is not closely related to either living group. They have been suggested as being ancestral to Cercopithecoidea, but

are specialized in their own ways while lacking the different specializations of the cercopithecoids (or the ateloids; see table). For example, some parapithecid species had complex teeth with extra cusps, while one had no lower incisor (front) teeth. Probably, they occupied ecological niches later exploited by cercopithecoids, but became extinct without descendants. Previously, it was thought that they represented the earliest known catarrhines, but it now seems more likely that they were part of a radiation of "primitive" anthropoids which preceded the split between catarrhines and platyrrhines. It appears that a number of isolated teeth belonging to early parapithecids have been recovered from the 42–46-Ma

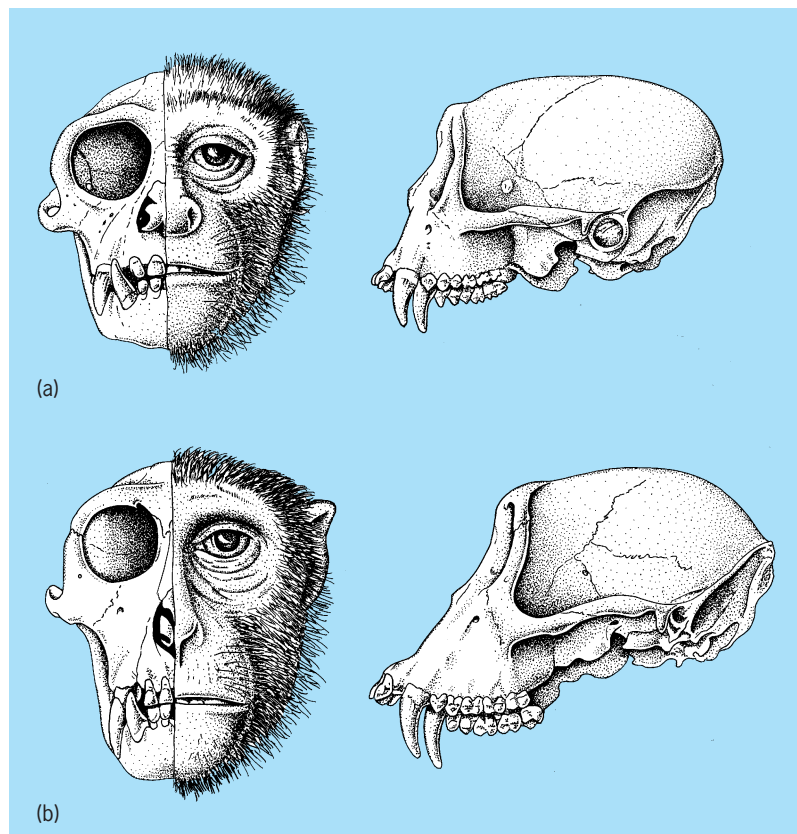


Fig. 2. Side and front views of the heads and skulls of adult male monkeys: (a) *Cebus* (New World) and (b) *Macaca* (Old World). (After A. H. Schultz, *The Life of Primates*, Universe Books, 1969)

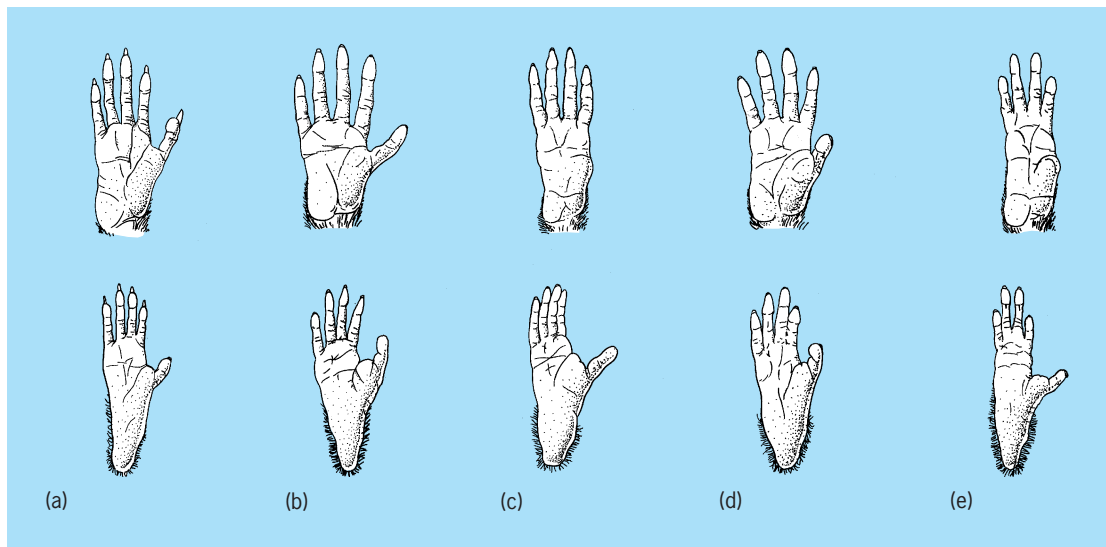


Fig. 3. Right feet (below) and hands (above) of (a) *Saguinus*, (b) *Cebus*, (c) *Ateles*, (d) *Macaca*, and (e) *Colobus*. (After A. H. Schultz, *The Life of Primates*, Universe Books, 1969)

site of Glib Zegdou, Algeria, which would make them the oldest recognized anthropoids.

Another group of Fayum primates which may belong to the same paracatarrhine radiation is the oligopithecids. Occurrence of these forms ranged around 36–34 Ma, and recently several partial damaged skulls have been recovered. They do not seem to have been monkeylike in adaptation and are really on the border of being anthropoids. Their teeth and lower jaw were conservative, but their skull presents the important anthropoid features of fusion in the middle of the frontal bone (forehead) and complete closure of the back of the orbit (eye socket). It is interesting that the oligopithecids are far more “primitive” than the contemporaneous (and even much older) parapithecids, suggesting that a yet-unknown species of oligopithecid must have lived even earlier than the first parapithecids, perhaps about 45–48 Ma.

The third group of Fayum primates is the Propithecidae (32–30 Ma). These species are monkeylike but also in some ways almost apelike; they may well have been close to the common ancestors of cercopithecoid monkeys and hominoids (apes and humans). Although some authors have suggested that the oligopithecids are close relatives of the propithecids, this appears unlikely—their similarities are convergent (independent evolutionary developments) rather than homologous (inherited from a recent common ancestor).

Old World species. Today, these animals are found throughout all the warmer regions of the Eastern Hemisphere except Australia and Madagascar. Many of the familiar monkeys are included in this family, such as the rhesus macaque, Barbary “ape,” proboscis monkey, baboon, and mandrill. The earliest known representatives of the Cercopithecidae, *Prohylobates* and *Victoriapithecus* (the Victoriapithecinae), are from the early Miocene in north and east Africa, some 20–12 Ma. They probably diverged from

their apelike relatives by adapting to a relatively more folivorous (leaf-eating) diet, involving reshaping of the dental chewing surfaces, and also by becoming more terrestrial. Victoriapithecines are often considered to be a common ancestral stock for all later cercopithecids, but new studies suggest they might instead be closely related to the cercopithecines; these alternative hypotheses need to be carefully evaluated.

By the late Miocene, some 10–9 Ma, colobines were present in east Africa (*Microcolobus*) as well as in Europe (*Mesopithecus*), where one line (*Dolichopithecus*) became highly terrestrial before dying out late in the Pliocene. Colobines probably entered Asia in the later Miocene as well, and many fossil forms are known in the African Plio-Pleistocene. Living (and presumably most extinct) members of this subfamily are primarily arboreal leaf eaters and have sharp teeth and specialized stomachs like ruminants to process difficult-to-digest leaf protein. The dispersal of the cercopithecines may have been somewhat later, with the oldest macaque-like forms appearing in north Africa about 7 Ma, whence they spread to Europe and Asia before 5 Ma. In sub-Saharan Africa, possible ancestors of the baboons and mangabeys (the extinct *Parapapio*) are present from the Pliocene, and early relatives of the modern geladas (*Theropithecus*) were widespread in the Pliocene and Pleistocene. The *Cercopithecus* group probably separated from the macaques and baboons in the late Miocene and entered the high forest, but fossils are very rare. All cercopithecines have cheek pouches for temporary food storage, and many are terrestrial.

Macaques. About 12–16 species of the genus *Macaca* are known as macaques, and with the exception of *M. sylvanus*, the Barbary “ape,” all are found in southern Asia. The Barbary ape, so called because it lacks an external tail like true apes, occurs in the wild in Algeria and Morocco, and is the only monkey now found in Europe, where it

was introduced to Gibraltar. It is thought to be the remains of a population that originally extended from Britain to the Caucasus, north Africa, and southwestern Asia. *Macaca sylvanus* is a large, robust, and rather terrestrial species with a thick coat that protects it from low temperatures in its natural habitat, where it roams through cedar forests and mountain gorges in large troops. Its diet consists of fruit, leaves, and other vegetation, in addition to small invertebrates such as insects; in fact, this diet is typical of most cercopithecine monkeys. Because these animals raid farm areas and beg for food, they have become a nuisance on Gibraltar. Their gestation period is about 30 weeks, with a single young born almost annually, as in most monkeys. The adult male stands over 2 ft (66 cm) at the shoulder and weighs up to 40 lb (18 kg).

Macaca mulatta is the rhesus monkey, which is found throughout southern continental Asia. Large numbers were previously exported from India each year for medical research, but restrictions have been adopted to protect the species. These monkeys have been used in studies on the effects of space travel and in the development of polio vaccine. The rhesus, or Rh, blood group was so named because the antigen was first found in the red blood cells of rhesus macaques. Further genetic studies have shown that the eastern rhesus of China is in some features more closely related to the macaques of Taiwan and Japan than to the western rhesus of India. This suggests that an offshoot of the eastern rhesus may have given rise to the island populations (now separate species) within the past million years while still remaining part of the same species as the rhesus of India. This is one of the few cases of one living species of primate (or any mammal) being the ancestor of another.

The rhesus is an agile, gregarious species with a short tail, robust limbs of almost equal length, and a stocky build. It is considered sacred in some parts of the Indian subcontinent and has become a scavenger in towns and cities, as well as living in larger troops in its natural forested habitat. Troops of 15–40 are led by an older, experienced male who is dominant to the others in terms of access to food, mates, and living space. These animals do not defend a true territory, but groups generally avoid each other. Macaques have large cheek pouches, which they fill with food. They then wander about eating directly, and when danger approaches, they scatter to the nearest place of safety with a ready supply of food. The adult male is about 2 ft (66 cm) high; he is sexually mature at 4 years and full grown at 5 years, with a maximum lifespan of about 20 years. Mating is promiscuous within a troop, with the dominant male probably siring most offspring. The gestation period is about 23 weeks. On average, males weigh 18 lb (8 kg) and females 11 lb (5 kg).

The natural history of some of the other species of macaques is less well known, although studies are proliferating. Some live at high altitudes, and one (the Japanese *Macaca fuscata*) roams in winter snows. The lion-tailed macaque (*M. silenus*) of southern India is one of the least well known and

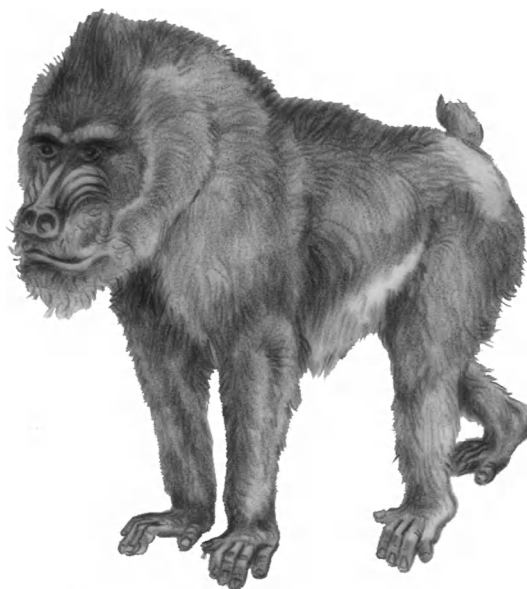


Fig. 4. Mandrill, a large Old World cercopithecoid monkey.

least typical of the group, being a shy animal that inhabits dense forests. The stump-tailed macaque (*M. arctoides*) is found at high altitudes from Assam to southeastern China. It has long fur, and the naked face is brightly colored, so it is also known as the red-faced monkey. Some researchers have suggested that it is the evolutionary result of hybridization between two other living species.

Mandrills, baboons, and geladas. These are terrestrial animals with long canines in the males, a long muzzle with terminal nostrils, and a great disparity in size between the sexes. Mandrills (*Mandrillus sphinx*) are found on the forest floor of west-central Africa. The males have brightly colored red and blue facial markings for recognition by subordinate troop members (Fig. 4). In both sexes, the body is robust, the legs stout, and the tail short. Males weigh up to 77–99 lb (35–45 kg), making them the heaviest of all extant monkeys; females weigh only about one-third as much as large males. Mandrills thus have the highest ratio of sexual dimorphism of all primates. The closely related drill (*M. leucophaeus*) is smaller and less brightly colored than the mandrill but otherwise quite similar. Troop size varies from a dozen (with one adult male) to hundreds.

Typical or “savannah” baboons of the genus *Papio* include at least five or six subspecies of *P. hamadryas*. They were previously thought to represent separate local species, but studies of contact zones have revealed significant interbreeding and hybridization. As full species are defined to be reproductively isolated, the baboon groups are now usually considered only subspecies. *Papio hamadryas* inhabits a “7”-shaped belt of mainly savannah and open woodland from Guinea to Ethiopia (and coastal Arabia) to south Africa. The body and limbs of *Papio* baboons are more slender than the mandrill’s (Fig. 5), and body size ranges widely: adult males average 29–64 lb (16–29 kg) in the various subspecies, while female averages are 22–34 lb



Fig. 5. Olive baboon (*Papio hamadryas anubis*). (© California Academy of Sciences; photograph by Dr. Robert Thomas and Margaret Orr)

(10–16 kg). Troops are often led by a coalition (central hierarchy) of high-ranking males and contain 30–200 members of all ages. In some varieties, especially *P. b. hamadryas*, large troops sleep together but divide during the day into one-male “harems” of 4–14 animals for feeding. Other varieties of social organization may depend on local environmental conditions. *Theropithecus*, commonly called the gelada baboon, is a very distinctive genus; its fossil record goes back to about 4 Ma. Today, it is restricted to the Ethiopian highlands, where it lives in one-male groups on rocky gorges. It eats mostly grasses.

Mangabeys. There are five species in this group, all from central Africa. Two sets of species are recognized, the *torquatus* and *albigena* groups, which were long placed in the genus *Cercocebus*. One feature which linked them to each other is the presence of deep depressions or fossae in the bone below the orbit (eye socket). Recent studies of molecular biology as well as of limb bones and teeth have argued that these two groups are not, in fact, each other’s closest relatives, and suggested persuasively that they be placed in separate genera. *Lophocebus albigena* and relatives are almost entirely arboreal, with long tails and long fore- and hindlimbs that make them excellent climbers. They are thought to be evolutionarily linked to the baboons and geladas. *Cercocebus torquatus* and relatives may be more terrestrial, using their strong forelimbs to search through leaf litter, much as do their close relatives, the mandrills. Both groups of mangabeys live in troops with one or several male leaders and are frugivorous to omnivorous. It is not yet clear whether the suborbital fossae found in mangabeys were evolved independently or inherited from a common ancestor whose other cercopithecine descendants lost this feature.

Guenons. These are the dozen or so species of monkey in the genus *Cercopithecus* and allies. The mona monkey (*C. mona*) is typical of guenons found in the forested areas of west and central Africa. Sev-

eral species may inhabit a single type of tree, avoiding competition by vertical spacing and slightly different diets, including fruits, some leaves and insects, as well as tree snails. These more arboreal guenons hardly ever come to the ground; they are agile climbers and leap between treetops. Most groups have a harem social organization, with additional males peripheralized. The vervet (*C. aethiops*) is one of the most common of all primates in Africa, where it is widely distributed throughout the grasslands, being equally at home on the ground and in trees. Its yellow-brown fur often has a greenish tinge, and the males engage in a “red, white, and blue” display of the perineal region. Recent studies of DNA (supported by revised work on limb bones) have demonstrated that the partly terrestrial vervets and Lhoest’s guenons are in fact close evolutionary relatives of the even more highly terrestrial patas monkey (previously placed in the genus *Erythrocebus*). As a result, these three animals are now placed in a separate genus, *Chlorocebus*, which may in turn be divided into subgenera for each species [for example, *Chlorocebus (Erythrocebus) patas*]. The tiny talapoin (*Miopithecus*) is the smallest Old World monkey, with most males and females averaging only 3 lb (1.3 kg). The little-known Congo forest form *Allenopithecus* (the swamp monkey) is in some ways a link between the guenon and the baboon-macaque groups; it retains some of the features of their common ancestor also still seen in the latter species while presenting a number of the special characters of the guenons as well.

Langurs. These animals are slender, long-tailed members of the subfamily Colobinae, which are found in forested areas of southern Asia and Indonesia. They may live in large troops or small ones with a single male leader. Two genera of “typical” langurs are now recognized: *Presbytis* for the smaller, mainly island species, and *Semnopithecus* (including *Trachypithecus*) for the larger, mostly mainland forms. Seven species of *Presbytis* occupy peninsular Malaysia, Sumatra, Borneo, eastern Java, and many smaller islands such as the Mentawaiis. They are excellent leapers and show little size difference between the sexes. *Semnopithecus* species range from Sri Lanka and India into China, through Malaysia and onto Sumatra, Java, Borneo, and some other Indonesian islands. They tend to leap less and are more quadrupedal, with moderate to strong sexual size dimorphism. Eight to ten of these species are placed in the subgenus *Trachypithecus*, while the subgenus *Semnopithecus* includes the Indian langurs, especially *S. entellus*. This is probably the best-known species of the group and the most terrestrial of living colobines. It superficially resembles the spider monkey but of course lacks the prehensile tail. The entellus langurs are common in India up to altitudes of 13,000 ft (4000 m) and are considered sacred by Hindus. It has been observed that males of *P. entellus* and other langurs may kill infants in a troop when ousting a previous leader male. The “adaptive” versus “aberrant” nature of this pattern is debated.



Fig. 6. Black-and-white Colobus monkey.

An unusual Asian colobine is the proboscis monkey (*Nasalis larvatus*) found only in Borneo. This is a powerfully built animal that stands about 2.5 ft (83 cm) tall; weight of a large male averages 44 lb (20 kg). The unusual nose, which may function as a resonating chamber during male calling, is a swollen pendant structure up to 3 in. (7.6 cm) long that, at rest, hangs down over the mouth. These animals live in troops of about 20, feeding on mangroves and palm leaves. They are expert swimmers and have been observed far out to sea. Other related species have upturned noses, such as *N. (Simias) concolor* of the Mentawai Islands and *Pygathrix (Rhinopithecus) roxellana*, the snub-nosed monkey, which is an inhabitant of Tibet and southern China and may spend part of its life in regions of perpetual snow. The douc langur (*Pygathrix nemaeus*) of Indochina completes the roster of the group often called the “odd-nosed” colobines.

Colobus. In Africa the Colobinae are represented by two genera, one of which has two marked subgenera. All are forest-dwelling leaf eaters with rudimentary external thumbs (those of the Asian colobines are short but present). The black-and-white *Colobus* (Fig. 6) has five species, including *C. polykomos* and *C. guereza*, which have long been hunted for their magnificent pelts. They live in troops of 5–15 with a single male leader and defend a small territory. The genus *Procolobus* includes the so-called red and olive colobus monkeys. *Procolobus (Piliocolobus) badius* and related species, the red colobus monkeys, does not exhibit territorial behavior but live in multimale groups of 20–50 in a larger home range and have more varied dietary preferences. All these forms range across equatorial Africa, and there is much uncertainty about how many species (or subspecies) to recognize. Less is known of the much smaller olive colobus, *P. verus*, restricted to west Africa.

New World species. The New World monkeys, or ateloids, occupy forested areas from southern Mexico to Argentina. They are divided into two main groups, or families (their major characteristics are given in the table). All are arboreal, including a few with prehensile tails; there is no living form, or any evidence of a fossil form, that has come to the ground habitually. The fossil record of platyrrhines is relatively limited, but documents early diversification

of the main lineages, as opposed to the successive replacements seen among catarrhines. The oldest known ateloid, *Branisella*, dates back 26 Ma and may be related to the Cebidae. A variety of monkeys inhabited southern Argentina between 16 and 20 Ma, including pitheciines related to owl monkeys (*Tremacebus*) and titis (*Homunculus*) and two cebines, one apparently close to squirrel monkey ancestry (*Dolichocebus*), and the more “generalized” *Killikaike*, described in 2006 from a partial cranium with a relatively large brain. *Chilecebus* was found in 20-Ma deposits in northern Chile and may be a relative of *Dolichocebus*. In Colombia, 13-Ma deposits at La Venta have yielded forms very close to the living night, squirrel, saki, spider, and howler monkeys and to marmosets. Brazilian cave deposits dated between 100,000 and 10,000 years ago yielded fossils of two large ateline species in the 1990s. Two or three distinctive genera also inhabited the islands of Jamaica, Cuba, and Hispaniola during the past 10,000 years.

Marmosets. The subfamily Callitrichinae consists of a number of small-sized species of four or five genera. Body weight (for males and the similarly sized or even slightly larger females) ranges across many species from 1 to 4.5 oz (110 to 620 g), making callitrichines the smallest of all living anthropoid primates. Most eat insects and fruits, and live in family, or pair-bonded, groups of male, female, and offspring. Births may occur twice a year, and twins are common, so that up to eight young may be born in the 2 years the offspring spend with their parents. Both sexes share equally in the care of the young, the male especially carrying them when traveling. Several forms, particularly the brightly colored golden lion marmoset (*Leontopithecus rosalia*), are severely endangered species. The marmosets were previously thought to be very primitive forms (with claws rather than nails and simple teeth) uniquely separate from all other ateloids. However, research has shown them to be specialized sap feeders, with secondarily evolved clawlike nails and reduced dentition. They share common ancestry with the cebines, which has resulted in the classification of Cebidae and Atelidae as employed here.

Capuchins. The family name Cebidae comes from the capuchin monkey, *Cebus*, which lives in forested areas from Honduras to Argentina in troops of up to 30 individuals (Fig. 7). Capuchins are extremely agile leapers and runners in their arboreal habitat. They are omnivorous, feeding on small birds, insects, grubs, leaves, and fruits, and often raid plantations for fruit and grain. Studied in captivity, they have been observed to use simple tools, make drawings, and paint. They have a partly prehensile tail, but it is rather different from that of the atelids.

The squirrel monkeys (*Saimiri*) are close relatives and some of the most widely distributed animals of South America. They are relatively small animals, living in troops of up to 100 individuals, apparently “dominated” by females, with older males peripheralized.

Titis. Three species of *Callicebus* are found in South America in the region of the Amazon, where they live in the finer branches of the high forest trees. These animals (male or female) weigh about 2–3 lb (0.8–1.3 kg), are omnivorous, and live in pair-bonded groups of five or six. They are strongly territorial, and the adults are very dependent on one another, often sitting (or sleeping) side by side on a branch with tails intertwined. A close relative is *Aotus*, the owl or night monkey, the only nocturnal platyrrhine (or anthropoid, for that matter).

Sakis and uakaris. This group of generally little-known pitheciines comprises fruit-eating rainforest dwellers. The sakis have long hair and a bushy tail. There are four species, found in the Guianas and the Amazon basin: *Pithecia monachus*, the hairy saki; *P. pithecia*, the pale-headed saki; *Chiropotes satanas*, the black-nosed saki; and *C. albinasus*, the white-nosed saki. Sakis do not thrive in captivity, and have been known to die simply from the shock of being moved from one cage to another.

The three species of uakari, genus *Cacajao*, are relatively rare and occur in restricted ranges. They are found high in the tops of trees and rarely descend to the ground. The head and face are bright scarlet, and the tail is short and of little use as a balancing organ.

Howler monkeys. All six species of the howler are included in the genus *Alouatta*. They range from South America into Central America and are among the largest of the New World monkeys. Males reach 15–22 lb (7–10 kg). They are primarily leaf-eaters that can hang by the prehensile tail to reach food, much of which they drop after eating partially, and they spend most of their day resting. This is now seen to be an adaptation to processing the large quantity of toxic secondary compounds in their food. Plants which taste bad (that is, are potentially poisonous) may be discarded, while long rest periods allow the howler's stomach to detoxify the hard-to-digest leaf proteins consumed. Their most unusual characteristic is the howling voice, which is perhaps used to space the troops of 20 to 30 apart from each other. It can be heard miles away. The enlarged, cavernous vocal apparatus is a trumpetlike bony box that is accommodated by the enlarged lower jaw and throat.



Fig. 7. Capuchin monkey (*Cebus*).



Fig. 8. Red spider monkey (*Ateles* sp.). (© California Academy of Sciences; photograph by Dr. Lloyd Glenn Ingles)

Spider and wooly monkeys. Comprising three genera, these monkeys also have prehensile tails but often lack an external thumb. Four species of spider monkey (*Ateles*; Fig. 8) are often found alongside howlers, but they are more active and concentrate on eating fruit, insects, and fewer leaves. They range between Mexico and Bolivia in groups of 15–50, with either one or several adult males. The wooly monkeys (*Lagothrix*) are like heavysset spider monkeys but do not lack a thumb, while the rare wooly spider monkey (*Brachyteles*) is similarly built and lacks a thumb, as do the spiders. Eric Delson

Bibliography. G. Davies and J. F. Oates (eds.), *Colobine Monkeys: Their Ecology, Behaviour and Evolution*, 1994; E. Delson et al., Body mass in Cercopithecidae (Primates, Mammalia): Estimation and scaling in extinct and extant taxa, *Anthropol. Pap. Amer. Mus. Nat. Hist.*, 83:1–159, 2000; E. Delson et al. (eds.), *Encyclopedia of Human Evolution and Prehistory*, 2d ed., 2000; J. G. Fleagle, *Primate Adaptation and Evolution*, 2d ed., 1998; C. Groves, *Primate Taxonomy*, 2001; W. C. Hartwig (ed.), *The Primate Fossil Record*, 2002; N. G. Jablonski (ed.), *The Natural History of the Doucs and Snub-nosed Monkeys*, 1998; N. G. Jablonski (ed.), *Theropithecus: Rise and Fall of a Primate Genus*, 1993; W. G. Kinzey (ed.), *New World Primates: Ecology, Evolution and Behavior*, 1997; R. M. Nowak, *Walker's Mammals of the World*, 6th ed., 1999; N. Rowe, *The Pictorial Guide to the Living Primates*, 1996; R. J. Smith and W. L. Jungers, Body mass in comparative

primatology, *J. Hum. Evol.*, 32:523–559, 1997; F. S. Szalay and E. Delson, *Evolutionary History of the Primates*, 1979; P. F. Whitehead and C. J. Jolly (eds.), *Old World Monkeys*, 2000.

Monocleales

An order of liverworts of the subclass Marchantiidae, consisting of a single genus (*Monoclea*). The gametophyte is among the largest of all liverworts; its thallus consists of homogeneous cells except for scattered oil cells. The rhizoids are smooth and both thick-walled and thin-walled. The sex organs are grouped but not elevated: the antheridia are sunken in cavities and grouped into receptacles, while the archegonia are enclosed in groups by involucre. The very long, massive seta considerably elevates the capsules which dehisce spoonlike by one slit. The lobing of spore mother cells is unique in the subclass Marchantiidae. See BRYOPHYTA; HEPATICOPSIDA; MARCHANTIIDA.

Howard Crum

Bibliography. E. O. Campbell, The structure and development of *Monoclea forsteri* Hook, *Trans Roy. Soc. N. Z.*, 82:237–248, 1954; Looking at *Monoclea* again, *J. Hattori Bot. Lab.*, 55:315–319, 1984.

Monoclonal antibodies

Antibody proteins that bind to a specific target molecule (antigen) at one specific site (antigenic site). Monoclonal antibodies are produced by a laboratory culture of cells (a cell line), grown from a single antibody-secreting cell. Development of the technology for making monoclonal antibodies has had an impact in both science and medicine; applications include experiments in basic immunology, medical diagnostics, and therapeutics.

Production. A cell line that produces monoclonal antibodies is created by a combination of immunological and tissue culture techniques (see **illus.**). First, a foreign substance (antigen) is injected into a mouse. The mouse immune system reacts to the antigen by producing antibodies that bind to it. An antibody is a protein molecule which binds to a specific conformational feature (antigenic site) of another molecule via a variable binding region. Immunization with a complex antigen results in the production of a number of different antibodies binding to its different conformational features (antigenic sites). In the spleen each antibody-producing cell makes only one type of antibody, which reacts with a single antigenic site. An immunized mouse will have many antibodies in the serum component of its blood (antiserum) which react with many different features of a complex antigen. Its spleen will contain as many different antibody-producing cells, each making one type of antibody with a single binding specificity for one conformational feature of the antigen. If these spleen cells were removed from the mouse and cultured in the laboratory with nutrients, they would survive

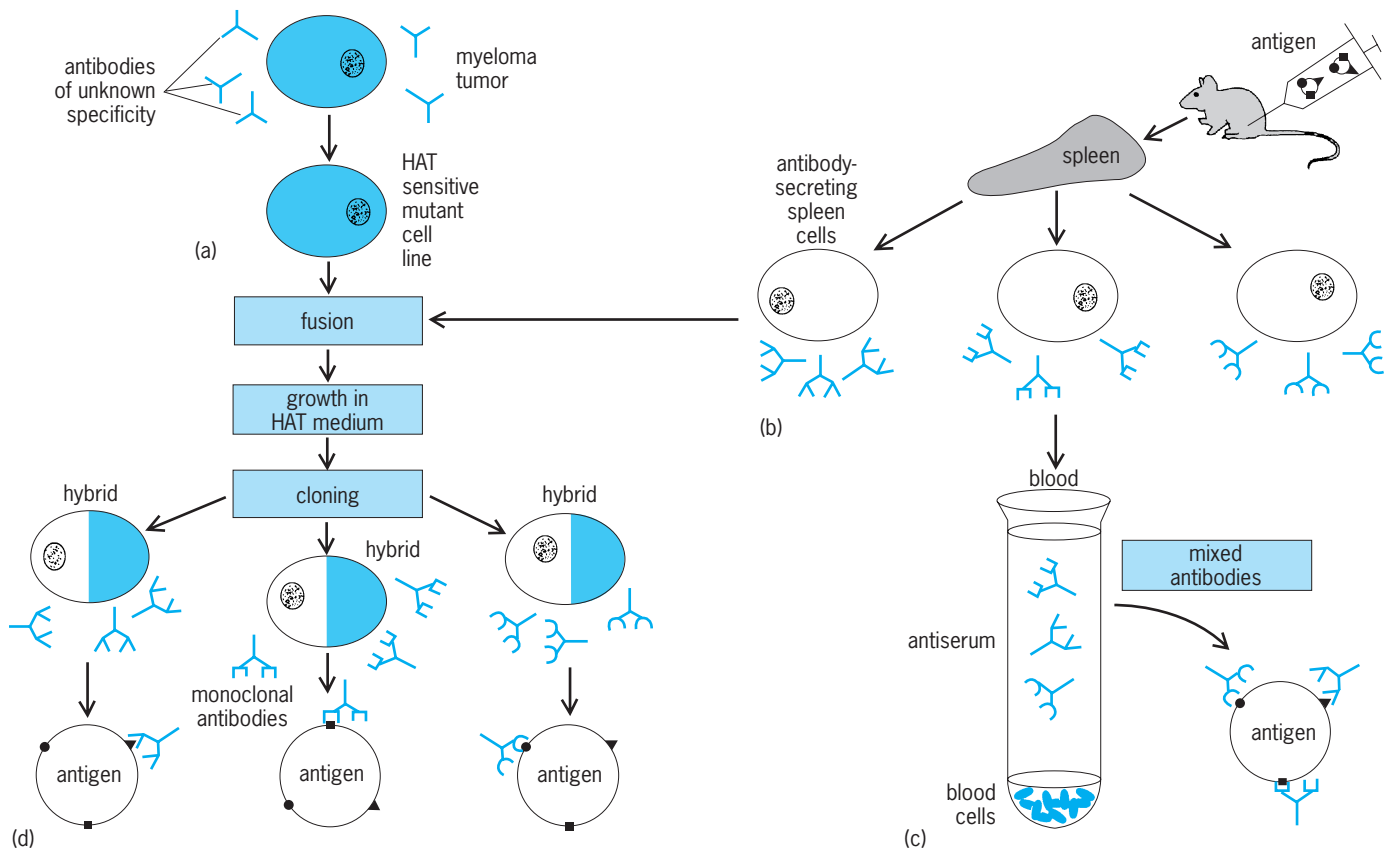
only for a few days. In 1975 G. Köhler and C. Milstein developed a technique for immortalizing spleen cells in laboratory culture and simultaneously retaining their antibody secretion ability. This allowed the isolation and culture of individual antibody-secreting cells and the subsequent production of monoclonal antibodies. Köhler and Milstein were awarded the Nobel prize in medicine in 1984 for this contribution.

Köhler and Milstein were working with an immortal cell line, grown from a mouse myeloma tumor which consisted of cancer cells that produced antibody of unknown specificity. They found that when a cell from the myeloma cell line was combined by fusion with a spleen cell from an immunized mouse, a unique type of hybrid cell (hybridoma) was formed. The hybridoma secretes antibody with binding specificity of the antibody made originally by the parent spleen cell. A hybridoma also retains the immortal growth properties of the parent myeloma cell line. Isolation of such hybrid cells is facilitated by a mutational defect in the myeloma cell line which prevents it from growing in a selective medium unless it can use metabolic enzymes provided by a normal cell fused to it. Growth in selective medium to isolate hybrid cells is a classic tissue culture technique, developed by J. Littlefield.

After antibody-secreting hybrid cells are produced, they are diluted in growth medium so that individual hybrid cells arising originally from individual spleen cells are separated from each other and will grow into individual colonies, called clones. Clones are then maintained as separate cultures, each producing only one type of antibody with specificity for a single antigenic site. The antibodies produced by these clones are monoclonal antibodies.

Applications. The ability to produce monoclonal antibodies has revolutionized immunology and the use of immunological techniques in other disciplines. Previously, antibodies could be isolated only from serum of an immunized animal, and analysis of an antibody response reflected the combination of activities and properties of all the different antibodies in that antiserum. But by making monoclonal antibodies from an immunized animal, it has been possible to analyze individual molecular properties of antibodies produced in a single immune response and to understand more about the mechanism by which antibodies are generated.

Production of monoclonal antibodies has also made it possible to take advantage of the exquisite specificity of individual antibodies. In cases where antisera had been used to identify a complex antigen, a monoclonal antibody can be used more confidently and precisely. With a monoclonal antibody, there is never the question of whether more than one antigenic site is being detected. In addition, there is a virtually unlimited supply of monoclonal antibody from cultured hybridoma cells, whereas antiserum from animals never has the same composition of antibodies, even when an immunization protocol is reproduced exactly. Cells are among the most complex antigens which have been used to elicit antibodies,



Monoclonal antibody production. (a) A myeloma-derived cell line is used to select a mutant that cannot metabolize the nutrients in HAT (hypoxanthine, aminopterin, thymidine) medium. The mutant cell line is fused with (b) an antibody-secreting spleen cell. (c) Antibodies from the spleen cells are secreted into the serum component of blood, forming an antiserum. (d) After fusion, a monoclonal antibody-producing cell line (hybrids) results. (After C. Milstein, *Monoclonal antibodies*, *Sci. Amer.*, 243:66-74, 1980)

and classifying cells by using antisera has been difficult. Monoclonal antibodies reacting with known molecular targets on the surface of cells have helped to distinguish between different types of cells when other criteria have failed. For example, they have been particularly useful in identifying functional subsets of leukocytes.

The potential ability of monoclonal antibodies to recognize structures specific to many different cell types has heightened research interest in the "magic bullet" concept. An antibody with specificity for a particular cell could possibly be used to deliver a toxic substance or nutrient to that cell without affecting surrounding cells. For this reason, monoclonal antibodies may become useful in cancer therapy. The possibility of using monoclonal antibodies for therapy has raised the question of whether the mouse-derived or rat-derived monoclonal antibodies will be tolerated by the human system. Early experiments with mouse monoclonal antibody therapy in humans demonstrated that the recipient does react to the administered antibody. For this reason, laboratories have attempted to produce human-derived monoclonal antibodies, but with only moderate success; no human equivalent of the mouse myeloma parent cell line has been found. There has been some success through using several techniques to immortalize

human antibody-secreting cells, but it is difficult to isolate human cells secreting antibody to a selected antigen, given the necessary limitations on immunizations.

While the use of monoclonal antibodies for therapy is still in early phases of development, their application in diagnostics has grown. Many diagnostic tests use antisera to detect cells, infectious organisms, or levels of blood and fluid components (such as hormones and proteins). Appropriate monoclonal antibodies can be used instead of an antiserum in all such tests, with the advantage of unlimited supply and greater reproducibility and specificity. See ANTIBODY; ANTIGEN; IMMUNOLOGY; TISSUE CULTURE.

Frances M. Brodsky

Bibliography. G. Köhler and C. Milstein, Continuous cultures of fused cells secreting antibody of predefined specificity, *Nature*, 256:495-497, 1975; J. W. Littlefield, Selection of hybrids from matings of fibroblasts *in vitro* and their presumed recombinants, *Science*, 145:709-710, 1964; C. Milstein, Monoclonal antibodies, *Sci. Amer.*, 243(4):66-74, 1980; J. H. Peters and H. Baumgarten (eds.), *Monoclonal Antibodies: A Practical Guide*, 1992; M. A. Ritter and H. M. Ladgman (eds.), *Monoclonal Antibodies: Production Engineering and Clinical Application*, 1994.

Monocotyledons

This group of flowering plants (angiosperms), with one seed leaf, was previously thought to be one of the two major categories of flowering plants (the other group is dicotyledons). However, deoxyribonucleic acid (DNA) studies have revealed that, although they do constitute a group of closely related families, they are closely related to the magnoliids, with which they share a pollen type with a single aperture. The other dicots, the eudicots, are much more distantly related. In general, monocots can also be recognized by their parallel-veined leaves and three-part flowers. Their roots have disorganized vascular bundles, and if they are treelike (yuccas, aloes, and dracaenas) their wood is unusually structured. Among the important monocots are grasses (including corn, rice, and wheat), lilies, orchids, palms, and sedges. See DICOTYLEDONS; EUDICOTYLEDONS; FLOWER; GRASS CROPS; LILIALES; MAGNOLIOPHYTA; ORCHID.

Mark W. Chase; Michael F. Fay

Monogenea

A class of the Platyhelminthes which are ectoparasites of the gills, skin, and orifices of fishes and, less frequently, of the esophageal tracts and bladders of amphibians and turtles. They have conspicuous anterior and posterior holdfasts, the latter usually armed. The terminal genitalia are frequently sclerotized. The

group is characterized by sexual reproduction, direct development, and a single host in the life cycle. Egg capsules have terminal filaments.

Taxonomy. The distinctive shapes and sclerotized holdfasts facilitate characterization and arrangement (Fig. 1). The most widely used classification employs two subclasses, the Monopisthocotylea, in which the posthaptor is without discrete multiple suckers or clamps, and the Polyopisthocotylea, with suckers or clamps on the posthaptor.

Morphology. Body shapes of the various genera are distinctive, sometimes bizarre, as in *Vallisia*, which is sickle-shaped. Paired external suckers or buccal cavity suckers and adhesive glands occur anteriorly. The posterior holdfast is either solid and armed with central anchors and marginal hooks (Gyrodactyloidea), sucker-shaped with anchors and hooks (Capsaloidea), or solid and bearing suckers or clamps (Polyopisthocotylea). The pharynx is muscular and protrusible, and the esophagus is short, often ramified. Multiple testes and numerous vitelline glands are usually present. A genitointestinal canal occurs in many, and the penes are usually armed.

Life history. Monogenea usually have direct development involving simple metamorphosis from the ciliated larval stage to the nonciliated juvenile. Juvenile anchors and hooks may be retained or replaced by adult suckers or clamps. Cross-fertilization or, perhaps less frequently, self-fertilization of hermaphroditic individuals resulting in egg capsules which hatch on the host or in its

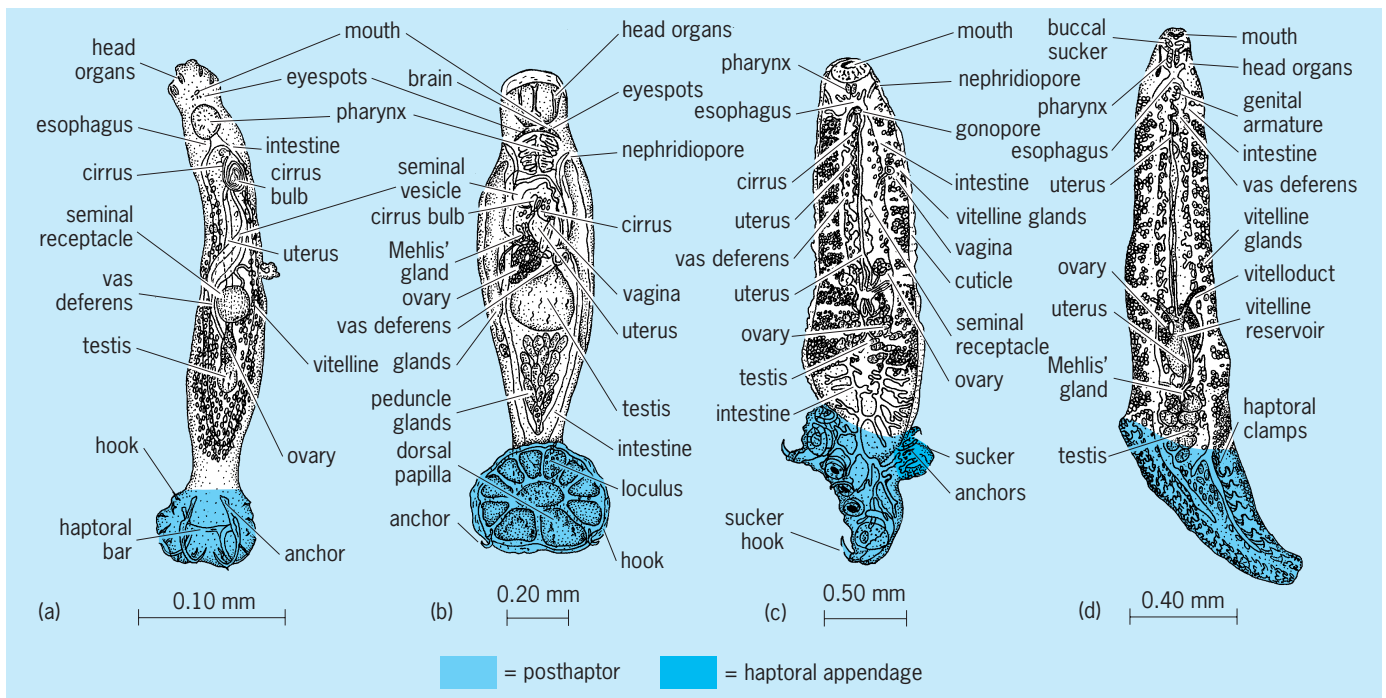


Fig. 1. Representative monogeneids. (a) Superfamily Dactylogyroidea, *Pseudohaliotrema carbunculus* from the pinfish (*Lagodon rhomboides*), shown in dorsal view. (b) Superfamily Capsaloidea, *Heterocotyle aetobatis* from the spotted eagle ray (*Aetobatis narinari*), ventral view. (c) Superfamily Polystomatoidea, *Heteronchocotyle leucas* from the bull shark (*Carcharhinus leucas*), ventral view. (d) Superfamily Microcotyloidea, *Heteraxinoides xanthophilis* from the spot (*Leiostomus xanthurus*), ventral view.

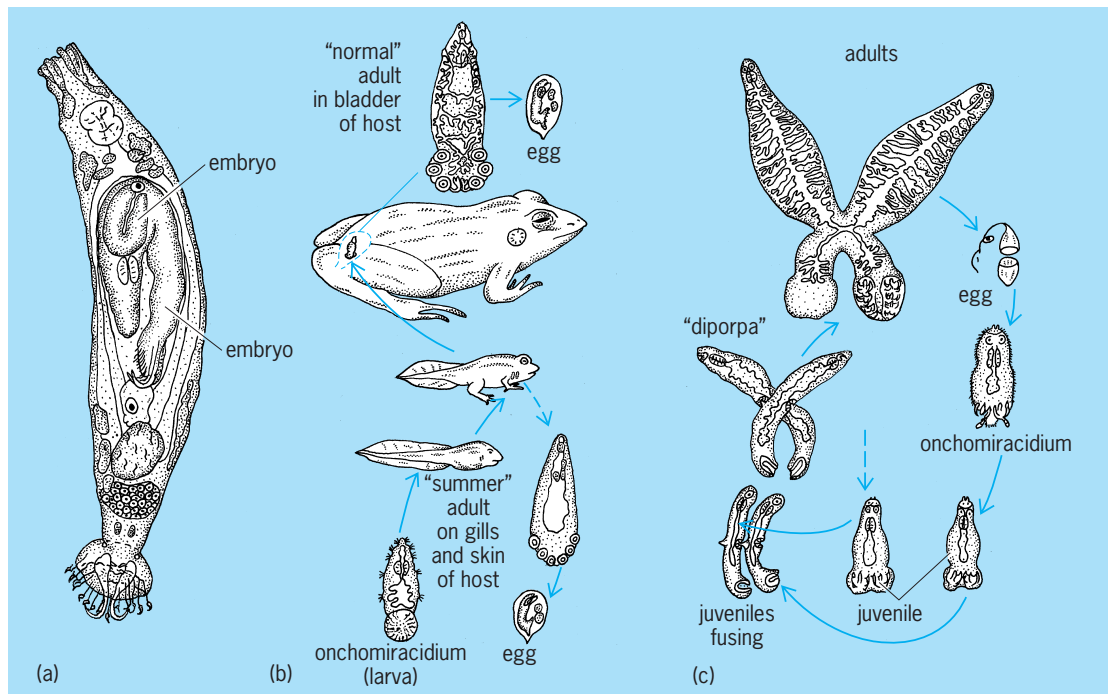


Fig. 2. Three unusual monogeneid life cycles. (a) *Gyrodactylus funduli* from the longnose killfish (*Fundulus similis*), dorsal view. In the uterus of the viviparous parent is an embryo within an embryo, thought by some to be the result of polyembryony. (b) *Polystoma integerrimum* from the frog bladder. The larva on tadpole gills grows as the tadpole develops. If, when the tadpole metamorphoses, the juvenile fluke does not enter the esophagus of the host, whence it can travel to the bladder, it matures neotenually into a "summer" adult on gills and skin of host. If it does reach the frog's bladder, the fluke matures normally and reproduces in the third year. (c) *Diplozoon* sp. The larvae undergo simple metamorphosis into nonciliated juveniles that attach to each other and become fused "diporpa larvae." Fusion of reproductive organs becomes complete as the flukes mature, and cross-fertilization occurs as the occasion demands.

environment is most common; however, interesting exceptions exist (Fig. 2).

Physiology. Monogeneids were at one time thought to feed on mucus, but there is good evidence that some are blood feeders. Most live in well-aerated microhabitats and probably require more oxygen than do entoparasites. The male and female organs of some species mature at different times, but the reproductive significance is uncertain.

Ecology. Monogeneids are capable of delicate ontogenetic responses to the microhabitat; for example, the direction of body asymmetry often depends upon which side of the host the worm inhabits. The direct life history seems to favor a higher degree of host specificity than in other trematodes. Often a parasite occurs on a single host species, and closely related fishes bear closely related parasites.

Though usually well accommodated by their hosts, monogeneids may cause intense discomfort and wasting and pave the way for invasion by bacteria, myceliated fungi, and protozoans. Little is known of their effects on healthy, natural populations. Heavy infections cause serious mortality in hatcheries, aquariums, and culture ponds, and sometimes occur in nature. See ASPIDOGASTREA; DIGENEA; PLATYHELMINTHES; TREMATODA.

William J. Hargis
Bibliography. N. G. Sproston, A synopsis of the monogenetic trematodes, *Zool. Soc. (London)*, 25(4):185-600, 1946.

Monogononta

A class of the phylum Rotifera which contains the majority of species in this invertebrate class. The organisms of this order are characterized by the presence of a single gonad in both males and females. There is a striking degree of sexual dimorphism, with the males being small and degenerate. The order is

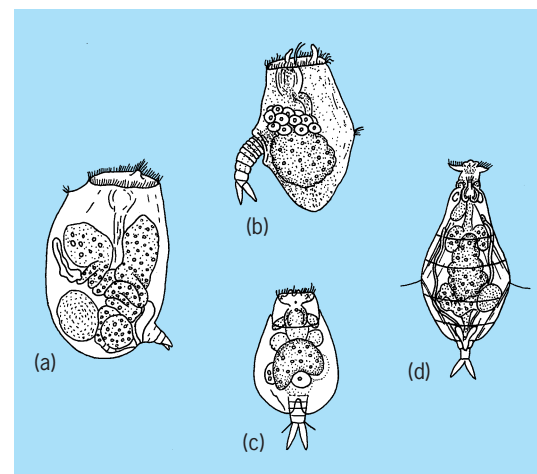


Fig. 1. Ploimates. (a) *Asplanchnopus* sp. (b) *Ploesoma* sp. (c) *Euchlanis dilatata*. (d) *Notomata copeus*.

made up of three suborders: Ploima, Flosculariacea, and Collotheceacea.

In suborder Ploima there is an exceptional diversity of form, varying from soft-bodied wormlike rotifers to species with variously ornamented, loricate shells (Fig. 1). Most of the free-swimming benthonic and pelagic rotifers belong to this suborder. Locomotion is by the ciliated corona. No fewer than five kinds of trophi occur: virgate (notommatids), cardate (*Lindia*), forcipate (dicranophorids), malleate (brachionids), and incudate (asplanchnids). Species range in size from 0.002 in. (0.05 mm) in *Colurella* to 0.06 in. (1.60 mm) in *Asplanchna*.

The suborder Flosculariacea contains the spectacular sessile rotifers formerly known as melicerateans of the family Flosculariidae, as well as a number of equally notable free-swimming forms included in the family Testudinellidae (Fig. 2). Among the latter are the globular *Trochosphaera*, hailed as a connecting link between rotifers and annelids, and the appendage-bearing *Pedalia*, once considered as transitional to crustacea. All members of this suborder have a malleoramate mastax. The flosculariids often have free-swimming young, and a few kinds are free-swimming as adults. Some of the sessile forms are encased in intricately constructed tubes. Some species occur in spherical or ellipsoidal colonies containing many individuals. Colonies may be over 0.08 in. (2.0 mm) in diameter, but individual floscularians are seldom over 0.04 in. (1.0 mm) in length.

The suborder Collotheceacea contains but a single family, the Collotheceidae, made up of five genera. Most species of Collotheceidae are sessile, and many are encased in gelatinous tubes (Fig. 3). The largest genus, *Collotheca*, with an expanded funnel-shaped anterior end, has a wide range of coronal shapes among its 50 or more species. In its simplest form,

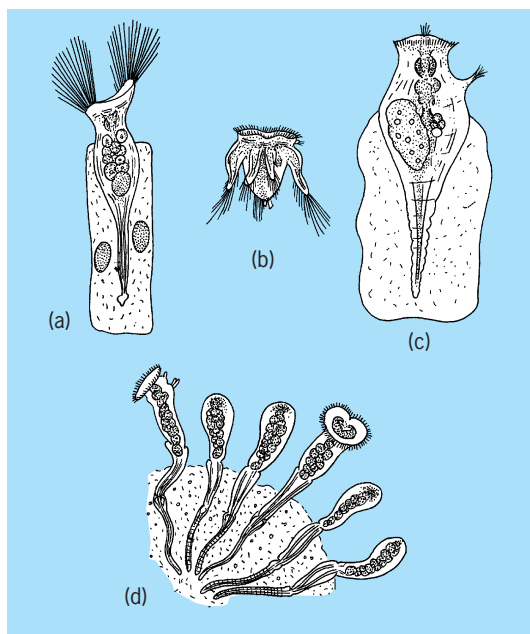


Fig. 2. Flosculariacea. (a) *Floscularia mutabilis*. (b) *Pedalia mira*. (c) *Conochiloides* sp. (d) *Lacinularia socialis*.

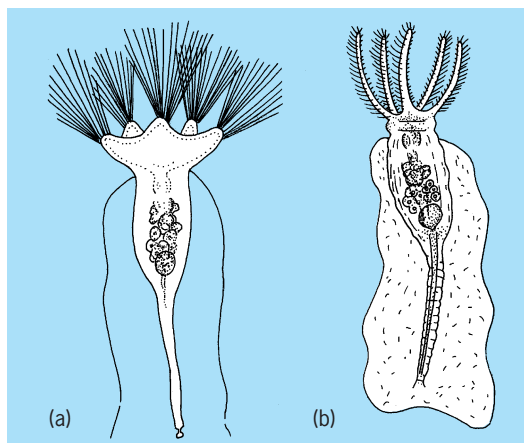


Fig. 3. Members of suborder Collotheceacea. (a) *Collotheca* sp. (b) *Stephanoceros* sp.

the corona of *Collotheca* is without lobes, but in some species the corona has 5-7 lobes, which may be long and fingerlike. Three genera lack a corona. All members of this suborder have an uncinata mastax. *C. boodi*, the largest rotifer known with the exception of *Seisonacea*, belongs to this group and reaches a length up to 0.1 in. (2.5 mm). See ROTIFERA.

Elbert H. Ahlstrom

Monomolecular film

A film one molecule thick. It is often referred to as a monolayer. Films that form at surfaces or interfaces are of special importance. Such films may reduce friction, wear, or corrosion, or may stabilize emulsions, foams, and solid dispersions. Thin films on water surfaces can reduce evaporation losses, though the spreading of thin films on such surfaces can represent a serious environmental problem. In some instances, the film spreads by itself and is essentially insoluble in the substrate (an insoluble monolayer). In other cases, the film molecules have a low substrate solubility and concentrate at the surface (a Gibbs monolayer).

In many cases, the film molecules are amphipathic, having a dual nature. On water surfaces, polar groups can anchor the film molecules, preventing evaporation, while the nonpolar sections prevent them from going into solution. On solid surfaces, polar interactions, or even chemisorption, may resist film removal and provide an essentially new surface. Based on the design of the film molecules, the newly created surface can be tailored to a number of different tasks. There is a wide range of applications of such films, but two areas stand out. On solid surfaces, the formation of multilayers from monolayers can lead to the creation of new materials, often having properties significantly different from that of typical three-dimensional (3D) solids. On water surfaces, the greatest number of studies relate to biological systems, with an emphasis on biological membranes. Many of the earlier studies were on long-chain fatty acids



Fig. 1. Film-balance apparatus. The barriers of the film balance are made from a combination of Teflon[®] and Delrin[®] (acetal resin) to prevent film and liquid leakage and are moved by a microstepping motor (on the right). The trough bottom is roughened to facilitate wetting and has thin Teflon rims to allow for a low volume subphase (55 milliliters with and 20 milliliters without the dipping well). The miniature dipper can be used for deposition on small size substrates. The Willhelmy plate consists of a 0.5-mm-diameter metal alloy wire coupled to a miniaturized microbalance. Connecting cables to the A/D board behind are not shown. (Courtesy of Kibron Inc.)

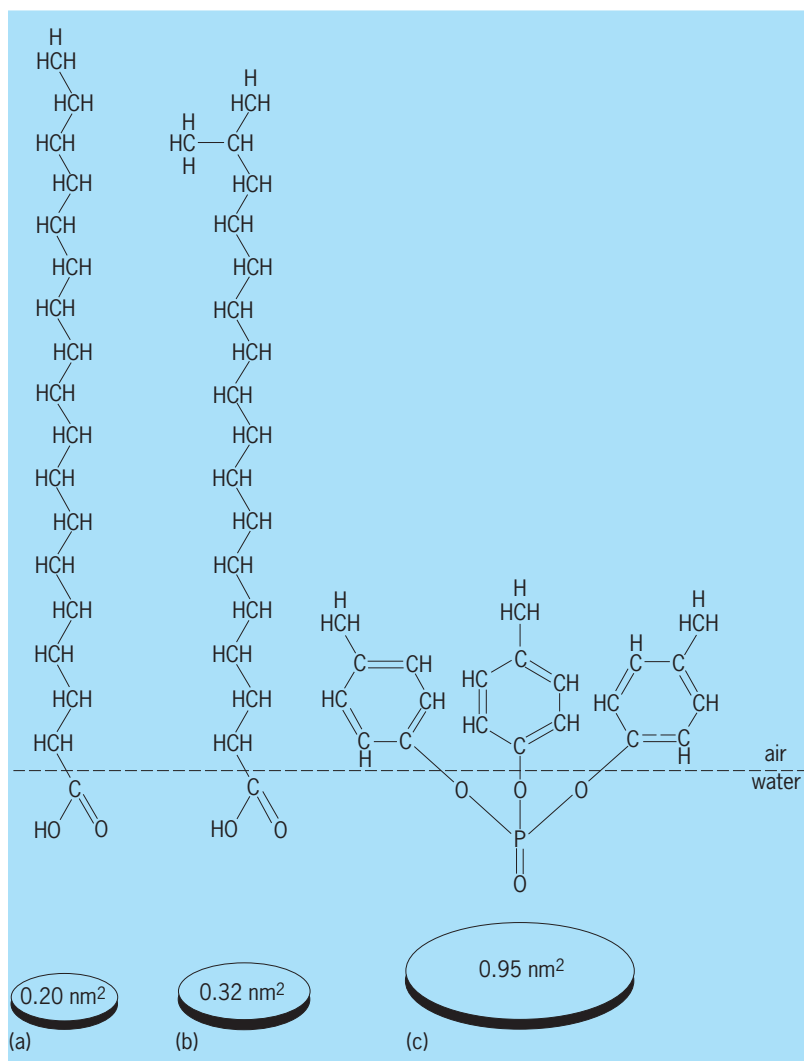


Fig. 2. Molecular orientation and cross-sectional areas of three representative polar organic molecules that are oriented at the water–air interface. (a) Stearic acid. (b) Isostearic acid. (c) Tri-*p*-cresyl phosphate.

and related substances at the air–water interface. Using typical membrane lipids, such as long-chain phospholipids, cholesterol, and membrane proteins, the monolayer constitutes an approximation to one-half of a biological membrane, and has provided considerable information on membrane behavior at the molecular level. See CELL MEMBRANES; MEMBRANE MIMETIC CHEMISTRY; POLAR MOLECULE.

Experimental techniques. The film balance provides basic information on molecular geometry and orientation, including the location and strength of polar groups, and the forces of cohesion and adhesion. The basic measurements are those of the surface pressure (or lowering of the surface tension) and the area per film molecule at the surface. Typically, the equipment consists of a thermostated trough having a hydrophobic surface, typically polytetrafluoroethylene (Fig. 1). The trough is filled to just above the rim with purified water, and all trace contaminants are carefully removed from the surface. Surface pressure is measured by immersing a thin rectangular plate in the surface prior to spreading the monomolecular film. The plate (a Willhelmy plate) consists of surface-treated platinum or glass or even filter paper, and should be wetted by the aqueous substrate. An ideal system will maintain the depth of the plate constant using an automatic nulling balance. Film molecules are spread from a dilute solution of known concentration using a volatile solvent. The film is confined between hydrophobic surface barriers (one static and one mobile) so that the film can be compressed continuously while the surface pressure and area per molecule isotherm can be recorded using the appropriate software. See SURFACE TENSION.

Stearic acid (Fig. 2a) is the classical compound in monolayer studies; it is the simplest structure representative of thousands of important film-forming compounds. The stearic acid molecule consists of a long, straight hydrocarbon chain and a polar group at one extremity. A very similar molecule, isostearic acid (Fig. 2b), and a very dissimilar molecule, tri-*p*-cresyl phosphate (Fig. 2c), are also used frequently in monolayer studies. The difference between isostearic and stearic acid is very slight—the displacement of the small methyl group (CH_3) at the end of the molecule opposite the polar group. Tri-*p*-cresyl phosphate is greatly different; it has a bulky three-ring hydrocarbon portion attached to a strongly polar phosphate group.

Monolayer studies include graphing pressure–area isotherms (Fig. 3). The surface pressure in millinewtons per meter (or dynes per centimeter) is plotted against the average area per molecule. Extrapolation or extension of the steepest part of an isotherm to zero pressure is often used as a measure of molecular area. The point at which the pressure falls or remains constant is called the collapse pressure. Compressibility of the monolayer may be calculated from the slope of the isotherm. Thickness of the monolayer, or the length of the vertically oriented molecules, may be estimated by assuming a density for the monolayer; the volume and area then yield the thickness.

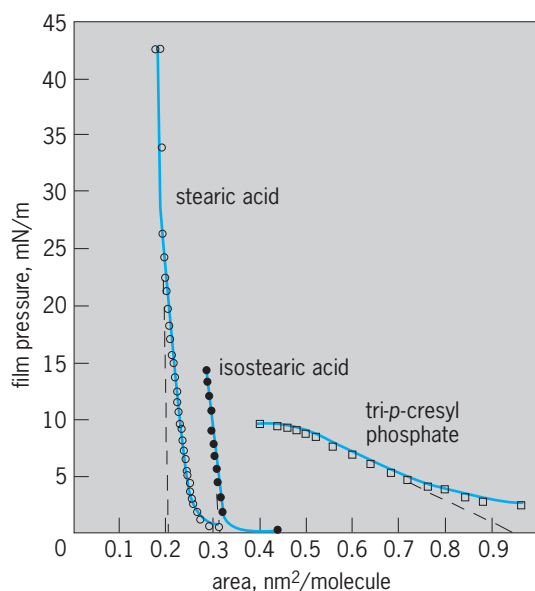


Fig. 3. Pressure-area isotherms for three compounds.

Comparison of the isotherms for stearic and isostearic acids (Fig. 3) demonstrates that the single, small side chain of isostearic acid has increased the cross section from 0.20 nm^2 for the stearic acid molecule to 0.32 nm^2 for isostearic acid, an increase of more than 50%. Collapse pressure falls from 42 mN/m to one-third of this value, 14 mN/m . These are striking differences between molecules that are extremely difficult to distinguish by most chemical methods.

The curve for tri-*p*-cresyl phosphate (Fig. 3) reflects a very different molecular structure. The extrapolated area, $0.95 \text{ nm}^2/\text{molecule}$, shows the bulkiness of the three-ring group held close to the surface. The low collapse pressure, 9 mN/m , reflects

the weakness of such a film. The gradual slope of the curve, or the high compressibility of the film, indicates poor packing of the molecules.

In addition to the basic surface pressure per area per molecule isotherm, a variety of other properties can be measured simultaneously. Among these are the surface (Volta) potential, the surface viscosity, and the surface elasticity. More recently, absorption/emission spectroscopy, fluorescence microscopy (Fig. 4), Brewster angle microscopy (BAM), and grazing incidence diffraction (GIDX) using an intense synchrotron x-ray source have been used. Such techniques are particularly useful where films undergo phase transitions from one state to another. These states range from a two-dimensional (2D) gaseous state, through differing liquid states, to a 2D and a 3D solid state. Thus, both fluorescence microscopy and BAM are capable of distinguishing the different condensed states and providing detailed information on the molecular packing of the film. GIDX, though requiring an intense x-ray source, also provides information on molecular packing and orientation. Following transfer to a solid substrate, several other techniques become available. Most noteworthy are atomic force microscopy (AFM) and, for conducting films, scanning tunneling microscopy. Scanning electron microscopy (SEM) has also proved useful in examining multiphase films deposited on solid surfaces. See FLUORESCENCE MICROSCOPE; INTERFACE OF PHASES; SCANNING ELECTRON MICROSCOPE; SPECTROSCOPY; SURFACE AND INTERFACIAL CHEMISTRY; SURFACE PHYSICS.

Transfer of spread monolayers. Once a monolayer has been formed at the water-air interface, it is possible to transfer it quantitatively to another surface, such as that of a smooth solid. Dipping a clean glass slide into and out of the film-balance trough while the monolayer on it is held at constant surface

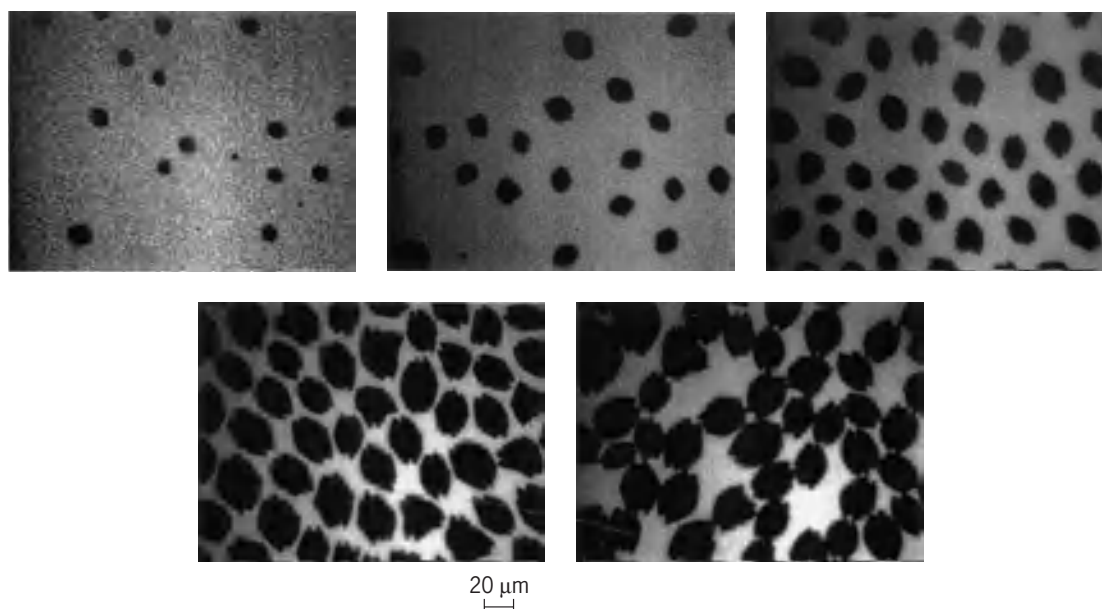


Fig. 4. Fluorescence micrographs of 6-hydroxyoctadecanoic acid (stearic acid) at 22°C during film compression (left to right). The dark regions are those of a liquid condensed phase over the lighter background of the liquid expanded phase. The fluorescence probe was at less than 1% concentration, and it concentrates in the liquid expanded region.

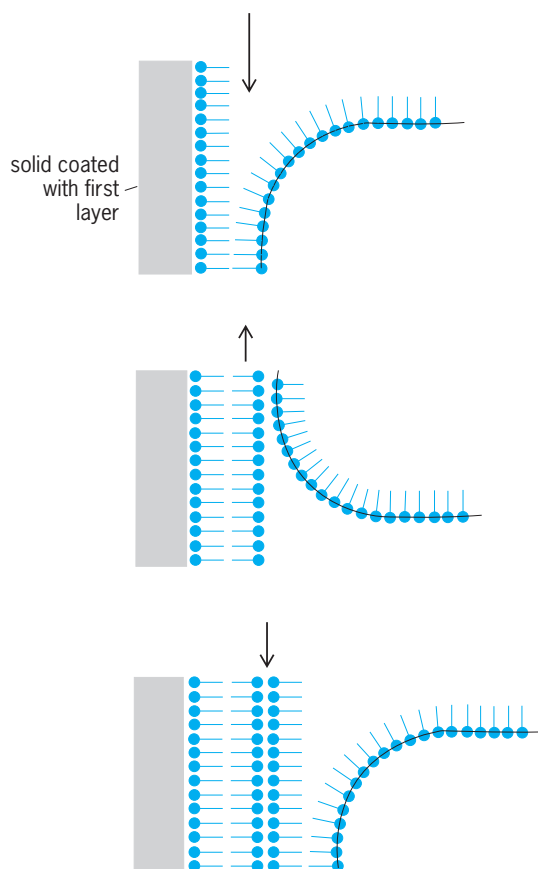


Fig. 5. Successive deposition of monolayers onto a solid plate as it is dipped into and out of the liquid in the Langmuir-Blodgett multilayer deposition method. (After G. L. Gaines, Jr., *Insoluble Monolayers at Liquid-Gas Interfaces*, Wiley, 1966)

pressure (by advancing the confining barrier) leads to such a transfer. The area of monolayer taken up is the same as the surface area of the dipped slide. Properties of the transferred layer, such as its wettability and electron diffraction pattern, indicate that the molecules on the solid surface retain the preferred orientation that they had on the water surface. Monolayers transferred to solids in this way bear close resemblance in structural details both to the precursor liquid-supported monolayer and to films formed on solids by other processes such as adsorption from solutions, but there are usually some differences. The differences, in structure, tightness of bonding, and so on depend on the nature of both the monolayer and the solid support. See ADSORPTION.

It is also possible to deposit certain types of monolayers (especially heavy-metal soaps of long-chain fatty acids) sequentially on solid surfaces to form built-up films or multilayers (Fig. 5). Since each monolayer is extremely thin but uniform (for barium stearate, for example, almost exactly 2.5 nm), such layer structures are very useful as spacers or thickness gages. Films more than 1000 layers thick have been made. Much interest has developed in the optical and electrical properties of such structures. It is possible to assemble multicomponent structures by using monolayers containing more than one chemi-

cal species and by changing from one kind of monolayer to another at different cycles in the dipping process. The location of different molecules at known small distances and in controlled relative orientation has permitted the study of such processes as energy transfer and electron transfers between them. While alterations of the packing may be small for the first few layers of the film, errors tend to increase as the number of layers also increases. It should also be realized that the transfer process depicted in Fig. 5 represents an idealized process, and not all film molecules fit the picture. This is particularly the case where complex molecules are involved. Thus, for polymer films the first layer will not have a perfect polar/nonpolar orientation, and subsequent layers will be less well oriented.

Applications. A large number of monolayer studies involve modification of the film molecules and interpretation of the resultant changes in the film behavior. Increasing the chain length of surfactants will lead to more condensed states. On the other hand, chain branching produces more expanded states, with the effect increasing as the branching entity moves toward the midpoint of the chain. The effects can be even more dramatic when the branching entity is polar, especially at low pressures.

Deuteration similarly leads to an expansion with the effect being roughly proportional to the degree of deuteration. It is clear that monomolecular films can exist in a number of physical states, depending on the nature of the film molecules and the surface pressure and temperature. This is illustrated in Fig. 6 for *L*- α -dipalmitoylphosphatidylcholine (dipalmitoyl lecithin, DPPC), a typical biological membrane lipid. Isotherms are shown for the temperatures in the insert. At very high areas per molecule,

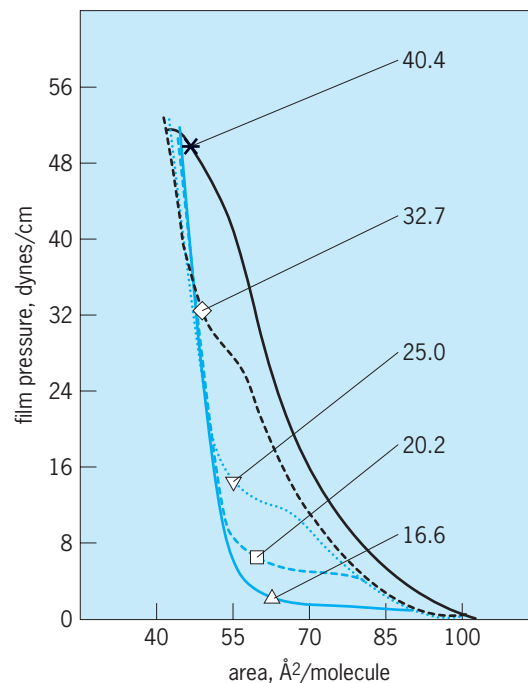


Fig. 6. Pressure-area isotherms for *L*- α -dipalmitoylphosphatidylcholine.

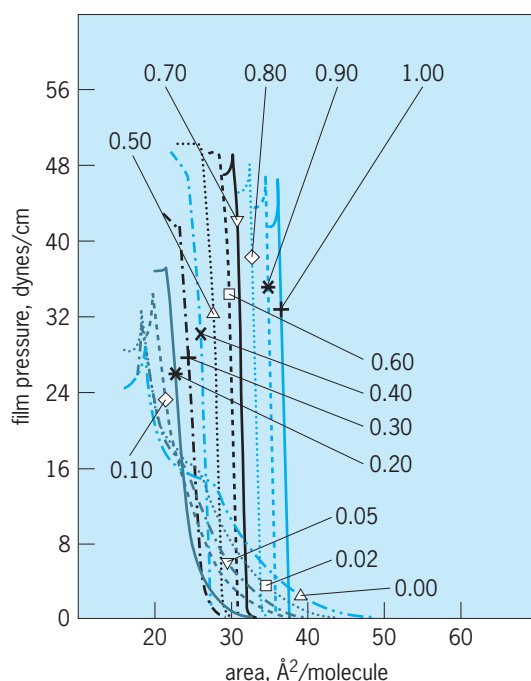


Fig. 7. Pressure-area isotherms for mixed films of myristic acid and cholesterol.

the film molecules are in a 2D gaseous state. While still at low pressures (<1 dyne/cm) and high areas (>100 $\text{\AA}^2/\text{molecule}$), all films are in an equilibrium between a liquid-expanded gaseous state and a gaseous state (these states are not visible in Fig. 6). At some point, around 100 $\text{\AA}^2/\text{molecule}$, all gaseous film has been converted to a liquid expanded (LE) state and the pressure begins to increase rapidly. At increasing pressures with increasing temperature, an inflection in the isotherm is found that denotes a first-order transition to a liquid condensed (LC) state, the inflection becoming less sharp as the temperature approaches 41°C (106°F). Above this temperature, no transition takes place and the film becomes unstable and collapses. The behavior shown is typical for most amphipathic molecules ranging from fatty acids to complex membrane lipids such as lecithins and phosphatidylethanolamines. However, not all surface-active molecules behave in this way, and one example is cholesterol. See SURFACTANT.

Mixed films. Typical mixed-film studies are usually limited to two components. Figure 7 illustrates a series of mixed-film isotherms of myristic acid and cholesterol at 23°C (73°F), the insert numbers indicating the mole % cholesterol. The cholesterol isotherm on the extreme right shows only a fully condensed state at all pressures, with a collapse pressure around 46 dynes/cm. This collapse pressure varies with composition, reaching a maximum around 50 mole % cholesterol. In contrast, myristic (hexadecanoic) acid has a much lower collapse pressure and shows expanded phases at lower pressure similar to those in Fig. 6 for DPPC. With increasing cholesterol concentration, both the LE/LC phase transition and the collapse pressure vary. Variation of a phase transition, including collapse, with composition is indica-

tive of at least partial miscibility of the components of the film. Cholesterol shows that it condenses the expanded myristic acid film but expands the condensed film. The ability of cholesterol to alter the fluidity of a monolayer is indicative of a similar role in biological membranes. At lower pressures, cholesterol is unaffected above about 20 mole %. For DPPC, this occurs at about 43 mole %. This suggests that cholesterol will segregate above these values in the respective mixed films. Mixed film studies can give information on both segregation and interaction of the two components.

Substrate interactions. In many instances, studies of film interactions have been made in which one component is present in the aqueous substrate. Fatty acids have been known to be encapsulated in β -cyclodextrin, and this has been illustrated using the effects of the β -cyclodextrin in the substrate on fatty-acid films. The influence of differing cations on fatty acids and on acidic phospholipids has had numerous studies; for example, the ability of lipases in the substrate to hydrolyze biological membrane lipids is an obvious case in point. Many of the hydrolyzed products are soluble and are lost from the monolayer, reducing the surface pressure and thus providing a way of following the reaction. Lipid-protein interactions may similarly be followed. Typical results indicate that in many cases the initial interaction is ionic and that it is often followed by insertion of a hydrophobic chain into the monolayer. Specific types of interactions may be studied by using polypeptides designed for the purpose. Other studies have included lipid-antibiotic interactions and the ability of local anesthetics to penetrate or fluidize monolayers as a model for the effects on membranes and a possible role in anesthesia. In such studies, it is assumed that the lipid monolayer model is sufficiently similar to the behavior of a bilayer and that it approximates a biological membrane.

Pulmonary surfactant. For some time, it has been known that pulmonary (lung) surfactant plays a vital role in reducing the surface tension in the lungs of mature humans. However, this surfactant is missing in the lungs of children born very prematurely and can cause serious lung problems or even death. The need for an artificial pulmonary surfactant has led to many studies of lung surfactant composition, and of mixed films of differing lipid components since these are thought to exist as a monolayer within the lung. It is now known that DPPC is fully capable of reducing the surface tension but has difficulty in respreading. The role of lung-surfactant-specific proteins has been investigated, and it is thought that these proteins play a role in the respreading of the surfactant.

D. Allan Cadenhead; George L. Gaines, Jr.

Self-assembled films. Self-assembled monolayers are monomolecular films formed by the immersion of an appropriate substrate into a solution of an active surfactant (Fig. 8). The molecular order in these 2D systems is produced by a spontaneous chemical synthesis at the interface as the system approaches

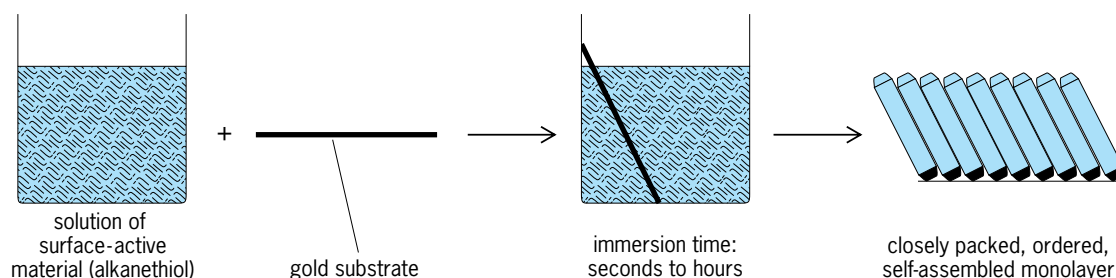


Fig. 8. Steps in forming a self-assembled monolayer.

equilibrium. In contrast to Langmuir-Blodgett monolayers, which require substrate pretreatment for significant bonding and are often metastable, self-assembled monolayers exhibit more uniform and stable behavior. The packing and ordering in self-assembled monolayers—both the conformation of the individual chains within the assembly and their packing and ordering with respect to each other—are determined by the interplay between interchain forces and the interactions with the surface, combined with entropic effects, and specific surface interactions of terminal groups.

There are two cases of monolayer formation. The first is on amorphous surfaces (such as glass), where intermolecular interactions define the packing and ordering; and the second is on crystalline surfaces (for example, the {111} face of gold), where the structure of the overlayer is determined by the surface. See CRYSTALLINE STRUCTURE.

Fatty acids. Long-chain carboxylic acids form monolayers on metal oxide surfaces such as aluminum, silver, and copper oxides via an acid-base reaction; on quartz and glass the mechanism may be ion exchange. They also form monolayers on oxide-free surfaces such as gold and copper in ultrahigh vacuum. The stability of the monolayers increases with increasing basicity of the substrate, in the order quartz \sim gold $<$ glass \sim aluminum \sim copper $<$ silver. In the case of carboxylic acid derivatives with nonpolar functionalities, highly ordered monolayers can be formed. This is the case where vinyl, acetylene, diacetylene, pyrene, benzene, naphthalene, and biphenyl are incorporated into the chain. Polar functionalities cannot be introduced because of their tendency to adsorb onto the surface, thus competing with the carboxylate group (COOH). See ACID AND BASE; ION EXCHANGE.

Very little is known about monolayer formation from the gas phase, and most studies have been carried out by using solution-adsorption. Monolayer formation is a function of both the substrate and the chain length. Thus, while hexanoic acid [$\text{CH}_3(\text{CH}_2)_5\text{COOH}$] forms ordered monolayers on silver oxide (AgO), for aluminum oxide (Al_2O_3) and copper(II) oxide (CuO) at least dodecanoic acid [$\text{CH}_3(\text{CH}_2)_{11}\text{COOH}$] is required. The chemisorption of alcanoic acids on amorphous metal oxide surfaces is not unique. On silver oxide, the two oxygen atoms of the carboxylate bind to the surface nearly symmetrically, and ordered monolayers are formed with

a chain tilt angle from the surface normal of $15\text{--}25^\circ$. On surfaces of copper oxide and aluminum oxide, the carboxylate oxygen atoms bind asymmetrically to the surface, displaying tilt angles close to zero. This suggests that controlling the chemisorption mode determines molecular orientation, and hence physical properties that are orientation-dependent.

A much higher concentration of surfactant is required to produce good monolayers on glass than is required on aluminum oxide. The kinetics is faster for adsorption on glass than on aluminum oxide. It is possible to calculate the Gibbs free energy of adsorption for dodecanoic acid on aluminum oxide and on glass, revealing a stronger binding interaction of the carboxylate to the glass surface. See FREE ENERGY.

Alkyltrichlorosilanes. Self-assembled monolayers of alkyltrichlorosilanes (R-SiCl_3) require hydroxylated surfaces as substrates [glass, silicon dioxide/silicon (SiO_2/Si), aluminum oxide, and others], since the formation of the self-assembled monolayer is essentially an in-place formation of polysiloxane, which is connected to surface silanol groups (Si-OH) via silicon-oxygen-silicon (Si-O-Si) bonds. The $\text{Si}\cdots\text{Si}$ spacing [0.44 nm] is determined by the formation of a polysiloxane chain. This spacing leaves little free volume, and the chains are locked perpendicular to the surface. During adsorption, silicon-chlorine (Si-Cl) bonds react with the hydroxyl (OH) groups present on the surface of the substrate, and with trace water, which is adsorbed on the surface, to form a network of (Si-O-Si) bonds.

Self-assembled monolayers from octadecyltrichlorosilane ($\text{C}_{18}\text{H}_{37}\text{SiCl}_3$) are temperature-dependent. Complete surface reaction of the $-\text{SiCl}_3$ groups occurs. The kinetics of the formation of long-chain alkyltrichlorosilane monolayers showed that monolayer formation is complete within ~ 40 min. Monolayers of alkyltrichlorosilane show remarkable stability.

Construction of multilayers from alkyltrichlorosilane derivatives requires that the monolayer surface be modified to a hydroxylated one. Such surfaces can be prepared by a chemical reaction and the conversion of a nonpolar terminal group to a hydroxyl group. Once a subsequent monolayer is adsorbed on the activated monolayer, multilayer films may be built by repetition of this process. There is a linear relationship between the film thickness and the layer number. However, building thin organic films with interesting physical properties by using

self-assembly is not straightforward. Alkyltrichlorosilane derivatives of large aromatic systems are difficult to purify, and their solutions in organic solvents have to be replaced frequently because of polysiloxane formation. This is especially true for polar systems that are required for building waveguides for nonlinear optical application. Therefore, the unique stability of self-assembled monolayers made from silane derivatives makes them ideal materials for surface modification and functionalization applications, for example, as adhesion promoters, boundary lubricants, and active layers in chemical and biological sensors. *See* HYDROXYL; LUBRICANT; ORGANOSILICON COMPOUND.

Alkanethiols. Organic molecules that contain sulfur or selenium atoms in a neutral or negative oxidation state adsorb spontaneously on metallic surfaces such as gold, silver, platinum, copper, iron, and mercury. However, gold surfaces are the preferred substrate, because they are easily prepared, are simple to handle, and do not have a stable oxide. Examples are alkanethiols, di-*n*-alkyl sulfides, thiophenols, mercaptopyridines, mercaptoanilines, thiophenes, cysteines, alkaneselenoles, xanthates, thiocarbaminates, thiocarbamates, mercaptoimidazoles, and thiourea derivatives. The materials that were first recognized as forming closely packed, ordered self-assembled monolayers on gold substrates were the alkyl disulfides.

Alkanethiolates coordinate very strongly to gold. In many cases, the structure formed by alkanethiolates on the {111} face of gold is commensurate with the underlying gold lattice. *See* ARTIFICIALLY LAYERED STRUCTURES.

Alkanethiolates may have two binding modes at the hollow site of the {111} face of gold. Molecular mechanics energy minimization indicates that these two modes lead to monolayers exhibiting different types of packing arrangements but comparable in their ground-state energies. Therefore, monolayers may consist of two different chemisorption modes ordered in different domains, that is, simultaneously coexisting homogeneous clusters, each characterized by a different conformer in its unit cell. Annealing may result either in a unique chemisorption mode or in a much more complex structure, as the different modes are redistributed. *See* MOLECULAR MECHANICS.

Because of the relatively facile synthesis and purification of alkanethiols and the dense, stable structure of their self-assembled monolayers, these monolayers have potential applications in corrosion prevention, wear protection, and biosensing devices. The ability to tailor both head and tail groups of the constituent molecules makes them ideally suited for a more fundamental understanding of phenomena affected by competing intermolecular, molecule-substrate, and molecule-solvent interactions such as ordering and growth, wetting, adhesion, lubrication, and corrosion. *See* BIOSENSOR.

Mixed self-assembled monolayers of alkanethiolates on gold surfaces are especially suited for the studies of interfacial phenomena because of the fine

control of the concentration of surface functional groups. This control is achieved by changing the concentration in the solution from which the mixed monolayers are adsorbed. Self-assembled monolayers are much better systems than polymer surfaces for studies of the effect of molecularly specific interfacial phenomena. The reason is that a polymer exposes a nonhomogeneous, molecularly rough surface because of different degrees of polymerization on different sites, while a self-assembled monolayer exhibits a smooth homogeneous surface. A large number of ω -terminated alkanethiols provide practically endless possibilities for surface engineering. For example, two self-assembled monolayer surfaces can be prepared with the same critical surface tension but with different degrees of hydrogen-bonding ability, for example, a pure chlorine surface and a mixed one of methyl and hydroxyl. Such surfaces, while exhibiting the same contact angle with different liquids, may show differences when contact angles are studied as a function of temperature and relative humidity.

Self-assembled multilayers of diphosphonates. Phosphate salts of tetravalent metal ions are extremely insoluble. These materials have a layered structure in which three of the four phosphate oxygen atoms are bonded to the transition-metal ion, and the fourth phosphorus-hydroxyl (P—OH) group is perpendicular to the sheet at both sides. Thus, replacing the hydroxyl group by an organic moiety should result in a lamellar structure of alkanephosphonates, cross-linked by the transition-metal ions. *See* FILM (CHEMISTRY); MATERIALS SCIENCE AND ENGINEERING.

Abraham Ulman

Bibliography. A. W. Adamson and A. P. Gast, *Physical Chemistry of Surfaces*, 6th ed., 1997; H. Bubert and H. Jenett (eds.), *Surface and Thin Film Analysis: A Compendium of Principles, Instrumentation, and Applications*, 2002; H.-J. Butt, K. Graf, and M. Kappl, *Physics and Chemistry of Interfaces*, 2003; G. L. Gaines, Jr., *Insoluble Monolayers at Liquid-Gas Interfaces*, 1966; R. Miller and D. Mobius (eds.), *Monolayers and Assemblies: Structure, Processes and Function*, 2002; G. A. Somorjai, *Introduction to Surface Chemistry and Catalysis*, 1994.

Mononchida

An order of nematodes having a full complement of cephalic sensilla on the lips in two circlets of 6 and 10. The amphids are small and cuplike, and are located just posterior to the lateral lips; the amphidial aperture is either slitlike or ellipsoidal. The stoma is globular and heavily cuticularized, and is derived primarily from the cheilostome. The stoma bears one or more massive teeth that may be opposed by denticles in either transverse or longitudinal rows. The esophagus is cylindrical conoid, with a heavily cuticularized luminal lining. The excretory system is atrophied. Males have ventromedial supplements and paired spicules. The gubernaculum may possess lateral accessory pieces. Females have one or two

ovaries. Caudal glands and a spinneret are common; however, they may be degenerate or absent.

There are three mononchid superfamilies, Mononchoidea, Bathyodontoidea, and Monochuloidea.

The superfamily Mononchoidea contains some of the most common and easily recognized free-living nonparasitic nematodes that occur in soils and fresh waters throughout the world.

The closely related Bathyodontoidea are inhabitants of soil or fresh water and prey on small microorganisms.

The nonparasitic Monochuloidea comprises both soil and fresh-water species, all of which are predators of microfauna. See NEMATA (NEMATODA).

Armand R. Maggenti

Monoplacophora

A class of the phylum Mollusca. Although fossil monoplacophorans had been known since the end of the nineteenth century, it was the discovery of a living species in deep water off Costa Rica in

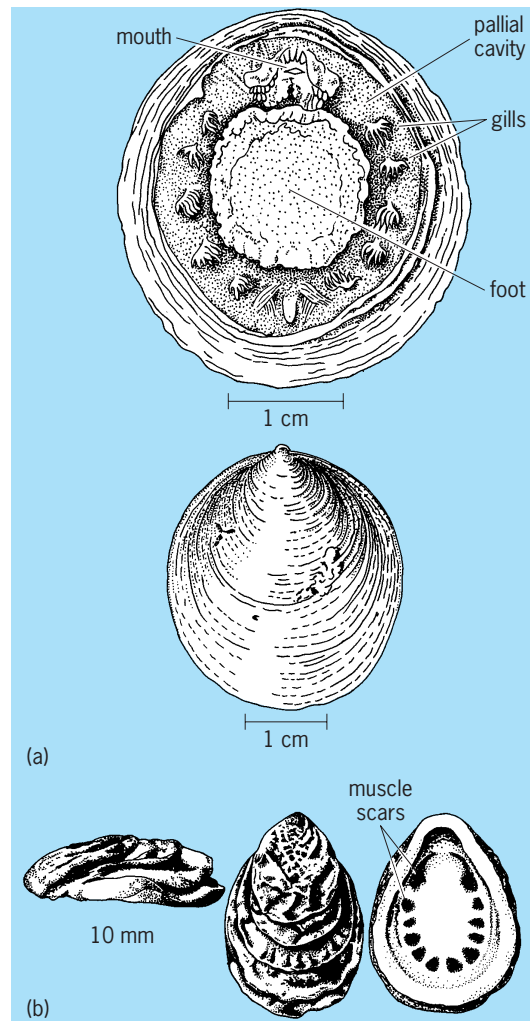


Fig. 1. Living and fossil Monoplacophora. (a) *Neopilina galathea*. (b) *Tryblidium reticulatum*. (After R. C. Moore, ed., *Treatise on Invertebrate Paleontology*, pt. I, 1957)

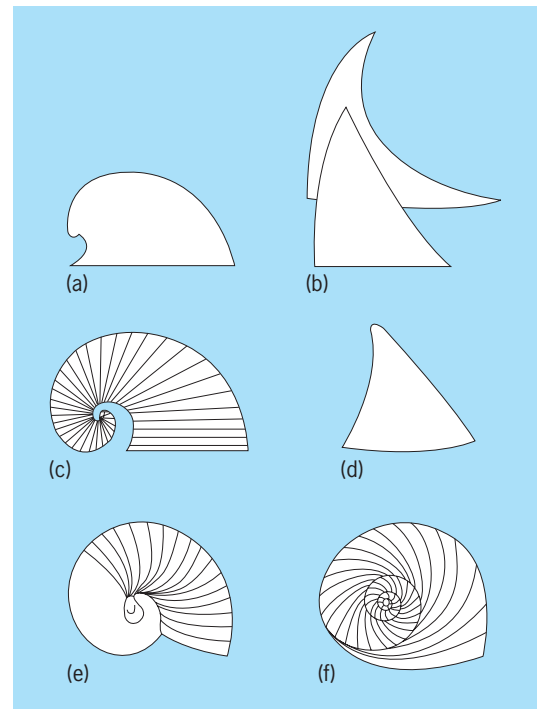


Fig. 2. Side views of various helcionellacean and bellerophon univalved mollusks. (a) *Cyrtoneopsis*; (b) *Hypseloconus*; (c) *Cyclocyrtoneilla*; (d) *Stenotheca*; (e) *Bellerophon*; (f) *Tropidodiscus*. (After B. Runnegar, *Class Monoplacophora*, in R. S. Boardman, A. H. Cheetham, and A. J. Rowell, eds., *Fossil Invertebrates*, Blackwell Scientific Publications, 1987)

the 1950s that led to universal acceptance of the class. Monoplacophorans are bilaterally symmetrical, univalved mollusks that vanish from the fossil record at the end of the Paleozoic, about 240 million years ago. The living species, such as *Neopilina galathea*, are rare and inhabit deep water, which may explain their absence from Mesozoic and Cenozoic rocks. Living monoplacophorans have a limpet-shaped shell, a circular foot attached by pairs of retractor muscles, and several gills on each side of the body (Fig. 1). An anterior mouth, but no distinct head, is evident, along with a radula for obtaining food, a central gut, and a posterior anus. Many of the organs are repeated as many as eight times down the length of the body, suggesting that the monoplacophorans are derived from some kind of segmented animal. An alternative view, which relates the monoplacophorans and other primitive mollusks to unsegmented flatworms (platyhelminths), has been challenged recently by both anatomical and molecular evidence.

Within the Mollusca, the Monoplacophora is regarded as the most primitive of the shelled, or conchiferan, classes. Viewed as a sister group of the eight-plated Polyplacophora (chitons), the Monoplacophora is seen as the ultimate source of the major molluscan classes, the Gastropoda, Bivalvia, and Cephalopoda. For this reason, much attention has been given to the fossil representatives of the Monoplacophora, and perhaps as a result

there is some controversy about which kinds of fossils should be placed within the class. See POLYPLACOPHORA.

All who have studied the group agree that the early Paleozoic cap-shaped shells such as *Tryblidium reticulatum* (Fig. 1) belong in the Monoplacophora. However, other bilaterally symmetrical, univalved mollusks from Paleozoic strata, including the early Paleozoic horn-shaped helcionellaceans and the planispirally coiled bellerophon mollusks, may also be monoplacophorans (Fig. 2). In fact, the bellerophonts may have been the most successful group of the Monoplacophora before they became extinct in the Early Triassic.

If bellerophonts were monoplacophorans, they show that gastropodlike features evolved independently in the group. The tight coiling of the shell may have been an adaptation to avoid predation on soft substrates: by being able to withdraw into a coiled shell, univalved mollusks were able to leave the security of rock substrates that acted as a second "valve" for limpet-shaped shells. Thus, many of the characters that are used to relate the bellerophonts to either the Monoplacophora or the Gastropoda are of uncertain phylogenetic significance, and the status of many extinct molluscan univalves remains unresolved.

Some early Paleozoic (Cambrian) monoplacophorans were laterally compressed, a feature that presumably facilitated slicing into soft substrates. Those forms gave rise to the Bivalvia by way of the pseudobivalved *Rostroconchia*. Other early monoplacophorans, ancestors of the Cephalopoda, had tall, partitioned, conical shells that were preadapted for flotation. A third Cambrian monoplacophoran developed an asymmetric shell and eventually gave rise to the Gastropoda. Thus, all of the major classes of the Mollusca were established within the Monoplacophora at or soon after the beginning of the Paleozoic. See GASTROPODA; MOLLUSCA. Bruce Runnegar

Bibliography. R. S. Boardman, A. H. Cheetham, and A. J. Rowell (eds.), *Fossil Invertebrates*, 1987; J. H. McLean, A new monoplacophoran limpet from the continental shelf off southern California, *Contrib. Sci. Nat. Hist. Mus. Los Angeles County*, 307:1-19, 1979; B. Runnegar and P. A. Jell, Australian middle Cambrian molluscs and their bearing on molluscan evolution, *Alcheringa*, 1:109-138, 1976; K. G. Wingstrand, On the anatomy and relationships of Recent Monoplacophora, *Galathea Rep.*, 16:7-94, 1985.

Monopulse radar

Radar capable of estimating target position based on the return from a single pulse. In many radars, precise angular position is estimated by conically scanning a single beam around the initial coarse angle estimate; the orderly amplitude variation of echoes during such scanning provides the refinement. Such measurement is limited, however, by pulse-to-pulse

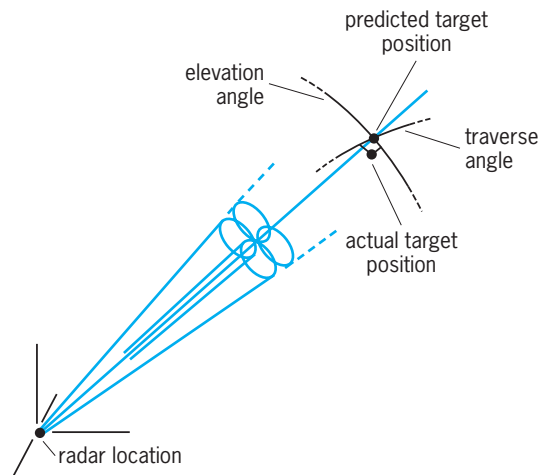


Fig. 1. Use of squinted beams in amplitude-comparison monopulse radar.

fluctuations in echo strength, a property quite common in radar targets.

Monopulse radars use antennas that provide a local cluster of simultaneous beams (instead of scanning just one beam) to make the same precise angle estimate with each pulse transmitted. Since angle information is contained in each return, fluctuations in echo strength do not significantly degrade the measurement. Monopulse radars with mechanically positioned antennas address only a single target, and average the measurements over many pulses for improved accuracy. Radars using electronic beam steering in stationary phased-array antennas may make such a measurement in a single-pulse "dwell," doing so on dozens of targets, returning to each several times a second if necessary.

Target position may be defined by two angles (azimuth and elevation) and range, as in a radar-centered spherical coordinate system. Local angle deviations are often sensed in elevation and traverse (with subsequent coordinate conversion) depending on the characteristics of the antenna used. An antenna producing a cluster of slightly squinted beams (that is, beams which are not parallel but point in directions that make small angles with each other) [Fig. 1] permits angle measurement by comparing the

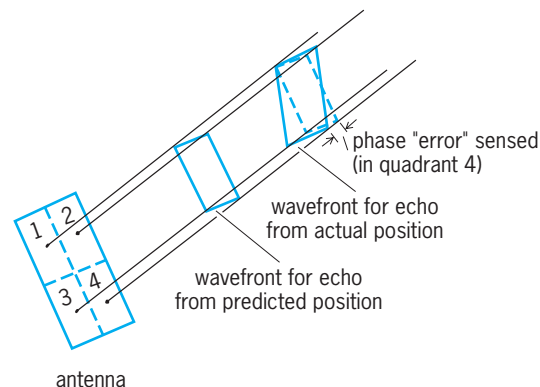


Fig. 2. Use of parallel beams in phase-comparison monopulse radar.

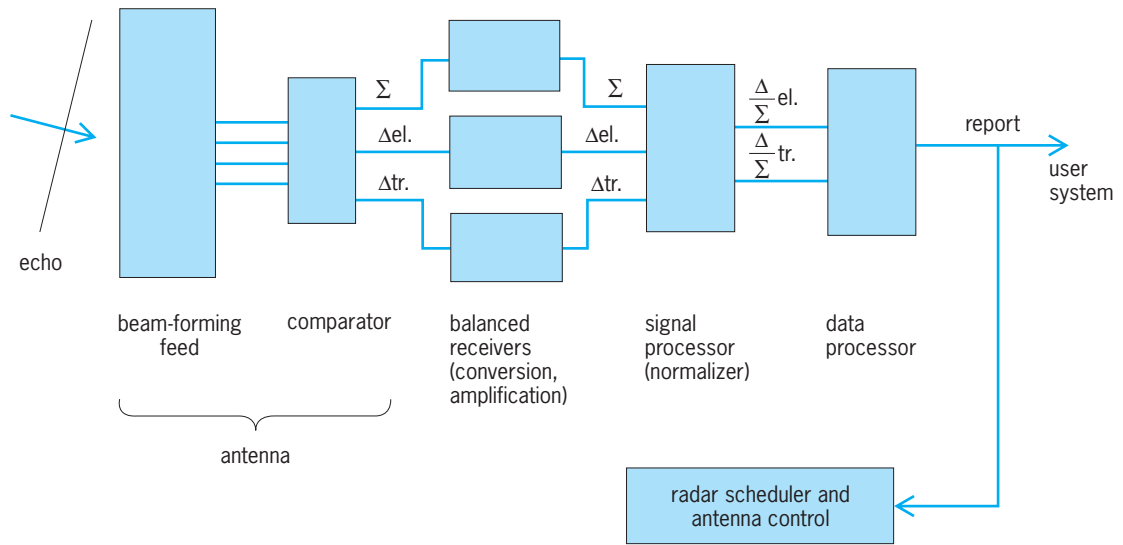


Fig. 3. Generic monopulse radar receiver. el. = elevation; tr. = traverse.

amplitudes of the echoes in the several beams; such is called amplitude-comparison monopulse. In phased-array antennas, it may be more convenient to form beams, one from each quarter of the antenna, all pointing in the same direction (not squinted, but parallel) and to use the relative phase of the echoes as the precise indication of angle. This approach (Fig. 2) is called phase-comparison monopulse.

In both approaches, the antenna feed includes a comparator, a device performing the appropriate comparison among the beams. Its outputs are the sum (Σ) of the echoes in the several beams and the differences (Δ) in elevation and traverse. The receiver (Fig. 3) processes these outputs to establish the quotient, Δ/Σ , indicative of the difference in each angle between the measurement and the prediction. This normalizing (division by Σ) makes the measurement nearly independent of target amplitude fluctuations.

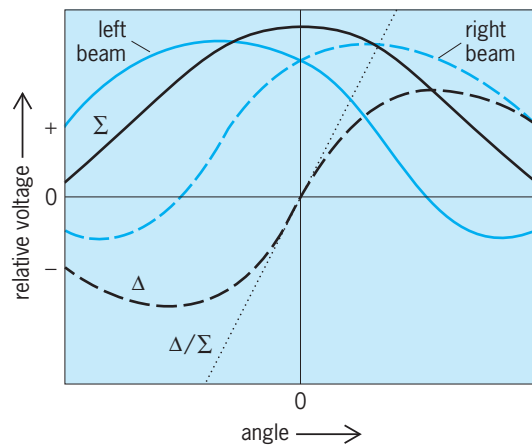


Fig. 4. Antenna patterns for amplitude-comparison monopulse radar in one measurement plane. Patterns of individual beams (left and right), sum (Σ) and difference (Δ) patterns, and the normalized calibration curve (Δ/Σ) are shown.

Antenna patterns formed at the comparator output ports comprise a sum pattern and two difference patterns, having, respectively, even and odd angle symmetry. Figure 4 illustrates, for amplitude comparison in just one measurement plane, the individual squinted beams, the sum and difference patterns, and the normalized calibration curve, Δ/Σ . A good monopulse radar requires (as does any radar) good antenna patterns—a sum pattern with a narrow main beam, a difference pattern with a steep central slope (a deep null in the power pattern), and low sidelobes in each. The feed system must be well designed and may be somewhat complicated to achieve these qualities jointly.

Angle accuracy may be severely degraded by unresolved multiple targets (targets very close in angle and range) because the signals add together coherently in the separate receiver channels before the normalization. Receivers using only the real part of the complex quotient Δ/Σ can incur large errors in this situation.

A similar single-pulse technique using pairs of range gates to refine range estimates is sometimes called range monopulse. Most radar engineers think of monopulse, however, as the angle-measuring technique described here.

The emerging technique of computer-based digital beamforming in radar antennas will change the physical nature of the antenna considerably, but the basic principle—that of interpolating among multiple simultaneous beams, however formed, to estimate angle—will remain. See ANTENNA (ELECTROMAGNETISM); RADAR.

Robert T. Hill
Bibliography. D. K. Barton and H. R. Ward, *Handbook of Radar Measurement*, Artech House, 1984; S. Sherman, *Monopulse Principles and Techniques*, Artech House, 1984; M. I. Skolnik, *Introduction to Radar Systems*, 3d ed., McGraw-Hill, 2001; M. I. Skolnik (ed.), *Radar Handbook*, 2d ed., McGraw-Hill, 1990.

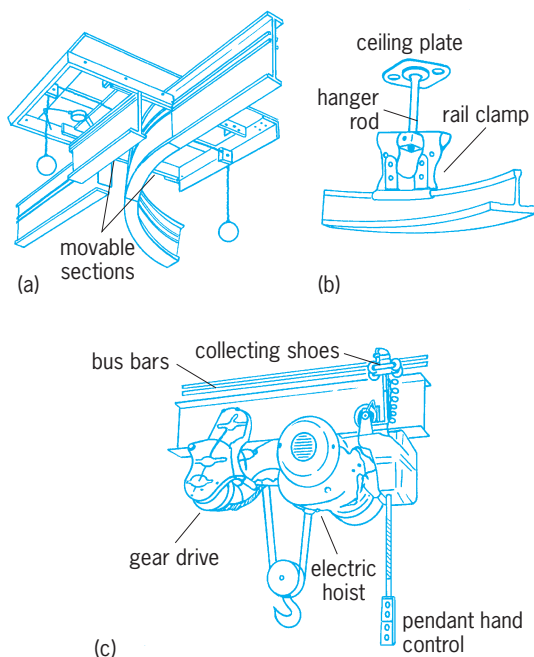
Monorail

A distinctive type of materials-handling machine that provides an overhead, normally horizontal, fixed path of travel in the form of a trackage system and individually propelled hand or powered trolleys which carry their loads suspended freely with an intermittent motion. Because monorails operate over fixed paths rather than over limited areas, they differ from overhead-traveling cranes, and they should not be confused with such overhead conveyors as cableways.

Relatively simple but adequately efficient monorail systems for specialized applications have flat steel bars or galvanized pipes for trackage. However, standard I beams or other similar shapes are used more and more frequently, with the wheels of the trolley bearing directly upon the lower flange of the beam. The latter type are used more for heavy-load service.

Garment manufacturers and cleaning establishments use the simple pipe-rail system for hanging garments on hangers or wheel-equipped trolleys. Switches, crossovers, and other components make setups flexible (see *illus.*). These systems can be arranged to run onto freight elevators, along loading platforms, and directly into carriers.

Of primary importance in a conventional monorail system are the rails. They are connected with butt or lap joints and must be smooth to allow the trolley to move freely. Clamps or brackets suspend the trackage from ceilings or walls. Because monorail systems do not operate with continuous motion, the tracks need not be arranged in self-closing lines. Spurs can be run into paint booths, cooling rooms, and similar work areas. These paths



Typical parts of overhead monorails. (a) Slide (glider) type. (b) I-beam track. (c) Powered trolley.

are selected by switches, such as the tongue or the slide (glider) varieties. Specially constructed sections permit 90° changes in trolley travel; turntables permit articles to be turned through 360°. Where monorails pass through fire doorways, liftout sections permit the door to close in the event of fire. Lift and drop sections shift the flow of traffic from one line to another at a different level, thus eliminating the need for inclined and declined tracks.

Wheels mounted in trolleys ride on the flanges of the track, in most varieties. Two-wheeled trolleys, connected by bars, combine to make up carriers. The number of wheels depends upon the desired load-carrying capacity. In many instances hand or powered hoists are suspended from the carriers. To provide current for powered hoists, special electrification equipment, such as bus bars and collecting shoes, are used. If powered carriers are required, the drive may be secured in one of several ways. Examples are the motor-driven rubber tire which contacts the under flange of the track, or the trolley wheels which are driven through gear trains by an electric motor. Trolleys of the first type operate without slippage and make possible the introduction of sections with slight inclines and declines. Special carriers are used with monorail systems for handling specific products, such as batches in bakeries and the movement of clothes from one process to another in laundries. Scale sections can be included in a trackage line to weigh such products as batches, textile beams, or rolls of paper without causing a delay in traffic flow. Many below-the-hook devices can be used with monorails. See HOISTING MACHINES.

Monorail elements are utilized in the construction of overhead-traveling cranes if distribution over an area rather than along a fixed path is required. Such cranes can not accommodate the heavy loads handled by cranes built with structural girders, but they are adequate for many industrial applications. See BULK-HANDLING MACHINES; MATERIALS-HANDLING EQUIPMENT.

Arthur M. Perrin

Bibliography. J. M. Apple, *Plant Layout and Materials Handling*, 1977, reprint 1991; F. E. Meyers, *Plant Layout and Material Handling*, 1993.

Monosaccharide

A class of simple sugars containing a chain of 3-10 carbon atoms in the molecule, known as polyhydroxy aldehydes (aldoses) or ketones (ketoses). They are very soluble in water, sparingly soluble in ethanol, and insoluble in ether. The number of monosaccharides known is approximately 70, of which about 20 occur in nature. The remainder are synthetic. The existence of such a large number of compounds is due to the presence of asymmetric carbon atoms in the molecules. Aldohexoses, for example, which include the important sugar glucose, contain no less than four asymmetric atoms, each

of which may be present in either D or L configuration. The number of stereoisomers rapidly increases with each additional asymmetric carbon atom. See CARBOHYDRATE; KETONE; OPTICAL ACTIVITY; STEREOCHEMISTRY.

A list of the best-known monosaccharides is given below:

- Trioses: $\text{CH}_2\text{OH} \cdot \text{CHOH} \cdot \text{CHO}$, glycerose
(glyceric aldehyde)
 $\text{CH}_2\text{OH} \cdot \text{CO} \cdot \text{CH}_2\text{OH}$, dihydroxy
acetone
- Tetroses: $\text{CH}_2\text{OH} \cdot (\text{CHOH})_2 \cdot \text{CHO}$, erythrose
 $\text{CH}_2\text{OH} \cdot \text{CHOH} \cdot \text{CO} \cdot \text{CHO}$, erythrulose
- Pentoses: $\text{CH}_2\text{OH} \cdot (\text{CHOH})_3 \cdot \text{CHO}$, xylose,
arabinose, ribose
 $\text{CH}_2\text{OH} \cdot (\text{CHOH})_2 \cdot \text{CO} \cdot \text{CH}_2\text{OH}$,
xylulose, ribulose
- Methyl pentoses (6-deoxyhexoses):
 $\text{CH}_4(\text{CHOH})_4 \cdot \text{CHO}$, rhamnose, fucose
- Hexoses: $\text{CH}_2\text{OH} \cdot (\text{CHOH})_4 \cdot \text{CHO}$, glucose,
mannose, galactose
 $\text{CH}_2\text{OH} \cdot (\text{CHOH})_3 \cdot \text{CO} \cdot \text{CHOH}$,
fructose, sorbose
- Heptoses: $\text{CH}_2\text{OH} \cdot (\text{CHOH})_5 \cdot \text{CHO}$,
glucoheptose, galamannoheptose
 $\text{CH}_2\text{OH} \cdot (\text{CHOH})_4 \cdot \text{CO} \cdot \text{CH}_2\text{OH}$,
sedoheptulose, mannoheptulose

Aldose monosaccharides having 8, 9, and 10 carbon atoms in their chains have been synthesized.

Reactions. As polyhydroxy aldehydes or ketones, the monosaccharides undergo numerous reactions.

Reduction and oxidation. On reduction, the aldoses take up two atoms of hydrogen and are converted to the corresponding sugar alcohols. A pentitol or pentahydric alcohol is obtained from a pentose, and a hexitol or hexahydric alcohol is obtained from a hexose. On oxidation, the monosaccharides yield carboxylic acids. Mild oxidation converts aldoses first into the corresponding monocarboxylic acids with the same number of carbon atoms; thus, aldopentoses are transformed into pentonic acids, $\text{CH}_2\text{OH} \cdot (\text{CHOH})_3 \cdot \text{COOH}$, and aldohexoses into hexonic acids, $\text{CH}_2\text{OH} \cdot (\text{CHOH})_4 \cdot \text{COOH}$. With stronger oxidizing agents, the process may proceed further, and hexoses, for example, may be oxidized to the corresponding isomeric, saccharic, or tetrahydroxyadipic acids, $\text{COOH} \cdot (\text{CHOH})_4 \cdot \text{COOH}$. The ketoses, on oxidation, yield acids containing a smaller number of carbon atoms. The reducing properties of the monosaccharides are shown by their behavior with ammoniacal silver nitrate solution, from which metallic silver is precipitated, and particularly with Fehling's solution, from which, on warming, a brick-red precipitate of cuprous oxide is formed. This behavior is characteristic of aldoses as well as ketoses.

Phenylhydrazine. Emil Fischer's introduction in 1884 of phenylhydrazine, $\text{C}_6\text{H}_5\text{NHNH}_2$, as a reagent in sugar chemistry has proved of the greatest value in the separation and identification of the various monosaccharides. When 1 mole of phenylhydrazine

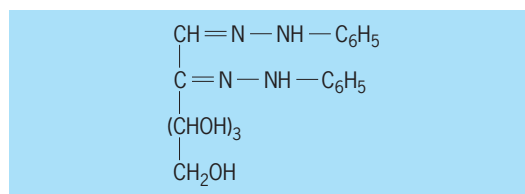
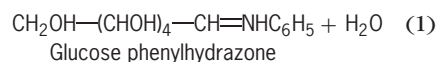
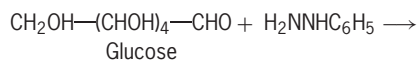
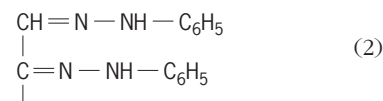


Fig. 1. Structural formula for hexosazone.

reacts with 1 mole of an aldose or ketose sugar, the first product is a hydrazone, as shown in reaction (1).



On warming with excess of phenylhydrazine, the hydrazone first formed is oxidized in such a way that the CHOH group adjacent to the original aldehydic or ketonic group is converted into a CO group. The latter then combines with another mole of phenylhydrazine to give a dihydrazone containing group (2). These compounds are termed osazones (Fig. 1).



The osazones are colored compounds which are difficult to purify. For this reason, their melting points and specific rotations cannot be relied upon. However, since the osazones produced by various sugars possess characteristic crystalline forms, they are frequently used for cursory identification of the parent sugar by examining the crystals under the microscope. The osotriazoles (Fig. 2) obtained by oxidation of the osazones with copper sulfate are to be preferred, because they are colorless crystalline compounds and have more definite melting points and higher specific rotations than the osazones.

Osazones, like all hydrazones, are hydrolyzed when heated with hydrochloric acid, resulting in regeneration of phenylhydrazine. The original sugar, however, is not recovered, because in the process

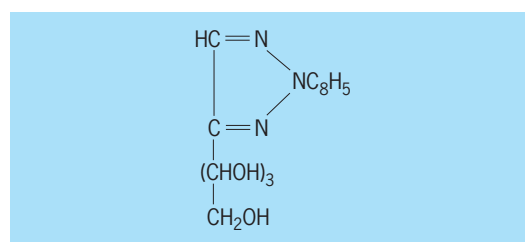
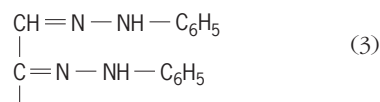


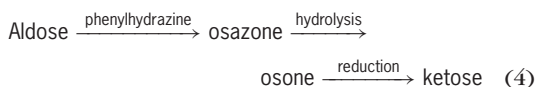
Fig. 2. Structural formula for hexosotriazole.

of regeneration of the sugar, group (3) is converted



into the group —CO—CHO. The highly reactive compound so formed is an oxidation product of the original sugar and is termed osone. In the example quoted above, the osazone of glucose yields glucosone, $\text{CH}_2\text{OH} \cdot (\text{CHOH})_2 \cdot \text{CO} \cdot \text{CHO}$.

The phenylhydrazine residues from sugar osazones may also be removed with a competing aldehyde, such as benzaldehyde or acetaldehyde, resulting in osone formation. On mild reduction of this compound with zinc dust and dilute acetic acid, the aldehydic group alone is attacked and converted into an alcoholic group, the keto group remaining unchanged. In the case of glucosazone, the sugar finally obtained is fructose, $\text{CH}_2\text{OH} \cdot (\text{CHOH})_3 \cdot \text{CO} \cdot \text{CH}_2\text{OH}$, in place of the glucose used as starting material. These reactions may be used as a general method of transforming an aldose into a ketose, according to scheme (4).



Cyanohydrin synthesis. Monosaccharides, such as aldehydes and ketones, react with hydrogen cyanide to form cyanohydrins. By the use of this reaction, which is due to H. Kiliani, the synthesis of a higher from a lower aldose can be effected. The cyanohydrins are first hydrolyzed to hydroxy acids, which are readily converted into lactones. The latter are then reduced to aldoses by means of sodium amalgam. Thus glucose, under these conditions, results in a new seven-carbon sugar, glucoheptose. This is shown in reaction sequence (5).

Similarly, by using the glucoheptose and continuing with the process of cyanohydrin synthesis, an octose can be obtained. The synthesis has been

carried as far as glucoheptose.

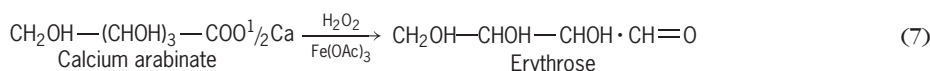
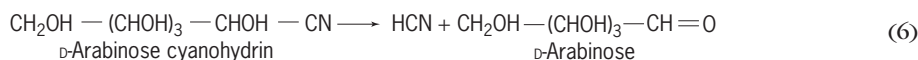
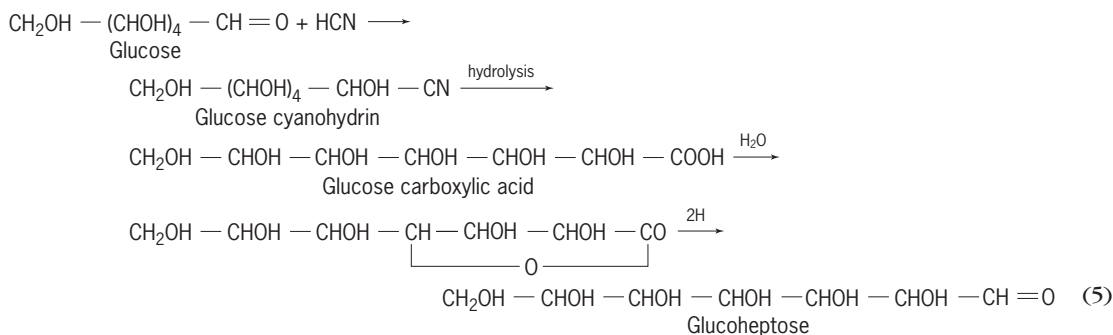
Since cyanohydrin synthesis introduces a new asymmetric carbon atom, two products are obtained from a single monosaccharide. As an example, both L-gluconic and L-mannonic acids are produced from L-arabinose by this process.

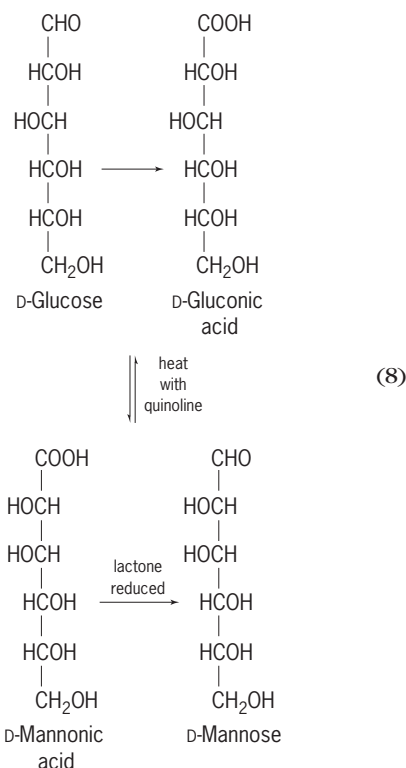
Hydroxylamine. Monosaccharides react with hydroxylamine to yield oximes, the aldehydic or ketonic oxygen being replaced by the group =N—OH. Using these compounds, A. Wohl devised a method for effecting degradation of an aldose to one of lower carbon content. Thus, when the aldoxime of D-glucose, $\text{CH}_2\text{OH}-(\text{CHOH})_4-\text{CH}=\text{NOH}$, is heated with acetic anhydride, it is converted into the acetyl derivative of the nitrile, $\text{CH}_2\text{OAc} \cdot (\text{CHOAc})_4 \cdot \text{CN}$. On treatment with ammoniacal silver nitrate, this compound is deacetylated and the cyanide is released, leaving the corresponding aldopentose, D-arabinose, as the free sugar, as shown in reaction (6).

It is also possible to reduce the number of carbon atoms in an aldose carbon chain by O. Ruff's method, in which the calcium salt of an aldonic acid is oxidized with hydrogen peroxide in the presence of ferric acetate. Thus, from calcium arabinatate, erythrose is obtained, as shown in reaction (7).

Epimerization. When two sugars or their derivatives, such as sugar acids, differ only in the configuration of the substituents on the carbon atom adjacent to the reducing group, they are called epimers. The aldonic acids are noteworthy for the ease with which they undergo epimerization or partial inversion of the asymmetry at carbon atom 2 upon heating with a weak base, such as pyridine or quinoline, to produce a mixture of the two epimers.

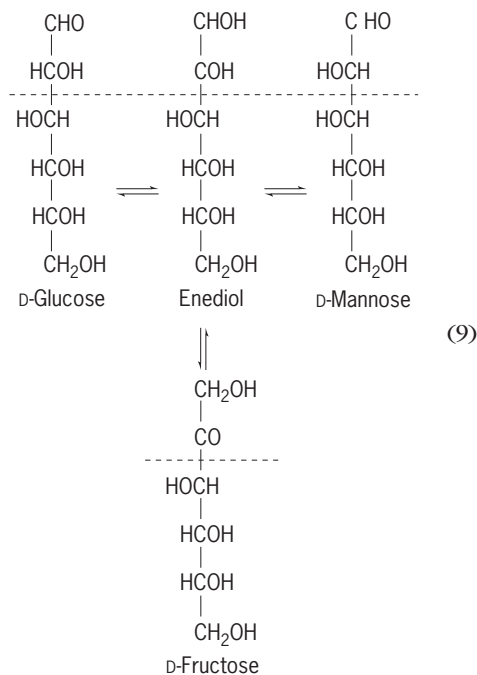
An aldose is first oxidized to the corresponding monocarboxylic acid, which is then heated with aqueous quinoline or pyridine, to yield a mixture of the epimeric aldohexonic acids. The latter may be separated by fractional crystallization of their lactones. Reduction of the lactone yields the corresponding aldose. This process may be illustrated by the transformation of D-glucose to D-mannose, as shown by reaction (8).





Enolization. Treatment of an aldose sugar with dilute alkali results in a mixture of an epimeric pair and 2-ketohexose. For example, when either D-glucose, D-mannose, or D-fructose is used, a mixture of these three sugars is obtained. The reaction, known as the Lobry de Bruyn-Ekenstein transformation, which is of general application, is due to the production of enolic forms in the presence of hydroxyl ions, followed by a rearrangement.

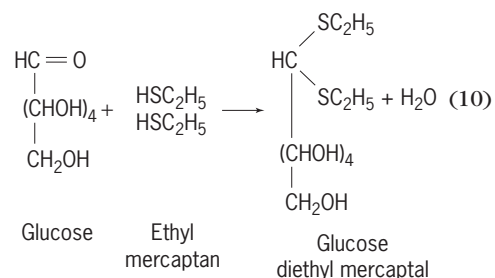
The transformation serves to show the close relationship between glucose, fructose, and mannose, as in notation (9). The structural representations of these three sugars are identical below the dotted line of the formula.



In a similar manner, the D-galactose series yields a mixture containing D-galactose, its epimer D-talose, and the ketose D-tagatose. When either D-xylose, D-lyxose, or D-xylulose is used, a mixture of the three sugars is obtained.

Derivatives. Derivatives of the monosaccharides are discussed in this section.

Sugar mercaptals. Reducing sugars, except ketoses, react with mercaptans in the presence of concentrated hydrochloric acid to form mercaptals. The reaction with glucose and ethyl mercaptan is given as an illustration in reaction (10).



The mercaptals of the sugars are well-defined crystalline compounds. They are of special interest because they are open-chain compounds and have been found useful for the preparation of other derivatives that have this structure.

Acetone sugars. The reducing sugars yield condensation products with aldehydes and ketones. Those with acetone have played an important role in solving problems of sugar structure. The acetone glucoses such as those shown in Fig. 3 are obtained by treating glucose in acetone with a condensing reagent such as zinc chloride or sulfuric acid.

Acetylated sugars. Acetylation of all free hydroxyls resulting in the formation of an ester may be accomplished by heating the monosaccharide with acetic anhydride in the presence of a catalyst such as anhydrous sodium acetate or zinc chloride, or by treating the sugar with acetic anhydride in a pyridine solution. When the hydroxyl at carbon 1 of an aldose is acetylated, two isomers (α, β) may be produced (Fig. 4). The directive influence of the catalyst used in the reaction is important. Glucose acetylated with

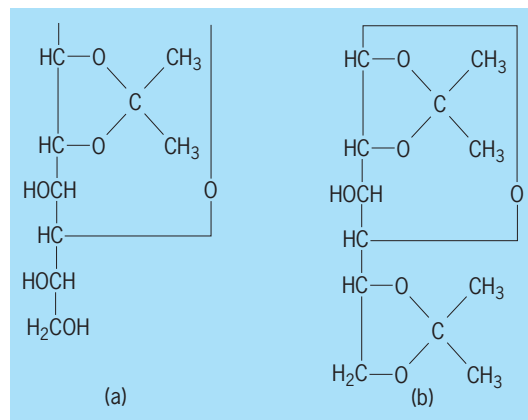


Fig. 3. Structural formulas for two acetone glucoses. Both are furanose derivatives. (a) 1,2-Monoacetone-D-glucofuranose. (b) 1,2-5,6-Diacetone-D-glucofuranose.

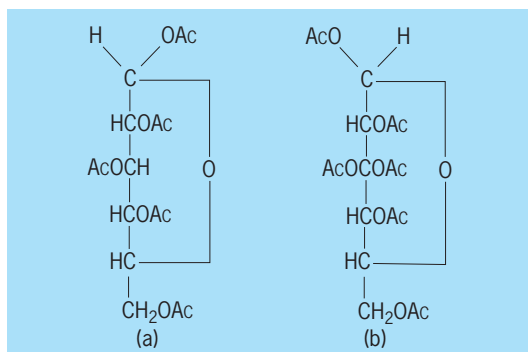
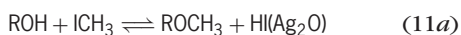


Fig. 4. Structural formulas for two isomers produced from acetylated D-glucose. (a) α -D-Glucose pentaacetate. (b) β -D-Glucose pentaacetate.

acetic anhydride in the presence of zinc chloride gives the α -pentaacetate, whereas with a sodium acetate catalyst the β form is produced. Furthermore, the β form, on heating with zinc chloride, is converted into the α isomer.

The free sugar may be recovered by deacetylation with dilute sodium hydroxide, or by catalytic deacetylation with barium or sodium methoxide.

Methylated sugars. Treatment of sugars with methyl iodide and silver oxide or with dimethyl sulfate and sodium hydroxide results in the formation of methylated derivatives (ethers) according to reactions (11).



Reaction (11a) is reversible; thus, in order to drive it to completion, the acid end product is removed by the addition of silver oxide.

The methyl ethers are the most common derivatives of this type and can be exemplified by fully methylated D-glucose, in which the hydrogens of all five free hydroxyls, including that of the glucosidic hydroxyl, are substituted by CH_3 groups (Fig. 5).

The glucosidic methoxyl group of the acetal is easily hydrolyzed with acid; however, the other methoxyls, which are true ethers, are resistant to acid as well as to alkali hydrolysis. The stability of the methoxyl group in these reagents makes the methylated sugars extremely useful in structural investigations. It is possible by appropriate methods to methylate selectively the hydroxyl groups in a monosaccharide. These partially methylated derivatives are utilized as reference compounds in the investigation of the constitution of oligosaccharides and polysaccharides.

Sugar phosphates. The naturally occurring phosphorylated sugars (including D-fructose-1,6-diphosphate, D-fructose-6-phosphate, D-glucose-6-phosphate, α -D-glucose-1-phosphate, D-glyceraldehyde-3-phosphate, and others) are of great metabolic importance. They function as intermediates in the processes of glycolysis, fermentation, and photosynthesis, and in most oxidative biological processes. Pentose phosphate esters occur as constituents of nucleic acids and a variety of coenzymes. See FERMENTATION; PHOTOSYNTHESIS.

Structurally the phosphorylated sugars are esters. The mono- and diesters of sugars are strongly acidic substances usually isolated as barium, calcium, lead, sodium, cyclohexylammonium, or alkaloid salts. D-Glucose-6-phosphate may be prepared by treating the 1,2,3,4-tetra-O-acetyl- β -D-glucose with diphenyl chlorophosphonate. The phenyl groups are removed from the resulting tetra-O-acetyl-D-glucose-6-diphenylphosphate by catalytic hydrogenolysis employing a platinum catalyst, and the product deacetylated by acid hydrolysis. Other phosphorylated sugars may be similarly prepared from the proper acetylated derivatives.

Aldose-1-phosphates may be prepared by reacting the poly-O-acetylglycosyl bromide with trisilver phosphate or silver diphenylphosphate. The resulting phosphate triester is simultaneously deacetylated and hydrolyzed under controlled conditions to give the aldose-1-phosphate. By this procedure, α -D-glucose-1-phosphate, α -D-galactose-1-phosphate, α -D-mannose-1-phosphate, and α -D-xylose-1-phosphate have been prepared. The β anomers may be prepared by reacting the poly-O-acetylglycosyl bromide with silver dibenzylphosphate, followed by catalytic hydrogenation to remove the benzyl groups and saponification of the acetyl groups. They also can be synthesized by coupling the poly-O-acetylglycosyl bromide with mono silver phosphate as the phosphorylating agent.

Sugar alcohols. These are acyclic linear polyhydric alcohols. They may be considered sugars in which the aldehydic group of the first carbon atom is reduced to a primary alcohol group. They are classified according to the number of the hydroxyl groups in the molecule. Thus, erythritol with four hydroxyls is considered to be a tetrose; arabitol with five hydroxyls is a pentose.

Sorbitol (D-glucitol, sorbite) is one of the most widespread of all the naturally occurring sugar alcohols (Fig. 6a). It is found in higher plants, especially in berries and also in algae (seaweeds). Mannitol (Fig. 6b), like sorbitol, is widespread among plants. However, unlike sorbitol, it is frequently found in exudates of plants.

As a group, the sugar alcohols are crystalline substances, having low specific rotations, and ranging in taste from faintly sweet to very sweet. Their

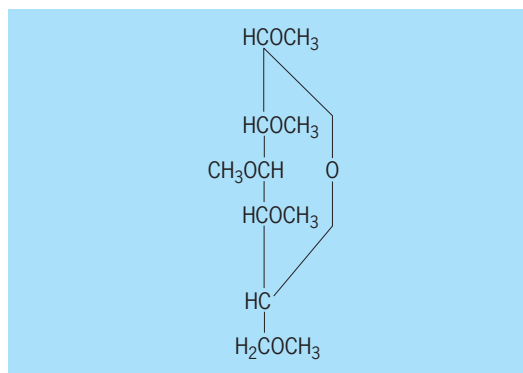


Fig. 5. Structural formula for methyltetra-O-methyl- α -D-glucose, a methylated ether.

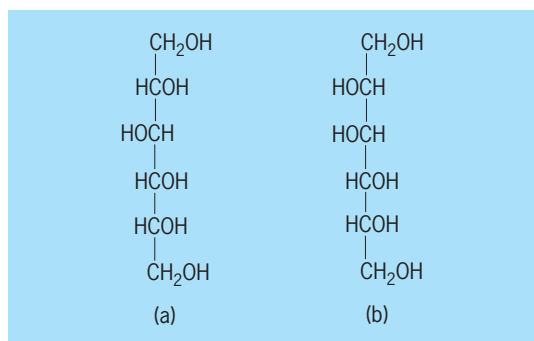


Fig. 6. Structural formulas for two common hexitols. (a) Sorbitol. (b) D-Mannitol.

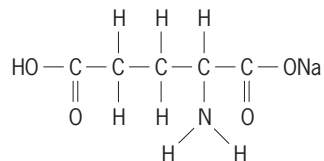
distribution in nature is limited exclusively to plants.

William Z. Hassid

Bibliography. H. S. Khadem, *Carbohydrate Chemistry: Monosaccharides and Their Oligomers*, 1988; W. W. Pigman, *The Carbohydrates*, 2d ed., vols. 1A, 1B, 2A, and 2B, 1970-1978.

Monosodium glutamate

The single sodium salt of glutamic acid used in foods to accentuate flavors. It is also known as MSG. The molecular structure is:



The crystal form available in commerce is the monohydrate, with structure as represented plus one molecule of water of hydration.

Glutamic acid is one of the more common of the amino acids. Its structural formula is the same as the monosodium salt excepting that the —COONa group on the right is replaced by a —COOH group, making the right end similar to the left. See AMINO ACIDS.

The formulas for glutamic acid and its salts show an asymmetric carbon atom. This is the fourth carbon atom from the left. It is attached to four entirely different groups. Therefore, the acid itself and each of its salts exist in three forms, the two isomers, L and D, and the racemic of D,L. The form is the so-called natural or active isomer, and its monosodium salt has the power of bringing out or emphasizing flavors, as distinguished from tastes, of certain foods, notably fish, fowl, meat, and vegetables. It is not a flavoring agent but, like salt, aids in developing the savor of foods. As a major constituent of all proteins, it participates in many of the metabolic processes.

Source. Originally produced from seaweed in the Orient, principal modern production is from cereal glutens, such as those of wheat, corn, and soybeans, from solutions evolved in the manufacture of beet sugar, and by microbiological fermentation of carbohydrates. To be commercially feasi-

ble as sources, the proteins from cereals must be concentrated and cheap. The two raw materials used for the greater proportion of commercial production are wheat gluten and desugared beet-sugar molasses. The world's largest single producer is a Japanese firm, principally using wheat gluten. In the United States, a number of factories exist; and several work solutions from beet-sugar molasses. The largest United States manufacturer uses a fermentation process and also what is commonly termed concentrated Steffen filtrate (CSF). This results from the multiple effect concentration at the sugar factories of the dilute waste liquor produced in recovering sugar from beet molasses by the Steffen process. Another factory uses liquor from both the Steffen and the barium (Deguide) process of sugar recovery.

Glutamic acid appears in the sugar beet as glutamine. During the sugar process, the glutamine changes to the internal anhydride of glutamic acid, pyrrolidone carboxylic acid. The latter is readily hydrolyzed by heating with either acid or alkali, and since it is not a protein, the glutamic acid which results is in the desired L form.

Although basically simple, all of the processes used are somewhat complex because of the many organic substances present in the raw materials. However, commercial production attains a high degree of purity of product, over 99.9% monosodium glutamate.

Glutamic acid produced by the microbiological method using a carbohydrate as raw material is manufactured by a United States pharmaceutical manufacturer and a Japanese firm.

Synthesis. A number of methods for synthesizing glutamic acid from several raw materials have been published in the scientific literature. Synthesis, however, invariably results in the racemized or D,L form. Methods for the resolution of this are known, but the few which have been made public are complex and costly. Commercially feasible methods of synthesis and resolution have been developed by two United States chemical manufacturers, one of which is a large producer of monosodium glutamate.

Uses. United States production includes glutamic acids, its hydrochloride, the mono-salts of sodium, potassium, ammonium, and calcium glutamates, all in the L forms. These find uses in medicine. By far the greatest proportion of glutamic acid production is used as raw material in making monosodium glutamate for the food industry. It is used both in the processing of foods and in the institutional and domestic fields. The Food and Drug Administration has approved it as an additive, and since in itself it is not a flavoring agent, its label designation in use is as a constituent and not as an artificial flavor. Originally, the latter was required because the product made then had a meatlike flavor due to impurities.

Monosodium glutamate is recognized as a standard of identity ingredient in several commercial food preparations. While it has been toxicologically cleared for use in food, there have been reports of incidences of allergic reactions. It is also contraindicated for people who must maintain low-sodium

diets; it is not permitted as an additive in baby food for infants under 12 weeks of age.

Its principal use is in the preparation of canned and dried soups, but it also enters into the production of some meat, vegetable, fowl, and fish products.

Monosodium glutamate has been and is of great importance in the Oriental diet. In the Orient, it has even been used as a medium of exchange and is often diluted with lactose or salt in order that it may be sold at a price within reach of the poor. See FOOD ENGINEERING.

Paul D. V. Manning

Bibliography. H. B. Heath and G. A. Reineccius, *Flavor Chemistry and Technology*, 1986; I. D. Morton and A. J. McLeod (eds.), *Food Flavors: pt.A: An Introduction*, 1982; N. D. Pintauro, *Sweeteners and Enhancers*, 1977.

Monotremata

The single order of the mammalian subclass Prototheria. Two living families, the Tachyglossidae and the Ornithorhynchidae, make up this unusual order of quasi-mammals, or mammal-like reptiles.

The known species are highly specialized for their mode of life, and monotreme anatomy is therefore a curious mixture of reptilian and mammalian features along with highly diagnostic monotreme features. The fossil record of the group is virtually unknown, and Pleistocene forms can be classified within the living genera. Although ties with Jurassic multituberculates, docodonts, and other groups have been advocated, monotreme relationships are obscure. It is likely that they represent a divergent form of therapsid reptiles, independent of all other known groups of mammals, without fully attaining a truly mammalian grade of biological organization.

Morphology. The rostrum on the birdlike, toothless (in the adult) skull is covered with a leathery beak; lacrimals and auditory bullae are absent; cervical ribs are present; all vertebrae, except caudals, are without epiphyses; and large epipubic bones are characteristic. The scapula spine is restricted to the anterior ridge and the suprascapular fossa is absent; large precoracoids, coracoids, interclavicles, clavicles, and scapulae form the pectoral girdle. Monotremes are oviparous with telolecithal and meroblastic eggs. The poorly developed uteri are unfused. A cloaca is present, and the ureters open into a papilla in the urogenital sinus. The mammae are teatless; the penis is attached to the ventral wall of the cloaca; the prostate and the seminal vesicle are absent; the testes are abdominal; and there is one pair of bulbourethral glands. There is no corpus callosum, and the non-glandular stomach is lined with stratified epithelium. Endothermy is poorly developed.

Tachyglossidae. The echidnas (spiny anteaters) have relatively large brains with convoluted cerebral hemispheres. The known genera, *Tachyglossus* and *Zaglossus*, are terrestrial, feeding on termites, ants, and other insects. They are capable diggers, both to obtain food and to escape enemies. Like hedgehogs, they can erect their spines and withdraw their limbs

when predators threaten. The long, sticky tongue can be extended 6 or 7 in. (15–18 cm) past the snout. Echidnas have well-developed senses of smell and hearing but their vision is poor. Commonly one egg, but occasionally two or even three, is laid directly into the marsupium (pouch) of the mother where it is incubated for up to 10 days. When the young hatch, they are poorly developed and remain in the pouch for about 2 weeks.

Species of *Tachyglossus* live in rocky areas, semideserts, open forests, and scrublands. They are found in Australia, Tasmania, New Guinea, and Salawati Island. Species of *Zaglossus* are found in mountainous, forested areas.

Ornithorhynchidae. The duck-billed platypus has a relatively small brain with smooth cerebral hemispheres. The young have calcified teeth, but in the adult these are replaced by horny plates which form around the teeth in the gums. The snout, as the name indicates, is duck-billed; the tongue is flattened out to work against the palate. The pelage is made of soft hairs, and the well-developed tail is flattened to aid in swimming.

The semiaquatic platypus is a capable swimmer, diver, and digger. When it is diving, its eyes and ears are closed by folds of fur. The bulk of its diet consists of aquatic vegetation, worms, aquatic insect larvae, crustaceans, and mollusks. Two eggs are usually laid by the female into a nest made of damp vegetation. After incubating the eggs for about 10 days the female leaves, and returns only when the eggs are hatched.

The platypus is found in Australia and Tasmania in almost all aquatic habitats, ranging from muddy streams and ponds to clear, rapid mountain streams. See MAMMALIA; PROTOTHERIA.

Frederick S. Szalay

Bibliography. A. F. DeBlase and R. E. Martin, *A Manual of Mammalogy*, 2d ed., 1980; R. M. Nowak and J. L. Paradiso (eds.), *Walker's Mammals of the World*, 2 vols., 6th ed., 1999.

Monsoon meteorology

The study of the structure and behavior of the atmosphere in those areas of the world that have monsoon climates. In lay terminology, monsoon connotes the rains of the wet summer season that follows the dry winter. However, for mariners, the term monsoon has come to mean the seasonal wind reversals.

In true monsoon climates, both the wet summer season that follows the dry winter and the seasonal wind reversals should occur. Winds from cooler oceans blow toward heated continents in summer, bringing warm, unsettled, moisture-laden air and the season of rains, the summer monsoon. In winter, winds from the cold heartlands of the continents blow toward the oceans, bringing dry, cool, and sunny weather, the winter monsoon.

Based on these criteria, monsoon climates of the world include almost all of the Eastern Hemisphere tropics and subtropics, which is about 25% of the surface area of the Earth (Fig. 1). As shown in Fig. 1, the

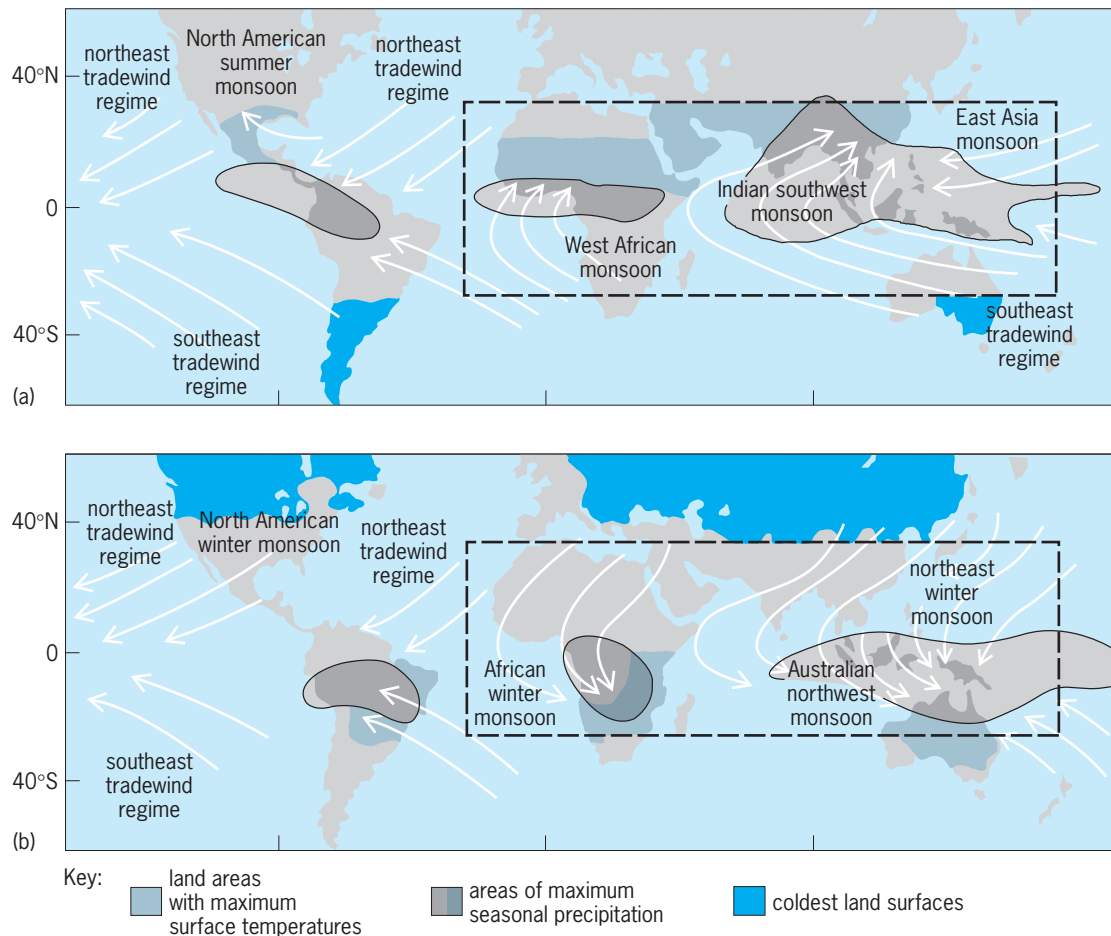


Fig. 1. Domains of the principal monsoon systems during the Northern Hemisphere (a) summer and (b) winter. Using the criteria of seasonal wind reversal and distinct wet summers and dry winters, the monsoon region is the region outlined by the broken rectangle. (After P. J. Webster, *The elementary monsoon*, in J. S. Fein and P. L. Stephens, eds., *Monsoons*, Wiley, 1987)

areas of maximum seasonal precipitation straddle or are adjacent to the Equator. Two of the world's areas of maximum precipitation (heavy rainfall) are within the domain of the monsoons: the central and south African region, and the larger south Asia-Australia region. The monsoon surface winds emanate from the cold continents of the winter hemisphere, cross the Equator, and flow toward and over the hot summer-hemisphere landmasses.

India presents the classic example of a monsoon climate region, with an annual cycle that brings southwesterly winds and heavy rains in summer (the Indian southwest monsoon; Fig. 1a) and northeasterly winds and dry weather in winter (the northeast winter monsoon; Fig. 1b).

Basic driving force. Like all weather systems on Earth, monsoons derive their primary source of energy from the Sun. About 30% of the Sun's energy that enters the top of the atmosphere is transmitted back to space by cloud and surface reflections. Little of the remainder is absorbed directly by the clear atmosphere; it is absorbed at the Earth's surface according to a seasonal cycle. The opposition of seasons in the Northern and Southern hemispheres leads to a slow movement of surface air across the

Equator from winter hemisphere to summer hemisphere, forced by horizontal pressure gradients and vertical buoyancy forces resulting from differential seasonal heating. Such a seasonally reversing rhythm is most pronounced in the monsoon regions. See ALBEDO; ATMOSPHERE; CLIMATOLOGY; EARTH ROTATION AND ORBITAL MOTION; INSOLATION.

Land and water respond differently to equal amounts of solar energy, for two reasons. First, the specific heat of water is twice that of dry soil; for the same amount of heating, the temperature of dry land increases twice as much as that of water. Second, and more important, the effective heat capacity, that is, the ability of a material to store heat, is many orders of magnitude larger for the oceans than for the land. Being translucent, the ocean transmits radiational energy through its surface layer to some depth. In addition, unlike a solid, which transfers heat solely by molecular diffusion, the oceans transfer heat by turbulent mixing to depths of about 50–100 m (150–300 ft). In contrast, the depth to which dry land transfers heat is about 1 m (3 ft). This difference in distribution of incoming energy in summer results in large horizontal surface-temperature gradients between hot land and cooler oceans within

the monsoon regions. *See* HEAT CAPACITY; SPECIFIC HEAT.

In winter, surfaces on Earth lose more heat through terrestrial radiation to space than they receive from the Sun. Heat stored at depth during the previous summer is transferred to surface layers, where it replaces part of the heat that has been lost. Because the heat stored in the oceans is much greater than that in the land, ocean surfaces cool much less than land surfaces. Thus the oceans act as a heat buffer, like a giant, sluggish flywheel, storing large quantities of heat in summer and releasing it in winter, thereby suppressing extreme seasonal temperature variations such as those that occur over neighboring landmasses. *See* HEAT; HEAT BALANCE, TERRESTRIAL ATMOSPHERIC; TERRESTRIAL RADIATION.

Annual cycle. The Indian monsoon is both the classical and dominant monsoon system of the world. The other monsoon systems shown in Fig. 1 operate under the same physical principles with variations that depend on geography and scale.

Monsoon summer circulation. As summer approaches in a hemisphere, the energy received from the Sun at the Earth's surface becomes greater than its radiant energy loss to space. Land surfaces heat rapidly and intensely. This is particularly true of the empty quarter of Saudi Arabia, one of the great subtropical deserts, and of the Tibetan Plateau, which, at an average elevation of 4 km (2.4 mi), dominates the Asian continent. These heat sources loosely define the western and northern boundaries of the Indian summer monsoon (Fig. 1). By May, northern India, dry since the previous summer, is scorched by air temperatures higher than 100°F (about 40°C). In contrast, land surfaces cool in the opposite (winter) hemisphere, because they radiate more energy than they receive. The oceans experience the same trends as the land in each hemisphere, but to a greatly reduced extent. Also, the maximum and minimum land surface temperatures occur very close to the solstices, while those of the ocean surface waters lag by about 2 months. The net result is intensely heated land surfaces over southern Asia and northern Africa adjacent to a cooler Indian Ocean. *See* SOLSTICE.

As the Asian monsoon regions approach their maximum temperatures, the density of the heated air over land decreases, and a horizontal gradient of surface pressure between ocean and land develops. The pressure gradient and buoyancy forces caused by the heated air produce near-surface convergent motion, which causes moisture-laden air to move across the Equator and Indian Ocean inward and upward over the hot, low-pressure area of southern Asia. Completing the circulation, the rising air diverges aloft out of a large high-pressure area that develops by late May at jet-stream level (about 10 km or 6 mi) over the Tibetan Plateau and south-central Asia. This diverging air then slowly subsides to lower levels, replacing the air being drawn to the continents. *See* JET STREAM.

The actual paths of the winds are curved because of the Coriolis acceleration. Low-level winds spiral

counterclockwise into the low-pressure centers over the Indian subcontinent (Fig. 1) and clockwise out of the upper-level high-pressure center over the Tibetan Plateau. These wind patterns are the classical cyclonic and anticyclonic wind circulations around lows and highs. *See* ATMOSPHERIC GENERAL CIRCULATION; CORIOLIS ACCELERATION; CYCLONE; WIND.

Monsoon rains. The rising air currents over the Indian subcontinent enter environments of reduced pressure aloft. The air thus expands and cools, and the moisture carried up from the surface layers in the form of water vapor condenses into clouds that ultimately produce the rains. The condensation process releases the latent heat stored in the water vapor molecules that were evaporated from the surface waters and carried aloft. This heat is a major source of added buoyancy for the monsoon circulation. The Ghats mountain range along the west coast of India and the massive Tibetan Plateau north of India provide orographic (mechanical) forcing, which further enhances condensation and precipitation processes.

The Asian summer monsoon rains last for about 100 days, starting in early to mid-June and ending in September. The date of the start of the rains over India—the onset—is fairly consistent from year to year, but it does vary infrequently by more than 30 days. The average onset date at Kerala, on the southern tip of India (8°N latitude), is June 1, with a standard deviation of about 1 week. The monsoon front advances slowly northwestward through the month of June; the average onset date is June 10 at Bombay (19°N) and June 15 at Delhi (28.5°N). By mid-July, the entire Indian subcontinent is under the influence of the summer monsoon. Water resources are so delicately balanced in India that distress situations occur if the onset is delayed by more than a week. Although onset dates have been observed to fluctuate by more than a month, studies of monsoon onsets over the west coast of India show the interesting and counterintuitive result that rainfall amounts, both for the month of June and for the total summer monsoon season, are not related to the date of onset. Rather, total monsoon seasonal rainfall amounts appear to be affected most by a phenomenon known as the break monsoon.

The phrase “break in the rains” refers to situations of reduced or no rainfall. During break conditions, reductions in rainfall over most of the Indian subcontinent are accompanied by enhanced rainfall over extreme northern India, in the foothills of the Himalayas. These shifts in rainfall patterns are associated with northward shifts of normal surface pressure and circulation patterns, but cause-and-effect relationships remain unknown. A study of 80 years (1888–1969) of monsoons showed that the range of the duration of break conditions was 3–21 days. Within the 80 years analyzed, there were 56 cases of break monsoons in July and 57 in August, with most lasting 3–5 days.

Figure 2 shows the seasonal (June–September) normal average rainfall for India and its standard

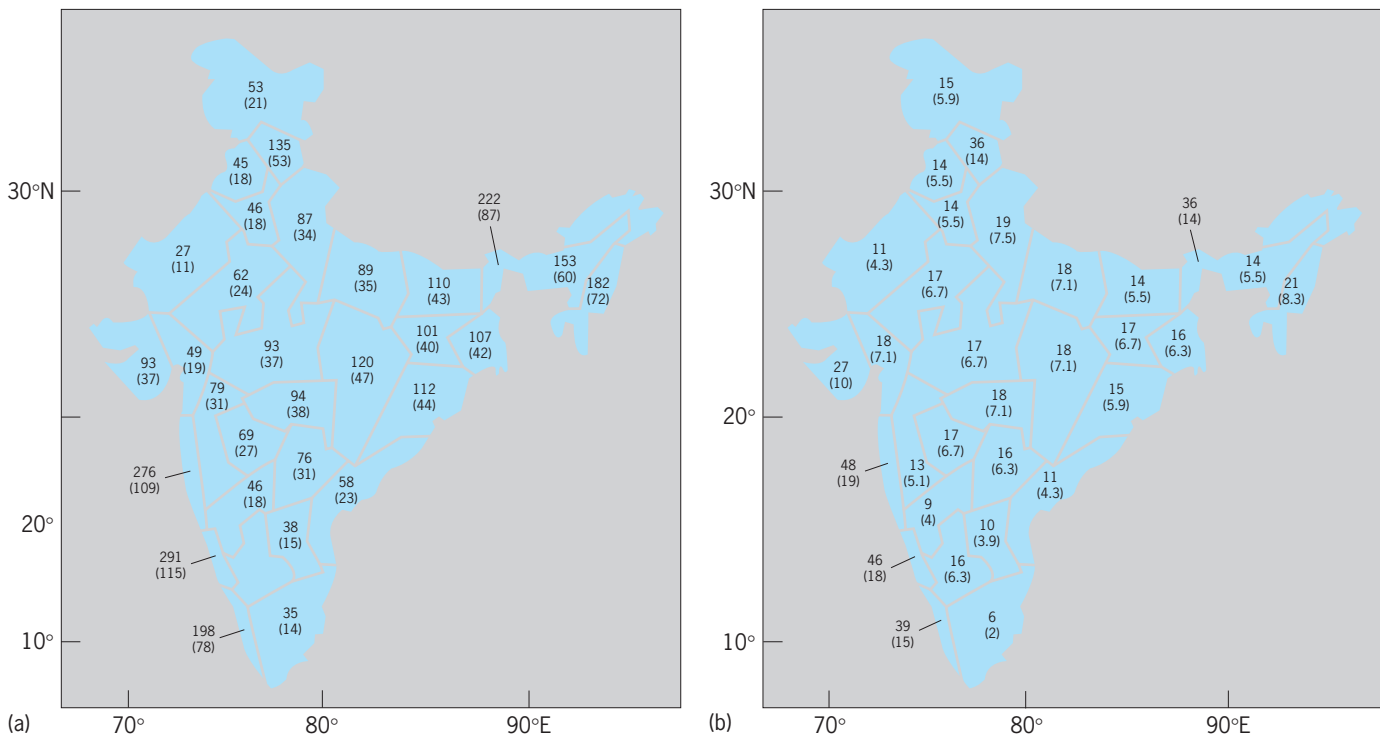


Fig. 2. Rainfall distribution in India, shown as number of centimeters (inches) in each subdivision (the value for one subdivision is not available). (a) Seasonal (June–September) mean rainfall calculated from 70 years (1901–1971) of data. (b) Standard deviation of seasonal mean rainfall calculated from 81 years (1901–1981) of data. (After J. Shukla, *Interannual variability of monsoons*, in J. S. Fein and P. L. Stephens, eds., *Monsoons*, Wiley, 1987)

deviation for the outlined subdivisions. Not shown are extreme local values, such as 800 cm (315 in.) at Cherrapunji (25.15°N, 91.44°E, elevation 1313 m or 4308 ft), which is the rainiest site measured on Earth, with more than 1000 cm (390 in.) of rain annually. The largest values of subdivisional mean rainfall occur over the west coast and northern regions of India, where orographic influences are important.

The total summer seasonal rainfall shown in Fig. 2a accounts for about 80% of the total annual rainfall for the Indian subcontinent. Of course, natural events are generally much more complex than models created on the basis of human perception. Few places in India are absolutely dry from October to May. Most areas receive up to a few centimeters of rain during these months. However, the state of Tamilnadu, in south India east of the coastal Ghats, and in a rain shadow region during the summer monsoon, receives most of its annual rainfall during October–December. In September, the summer monsoon rains withdraw southward across India. This retreat accompanies the equatorward movement of the regions of lowest surface pressure, the monsoon trough, which during the summer monsoon season is oriented east-west at about 25°N over north-central India. The monsoon trough reaches Tamilnadu in late October, concurrent with the start of its rainy season and, for that region, a wet winter monsoon. See RAIN SHADOW.

At this time, the rest of India begins its transition from wet to dry, from southwest to northeast winds,

from summer to winter monsoon. By winter solstice, the subcontinent is uniformly cool and dry with northerly winds. These conditions prevail through February, a period of cool, dry, sunny weather over virtually all of India. The transition from March to early June brings hot weather, broken only occasionally by premonsoon thunderstorms. Finally, by early June the southwesterlies invade the southern coast of India and advance northward, once again bringing the rains of the wet season, the summer monsoon, to India. This annual cycle of wet and dry weather, profoundly affects the people of the Eastern Hemisphere. See ASIA; METEOROLOGY; PRECIPITATION (METEOROLOGY); TROPICAL METEOROLOGY.

Jay S. Fein

Bibliography. C. P. Chang (ed.), *East Asian Monsoon*, 2004; C. P. Chang and T. N. Krishnamurti (eds.), *Monsoon Meteorology*, 1987; J. S. Fein and P. L. Stephens (eds.), *Monsoons*, 1987; S. Hastenrath, *Climate Dynamics of the Tropics*, 1991; R. N. Keshavamurty, *Physics of Monsoons*, 1992; M. J. Lighthill and R. P. Pearce (eds.), *Monsoon Dynamics*, 1981.

Monstrilloida

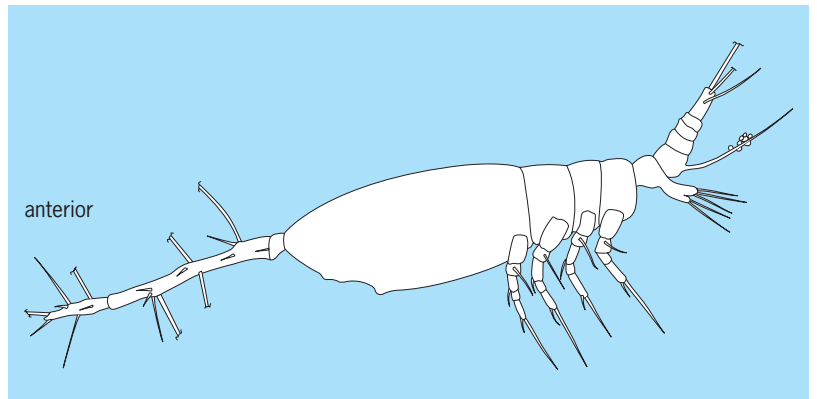
A copepod order of crustaceans containing 122 species in five genera, *Cymbasoma*, *Guanabaraenia*, *Monstrilla*, *Monstrillopsis*, and *Thaumatobesia*; Monstrillidae is the only family. Previously, the

five species of the family Thaumatopsyllidae had been included as a second family of Monstrilloida, but in 2003 this family was removed to its own order, Thaumatopsyllida, and unpublished observations of the development support this decision. See COPEPODA; CRUSTACEA.

Description. Adult females (1.5–2.5 mm or 0.07–0.11 in.) and males (1.0–1.5 mm or 0.04–0.07 in.) of the Monstrilloida are free-swimming and are often collected in plankton samples taken near the surface of shallow marine waters. The body has a podoplean architecture (the last two thoracic somites with the posterior part of the body), although the broader anterior part merges indistinctly with the narrower posterior part (see **illustration**). The second thoracic somite, bearing the first swimming leg, does not articulate with the first thoracic somite. Although antenna 1, the four swimming legs, and legs 5 and 6 are present, the feeding limbs, which include antenna 2, mandible, maxilla 1, maxilla 2, and the maxilliped, fail to form at this stage, and adults are assumed not to feed. Individual embryos often are attached to a long, spinelike attenuation of the female's genital somite. An embryo hatches as a feeble-swimming, simple nauplius (a crustacean larval stage) with well-developed antenna 1, antenna 2, and mandible. This nauplius penetrates the body of a polychaete or a mollusk; a hooklike structure on the mandibular endopod facilitates initial attachment. The next phase of the life cycle is passed within the blood system of the host as a sac-like form with several processes; this endoparasitic phase cannot be identified as a copepod. The monstrilloid leaves its host as an immature copepodid, apparently the fifth copepodid, and molts once to the adult. Multispecific aggregations of adults may occur at dusk, although the advantage to aggregating is not clear. This life cycle has been reported for only a few species and early in the last century; Alphonse Malaquin provides the most complete description. Monstrilloids appear to be found throughout the world's oceans, so it is surprising that their interesting life cycle has not attracted contemporary studies.

Taxonomy. Descriptions of new species continue to remain of interest to taxonomists. For example, 18 of the 122 species, almost 15%, have been described since 1999. This increase has been stimulated, in part, by recent revisions of several genera. Unfortunately, contemporary systematic analyses have not been applied to generic diagnoses, so the number of genera has remained static. The oldest genus, *Monstrilla*, was diagnosed by James Dwight Dana in 1849 and the youngest, *Guanabaraenia*, by Lejeune P. H. de Oliveira in 1945.

Phylogeny. The order Monstrilloida often is proposed as a close relative of the order Siphonostomatoida, despite the fact that the former are missing five limbs which are present on the latter. Closer inspection of this proposal, however, shows that this relationship is based on shared characters that have evolved many times among other copepods, and so do not provide the best evidence of common an-



Side view of a monstrilloidan.

cestry. Phylogenetic relationships of Monstrilloida with other copepod orders may remain problematical until an analysis of segment homologies of swimming legs and antenna 1 is undertaken.

Frank D. Ferrari

Bibliography. M. J. Grygier, Annotated chronological bibliography of Monstrilloida (Crustacea: Copepoda), *Galaxea*, 12, 1995; M. J. Grygier, Nomenclature, redescription, and new record from Okinawa of *Cymbasoma morii* Sekiguchi, 1982 (Monstrilloida), *Hydrobiologia*, 292/293, 1994; M. J. Grygier and S. Ohtsuka, SEM observation of the nauplius of *Monstrilla hamatapex*, new species, from Japan and an example of upgraded descriptive standards for monstrilloid copepods, *J. Crust. Biol.*, 15, 1995; A. Malaquin, Le parasitisme évolutif des Monstrillides (Crustacés Copépodes), *Arch. Zool. Exp. Gén.*, Série 3, vol. 9, 1901; E. Suárez-Morales, An aggregation of monstrilloid copepods in a western Caribbean reef area: Ecological and conceptual implications, *Crustaceana*, vol. 74, 2001; E. Suárez-Morales and C. Diaz, A new species of *Monstrilla* (Crustacea: Copepoda: Monstrilloida) from Brazil with notes on *M. brevicornis* Isaac, *Proc. Biol. Soc. Wash.*, vol. 114, 2001.

Monte Carlo method

A technique for estimating the solution, x , of a numerical mathematical problem by means of an artificial sampling experiment. The estimate is usually given as the average value, in a sample, of some statistic whose mathematical expectation is equal to x . In many of the useful applications, the mathematical problem itself arises in a problem of probability in physics or other sciences, operational research, image analysis, general statistics, mathematical economics, or econometrics. The importance of the method arises primarily from the need to solve problems for which other methods are more expensive or impracticable, and from the increased importance of all numerical methods because of the development of the electronic digital computer.

The method as described above is identical with the earlier method known as artificial or model

sampling, or simulation. In fact, the term “simulation” is appropriate when the mathematical problem arises directly from a model of a real-world situation. The main justification for the name Monte Carlo is that since the mid-1940s several tricks have been introduced for improving the efficiency of the method, so that the subject has acquired a special flavor.

One of the earliest examples of artificial sampling was the experimental estimation of π by the French naturalist G. L. L. Buffon in 1773. The method is to throw a needle on a striped tablecloth and see how often it falls touching more than one stripe. If the width of each stripe is equal to the length of the needle, then the proportion of “successes” will be close to $2/\pi$ for a long series of trials.

Classification of Monte Carlo methods. The usual method of applied mathematics is to replace a physical problem P by a mathematical problem M , by assuming an adequate mathematical model; and then to solve the mathematical problem, and thus to solve P (perhaps only approximately). However, sometimes the mathematical problem, when replaced by a numerical problem, is solved with the aid of a calculating machine, so M is replaced by a new physical system, P' . One may think of the matter in this way, especially if the calculating machine is analog, that is, a machine using nondigital arithmetic. The method then becomes $P \rightarrow M \rightarrow P'$, where the arrow means “is replaced by.” If $P = P'$, it can be said that P is solved by the crude method or direct experimental method, in which M is inessential. See ANALOG COMPUTER.

A special type of experiment with physical apparatus is a statistical experiment S , which makes use of dice, coins, needles, or Geiger counters. Sometimes random sampling numbers are used, that is, digits generated in such a manner that each selection of a digit has an independent chance of 0.1 (or 0.5) of giving each of the digits 0, 1, 2, . . . , 9 (or 0,1). More often, the digits are pseudorandom; that is, they are produced deterministically but have the appearance of being flat-random, like the digits in the decimal or binary expansion of $\sqrt{2}$. If a distinction is made between physical and other statistical experiments, there are, among others, the following methods of solving problems: $S \rightarrow S$, $S \rightarrow S'$, $S \rightarrow M$, $M \rightarrow S$, $S \rightarrow M \rightarrow S'$.

The method $S \rightarrow S$ involves estimating the solution of a real-life statistical problem by direct sampling. It may be regarded as a crude form of the Monte Carlo method. In $S \rightarrow S'$, a real-life statistical problem is replaced by a simpler model from which estimation is made by sampling; this method is perhaps best regarded as simulation. $S \rightarrow M$ is the ordinary form for the application of mathematical statistics. $M \rightarrow S$ is exemplified by Buffon's needle problem and was suggested for serious applications by Enrico Fermi, John von Neumann, Stanislaw Ulam, and Nicholas Metropolis in the 1940s. At that time it might have been reasonably called the true Monte Carlo method, but this name is no longer appropriate because of a change of emphasis. It is the method $S \rightarrow M \rightarrow S'$

which has been of most interest. This method can be of varying degrees of sophistication, depending on the amount of mathematical ingenuity required in order to transform S into S' via M .

Advantages and disadvantages. The main advantage of Monte Carlo is that other methods can be more costly or impracticable. A familiar example is the estimation of the probability of winning a game of pure chance: Sometimes the only reasonably simple method of estimation is to play the game several times. There are also numerical problems that can be solved by deterministic methods but can be more simply solved approximately by the Monte Carlo method. Sometimes poor approximations are satisfactory because the aim is merely to determine the strategic variables of a problem. This is likely to be a fruitful technique in mathematical economics. To simplify a complicated economic model without losing touch with reality, one could carry out an approximate Monte Carlo solution for the model, find out which of the variables are important, and then perhaps attempt a mathematical solution using only the important variables.

Another situation where a poor approximation is satisfactory occurs when there is available an iterative method of calculation, that is, a method of successive approximation, which converges closely to the right answer in a reasonable time provided that the first trial solution is not too far from the truth. The Monte Carlo method may then perhaps be used for obtaining a first trial solution. Modern Monte Carlo techniques are themselves usually iterative.

Sometimes the expense of a Monte Carlo method does not increase as fast as that of other methods when the dimensionality of a problem is increased. This seems to be true for multiple integration when it cannot be done analytically, and for the solution of Schrödinger's equation for several particles. See SCHRÖDINGER'S WAVE EQUATION.

The main disadvantage of some Monte Carlo methods is that for each extra decimal place required, it is necessary to multiply the sample size by 100. Thus, to calculate π to five decimal places by throwing a needle would require about 10^{10} throws, or 1 throw per second for about 300 years.

Applications. The method has been applied to the following problems, among others:

1. Size of cosmic-ray showers.
2. Critical size of nuclear reactors.
3. Other neutron transport problems, concerning, for example, the shielding properties of water or graphite. The probability that a neutron will cause a branching process that penetrates a shield may be as low as 10^{-10} . In a sample of random walks of reasonable size, there will be no “successes” unless some tricks are used.
4. Enumeration of high-polymer molecules or number of self-avoiding walks on a diamond lattice.
5. Percolation of a liquid through a solid.
6. Brownian motion and diffusion.

7. Birth-and-death branching stochastic processes.
8. Autoregressive time series.
9. Theory of queues, and other problems of commercial importance, such as storage, equipment replacement and maintenance, and insurance problems.
10. Laplace's partial differential equation.
11. Schrödinger's partial differential equation.
12. Integral equations.
13. Inversion of matrices.
14. Evaluation of definite integrals.
15. Random rounding-off.
16. Discovery of the t -distribution by Student.

These applications were all well studied before 1960. Since then the range of applications has expanded to include nearly every current numerical topic. See LATTICE (MATHEMATICS); QUEUEING THEORY.

Techniques. When a Monte Carlo method is performed on an electronic computer, the random sampling numbers may be replaced conveniently by pseudorandom sampling numbers. To do so has the advantage that every job can be checked precisely, even by other scientists if the "seed" is published. Having generated pseudorandom or random sampling digits, it is trivial to produce random variables that are uniformly distributed in an interval. There are then techniques for producing random variables having other probability distributions. Other techniques deal with the reduction of the variance of estimates. Example 3 above shows the need for such techniques. If x is a random variable with probability density function $p(x)$, the mathematical expectation of a function $f(x)$ may be estimated by replacing f and p by new functions, f^* and p^* , in such a way that $f^*(x)p^*(x) = f(x)p(x)$, where the integral of $p^*(x)$ is unity (so that p^* can be regarded as a probability point function of a new random variable), and where $p^*(x)$ is not small when x is in a region of importance. It is then adequate to estimate the expectation of f^* for the new random variable by averaging values of $f^*(x)$ where x is chosen with probability density $p^*(x)$. This technique, importance sampling, is useful where $f(x)$ is very small. Another technique, called acceptance or rejection sampling, also uses sampling from an adjunct distribution. The necessary adjustment is done by accepting each sampled observation only with an appropriate probability. Some other techniques are familiar in sampling survey work, such as the use of statistical regression and stratified sampling.

Markov-chain Monte Carlo (MCMC). This technique concerns a random vector X which has a known density function $\pi(\cdot)$. The goal is to estimate the expectation of $f(x)$. One way to do this is first to find a Markov chain that has the stationary distribution $\pi(\cdot)$. At first sight it is surprising that this can be done. Then, starting from a somewhat arbitrary initial vector X_0 , successive elements (vectors) of the Markov chain can be generated, say X_1, X_2, X_3, \dots . This sequence will usually tend to the stationary dis-

tribution $\pi(\cdot)$ from which the expectation of $f(X)$ can be estimated, say by averaging every t -th element in the sequence X_1, X_2, X_3, \dots , where t is large enough to "forget the past." A method for constructing an appropriate Markov chain was provided by W. K. Hastings in 1970, a generalization of a method due to Metropolis in 1953. The applications of interest to Metropolis were in statistical mechanics, but since 1990 the generalization has had numerous applications in statistics, when multiple integrals occur. This often happens in Bayesian statistics, including hierarchical Bayes. The Markov-chain Monte Carlo method can, however, take hours to converge, even on a modern computer. See BAYESIAN STATISTICS; STOCHASTIC PROCESS.

Gibbs sampling. This is a special but important case of Markov-chain Monte Carlo. It concerns a known density function f of several variables (x_1, x_2, \dots, x_n) , some of which may be parameters. The goal is to estimate the expectations and variances of the marginal distributions of the n variables. If the conditional distribution of each variable is known, given the others, this goal can be achieved by the following iterative technique known as Gibbs sampling. Start from an arbitrary sequence (x_2, x_3, \dots, x_n) and sample x_1 from the conditional distribution of x_1 , given the other $n - 1$ variables. Then do the same for x_2 given x_1, x_3, \dots, x_n , after updating x_1 . Then proceed to x_3 given $x_1, x_2, x_4, \dots, x_n$, and so forth, going around cyclically for many iterations. Then average all the values obtained for x_1 and for x_2 and so on. This gives estimates for the marginal expectations of each of the n components. This description is deliberately oversimplified for the sake of brevity. Gibbs sampling resembles the EM algorithm, in which expectations and maximum likelihood are alternated iteratively.

Random digits. Monte Carlo methods depend on long sequences of random or pseudorandom digits. It is advisable to apply tests of the sequence for apparent flat-randomness. One useful test is the generalized serial test which tests the equiprobability of all m -plets, for any fixed value of m where m is not too large. See OPERATIONS RESEARCH; PROBABILITY; PROBABILITY (PHYSICS); STATISTICS.

Irving J. Good

Bibliography. G. S. Fishman, *Monte Carlo: Concepts, Algorithms, and Applications*, 2d ed., Springer-Verlag, 1996; J. E. Gentle, *Random Number Generation and Monte Carlo Methods*, Springer-Verlag, 1998; W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, *Markov Chain Monte Carlo in Practice*, CRC Press, 1996; J. M. Hammersley and D. C. Handscomb, *Monte Carlo Methods*, Chapman and Hall, 1965; D. E. Knuth, *The Art of Computer Programming*, vol. 2: *Sem numerical Algorithms*, 3d ed., Addison-Wesley, 1997; D. P. Landau and K. Binder, *A Guide to Monte Carlo Simulations in Statistical Physics*, Cambridge University Press, 2000; M. E. J. Newman and G. T. Barkema, *Monte Carlo Methods in Statistical Physics*, Oxford University Press, 1999; C. P. Robert and G. Casella, *Monte Carlo Statistical Methods*, Springer-Verlag, 1999.

Month

Any of several units of time based on the revolution of the Moon around Earth.

The calendar month is one of the 12 arbitrary periods into which the calendar year is divided. See CALENDAR.

The synodic month, the period of the lunar phases, is the average period of revolution of the Moon with respect to the Sun, the same as the average interval between successive full moons. Its duration is 29.531 days. See PHASE (ASTRONOMY).

The tropical month is the period required for the mean longitude of the Moon to increase 360° , or 27.322 days.

The sidereal month, 7 s longer than the tropical month, is the average period of revolution of the Moon with respect to a fixed direction in space.

The anomalistic month, 27.555 days in duration, is the average interval between closest approaches of the Moon to Earth. The variation in the Moon's distance from the Earth causes a variation in the apparent size of the Moon and thus in the duration of solar eclipses.

The nodical month, 27.212 days in duration, is the average interval between successive northward passages of the Moon across the ecliptic, points known as nodes. Since eclipses can occur only when the Sun and Moon are near such nodes, this period is also known as a draconic month, after the Chinese mythical dragon that supposedly ate the Sun to cause a solar eclipse.

The fortuitous near-equality of 223 synodic months (6585.32 days), 242 nodical months (6585.36 days), 239 anomalistic months (6585.54 days), and 19 eclipse years gives a repetitive period of similar eclipses, called the saros, of 18 years, $11\frac{1}{3}$ days (the date varies ± 1 day, depending on leap years). See ECLIPSE; MOON; TIME; YEAR.

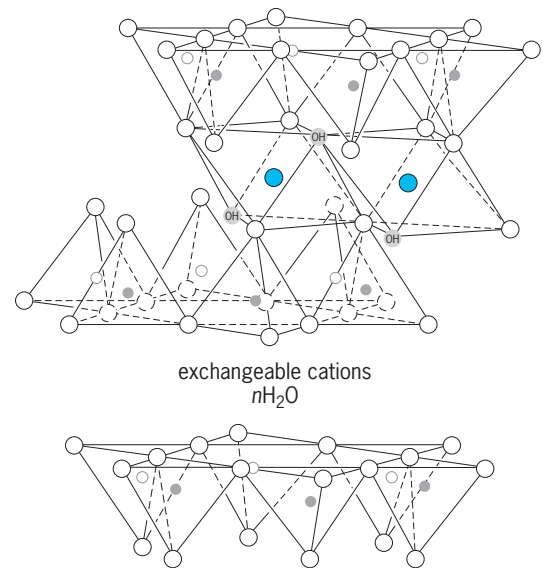
Gerald M. Clemence; Jay M. Pasachoff

Bibliography. J. Mitton, *Cambridge Dictionary of Astronomy*, Cambridge University Press, 2001; J. M. Pasachoff, *The Cosmos: Astronomy in the New Millennium*, 3d ed., Brooks/Cole, Belmont, CA, 2007; P. K. Seidelmann (ed.), *Explanatory Supplement to the Astronomical Almanac*, University Science Books, Mill Valley, CA, 1992.

Montmorillonite

A group name for all clay minerals with an expanding structure, except vermiculite, and also a specific mineral name for the high alumina end member of the group. See VERMICULITE.

Montmorillonite clays have wide commercial use. The high colloidal, plastic, and binding properties make them especially in demand for bonding molding sands and for oil-well drilling muds. They are also widely used to decolorize oils and as a source of petroleum cracking catalysts. See CLAY, COMMERCIAL.



Key:

- oxygen
- ⊕ hydroxyl
- aluminum, iron, magnesium
- silicon, occasionally aluminum

Diagrammatic sketch of structure of montmorillonite. (After R. E. Grim, *Clay Mineralogy*, McGraw-Hill, 1953)

Structure. Because of the extremely small particle size of the montmorillonite minerals, there is still some uncertainty regarding details of their structure. According to the structural concept which is currently accepted, montmorillonite is composed of units made up of two silica tetrahedral sheets with a central alumina octahedral sheet. The atoms in these layers which are common to both sheets become O instead of OH. Montmorillonite is thus referred to as a three-layer clay mineral with tetrahedral-octahedral-tetrahedral layers making up the structural unit (see *illus.*).

These silica-alumina-silica units are continuous in the *a* and *b* crystallographic directions and are stacked one above the other in the *c* direction. In the stacking of these units, the oxygen layers of neighboring units are adjacent. This causes a very weak bond and an excellent cleavage between the units. Water and other polar molecules can enter between the unit layers and cause an expansion of the structure in the *c* direction. Thus montmorillonite does not have a fixed *c*-axis dimension, but can vary considerably depending on the absence or presence of interlayer molecules. The *c*-axis spacing also varies with the nature of the interlayer cation present between the silicate layers. A montmorillonite in an air-dried condition with sodium as the exchange ion frequently has one molecular water layer and a *c*-axis spacing of about 1.25 nm. Under similar conditions there are two molecular water layers with calcium, giving a *c*-axis spacing of about 1.55 nm. The expansion properties of montmorillonite are reversible; however, reexpansion may be difficult after complete structural collapse by removal of all interlayer polar molecules.

Atomic substitution. The theoretical formula for montmorillonite without structural substitutions is $(\text{OH})_4\text{Si}_8\text{Al}_4\text{O}_{20} \cdot n\text{H}_2\text{O}$ (interlayer). However, montmorillonite always differs from the above theoretical formula because of structural substitution. In the tetrahedral sheet, aluminum and possibly phosphorus substitute for silicon, whereas ions such as magnesium, iron, and lithium substitute for aluminum in octahedral coordination. Total replacement of aluminum by magnesium yields the mineral saponite; replacement of aluminum by iron yields nontronite. If all octahedral positions are filled by ions, the mineral is trioctahedral; if only two-thirds are occupied, the mineral is dioctahedral.

The montmorillonite structure is always unbalanced by the substitutions noted above. The resulting positive net charge deficiency is balanced by exchangeable cations adsorbed between the unit layers and around their edges. The cation-exchange capacity of montmorillonite is normally quite high (100± milliequivalents per 100 g) and is not appreciably affected by particle size. Substitutions within the structure cause about 80% of the total exchange capacity, and broken bonds are responsible for the remainder.

Other properties. Montmorillonite particles are extremely small and may further disperse in water to units approaching single-cell-layer dimensions. Most montmorillonite units are equidimensional flakes. However, nontronite tends to occur in elongate lath-shaped units, and hectorite, the fluorine-bearing magnesium-rich montmorillonite, is found in thin laths.

There is general agreement that the adsorbed interlayer water between the silicate layers has some sort of definite configuration, but the precise nature of this configuration is not agreed upon. The extent and nature of the orientation of the adsorbed water varies with identity of the adsorbed cations.

When montmorillonite is dehydrated, the interlayer water is lost at a relatively low temperature (212–390°F or 100–200°C). The loss of structural (OH) water begins gradually at about 840–930°F (450–500°C), ending at 1110–1290°F (600–750°C). These temperatures vary with the type and amount of structural substitution. The structure of montmorillonite usually persists to temperatures of the order of 1470–1650°F (800–900°C). On further heating montmorillonite, a variety of phases form, such as mullite, cristobalite, and cordierite, depending on the composition and structure prior to fusion at 1830–2700°F (100–1500°C).

Organic ionic compounds enter into cation-exchange reactions with montmorillonite. Polar organic compounds, like glycerol, react by replacing the interlayer water, causing a shift in the *c*-axis spacing of the montmorillonite units. Thus, the identification of montmorillonite by x-ray diffraction is greatly simplified by preliminary treatment with certain organic reagents. The reaction of montmorillonite and organic material is the base of considerable economic use of montmorillonite clays.

Occurrence. Members of the montmorillonite group of clay minerals vary greatly in modes of for-

mation. Alkaline conditions and the presence of magnesium particularly favor the formation of these minerals. Montmorillonites are stable over a wide temperature range and have formed by low-temperature hydrothermal processes, as well as by weathering processes. Several important modes of occurrence are in soils, in bentonites, in mineral veins, in marine shales, and as alteration products of other minerals. Recent sediments have a fairly high montmorillonite content. See BENTONITE; CLAY MINERALS; MARINE SEDIMENTS. Floyd M. Wahl; Ralph E. Grim

Moon

The Earth's natural satellite (**Table 1**). United States and Soviet spacecraft have obtained lunar data and samples, and Americans have orbited, landed, and roved upon the Moon (**Fig. 1**). After a long hiatus, lunar exploration by spacecraft has resumed in the first decade of the twenty-first century and there is even the prospect that humans may revisit and eventually reside upon the Moon. Though the first wave of human exploration has passed, it left a store of information whose meanings are still being deciphered. Many of the Moon's properties are now well understood, but its origin and relations to other planets remain obscure. Theories of its origin include independent condensation and then capture by the Earth; formation in the same cloud of preplanetary matter with the Earth; fission from the Earth; and formation after the impact of a Mars-sized body on the proto-Earth. Because many of the Moon's geologic processes stopped long ago, its surface preserves a record of very ancient events. However, because the Moon's rocks and soils were reworked by geochemical and impact processes, their origins are partly obscured, so that working out the Moon's early history remains a fascinating puzzle.

The apparent motions of the Moon, its waxing and waning, and the visible markings on its face (**Fig. 1**), are reflected in stories and legends from every early civilization. At the beginning of recorded history on the Earth, it was already known that time could be reckoned by observing the position and phases of the Moon. Attempts to reconcile the repetitive but incommensurate motions of the Moon and Sun led to the construction of calendars in ancient Chinese and Mesopotamian societies and also, a thousand years later, by the Maya. By about 300 B.C., the Babylonian astronomer-priests had accumulated long spans of observational data and so were able to predict eclipses (**Table 2**).

Motions. The Earth and Moon now make one revolution about their barycenter, or common center of mass (a point about 2900 mi or 4670 km from the Earth's center), in $27^d 7^h 43^m 12^s$ (**Table 3**). This sidereal period is slowly lengthening, and the distance (now about 60.27 earth radii) between centers of mass is increasing, because of tidal friction in the oceans of the Earth. The tidal bulges raised by the Moon are dragged eastward

TABLE 1. Characteristics of the Moon

| Characteristics | Values and remarks |
|------------------------------------|---|
| Diameter (approximate) | 2160 mi (3476 km) |
| Mass | 1/81.301 Earth's mass, or 1.62×10^{23} lb (7348×10^{22} kg) |
| Mean density | 0.604 Earth's, or 209 lb/ft ³ (3.34 g/cm ³) |
| Mean surface gravity | 0.165 Earth's, or 5.3 ft/s ² (162 cm/s ²) |
| Surface escape velocity | 0.213 Earth's, or 1.48 mi/s (2.38 km/s) |
| Atmosphere | Surface pressure 10^{-12} torr (1.3×10^{-10} Pa); hints of some charged dust particles and occasional venting of volatiles; tenuous sodium and potassium clouds observed |
| Magnetic field | Dipole field less than 0.5×10^{-5} Earth's; remanent magnetism in rocks shows past field was much stronger |
| Dielectric properties | Surface material has apparent dielectric constant of 2.8 or less; bulk apparent conductivity is 10^{-5} mho/m or less |
| Natural radioactivity | Mainly due to solar- and cosmic-ray-induced background (about 1 milliroentgen per hour for quiet Sun) |
| Seismic activity | Much lower than Earth's; deep moonquakes occur more frequently when the Moon is near perigee; subsurface layer evident |
| Heat flow | 3×10^{-2} W/m ² (Apollo 15 site) |
| Surface composition and properties | Basic silicates, three sites (Table 4); some magnetic material present; soil grain size is 2–60 μ m and 50% is less than 10 μ m; soil-bearing strength 15 lb/in. ² (1 kg/cm ²) at depth of 1–2 in. (a few centimeters) |
| Rocks | All sizes up to tens of meters present, concentrated in strewn fields; rock samples from Mare Tranquillitatis include fine- and medium-grained igneous and breccia |
| Surface temperature range | At equator 260° F (400 K) at noon; –315 to –280° F (80–100 K) night minimum; 3 ft (1 m) below surface, –45° F (230 K); at poles –280° F (–100 K) |

by the Earth's daily rotation. The displaced water masses exert a gravitational force on the Moon, with a component along its direction of motion, causing the Moon to spiral slowly outward. The Moon, through this same tidal friction, acts to slow the Earth's rotation, lengthening the day. Tidal effects on the Moon itself have caused its rotation to become synchronous with its orbital period, so that it always turns the same face toward the Earth.

Tracing lunar motions backward in time is very difficult, because small errors in the recent data propagate through the lengthy calculations, and because the Earth's own moment of inertia may not have been constant over geologic time. Nevertheless, the attempt is being made by using diverse data sources, such as the old Babylonian eclipse records and the growth rings of fossil shellfish. At its present rate of departure, the Moon would have been quite close to the Earth about 4.6×10^9 years ago, a time which other evidence suggests as the approximate epoch of formation of the Earth.

The Moon's present orbit (Fig. 2) is inclined about 5° to the plane of the ecliptic (Table 3). As a result of differential attraction by the Sun on the Earth-Moon system, the Moon's orbital plane rotates slowly relative to the ecliptic (the line of nodes regresses in an average period of 18.60 years) and the Moon's apogee and perigee rotate slowly in the plane of the orbit (the line of apsides advances in a period of 8.850 years). Looking down on the system from the north, the Moon moves counterclockwise. It travels along its orbit at an average speed of nearly 0.6 mi/s (1 km/s) or about 1 lunar diameter per hour; as seen from Earth, its mean motion eastward among the stars is 13°11' per day.

As a result of the Earth's annual motion around the Sun, the direction of solar illumination changes about 1° per day, so the lunar phases do not repeat in the sidereal period given above but in the

synodic period, which averages 29^d 12^h 44^m and varies some 13 h because of the eccentricity of the Moon's orbit. See EARTH ROTATION AND ORBITAL MOTION; ORBITAL MOTION.

When the lunar line of nodes (Fig. 2) coincides with the direction to the Sun, and the Moon happens to be near a node, eclipses can occur. Because of the 18.6-year regression of the nodes, groups of eclipses recur with this period. When it passes through the Earth's shadow in a lunar eclipse, the Moon remains dimly visible because of the reddish light scattered through the atmosphere around the limbs of the Earth. When the Moon passes between the Earth and Sun, the solar eclipse may be total or annular. As seen from Earth, the angular diameter of the Moon (31') is almost the same as that of the Sun, but both apparent diameters vary because of the eccentricities of the orbits of Moon and Earth. Eclipses are annular when the Moon is near apogee and the Earth is near perihelion at the time of eclipse. A partial solar eclipse is seen from places on Earth that are not directly along the track of the Moon's shadow. See ECLIPSE.

The Moon's polar axis is inclined slightly to the pole of the lunar orbit (Fig. 2) and rotates with the same 18.6-year period about the ecliptic pole. The rotation of the Moon about its polar axis is nearly uniform, but its orbital motion is not, owing to the finite eccentricity and Kepler's law of equal areas, so that the face of the Moon appears to swing east and west about 8° from its central position every month. This is the apparent libration in longitude. The Moon does rock to and fro in a very small oscillation about its mean rotation rate; this is called the physical libration. There is also a libration in latitude because of the inclination of the Moon's polar axis. The librations make it possible to see about 59% of the Moon's surface from the Earth.

The lunar ephemeris, derived from precise astronomical observations and refined through lengthy

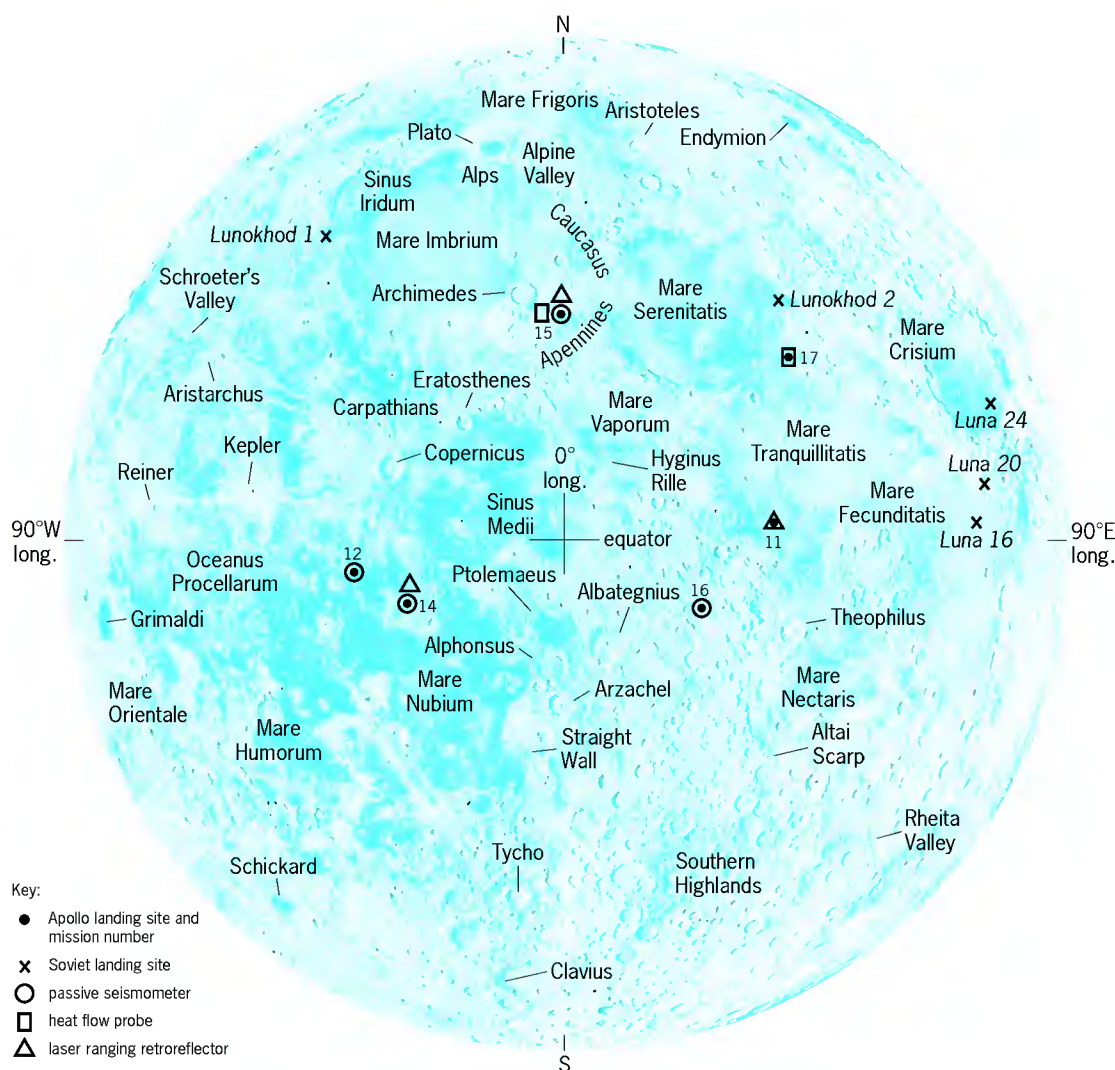


Fig. 1. Map of near side of Moon, showing principal features and some American and Soviet landing sites.

computations of the effects perturbing the movements of the Moon, has now reached a high degree of accuracy in forecasting lunar motions and events such as eclipses. Laser ranging to retroreflectors landed on the Moon, aided by radio ranging to spacecraft, provides measurements of Earth-Moon distances to a precision of the order of meters. See PERTURBATION (ASTRONOMY).

Selenodesy. The problem of determining the Moon's true size and shape and its gravitational and inertial properties has been under attack by various methods for centuries (Tables 1 and 2). However, results from space flights have invalidated some of the premises on which the earlier methods were based, and have revealed discrepancies in the older data. The relation between the Moon's shape and its mass distribution is very important to theories of lunar origin and the history of the Earth-Moon system. Radio-tracking data from lunar orbiters indicate that the Moon's gravitational field is ellipsoidal, with the short axis being the polar one (as expected for any rotating body), and with the equatorial section being an ellipse possibly slightly elongated in

the Earth-Moon direction. But the Earth-based radar measurements and tracking data from Rangers and Surveyors showed that the Moon's actual surface at the points of landing is about 1.2 mi (2 km) farther from the Earth than expected. Further evidence of an anomalous relationship between mass and shape for the Moon is provided by the mass concentrations in circular maria, discovered through analysis of short-term variations in the 1966 Lunar Orbiter tracking data and then mapped in detail by Apollo tracking. Measurements by *Galileo*, *Magellan*, and *Clementine* have provided highly accurate knowledge of the shape of the Moon's gravity potential field. By radio altimetry, Apollo confirmed that the Moon's surface on the far side is higher on the average than the near side; that is, the center of mass is offset from the center of figure. The offset is about 1.2 mi (2 km) toward the Earth. These observations suggest that the Moon's crust is thicker on the far side than on the near side. The *Clementine* mission (Table 2) extended measurements to nearly the whole Moon and revealed the depth of a huge basin on the southern far side (Fig. 3). See SPACE PROBE.

TABLE 2. Growth of human understanding of the Moon

| | | | |
|------------|---|-----------|--|
| Prehistory | Markings and phases observed, legends created connecting Moon with silver, dark markings with rabbit (shape of maria) or with mud | 1965 | Western far side photographed by <i>Zond 3</i> . |
| ~ 300 B.C. | Apparent lunar motions recorded and forecast by Babylonians and Chaldeans. | 1966 | Surface pictures produced by <i>Luna 9</i> and <i>Surveyor 1</i> . Radiation dose at surface measured by <i>Luna 9</i> . Gamma radioactivity measured by <i>Luna 10</i> . High-resolution, broad-area photographs taken by <i>Lunar Orbiter 1</i> . Surface strength and density measurements made by <i>Luna 13</i> . |
| ~ 150 B.C. | Phases and eclipses correctly explained, distance to Moon and Sun measured by Hipparchus. | 1967 | Mare soil properties and chemistry measured by <i>Surveyor 3, 5, and 6</i> . Whole front face mapped by <i>Lunar Orbiter 4</i> ; sites of special scientific interest examined by <i>Lunar Orbiter 5</i> . Particle-and-field environment in lunar orbit measured by <i>Explorer 35</i> . |
| ~ A.D. 150 | Ancient observations compiled and extended by C. Ptolemy. | 1968 | Highland soil and rock properties and chemistry measured by <i>Surveyor 7</i> . Mass concentrations at circular maria discovered. |
| ~ 700 | Ephemeris refined by Arabs. | 1968 | Astronauts orbit Moon, return with photographs. |
| ~ 1600 | Empirical laws of planetary motion derived by J. Kepler. | 1969 | Astronauts land and emplace instruments on Moon, return with lunar samples and photographs. |
| 1620 | Kepler's dream story, <i>Somnium</i> , uses correct description of lunar temperatures. | 1969–1972 | Lunar seismic and laser retroreflector networks established. Heat flow measured at two sites. Remanent magnetism discovered in lunar rocks. Geologic traverses accomplished. Orbital surveys of natural gamma radioactivity, x-ray fluorescence, gravity, magnetic field, surface elevation, and subsurface electromagnetic properties made at low latitudes. Metric mapping photos obtained. Samples returned by both piloted (United States) and automated (Soviet) missions; sample analyses confirmed early heating and chemical differentiation of Moon, with surface rocks enriched in refractory elements and depleted in volatiles. Age dating of lunar rocks and soils showed that most of the Moon's activity (meteoritic, tectonic, volcanic) occurred more than 3×10^9 years ago. |
| 1609 | Lunar craters observed with telescopes by T. Harriot and Galileo. | 1975 | Giant-impact hypothesis for lunar origin advanced by W. K. Hartmann and D. R. Davis. |
| 1650 | Moon mapped by J. Hevelius and G. Riccioli; features named by them in system still in use. | 1982 | Earth-based spectrometry reveals mineral variations over Moon's near side; central peaks of crater Copernicus found to be rich in olivine. |
| 1667 | Experiments by R. Hooke simulating cratering through impact and volcanism. | 1983 | Antarctic meteorite, ALHA 81005, proved to have come from the Moon. |
| 1687 | Moon's motion ascribed to gravity by I. Newton. | 1988–1992 | Thin, extended atmosphere of sodium and potassium discovered. Earth-based multispectral observations yield compositional mapping of near side. <i>Galileo</i> images Moon during Earth flybys. |
| 1692 | Empirical laws of lunar motion stated by J. D. Cassini. | 1994 | <i>Clementine</i> maps entire Moon in 11 visible and near-infrared bands. Topography measured by laser altimetry. |
| 1700–1800 | Lunar librations measured, lunar ephemeris computed using perturbation theory by T. Mayer. Secular changes computed by J. L. Lagrange and P. S. de Laplace. Theory of planetary evolution propounded by I. Kant and Laplace. Many lunar surface features described by J. H. Schroeter and other observers. | 1998 | Orbital geochemical and gravity mapping by Lunar Prospector spacecraft begins. Increased concentration of near-surface light elements observed at lunar poles, possibly indicating presence of hydrogen in cold-trapped water ice. |
| 1800–1920 | Lunar motion theory and observations further refined, leading to understanding of tidal interaction and irregularities in Earth's rotation rate. Photography, photometry, and bolometry applied to description of lunar surface and environment. Lunar atmosphere proved absent. New disciplines of geology and evolution applied to Moon, providing impetus to theories of its origin. | 2004 | Orbital imaging, mineral, chemical, and environmental particles and fields observation by <i>SMART 1</i> spacecraft begins. |
| 1924 | Polarization measured by B. F. Lyot, showing surface to be composed of small particles. | | |
| 1927–1930 | Lunar day, night, and eclipse temperatures measured by E. Pettit and S. B. Nicholson. | | |
| 1946 | First radar return from Moon. | | |
| 1950–1957 | New photographic lunar atlases and geologic reasoning; renewed interest in theories of lunar origin by G. P. Kuiper, H. C. Urey, and E. M. Shoemaker. New methods (for example, isotope dating) applied to meteorites; concepts extended to planetology of Moon. Low subsurface temperatures confirmed by Earth-based microwave radiometry. | | |
| 1959 | Absence of lunar magnetic field (on sunlit side) shown by <i>Luna 2</i> . | | |
| 1960 | Eastern far side photographed by <i>Luna 3</i> . Slower cooling of Tycho detected during lunar eclipse. | | |
| 1961 | United States commitment to human lunar flight. | | |
| 1962 | Earth-Moon mass ratio measured by <i>Mariner 2</i> . | | |
| 1964 | High-resolution pictures sent by <i>Ranger 7</i> . Surface temperatures during eclipse measured by Earth-based infrared scan. | | |

TABLE 3. Dimensions of Moon's orbit*

| Characteristics | Values |
|--|--|
| Sidereal period (true period of rotation and revolution) | (27.32166140 + 0.000000167) ephemeris days, where T is in centuries from 1900 |
| Synodic period (new Moon to new Moon) | (29.5305882 + 0.000000167) ephemeris days |
| Apogee | 252,700 mi or 406,700 km (largest); 251,971 mi or 405,508 km (mean) |
| Perigee | 221,500 mi or 356,400 km (smallest); 225,744 mi or 363,300 km (mean) |
| Period of rotation of perigee | 8.8503 years direct ("direct" meaning that the motion of perigee is in the direction of Moon's motion about the Earth) |
| Period of regression of nodes | 18.5995 years |
| Eccentricity of orbit | 0.054900489 (mean) |
| Inclination of orbit to ecliptic | 5° 8' 43" (oscillating $\pm 9'$ with period of 173 days) |
| Inclination of orbit to Earth's Equator | Maximum 28° 35', minimum 18° 21' |
| Inclination of lunar equator to ecliptic | 1° 32' 40" |
| to orbit | 6° 41' |

* In conventional coordinates with origin at the center of the Earth, rather than the Earth-Moon barycenter.

Body properties. The Moon's small size and low mean density (Table 1) result in surface gravity too low to hold a permanent atmosphere, and therefore it was to be expected that lunar surface characteristics would be very different from those of Earth. However, the bulk properties of the Moon are also quite different—the density alone is evidence of that—and the unraveling of the Moon's internal history and constitution is a great challenge to planetologists. See EARTH.

The Earth, with its dense metallic fluid core, convective mantle, strong and variable magnetic field with trapped radiation belts, widespread seismic tremors, volcanoes and folded mountain ranges, moving lithospheric plates, and highly differentiated

radioactive rocks, is plainly a planet seething with inner activity. Is the Moon also an active, evolving world or is it something very different? The answer lies in a group of related experiments: seismic investigations, heat-flow measurements, surface magnetic and gravity profiles, determination of abundances and ages of the radioactive isotopes in lunar material, and comparison of the latter with those found in the Earth and meteorites. Present theories and experimental data yield the following clues to the problem.

1. The Moon is too small to have compressed its silicates into a metallic phase by gravity; therefore, if it has a dense core at all, the core should be of nickel-iron. But the low mass of the whole Moon does not permit a large core unless the outer layers are of very light material; available data suggest that the Moon's iron core may have a diameter of at most a few hundred kilometers.

2. The Moon has no radiation belts, and behaves as a nonconductor in the presence of the interplanetary field. Moon rocks are magnetized, but the source of the magnetism remains a mystery as there is now little or no general lunar magnetic field.

3. The Moon's natural radioactivity from long-lived isotopes of potassium, thorium, and uranium, expected to provide internal heat sufficient for partial melting, was roughly measured from orbit by *Luna 10*, and the component above the cosmic-ray-induced background radiation was found to be at most that of basic or ultrabasic earthly rock, rather than that of more highly radioactive, differentiated rocks such as granites. *Apollo 11* and *12* rock samples confirmed this result; *Apollo 15*, *16*, and *17* mapped lunar composition and radioactivity from orbit (Fig. 3). X-ray experiments showed higher aluminum-silicon concentration ratios over highland areas and lower values over maria, while magnesium-silicon ratios showed a converse relationship—higher values over maria and lower values over highlands.

4. The Moon is seismically much quieter than the Earth. Moonquakes are small, many of them originate

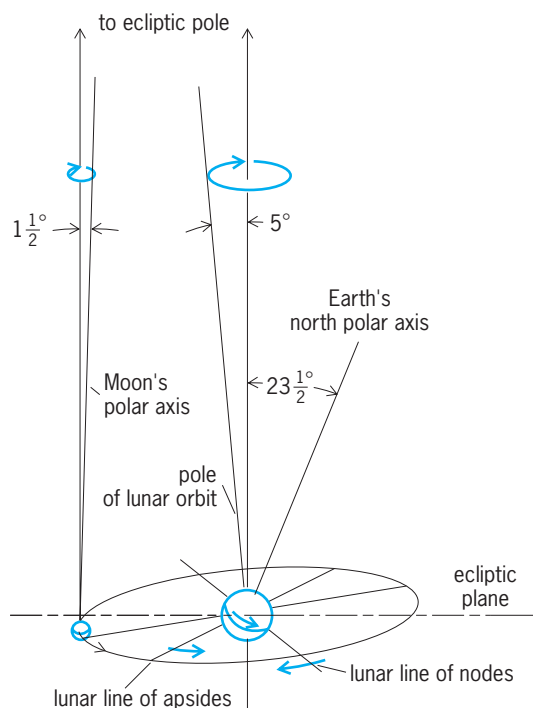


Fig. 2. Sketch of Moon's orbit.

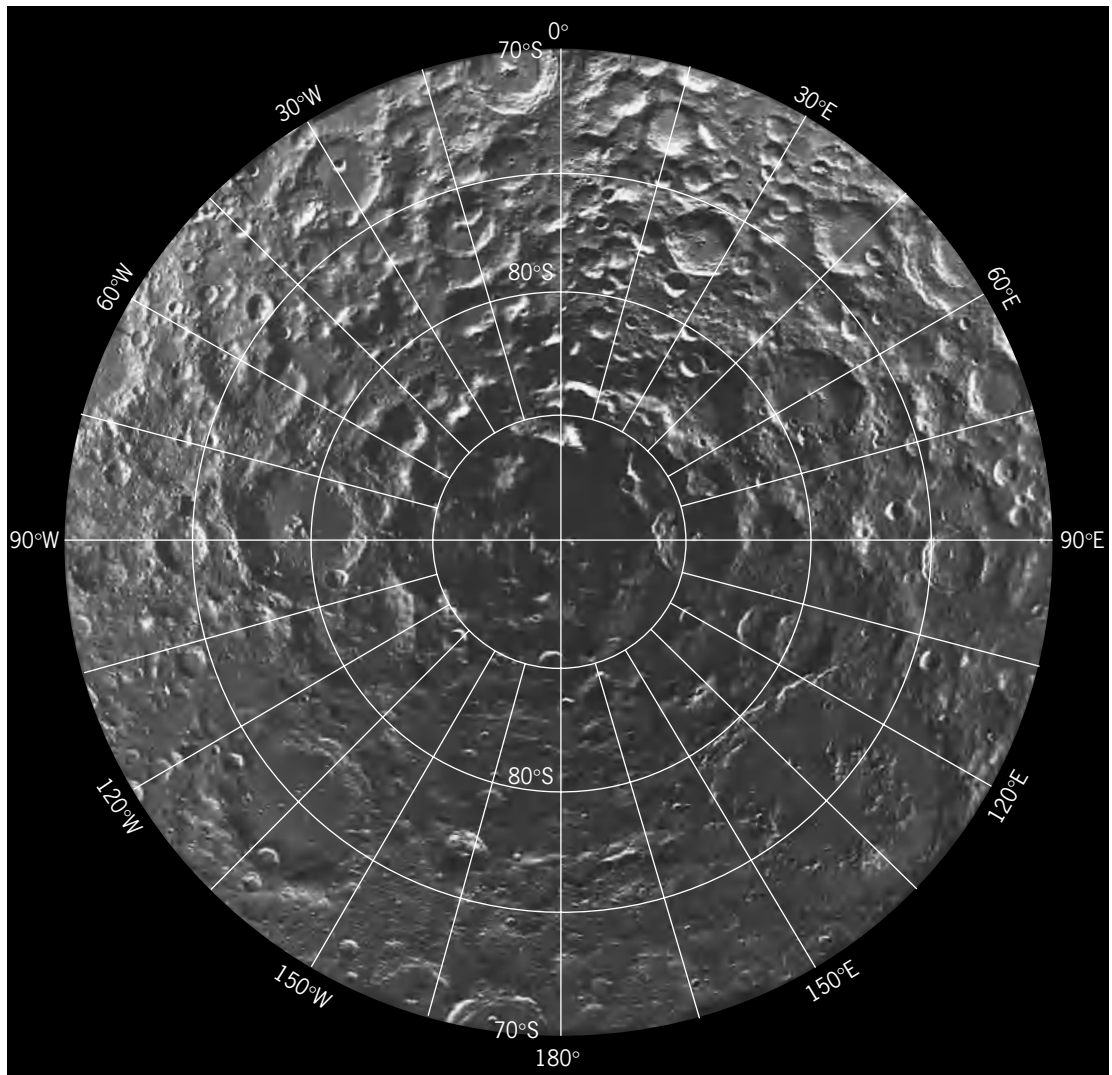


Fig. 3. Mosaic of 1500 *Clementine* images, taken through a red filter, of lunar south polar region. (Naval Research Laboratory)

deep in the interior, and activity is correlated with tidal stress: more quakes occur when the Moon is near perigee.

When all of the Apollo observations are taken together, it is evident that the Moon was melted to an unknown depth and chemically differentiated about 4.5×10^9 years ago, leaving the highlands relatively rich in aluminum and an underlying mantle relatively rich in iron and magnesium, with all known lunar materials depleted in volatiles. The subsequent history of impacts and lava flooding includes further episodes of partial melting until about 3.9×10^9 years ago, with the final result being a thick, rigid crust with only minor evidence of recent basaltic extrusions. The temperature profile and physical properties of the Moon's deep interior are, despite the Apollo seismic and heat-flow data, under active debate.

Large-scale surface features. As can be seen from the Earth with the unaided eye, the Moon has two major types of surface: the dark, smooth maria and the lighter, rougher highlands (Fig. 1 and Fig. 4). Photography by spacecraft shows that, for some

unknown reason, the Moon's far side consists mainly of highlands. Both maria and highlands are covered with craters of all sizes. Craters are more numerous in the highlands than in the maria, except on the steeper slopes, where downhill movement of material apparently tends to obliterate them. Numerous different types of craters can be recognized. Some of them appear very similar to the craters made by explosions on the Earth; they have raised rims, sometimes have central peaks, and are surrounded by fields of hummocky, blocky ejecta. Others are rimless and tend to occur in lines along cracks in the lunar surface. Some of the rimless craters, particularly those with dark halos, may be gas vents bringing dark pyroclastic material to the surface. Most prominent at full moon are the bright ray craters (Fig. 1) whose grayish ejecta appear to have traveled for hundreds of kilometers across the lunar surface. Observers have long recognized that some erosive process has been and is probably still active on the Moon. For example, when craters overlap so that their relative ages are evident, the younger ones are seen to have sharper outlines than the older ones.

Bombardment of the airless Moon by meteoritic matter and solar particles, and extreme temperature cycling, are now considered the most likely erosive agents, but local internal activity is also a possibility. Rocks returned by the Apollo astronauts are covered with tiny glass-lined pits, confirming erosion by small high-speed particles.

The lunar mountains, though very high (26,000 ft or 8000 m or more), are not extremely steep, and lunar explorers see rolling rather than jagged scenery (Fig. 5). There are steep slopes (30–40°) on the inside walls and central peaks of recent craters, where the lunar material appears to be resting at its maximum angle of repose, and rocks can be seen to have rolled down to the crater bottoms.

Though widespread networks of cracks are visible, there is no evidence on the Moon of the great mountain-building processes seen on the Earth. There are some low domes suggestive of volcanic activity, but the higher mountains are all part of the gently rolling highlands or the vast circular structures surrounding major basins. One of these is the Mare Orientale (Fig. 6), a large concentric structure which is almost invisible from Earth because it lies just past the Moon's western limb; at favorable librations, parts of its basin and mountain ramparts can be seen. A great region of radial sculpture surrounds the Orientale basin, strongly suggesting a catastrophic origin, with huge masses of matter thrown outward from the center. However, the lowest parts of the concentric rings are flooded by dark mare material which displays a gentle appearance. Other basins, namely, Imbrium, Serenitatis, and Crisium, appear more fully flooded (Fig. 1). These maria were created by giant impacts, followed by subsidence of the ejecta and (probably much later) upwelling of lava from inside the Moon. Examination of small variations in Lunar Orbiter motions has revealed that each of the great circular maria is the site of a positive gravity anomaly (excess mass). The old argument about impact versus volcanism as the primary agent in forming the lunar relief, reflected in lunar literature over the past 100 years, appears to be entering a new, more complicated phase with the confirmation of extensive flooding of impact craters by lava on the Moon's near side, while on the far side, where the crust is thicker, the great basins remain mostly empty.

In some of the Moon's mountainous regions bordering on the maria are found sinuous rilles (Fig. 7). These winding valleys, some of them known since the eighteenth century, were shown in Lunar Orbiter pictures to have an exquisite fineness of detail. Some of them originate in small circular pits and then wriggle delicately across the Moon's gentle slopes for hundreds of kilometers, detouring around even slight obstacles, before vanishing on the plains. Though their resemblance to meandering rivers is strong, the sinuous rilles have no tributaries or deltas. No explanation for them yet offered has proved entirely convincing.

Other strange large-scale features, observed by telescope and then revealed in more detail by space-

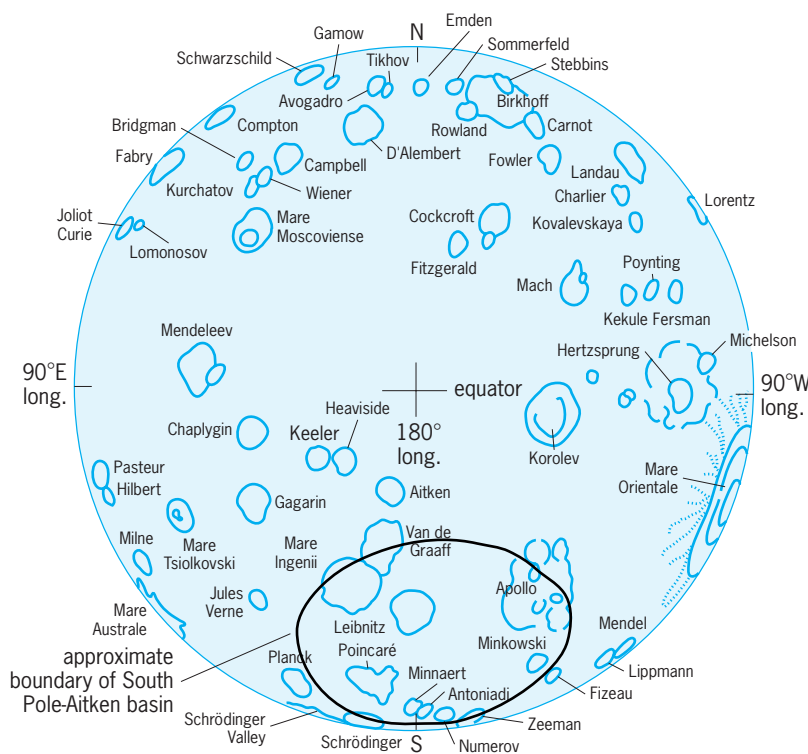


Fig. 4. Map of far side of Moon.

craft cameras, are the ghost craters, circular structures protruding slightly from the maria, and the low, ropy wrinkle ridges that stretch for hundreds of kilometers around some mare borders. Large, light-toned swirly features, with no topography but sometimes with magnetic anomalies, are unexplained (Fig. 8).

Small-scale surface features. Careful observations, some of them made decades before the beginning of space flight, revealed much about the fine-scale nature of the lunar surface. Since the smallest lunar feature telescopically observable from the Earth is some hundreds of meters in extent, methods other than direct visual observation had to be used. Photometry, polarimetry, and later radiometry and radar



Fig. 5. The crater Copernicus, showing the central peaks, slump terraces, patterned crater walls, and (background) slopes of the Carpathian Mountains. (Langley Research Center, NASA)



Fig. 6. Mare Orientale. (Langley Research Center, NASA)

probing gave the early fine-scale data. Some results of these investigations suggested bizarre characteristics for the Moon. Nevertheless, many of their findings have now been confirmed by spacecraft. The Moon seems to be totally covered, to a depth of at least tens of meters, by a layer of rubble and soil with very peculiar optical and thermal properties. This layer is called the regolith. The observed optical and

radio properties are as follows. See PHOTOMETRY; POLARIMETRY; RADAR; RADIOMETRY.

1. The Moon reflects only a small portion of the light incident on it (the average albedo of the maria is only 7%). See ALBEDO.

2. The full moon is more than 10 times as bright as the half moon.

3. At full moon, the disk is almost equally bright all the way to the edge; that is, there is no "limb darkening" such as is observed for ordinary spheres, whether they be specular or diffuse reflectors.

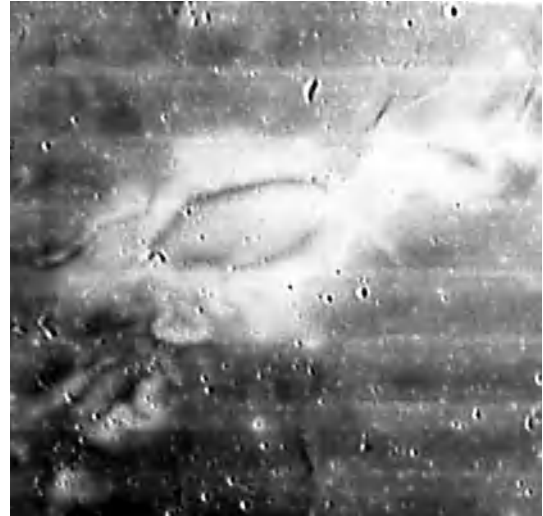


Fig. 8. *Lunar Orbiter* image of Reiner Gamma, the only example of a swirl feature on the Moon's near side. (Only three other such features are known to exist, two on the Moon's far side and one on Mercury.) The feature has an overall dimension of about 43 mi (70 km). It exhibits an almost total lack of elevation and other forms of relief. A relatively strong field of magnetization has been discovered to coincide with the feature. (NASA)



Fig. 7. Aristarchus-Harbinger region of the Moon, photographed from the *Apollo 15* spacecraft in lunar orbit, with the craters Aristarchus and Herodotus and Schroeter's Valley, the largest sinuous rille on the Moon. The impact crater Aristarchus, about 25 mi (40 km) in diameter and more than 2.5 mi (4 km) deep, lies at the edge of a mountainous region that shows evidence of volcanic activity. (NASA)

4. Color variations are slight; the Moon is a uniform dark gray with a small yellowish cast. Some of the maria are a little redder, some a little bluer, and these differences do correlate with large-scale surface morphology, but the visible color differences are so slight that they are detectable only with special filters. Infrared spectral differences are more pronounced and have provided a method for mapping variations in the Moon's surface composition. Multi-spectral remote sensing, both from Earth and from the *Clementine* and *SMART 1* spacecraft in polar lunar orbits has greatly advanced the study of compositional variations over the lunar surface.

5. The Moon's polarization properties are those of a surface covered completely by small, opaque grains in the size range of a few micrometers.

6. The material at the lunar surface is an extremely good thermal insulator, better than the most porous terrestrial rocks. The cooling rate as measured by infrared observations during a lunar eclipse is strongly variable; the bright ray craters cool more slowly than their surroundings.

7. The Moon emits thermal radiation in the radio wavelength range; interpretations of this and the

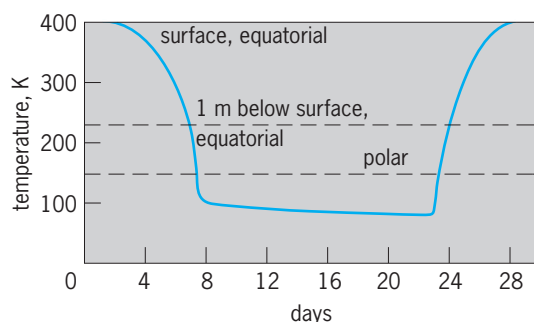


Fig. 9. Lunar surface and subsurface temperatures.

infrared data yield estimates of surface and shallow subsurface temperatures (Fig. 9).

8. At wavelengths in the meter range, the Moon appears smooth to radar, with a dielectric constant lower than that of most dry terrestrial rocks. To centimeter waves, the Moon appears rather rough, and at visible light wavelengths it is extremely rough (a conclusion from observations 1–5 above).

These observations all point to a highly porous or underdense structure for at least the top few millimeters of the lunar surface material. The so-called backscatter peak in the photometric function, which describes the sudden brightening near full Moon, is characteristic of surfaces with deep holes or with other roughness elements that are shadowed when the lighting is oblique. However, *Clementine* observations showed that, in addition to shadowing, light reflections within soil particles may contribute to the brightening.

The Ranger, Luna, Surveyor, and Lunar Orbiter missions made it clear that these strange electromagnetic properties are generic characteristics of the dark-gray, fine soil that appears to mantle the entire Moon, softening most surface contours and covering everything except occasional fields of rocks (Figs. 10 and 11). This soil, with a slightly cohesive character like that of damp sand and an average chemical composition similar to that of some basic silicates on the Earth, is a product of the radiation, meteoroid, and thermal environment at the lunar surface. A surface texture called patterned ground (Fig. 10) is common on the moderate slopes of the Moon. This widespread phenomenon is unexplained, though there are some similar surfaces developed on the Earth when unconsolidated rock, lava, or glacial ice moves downhill beneath an overburden. At many places on the Moon, there is unmistakable evidence of downward sliding or slumping of material and rolling rocks. There are also a few instances of apparent upwelling, as well as numerous “lakes” where material has collected in depressions.

Magnets on the Surveyors collected magnetic particles from the soil, demonstrating the presence of either meteoritic or native iron minerals at the sites examined. Meteoroid experiments on the Lunar Orbiters showed about the same flux of small particles as is observed at the Earth, so that the lunar soil would be expected to contain a representative sample of meteoritic and possibly also cometary mat-

ter. Apollo results confirmed and extended the Surveyor data and also indicated that glassy particles are abundant in and on the soil. Evidence of micrometeoroid bombardment is seen in the many glass-lined microcraters found on lunar rocks. See COMET; METEOR.

Chemical, mineral, and isotopic analyses of minerals from the *Apollo 11* site showed that mare rocks there are indeed of the basic igneous class and are very ancient ($3\text{--}4 \times 10^9$ years). The *Apollo 12* samples are significantly younger, suggesting that Mare Tranquillitatis and Oceanus Procellarum were formed during a long and complex lunar history. The *Apollo 12* astronauts visited *Surveyor 3* and brought back parts of that spacecraft to permit analysis of the effects of its $2\frac{1}{2}$ -year exposure on the surface of the Moon. The lunar rock and soil samples returned by the Apollo and Luna missions have yielded much new information on the composition and



Fig. 10. Lunar soil and rocks, and the trenches, 2 in. (5 cm) wide, made by *Surveyor 7*. (Jet Propulsion Laboratory, NASA)

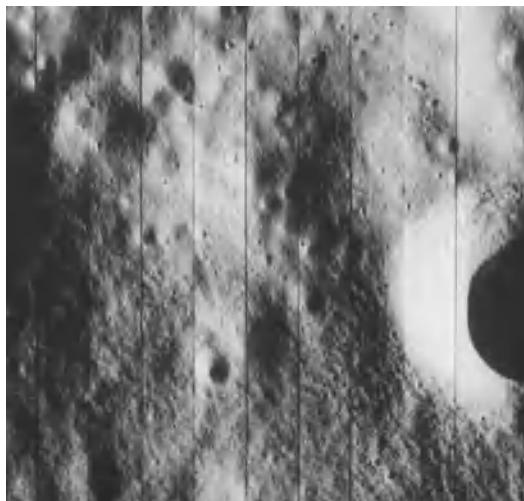


Fig. 11. Lunar patterned ground, a common feature on moderate slopes. (Langley Research Center, NASA)

TABLE 4. Some selected data from Apollo and Luna missions

| Mission | Main sample properties | Other data |
|------------------|--|---|
| <i>Apollo 11</i> | Mare basalts, differentiated from melt at depth 3.7×10^9 years ago. Some crystalline highland fragments in soils. Unexpected abundance of glass. Much evidence of impact shock and microcratering. No water or organic materials. | Study of seismic properties showed low background, much scattering, and low attenuation. |
| <i>Apollo 12</i> | Basalts 3.2×10^9 years old. One sample 4.0×10^9 years old includes granitic component. Some samples with high potassium, rare-earth elements, and phosphorus (KREEP) may be Copernicus crater ejecta. | Surveyor parts returned showed effects of solar and cosmic bombardment. |
| <i>Apollo 13</i> | Spacecraft failure—no samples. | Despite emergency, some lunar photos returned. |
| <i>Luna 16</i> | Basalt 3.4×10^9 years old, relatively high Al content. | Deep moonquakes. |
| <i>Apollo 14</i> | Shocked highland basalts, probably Imbrium ejecta, 3.95×10^9 years old, higher Al and lower Fe than mare materials. | |
| <i>Apollo 15</i> | Highland anorthosites including one sample 4.1×10^9 years old, mare basalts similar to <i>Apollo 11</i> samples. | Orbital remote sensing began mapping of surface compositions. |
| <i>Luna 20</i> | Possibly Crisium ejecta, 3.9×10^9 years old. | Seismic network began recording locations of impacts and deep moonquakes; orbital compositional mapping extended. |
| <i>Apollo 16</i> | Highland anorthosite breccias $3.9\text{--}4 \times 10^9$ years old, also possibly Imbrium ejecta. | |
| <i>Apollo 17</i> | Variety of basalts and anorthosites $3.7\text{--}4 \times 10^9$ years old, possibly volcanic glass, few dunite fragments 4.48×10^9 years old, possibly surviving from before the great highland bombardment. | Orbital mapping, and study of seismic particle-and-field, and subsurface electrical properties yielded comprehensive (but still unexplained) picture of Moon. |
| <i>Luna 24</i> | Very low titanium basalts from Mare Crisium. Sample includes rock 3.3×10^9 years old. | |

history of the Moon (Table 4). Among the dominant characteristics of these rocks are enrichment in refractories, depletion in volatiles, much evidence of repeated breaking up and rewelding into breccias, and ages since solidification extending back from the mare flows of $3\text{--}4 \times 10^9$ years ago into the period of highland formation more than 4×10^9 years ago, but not as yet including the time of the Moon's original accretion. Soil characteristics, through the analysis of embedded solar-wind atoms, have also yielded information on the ancient history of the Sun.

In 1983 an Antarctic meteorite was found to resemble some of the lunar samples, and it was determined by analysis to have come from the Moon. Many more lunar and Martian meteorites are now known and have been added to the *Apollo* and *Luna* sample collections. See METEORITE.

As the Apollo missions progressed, each new landing site was selected with the aim of elucidating more of the Moon's history. A main objective was to sample each of the geologic units mapped by remote observation, either by landing on it or by collecting materials naturally transported from it to the landing site. Although this process did result in collection of both mare and highland materials with a wide range of ages and chemical compositions, it did not result in a complete unraveling of the history of the Moon. Apparently, the great impacts of $3\text{--}4 \times 10^9$ years ago erased much of the previous record, resetting radioactive clocks and scrambling minerals of diverse origins into the complicated soils and breccias found today. See SPACE FLIGHT.

Atmosphere. Though the Moon may at one time have contained appreciable quantities of the volatile elements and compounds (for example, hydrogen, helium, argon, water, sulfur, and carbon compounds) found in meteorites and on the Earth, its high daytime

surface temperature and low gravity would cause rapid escape of the lighter elements. Solar ultraviolet and x-ray irradiation would tend to break down volatile compounds at the surface, and solar charged-particle bombardment would ionize and sweep away even the heavier gas species. Visual observations from the Earth, looking for a twilight glow of the lunar atmosphere just past the terminator on the Moon, and watching radio-star occultations have all been negative, setting an upper limit of 10^{-12} times the Earth's sea-level atmospheric density for any lunar gas envelope. However, subsequent ground-based spectroscopic observations have revealed a tenuous cloud of sodium and potassium atoms extending hundreds of kilometers above the lunar surface. Except for this cloud, either the lunar volatile compounds have vanished into space or they are trapped beneath the surface. The samples returned by Apollo are enriched in refractory elements, depleted in volatiles, and impregnated with hydrogen, helium (including helium-3, a possible nuclear fusion fuel), and other gases from the solar wind. No water appears to have ever been present at any Apollo site, and carbonaceous materials were present, if at all, only in very small amounts.

Occasional luminescent events reported by reliable observers suggest that some volcanic gases are vented from time to time on the Moon, particularly in the regions of the craters Aristarchus and Alphonsus. The *Apollo 17* samples included glassy particles indicative of pyroclastic eruptions. A slight, transient atmosphere does exist on the night side of the Moon as a result of the trapping and release of gas molecules at the very low temperatures prevailing there; also frozen liquids or gases could exist in permanently shadowed crater bottoms near the lunar poles. Data from *Clementine* suggest that favorable trapping conditions may exist near the south pole

(Fig. 3), and neutron spectrometry by the *Lunar Prospector* orbiter confirms concentrations of a light element, most likely hydrogen, within a meter of the surface at both poles. While opinions differ as to the source of this hydrogen, it strongly suggests the presence of cold-trapped water ice. *Lunokhod 2*, a Soviet roving spacecraft, measured a slight glow attributed to a very thin cloud of small dust particles within meters of the surface, which could explain the Surveyor observations of a slight horizon glow after sunset. Also, the ALSEP (Apollo lunar surface experiments packages) experiments landed by Apollo have occasionally detected small gas emanations, including water, from unknown sources; contamination from the landed spacecraft is hard to exclude from these measurements.

Lunar resources. Enough is known about the Moon to show that it is a huge storehouse of metals, oxygen (bound into silicates), and other materials potentially available for future human use in space. Because of the Moon's weak gravity, lunar materials could be placed into orbit at less than one-twentieth of the energy cost for delivering them from Earth. An objective of several lunar missions now planned in the United States, Europe, India, China, and Japan is to augment lunar scientific knowledge and thus add to understanding the possible uses of the natural resources of the Moon.

James D. Burke

Bibliography. D. Baker, *The Moon*, vol. 1: *Physics, Geology and Evolution*, 1992; P. Eckart, *The Lunar Base Handbook*, McGraw-Hill, 2000; W. K. Hartmann, R. J. Phillips, and G. J. Taylor (eds.), *Origin of the Moon*, 1986; G. Heiken, *Lunar Sourcebook: A User's Guide to the Moon*, Cambridge University Press, 1991; M. W. Huddleston, *Lunar Swirls, Magnetic Anomalies, and Reiner Gamma Formation*, Association of Lunar & Planetary Observers, 1994; Lunar and Planetary Institute, *Proceedings of the Lunar and Planetary Science Conferences*, 1970–2006; W. W. Mendell (ed.), *Lunar Bases and Space Activities of the 21st Century*, Lunar and Planetary Institute, 1985; A. Rukl, *Atlas of the Moon*, 1992; Special section on the *Clementine* mission, *Science*, 266:1835–1862, December 16, 1994; H. H. Schmitt, *Return to the Moon: Exploration, Enterprise, and Energy in the Human Settlement of Space*, Copernicus Books, 2006; P. D. Spudis, *The Once and Future Moon*, Smithsonian Press, 1997; S. R. Taylor, *Planetary Science: A Lunar Perspective*, 1982; D. Wilhelms et al., *Geologic History of the Moon*, U.S. Geol. Sur. Prof. Pap. 1348, 1988.

Moose

An even-toed ungulate (Artiodactyla) which is a member of the deer family, Cervidae. *Alces alces* is the largest member of the family and ranges in the boreal forested areas throughout North America and in northern Eurasia. The moose is known as the elk in Europe and is believed by some authorities to be a race of the American moose (*A. americana*). The adult male is 6 ft (1.8 m) high, weighs over 1200 lb



The moose, a member of the deer family, Cervidae.

(550 kg), and has spatulate antlers which may be over 6 ft (1.8 m) in width (see *illus.*). The legs are long, making the animal well-adapted for its feeding habits of wading for aquatic plants and browsing on trees and bushes. The moose live in small groups during the summer but tend to form larger groups for defense during the winter, since they are susceptible to predation from wolves and even wolverines. During the rutting season in the early fall, the male gathers a number of cows together, and mating takes place. After a gestation period of about 37 weeks, one or two calves are born. The moose is a big-game animal, but hunting restrictions have helped to maintain its numbers; it is abundant in Canada and the northern United States. See ARTIODACTYLA.

Charles B. Curtin

Moraxella

A genus of bacteria that are parasites of mucous membranes. Subgenus *Moraxella* is characterized by gram-negative rods that are often very short and plump, frequently resembling a coccus, and usually occurring in pairs. Subgenus *Branhamella* has gram-negative cocci occurring as single cells or in pairs with the adjacent sides flattened. They are usually harmless parasites of humans and other warm-blooded animals and are generally considered not to be highly pathogenic. Most species may be opportunistic pathogens in predisposed or debilitated hosts.

General characteristics. Members of the genus *Moraxella* are generally characterized as being non-spore-forming, nonmotile cells, often demonstrating a tendency to resist Gram decolorization. They are chemoorganotrophic and nutritionally fastidious, strictly aerobic, indophenol oxidase-positive, and usually catalase-positive. They do not produce acid from carbohydrates. Colonies are not pigmented. All species will grow on blood agar media commonly used in clinical laboratories, but some strains of *Moraxella* (*Moraxella*) *lacunata* may require media containing heated blood, such as chocolate agar. Most strains, with a few exceptions of *M. (M.) lacunata*, will grow on blood agar base media without

blood, but the addition of serum may improve growth of such strains. *Moraxella (M.) lacunata* and most *M. (M.) bovis* strains are the only *Moraxella* species that can liquefy the heat-coagulated serum in Loeffler coagulated serum agar slants.

There are presently six species in the subgenus *Moraxella*: *M. (M.) lacunata* (also known as *Diplobacillus moraxaxenfeld* and *liquefaciens*), *M. (M.) bovis*, *M. (M.) nonliquefaciens* (also known as *Bacillus duplex nonliquefaciens*), *M. (M.) atlantae*, *M. (M.) phenylpyruvica*, and *M. (M.) osloensis*. The different species are recognized on the basis of phenotypic properties, including liquefaction of coagulated serum, hemolysis of human blood in blood agar media, nitrate reduction, phenylalanine deaminase activity, urease activity, and growth on mineral salts medium with ammonium ion and acetate as the sole carbon source.

The subgenus *Branbamella* presently contains four species: *M. (B.) catarrhalis*, *M. (B.) caviae*, *M. (B.) ovis*, and *M. (B.) cuniculi*. The different species are recognized on the basis of hemolysis of human blood in blood agar media, and nitrate and nitrite reduction, among other properties.

Phenotypic relations. Members of *Moraxella* must be distinguished from species of *Neisseria* since both genera have similar phenotypic properties. The latter bacteria are characterized by cocci occurring singly but more often in pairs with adjacent sides flattened. They reduce nitrite, but usually not nitrate, and demonstrate carbonic anhydrase. Many species produce acid from carbohydrates, and some species produce a greenish-yellow carotenoid pigment.

The subgenus *Branbamella* is differentiated from *Neisseria* species by its frequent reduction of nitrate and negative reactions for carbohydrate acidification, pigment production, and carbonic anhydrase synthesis. The subgenus *Moraxella* is distinguished from *Neisseria* species by its rod-shaped appearance. Additionally, species of the subgenus *Moraxella* produce negative reactions for nitrite reduction, carbohydrate acidification, pigment production, and carbonic anhydrase synthesis.

Distinguishing short rods (subgenus *Moraxella*) from cocci (*Branbamella* and *Neisseria*) by microscopic observation sometimes is difficult. A reliable test for differentiation is to culture the organism in the presence of subinhibitory levels of penicillin: rod-shaped organisms form long, stringlike cells, but cocci retain their coccal morphology.

Pathogens. *Moraxella (M.) lacunata*, the type species of the subgenus *Moraxella*, was described independently by V. Morax in 1896 and T. Axenfeld in 1897, thus the name *Morax-Axenfeld bacillus*. It was a significant causative agent of human conjunctivitis and keratitis in the past but is only rarely isolated at present. Infectious keratoconjunctivitis in cattle, called pinkeye, is caused by *M. (M.) bovis*. The natural habitats of both organisms are healthy conjunctiva and sites in the upper respiratory tract of humans and cattle, respectively. *Moraxella (M.) nonliquefaciens* is most frequently isolated from the nasal cavity of humans, which is probably the natural habitat,

as well as other sites of the respiratory tract. It is considered to be a well-established parasite of humans and rarely causes disease, but it has been associated with endophthalmitis and pneumonitis with pulmonary abscess. *Moraxella (M.) atlantae*, only rarely isolated, has been found in human blood, cerebrospinal fluid, and abscess. Its natural habitat and clinical significance are unknown. *Moraxella (M.) phenylpyruvica* is isolated from human sources, including urine, blood, cerebrospinal fluid and the genitourinary tract, and from sheep, cattle, goats, and pigs. Its clinical significance is uncertain. *Moraxella (M.) osloensis* is isolated from the upper respiratory tract, genitourinary tract, blood, and cerebrospinal fluid of humans. Usually a harmless parasite, this organism has been frequently associated with such human infections as osteomyelitis, endocarditis, septicemia, meningitis, stomatitis, and septic arthritis.

Moraxella (B.) catarrhalis, the type species of *Branbamella*, is the only species of this subgenus recovered from humans. Its natural habitat is the nasal cavity. The organism is considered to be a well-adapted parasite but has been judged the etiologic agent of middle-ear infection, maxillary sinus infection, bronchitis, tracheitis, conjunctivitis, pneumonia, otitis media of infants, respiratory disease in the compromised host, septicemia, meningitis, and endocarditis. The remaining *Branbamella* species (*caviae*, *ovis*, and *cuniculi*) are parasites of guinea pigs, sheep, cattle, and rabbits.

Moraxella species are susceptible to most antimicrobial agents with the exception of the lincosamides. Their usually high susceptibility to the penicillins is a feature that separates them from most other gram-negative rods. See CLINICAL MICROBIOLOGY. Gerald Gilardi

Bibliography. W. A. Clark et al., *Identification of Unusual Pathogenic Gram-Negative Aerobic and Facultatively Anaerobic Bacteria*, 1985; S. D. Henriksen, *Moraxella, Acinetobacter*, and the Mimeae, *Bacteriol. Rev.*, 37:522-561, 1973; N. R. Kreig (ed.), *Bergey's Manual of Systematic Bacteriology*, vol. 1, 1984.

Mordant

A chemical substance, usually containing a metallic ion, used to facilitate the fixing of a dye to a fiber. From the earliest days of dyeing, it was observed that for many natural dyes the color imparted to a fabric was enhanced and the fastness (resistance to fading or running) improved if the fabric was first treated with a solution containing a soluble metallic salt of chromium, aluminum, or copper. In the simplest process model, metal ions are incorporated on the surface and within the fibers of the fabric and, upon exposure to dye molecules at the appropriate pH and temperature, form colored metal-dye complexes. The metal ion is called a mordant. The complex formation process is called chelation, and can result in a fast and deeply colored fabric. One

of the oldest mordant dyes is alizarin, derived from the root of the madder plant. Chemically, it is an anthraquinone derivative, and the dye color depends upon the metal with which it is complexed; with barium, a blue color is obtained, and with aluminum a rose-red. See CHELATION; CHROMIUM; PH; QUINONE.

In the dyeing process, an organic compound containing a chromogenic structural group is used to impart color to a substrate—fabric, paper, or other object. In order that the compound be useful as a dye, it must be bound to the fabric or surface in such manner that it will remain fixed during washing and cleaning; that is, it must be fast. The actual mechanism of the bonding process may be quite complex, and depends upon the structure of the dye and upon the chemical nature of the surface.

Mordant action is a complex chemical process, and a general mechanism is not applicable. As an example, chrome dyes form colored chelates with Cr(III) ions [Cr(III) represents chromium atoms in an oxidation state of +3]. Chromium is applied to a wool fabric in the form of dichromate ions in which the oxidation state is +6, Cr(VI). Adsorption occurs by the attraction of the negatively charged dichromate ions to the positively charged base groups in the wool. Before chelation with a dye molecule, the Cr(VI) atoms must be reduced to the Cr(III) state. This occurs by the oxidation of the sulfide groups in the fiber by dichromate and the simultaneous reduction of Cr(VI). The resulting metal-dye chelate and color depend upon pH, temperature, and concentration.

When groups on dye molecules can react at sites on the substrate's surface to form covalent bonds, direct dyeing is accomplished. Wool and silk fibers contain molecules comprising chain structures linked together by peptide units. These units contain —C=O (carbonyl) groups, —S—S— (disulfide) linkages, and basic —NH_2 (amino) groups. A dye with a reactive acidic group may form a chemical bond through a direct reaction at the fiber surface. Martius Yellow, for example, is a dye that contains acidic OH (hydroxyl) groups. It is bonded to wool or silk through reaction with the amino groups. See PEPTIDE; SILK; WOOL.

In a vat dyeing process, the fabric is immersed in a solution containing dye in a soluble form. Through an oxidation or other chemical transformation, the dye molecules are converted to an insoluble form and dispersed as a precipitate throughout the fabric. A true chemical bonding is not believed to be involved.

The science and practice of mordant dyeing draws heavily upon metal-chelate chemistry. Many synthetic dyes today incorporate metal atoms in the dye structure and are used without mordants. See DYE; DYEING.

Francis Johnston

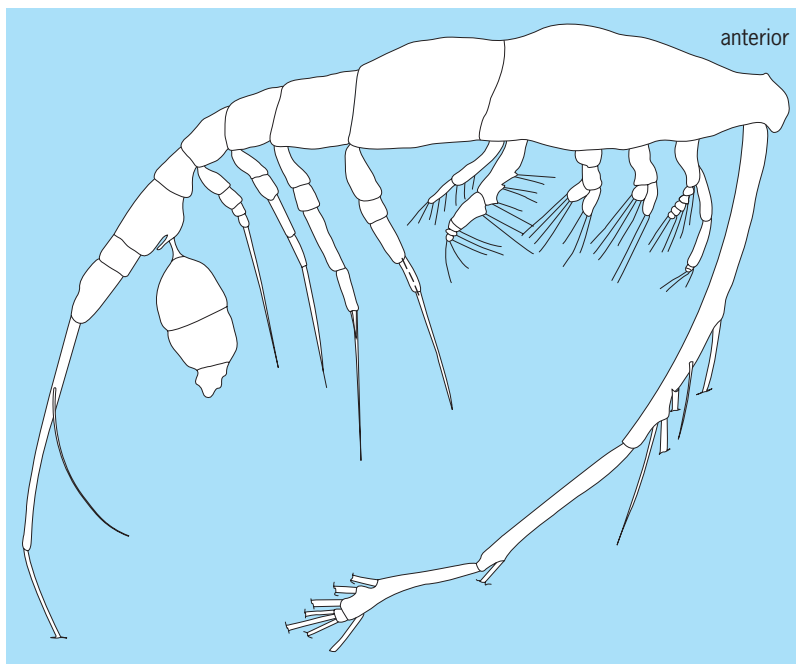
Bibliography. A. Johnson (ed.), *The Theory of Coloration of Textiles*, Society of Dyers and Colourists, West Yorkshire, U.K., 1989; D. Waring and G. Hallas (eds.), *The Chemistry and Application of Dyes*, Plenum Press, New York, 1990.

Mormonilloida

A copepod order of crustaceans containing four species. *Mormonilla minor* and *M. phasma* were described by Wilhelm Giesbrecht in the late 1800s; *M. polaris* and *M. atlantica* were described by Georg Ossian Sars and Richard Norris Wolfenden, respectively, in the early 1900s. The status of a fifth species in a second genus, Thomas Scott's *Corynuropis tenuicaudatus*, remains uncertain. In 1992, males of the order were collected and described for the first time. See COPEPODA; CRUSTACEA.

Description. Lengths range from 0.2–1.7 mm (0.01–0.07 in.) for adult females to 0.9–1.4 mm (0.04–0.06 in.) for adult males. The second thoracic somite, bearing the first swimming leg, articulates with the first, and the last two thoracic somites are associated with the posterior part of the body (podoplean architecture) [see **illustration**]. The only internal organ system that has been studied is the musculature; it is simpler than the musculature of copepods from other orders. One or two small embryos are carried in paired embryo sacs attached to the female; nothing else is known of the development. Species of Mormonilloida are deep-water, free-swimming copepods reported from the Arctic, eastern Indian, and eastern North Atlantic oceans, and the Mediterranean, Red, and Arabian seas. Specimens have been collected below 250 m (820 ft) in the Arctic Ocean, between 410 and 700 m (1350 and 2300 ft) in the eastern North Atlantic Ocean, and between 125 and 1850 m (410 and 6080 ft) in the Mediterranean, Red, and Arabian seas and in the eastern Indian Ocean.

Phylogeny. Phylogenetic relationships of Mormonilloida to the other copepod orders remain uncertain because body architecture, as described



Side view of a mormonilloidan.

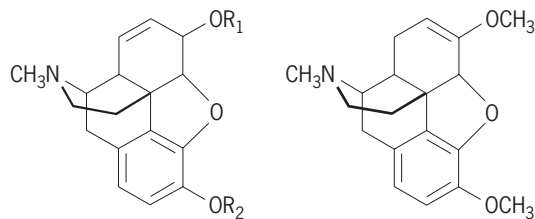
by the association of somites along the anterior-posterior axis, places *Mormonilloida* close to the podoplean copepods, such as cyclopoids and harpacticoids, while the multisegmental configuration of several limbs suggests a close relationship with gymnoplean calanoids (the last thoracic somite with the posterior part of the body). The phylogeny appears unclear in the analyses because neither body architecture nor limb structure is assumed to converge. See CALANOIDA; CYCLOPOIDA; HARPACTICOIDA.

Frank D. Ferrari

Bibliography. G. A. Boxshall, The comparative anatomy of two copepods, a predatory calanoid and a particle-feeding mormonilloid, *Phil. Trans. Roy. Soc. London, Ser. B (Biol. Sci.)*, vol. 311, 1985; G. A. Boxshall, The planktonic copepods of the northeastern Atlantic Ocean: Harpacticoida, Siphonostomatoida and Mormonilloida, *Bull. Brit. Mus. Nat. Hist. (Zool.)*, vol. 35, 1979; R. A. Buchanan and A. D. Sekerak, Vertical distribution of zooplankton in eastern Lancaster Sound and western Baffin Bay, July-October 1978, *Arctic*, vol. 35, 1982.

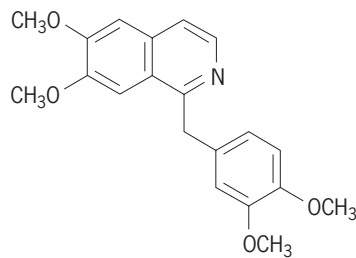
Morphine alkaloids

A group of closely related compounds obtained from the dried latex that exudes from the seed capsule of the opium poppy, *Papaver somniferum*. The major alkaloid is morphine [structure (1)], which constitutes about 12% of the weight of the crude opium resin. Some 20 minor alkaloids have been isolated from opium, among them codeine (1), thebaine (2), and papaverine (3). See ALKALOID; POPPY.



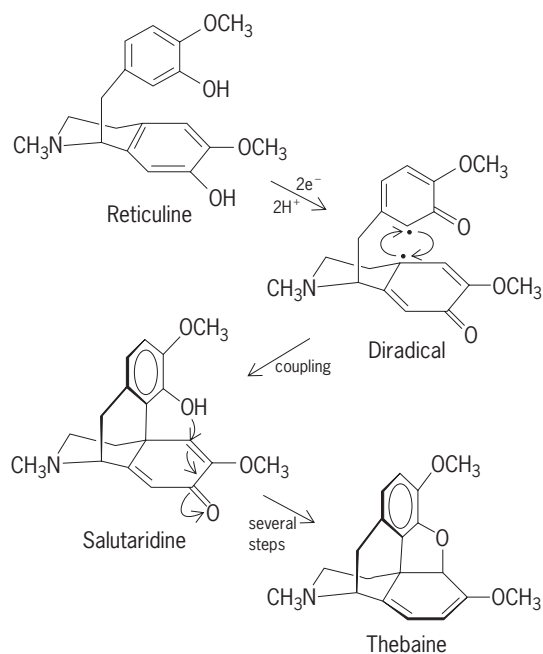
Morphine: $R_1=R_2=H$
Codeine: $R_1=H, R_2=CH_3$
Heroin: $R_1=R_2=COCH_3$

(1)



(3)

The chemistry of the morphine alkaloids is extremely complex, and the structure of these compounds was not finally established until the early 1950s, when the synthesis of morphine was accomplished. The unique benzomorphan ring system occurs in the plant as a result of oxidative phenol

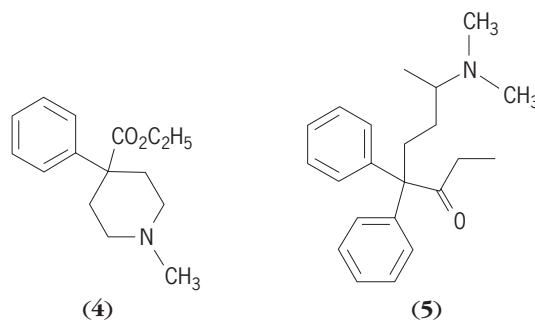


Biosynthesis of thebaine.

coupling of a benzylisoquinoline, as shown in the illustration for the formation of thebaine from reticuline. The discovery of a biosynthetic relationship between the benzylisoquinoline and the morphine ring systems provided a key insight to the structures of these and several other groups of alkaloids. The biosynthetic pathway, beginning with the amino acid tyrosine and proceeding via the steps in the illustration, has been established by the incorporation of isotopic labeling at specific positions in feeding experiments with labeled precursors. The final steps are successive demethylations of thebaine and codeine to the end product morphine. See BIOSYNTHESIS.

Morphine is an analgesic and central nervous system depressant, and is one of the most effective and widely used drugs for the relief of severe or prolonged pain. It is usually administered intramuscularly as the calcium salt. Morphine is also an addicting narcotic; prolonged use leads to dependence and physical addiction.

A number of compounds have been synthesized in an effort to obtain a nonaddicting substitute. Among these are several modifications of the morphine structure, including the simpler systems meperidine (Demerol; 4) and methadone (5). These two compounds have less addiction liability but also have less analgesic potency.



(4)

(5)

Heroin is prepared readily by acetylation of morphine. This compound is absorbed parenterally much more rapidly than morphine, and the psychological effects and euphoria of opium are more quickly achieved. Heroin is therefore a highly addictive substance, and it causes the most serious problem in illicit opium usage.

Codeine is a somewhat less active analgesic than morphine and is usually used orally in combination with a salicylate such as aspirin. Most of the codeine in medical use is made by methylation of morphine, since the amount present in opium is insufficient. Codeine and derivatives such as dihydro and hydroxy compounds are effective antitussive agents, widely used in cough syrups. See ADDICTIVE DISORDERS; ANALGESIC; NARCOTIC; OPIATES. James A. Moore

Bibliography. D. R. Dalton, *The Alkaloids*, 1979; *Kirk-Othmer Encyclopedia of Chemical Technology*, 4th ed., 1999; R. N. F. Manske (ed.), *Alkaloids*, vols. 6 and 13, 1960, 1971.

Morphogenesis

The development of form and pattern in animals. Animals have complex shapes and structural patterns that are faithfully reproduced during the embryonic development of each generation. Morphogenesis is a higher outcome of the process of differentiation, defined as the progressive structural and functional diversification and specialization of cells, and recognizable by specific molecular markers and morphological and histological organization. Differentiation is precisely patterned in space to create a three-dimensional functional tissue organization. The process by which this takes place is called pattern formation. Morphogenesis, differentiation, and pattern formation are related: Morphogenesis is accomplished via a complex series of individual and group cell movements called morphogenetic movements. The cell differentiations that lead to these movements are preterminal differentiations that result in coordinated changes in individual cell shape, adhesion, and motility, as well as production of extracellular matrix. The terminal differentiation of cells in the spatial patterns that define tissue relationships within organs and appendages takes place after morphogenetic movements cease, and serves to further refine the shapes generated by these movements. The starting point for these events is a single cell, the fertilized egg, or zygote, that divides repeatedly to form a multicellular embryo capable of carrying out the series of patterned differentiations leading to the morphogenetic movements that shape the embryo. See ANIMAL GROWTH; DEVELOPMENTAL BIOLOGY; EMBRYONIC DIFFERENTIATION; MOLECULAR BIOLOGY.

Differentiation. Cell differentiation involves the differential expression of genes in the nuclear DNA that encode proteins specifying the structure and function of each cell type. During cleavage, nuclei divide equivalently so that all cells of the embryo receive the total complement of genes contained in the DNA of the zygote nucleus (Fig. 1). However, cells in differ-

ent regions of the embryos of many animals contain cytoplasm which differs in composition. The cytoplasmic composition of a cell controls which genes will be expressed in each region of the embryo, resulting in a patterned differentiation. Regional cytoplasmic differences are generated during embryonic development by several means. In most animals, regional differentiation of the egg's cytoplasm takes place during its development in the ovary and/or after fertilization. Cleavage segregates these regional differences into separate cells. These differences produce a small number of cell types during cleavage of the zygote. In other animals, mammals being a prime example, the cytoplasm of the egg appears to be homogeneous. Initial differences in cells arise because of the positions they occupy: cells on the outside of the early mammalian embryo become trophoblast cells, whereas cells on the inside become the inner cell mass from which the embryo develops. In all developing multicellular organisms, interactions between these early cell types result in new patterns of gene activity that further increase the diversity of cell types. By means of cascades of such interactions, all of the different cell types of an animal body gradually emerge in the proper spatial patterns. See CELL DIFFERENTIATION; CLEAVAGE (EMBRYOLOGY); DEOXYRIBONUCLEIC ACID (DNA); GENE ACTION.

Axial polarity. The structural patterns of animals and their parts exhibit polarity; that is, they display structural differences along anteroposterior (AP), dorsoventral (DV), and right-left axes. In many animals, the axial polarities of the embryo are established during oogenesis and/or the period between fertilization and the first cleavage, by the localization of maternal determinants into specific parts of the egg. For example, in the nematode worm *Caenorhabditis elegans*, the sperm nucleus moves from its locus of entry to the nearest end of the oblong egg. This end of the egg becomes the posterior pole of the AP axis, and the opposite end becomes the anterior pole. Microtubules organized by the sperm centriole act as tracks to position partitioning (PAR) proteins at the anterior and posterior poles of the AP axis. PAR 1 and 3 localize to the posterior pole, and PAR 2 localizes to the anterior pole. The PAR proteins act to localize maternal differentiation factors into different regions of the egg, where

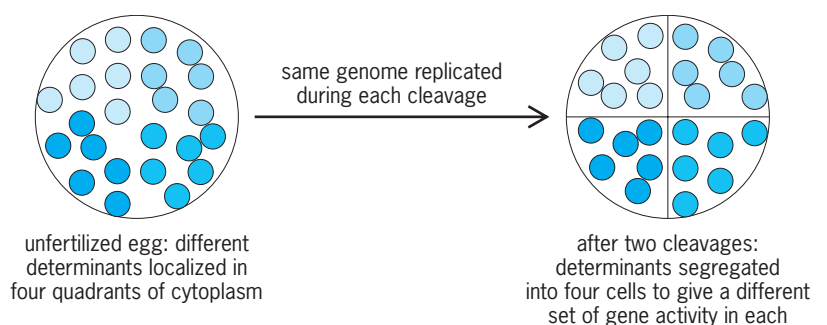


Fig. 1. General concept of how localized maternal determinants of the unfertilized egg give rise during cleavage to differential gene activity in different cells. Left, four types of determinants are each localized to a different quadrant of the egg. During cleavage, the same genome is copied to each cell. Right, two cleavages have partitioned each set of determinants into each of four cells, where they activate different sets of genes.

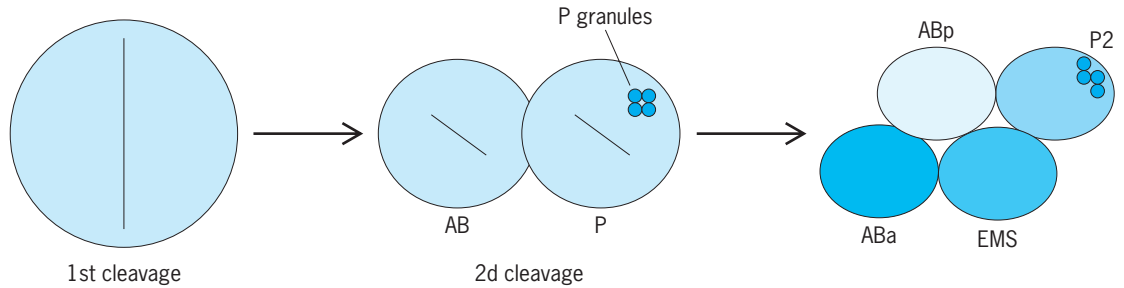


Fig. 2. Cleavage of the nematode worm *Caenorhabditis elegans*. Lines indicate cleavage planes. The first cleavage divides the egg into anterior and posterior blastomeres reflecting the anterior-posterior axial polarity of the fertilized egg. The posterior blastomere contains polar granules that will eventually become germ cells. The second cleavage is at a 45° angle to the first one, creating two dorsal and two ventral cells offset from one another, and establishing the dorsal-ventral axis of the worm.

they will be separated by cleavage. One set of such factors is the P granules, which become localized at the posterior pole. The first cleavage divides the egg into an anterior AB blastomere and a posterior P blastomere, which contains the P granules (Fig. 2). Those descendants of the P cells receiving P granules will become germ cells. The remaining descendants of the P cell and those of the AB cell will become the somatic cells of the worm. The DV axis of the nematode is established by the division of the AB and P cells in the plane of the AP axis, but at a 45° angle to that axis, placing two cells with different morphology perpendicular to the AP axis.

The AP and DV axial polarity of the eggs of anamniote and amniote vertebrates is established somewhat differently. The amphibian is the classic example of anamniote development. As the amphibian egg grows in the ovary, one half (the “vegetal” half) becomes laden with yolk, and a dark pigment is deposited in the cortical cytoplasm of the opposite half (the “animal” half) to form an animal-vegetal axis around which the egg is radially symmetrical.

The cells derived from the animal half during cleavage will become the ectoderm and mesoderm of the embryo, and the cells derived from the vegetal half will become the endoderm. The animal-vegetal axis approximates the AP axis of the embryo. During oogenesis, the signaling factor Vg1, the receptor for the Wnt signaling pathway (Disheveled) and the transcription factor VegT become localized to the cortical cytoplasm of the vegetal half and will be partitioned to the endodermal cells during cleavage. When the egg is fertilized and shed into the water, it orients with the heavier yolk half down. See FERTILIZATION; OOGENESIS; OVUM.

The DV axis of the amphibian embryo is established by a reorganization of the zygote cytoplasm initiated by the events of fertilization (Fig. 3). This axis can form in any of the planes that contain the animal-vegetal axis. The plane in which the DV axis actually forms is determined by the meridian at which the sperm enters the egg. This meridian is coincident with the ventral midline of the embryo; the opposite meridian coincides with the dorsal midline.

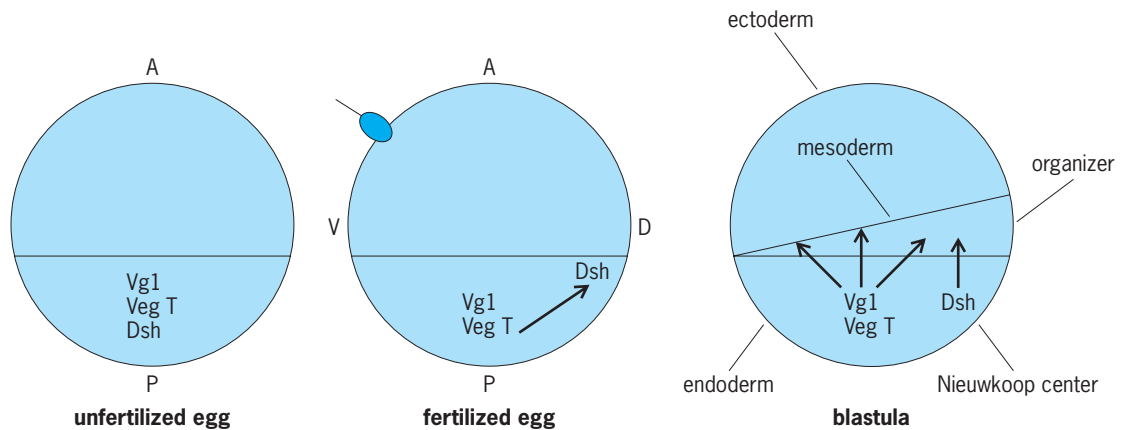


Fig. 3. Summary of axis formation and mesoderm induction during frog development. The unfertilized egg: the maternal determinants vegetal 1 (Vg1), vegetal T (Veg T), and Disheveled (Dsh) are localized to the vegetal cytoplasm of the egg. Fertilization establishes the dorsal-ventral axis. The side of sperm entry becomes the ventral (belly) side of the embryo, and the opposite side becomes the dorsal (back) side. Dorsality is conferred by a 30° upward rotation of the cortical cytoplasm that carries Dsh with it. Blastula stage formed by cleavage. Dsh is contained in the endodermal cells of a signaling center called the Nieuwkoop center and is part of a cascade of gene activation that induces the overlying ectoderm cells as the dorsal mesoderm of the embryo, while Vg1 and Veg T, localized into other endodermal cells, induce other parts of the mesoderm along the dorsal-ventral axis. The three germ layers—ectoderm, mesoderm, and endoderm—are now ready for the morphogenetic movements of gastrulation.

The establishment of the dorsal pole of the DV axis is associated with the upward contraction of the pigmented cortical cytoplasm toward the animal pole (Fig. 3). This contraction is asymmetrical, being greatest dorsally and least ventrally, thus creating a crescent-shaped band of lightly pigmented cytoplasm around the egg, called the gray crescent, the broadest part of which coincides with the dorsal midline. DV polarity is at first labile because it can be wiped out and reestablished in a new plane by tilting the AP axis of the zygote 90° off the vertical so that its ventral side is up. This procedure mimics the cytoplasmic reorganization which occurs after fertilization. A tongue of yolky cytoplasm is formed on the elevated side as the heavy yolk flows down under the influence of gravity, and a new gray crescent appears on this side, its dorsal midline coinciding with the plane in which the AP axis was tilted. The DV axis established by sperm entry becomes determined shortly before the first cleavage, but tilting the egg at this time can still lead to the establishment of a second DV axis. When this is done, a double embryo develops.

Regardless of how it is accomplished or its geographical location, the cortical rotation localizes dorsal maternal determinants to the dorsal pole of the DV axis (Fig. 3). These determinants will be incorporated during cleavage into the dorsalmost blastomeres of the endoderm, a region of the blastula called the Nieuwkoop center. This center plays an important role in inducing formation of the mesoderm, the dorsalmost region of which is called Spemann's organizer because it subsequently determines the AP and DV axial pattern of the embryo. The major determinant that is localized by the cortical rotation into the region that will become the Nieuwkoop center is Disheveled. Disheveled protects the transcription factor β -catenin from being degraded in this region, allowing it to act in combination with another transcription factor, Tcf3, to activate genes critical to the function of the Nieuwkoop center.

In birds, the AP axis is determined during early cleavage stages by gravity. The early embryo of the bird is a disc that lies atop a large mass of yolk. The egg is fertilized, and the cytoplasm on top of the yolk begins to divide as the egg passes down the oviduct. As it moves downward, the egg spins slowly, shifting the yolk so that one side of the radially symmetrical blastoderm is tilted upward. This side becomes the posterior pole of the AP axis. It is the equivalent of the Nieuwkoop center and expresses β -catenin. The equivalent of Spemann's organizer lies just anterior to the Nieuwkoop center. The AP axis in mammals also appears to be determined by gravity. Microgravity environments, such as space, interface with both early avian and early mammalian development. The DV axis of both birds and mammals appears to be determined by a pH and membrane potential difference above and below the blastodisc. The albumin above the blastodisc is basic, and the fluid of the subgerminal cavity below the blastodisc is acidic. The

blastodisc pumps water and sodium ions from the albumin into the subgerminal cavity, establishing a potential difference of ~25 mV (positive on the subgerminal cavity side) across the epiblast. The negative basic side of the blastodisc becomes dorsal and the opposite side ventral.

Left-right asymmetry in both invertebrates and vertebrates appears to be the result of a mechanism in which the Notch signal transduction pathway is activated on the left side of the embryo, leading to the expression of a Nodal-related gene in the mesoderm-forming cells on that side. Nodal is a signaling molecule that activates the gene for the transcription factor Pitx2, which specifies the direction of intestinal looping and heart development. In *C. elegans* and the frog, the asymmetric expression of the Nodal-related gene appears to be determined at fertilization by the localization of maternal determinants. In the frog, the determinant may be Vg1, which is activated to a higher degree on the left side of the embryo. In amniotes like birds and mammals, the Notch pathway and *nodal* expression are activated on the left side during gastrulation. The trigger for the asymmetric activation of Notch in the chick appears to be a transient accumulation of extracellular calcium resulting from membrane potentials set up by differential activity of ion pumps.

Determination of pattern. Once axial polarities are established, the pattern of structural differentiation along those axes is elaborated. In *C. elegans*, patterning relies heavily on further segregation of maternal determinants as cleavage proceeds. However, interactions between different cells are also required. For example, if the AB and P cells are isolated from one another, the P cell makes all of the cells it would normally make, but the AB cell makes only some of the cells it would normally make, indicating that interaction with the P cell is required for the AB cell to make a full complement of cell types. Gastrulation begins at the 24-cell stage, when the two cells that will form the endoderm migrate into the center of the embryo, creating a small blastopore. The cells that will form the germ line, the prospective mesenchymal and muscle cells, and the prospective pharyngeal cells migrate in succession through the blastopore into the interior of the embryo, where they take up their appropriate positions relative to one another and their descendants differentiate into the various parts of the worm.

In vertebrates, patterning relies mainly on cell interactions. The first step in patterning is to induce the formation of the mesoderm. At the mid-blastula stage of the frog embryo, when the embryo consists of several thousand cells, an inductive interaction takes place between the endoderm cells and the cells derived from the gray crescent region of the egg, causing the latter to become mesoderm cells. In the endoderm, VegT, Vg1, and β -catenin collaborate to activate the expression of Nodal-related genes. Nodal-related proteins are expressed in a gradient

that is highest dorsally in the Nieuwkoop center and lowest ventrally. β -Catenin and Tcf3 also activate the *siamois* gene in the Nieuwkoop center. Together, the Nodal-related and *siamois* proteins induce the dorsal mesoderm of Spemann's organizer by activating the *gooseoid* gene. Gooseoid is a transcription factor that activates genes, conferring on the organizer the ability to organize the AP and DV pattern of the mesoderm and initiate the morphogenetic movements of gastrulation. The cells of the organizer form the circular blastopore of the gastrula, through which the prospective endoderm, mesoderm, and endoderm migrate to form three concentric layers, with ectoderm on the outside, endoderm on the inside, and mesoderm in between. The ectoderm develops into the epidermis and nervous system, the endoderm into the organs of the alimentary tract, and the mesoderm into muscles, skeleton, heart, kidneys, and connective tissue. See BLASTULATION; GASTRULATION; GERM LAYERS.

Prior to gastrulation, the prospective organ regions of the ectoderm and endoderm are not yet determined, as shown by the fact that they are unable to differentiate in isolation, and that they can develop in ways different from their normal fate when transplanted elsewhere in the embryo. Mesodermal organ regions, however, are highly self-organizing under these conditions. Ectodermal and endodermal organ regions become determined during and after gastrulation by the inductive action of the mesoderm. For example, dorsal mesoderm normally invaginates and stretches out along the dorsal midline where it differentiates as notochord and trunk muscles. The ectoderm overlying the dorsal mesoderm differentiates as the central nervous system. When dorsal mesoderm is transplanted to the flank of an early gastrula, it invaginates and differentiates according to its normal fate. The overlying flank ectoderm, which normally would become skin epidermis, forms a secondary nervous system, indicating that the nervous system is normally determined by dorsal mesodermal factors acting on the overlying ectoderm. Surprisingly, this is not a stimulatory action but an inhibitory one in which the neural default state of the ectoderm is preserved by several proteins inhibitory to the action of bone morphogenetic proteins that induce ectoderm elsewhere to become epidermis. The dorsal mesoderm of the gastrulating embryo also specifies the AP and DV pattern of the neural tube by the region-specific expression of other inducing proteins.

In amniote embryos, such as birds and mammals, the Nieuwkoop center functions in the same way to induce an organizer. In birds, the blastula is a flat disc with an upper layer consisting mostly of prospective ectoderm with prospective mesoderm at its posterior pole, and a lower layer of prospective endoderm. The blastopore in amniote embryos is an elongated structure called the primitive streak, but it functions in the same way as the circular amphibian blastopore to create a three-layered gastrula. In this case, prospective mesoderm is internalized between the

prospective ectoderm and endoderm by ingression through the primitive streak. While not as thoroughly studied as the frog embryo, it is clear that many of the genes involved in Nieuwkoop center formation, organizer induction, and axial patterning are the same as in the frog. See FATE MAPS (EMBRYOLOGY).

Once induced, organ regions of vertebrate embryos can themselves induce additional organs from undetermined tissue. For example, the retina and iris of the eye develop from a vesicle growing out of the forebrain. This vesicle induces a lens from the overlying head ectoderm, and the lens then induces the cornea from head ectoderm to complete the eye. By means of such cascades of inductive interactions, all the organs of the body are blocked out. See EMBRYONIC INDUCTION.

Morphogenetic fields and positional information. Once determined, an organ region constitutes a developmental system, called a morphogenetic field, which specifies the detailed pattern of cell differentiation within the organ. Cells differentiate in patterns dictated by their relative positions within the field. It is generally accepted that graded molecular signals, or cues, are the basis of this positional information. It is proposed that the source of the signals is a set of boundary cells which define the limits of the field. All the cells of a field thus derive their positional information from a common set of boundary cells. The signals can be propagated by extracellular diffusion (paracrine action between different cells or autocrine action of a cell on itself), by intercellular contact via specialized connections between cells (called gap junctions), or by interactions between molecules located on the surfaces of adjacent cells (juxtacrine actions). These signals exert their effect by binding to receptors that transduce the signal to the nucleus by one of several signal transduction pathways. The cell integrates these signals to interpret how it should differentiate, and activates transcription factors that stimulate the appropriate gene activity to carry out the differentiation. Among the most important transcription factors involved in axial patterning are the homeodomain proteins encoded by the homeobox (*Hox*) genes. There are four clusters of 13 *Hox* genes each (A, B, C, and D). Within each cluster, the genes are numbered from 1 to 13, and they are expressed in that order from anterior to posterior and from dorsal to ventral.

Cells from different regions of a field can be interchanged within the field boundaries, and the pattern they would have formed is respecified to the pattern dictated by the positional information of their new location. For example, the proximal mesoderm of an embryonic chick leg bud normally forms thigh muscles, but will form toes if grafted to the distal end of the bud. Also, if part of the field boundary is removed, it is restored by some of the remaining cells. For example, if the anterior half of the limb field of a salamander embryo is removed, the posterior half regulates to form a whole limb. Some of the posterior cells take on anterior boundary

properties; the whole positional information gradient is then reestablished over the smaller number of cells, and the half pattern is adjusted to a whole. AP regulation of the limb field does not occur in some other embryos such as birds, which have a fixed determination.

Most fields become inactive after the pattern they specify begins to differentiate; the ability to form normal organs after removal or interchange of cells is then lost. However, in some animals, such as salamanders, the fields of certain organs can be reactivated by loss of a part, even in the adult. The missing part is then redeveloped in a process called regeneration. See REGENERATION (BIOLOGY).

Cell matrix and adhesivity. Each cell type of an animal is characterized by a specific set of gene-encoded proteins which determine the structure, shape, mobility, and function of the cell. In addition, different kinds of cells secrete molecules of protein and protein complexed with carbohydrate which constitute equally specific types and patterns of extracellular matrix. The matrix stabilizes tissue and organ structure, guides migrating cells to their proper locations, and is a medium through which cell interactions take place. For example, the supporting strength of bone is due to a compact, mineralized matrix of protein-carbohydrate fibers, whereas the matrix of muscle is a nonmineralized, loose network of fibers which can stretch and contract with the muscle cells.

Cell interactions take place at cell surfaces, the molecular composition of which is distinct from one cell type to another, allowing them to recognize one another. These differences are reflected in varying degrees of adhesivity between different kinds of cells, and between cells and different kinds of extracellular matrix. Differential adhesivity is the property upon which cell migration, clustering, and rearrangement is based, and is thus important in creating the conditions for cell-cell and cell-matrix interactions. This is shown by an experiment in which endoderm, mesoderm, and ectoderm cells of amphibian gastrulae were dissociated, then mixed together in a mass. Each of these cell types sorted out from one another so that their normal concentric spatial relationships were restored. In embryonic stages where organs have begun differentiation, the same sorting behavior is observed among cartilage, heart, retina, epidermis, brain, and liver cells. See CELLULAR ADHESION.

Differential growth. Another important mechanism for the development of complex patterns and shapes is the differential growth of organs and their parts. Differential growth is evident as soon as embryonic organ regions begin to develop. During the early part of their development, the growth rates of organs are controlled by intrinsic factors. For example, the vertebrate limb region is at first indistinguishable from the surrounding flank, and the cells of the two regions are dividing with the same frequency. Subsequently, the flank cells divide less frequently than the limb cells, resulting in the elevation and outgrowth of

the limb region from the body wall. At later stages of development, including postnatal life, organ growth is largely under the control of hormones secreted by cells of the endocrine glands. When growth ceases at maturity, the relative sizes of the body parts constitute distinctive features of individual and species-specific form.

Cell death. Apoptosis, or cell death by suicide, is an important feature in the refinement of final cell number and shape of developing organs and appendages. For example, a mature *C. elegans* worm has exactly 959 somatic cells, but during its development an additional 115 cells formed that were pruned out by apoptosis. Striking examples of the role of apoptosis in developing vertebrates are the elimination of excess neurons in the spinal cord, and foot development in ducks and chickens. Ducks have webbed toes while chickens do not. This difference results because as the chicken leg bud grows and forms the toes, the cells between the developing digits undergo apoptosis. Various grafting and culturing experiments have indicated that it is a cell's relative position in the limb bud which establishes its fate to die at some later time in development. Apoptosis is either triggered by a signal from other cells or is cell-autonomous, but is preventable if a rescue signal is available. Thus within each cell, there are genes that protect against apoptosis and genes that carry out the death sentence. The protective genes belong to the *Bcl-2* family. The genes that destroy the cell are *Apaf1* (apoptotic protease activating factor 1), which activates the proteases caspase 3 and 9. The caspases digest the cell from within.

David L. Stocum

Bibliography. S. F. Gilbert, *Developmental Biology*, 7th ed., Sinauer, 2003; K. Kalthoff, *Analysis of Biological Development*, McGraw-Hill, 1996; C. D. Stern (ed.), *Gastrulation: From Cells to Embryo*, Cold Spring Harbor Laboratory Press, 2004.

Mortar

A binding agent used in construction of clay brick, concrete masonry, and natural stone masonry walls and, to much less extent, landscape pavements. Modern mortars are improved versions of the lime and sand mixtures historically used in building masonry walls. See BRICK; MASONRY.

Mortar types are identified by property or performance specifications of the American Society for Testing and Materials. Selection of the correct type is based on the anticipated exposure—interior or exterior—and structural requirements. Type O, for example, is not designed for exterior load-bearing applications, but may be used in non-load-bearing interior positions.

Masonry mortar is composed of one or more cementitious materials, such as masonry cement or portland cement and lime, clean sand, and sufficient water to produce a plastic, workable mixture. Masonry cements are prepackaged blends of hydraulic

cementitious materials (including portland cement or blended cement), plasticizing additives, and sometimes pigments. They require only the addition of sand and water at the job site. Portland cement-lime mortars are proportioned on the job site from bulk or bagged cement and lime, along with sand, water, and any admixtures. Proportions differ greatly; however, a sample mortar might have one part cementitious material and three parts sand, with water quantity sufficient to create a workable mixture. *See* CEMENT; LIME (INDUSTRY).

Mortars are batched in machine mixers, in larger, mobile plants, or at a central mixing plant. The plastic mixture is then spread over masonry units to form horizontal and vertical joints, typically $\frac{3}{8}$ in. (1 cm) wide. Joints are finished (tooled) in a vee, concave, flush, or other desired profile.

Mortars are closely related to concrete but, like grout, generally do not contain coarse aggregate. (Mortar and grout differ, however, and are not interchangeable.) Mortars function with the same calcium silicate-based chemistry as concrete and grouts, bonding with masonry units into a contiguous, weatherproof surface in the process. Masonry cement or portland cement-lime mortars can be formulated not only to meet guidelines for types M, S, N, and O, but also to address job-specific requirements including setting time, rate of hardening, water reactivity, and extended workability. *See* CONCRETE; GROUT.

John Melander

Mosaicism

The condition in which more than one genetically distinct population of cells coexists within one individual. The term mosaic is used for an individual composed of two or more cell lines of different genetic or chromosomal constitutions, where the cell lines originate from one zygote. Individuals derived from more than one zygote or from more than two gametes are called chimeras. However, because the etiology of mosaicism cannot always be established, the term mosaic often is used to refer to any kind of genetic multiplicity within an individual. Mosaicism is a common phenomenon in both the plant and animal kingdoms; it may originate at any state in the course of ontogeny, and in any tissue in which cells proliferate. Cells with a particular genetic constitution form a clone, and this may appear as a mosaic spot embedded in a background containing cells of different genotype. The several types of mosaicism include chromosomal mosaicism, mosaicism due to mitotic recombination or gene mutation, and functional mosaicism.

Chromosomal mosaicism. Abnormal segregation of chromosomes in the course of mitosis results in progeny cells with irregular numbers of chromosomes. Chromosome nondisjunction brings about aneuploidy, a condition in which the cell bears more or fewer chromosomes than normal (**Fig. 1**). Such a condition is often lethal to the cell due to genetic imbalance, especially when more than one large auto-

some is involved. In humans, trisomic mosaicism for several autosomes has been reported, and trisomy 21, that is, mosaicism for Down syndrome, is the most common. Depending on the ratio and distribution of trisomic and normal diploid tissues, the clinical features can range from syndromes characteristic of complete trisomy to a normal phenotype. Mosaicism can also involve germ cells; such individuals, although they may not manifest symptoms of trisomy, can produce aneuploid progeny. *See* CHROMOSOME ABERRATION; DOWN SYNDROME.

Abnormal chromosome segregation can involve not only one or a few chromosomes but also an entire genome. In these instances, diploid/triploid or diploid/polyloid mosaicism develops. This type of mosaicism can occur spontaneously, or it can be induced by drugs. In *Drosophila* and a number of plant species the mitotic poisons colcemid and colchicine can induce mosaics of this type. Diploid/polyloid mosaicism can involve somatic as well as germ cells. *See* POLYPLOIDY.

The elimination of complete chromosomes or chromosome fragments is a frequent source of mosaicism. As a rule, in some organisms (for instance, the gall midge *Wachtiella* and the nematode *Ascaris*) several chromosomes or fragments are eliminated during embryogenesis from precursors of somatic cells but not from germ cells.

Elimination of sex chromosomes is often the source of sex mosaicism. In *Drosophila* XX/XO female/male sex mosaic gynandromorphs develop from XX zygotes through elimination of one of the X chromosomes (here O represents an absent sex chromosome). Gynandromorphs of other insect species (butterflies, moths, and beetles) are also known. Many hymenopterans, such as the wasp *Habrobracon juglandis*, possess a method of sex determination where unfertilized eggs produce haploid males and fertilized eggs produce diploid females. Gynandromorphs in these species are diploid/haploid mosaics, where the female part is diploid and the male is haploid.

In humans, XX/XO mosaics are females who usually possess symptoms of Turner's syndrome, but some have normal, XX-bearing ovarian cells and are fertile. Besides XX/XO mosaics in humans and in other mammals, XX/XY, XY/XO, XXY/XO, XO/XY/YYY, XYY/XO, XXX/XO, and other mosaics have also been reported. Most of them were noticed as a consequence of the abnormal sexual development of the mosaic individual. Such mosaics may have originated after aggregation of two embryos or as a consequence of "double fertilization," where one sperm nucleus fused with the oocyte and another with a polar body nucleus. In humans most of the XX/XY mosaics are hermaphrodites (persons with both ovarian and testicular tissues). *See* HUMAN GENETICS.

In plants, elimination of chromosomes or chromosome segments that code for the synthesis of chlorophyll or other colored substances results in a colorless cell, and the clonal descendants of this cell form a lightly colored or noncolored patch on the leaves,

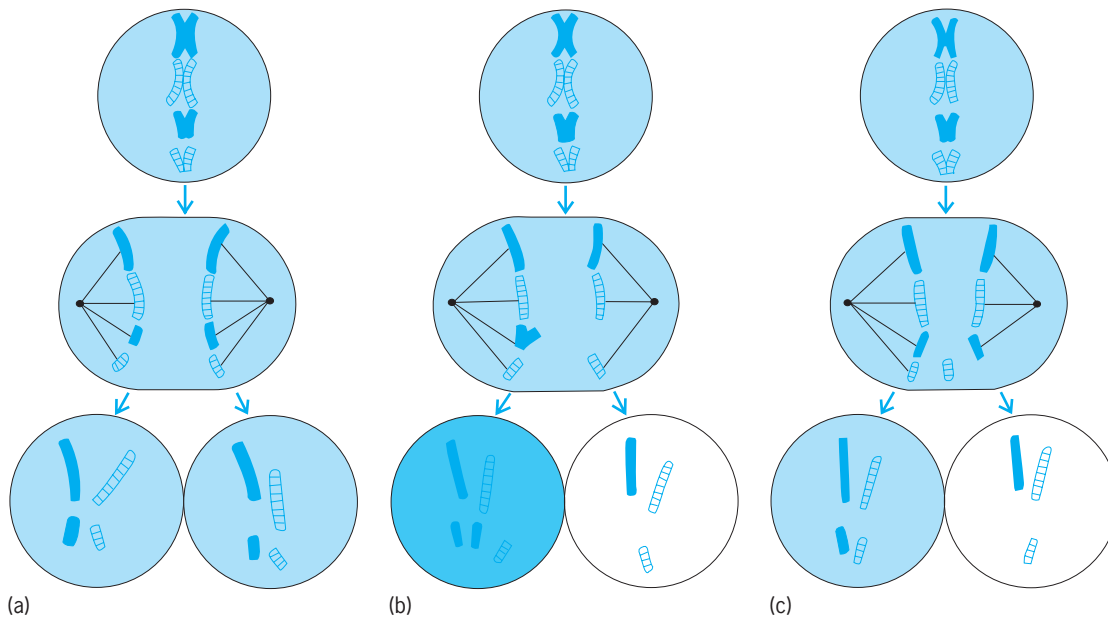


Fig. 1. Chromosomal mosaicism. (a) Normal mitosis with two pairs of homologous chromosomes. (b) Somatic nondisjunction resulting in aneuploid daughter cells: one bears an additional chromosome (trisomy) and the other lacks one chromosome (monosomy). (c) Chromosome elimination leading to monosomy.

flowers, or fruits of the plant. This type of pattern, called variegation, can also occur when normal green and mutant colorless plastids sort out in the course of successive cell divisions. This type of mosaicism is quite distinct in pattern and inheritance from those attributable to any kind of nuclear mutation. It is inherited maternally rather than in mendelian fashion, because plastids are passed to subsequent generations through the female germ line via egg cytoplasm. See MATERNAL INFLUENCE.

Loss of chromosomes or chromosome fragments can occur spontaneously (with a low frequency), but it can also be induced by several chemicals, and by x-rays and gamma rays. The effect of radiation is that abnormal chromosomes often form rings or dicentrics which are eliminated during subsequent mitoses.

Mitotic recombination. The exchange of parts of homologous chromosomes often results in cells with a different genetic constitution from that of the mother and the surrounding cells (Fig. 2). Progeny of the daughter cells, when properly marked genetically, can be recognized after cell proliferation in both somatic and germ cells. Mitotic recombination (also called somatic crossing-over) is a relatively rare event, but has been known to be the source of spontaneous mosaicism in *Drosophila* and some other insects, in the nematode *Caenorhabditis elegans*, in mice, and in a number of plant species (soybean, tobacco). The frequency of somatic crossing-over can be greatly increased by ultraviolet, x, and gamma rays and also by chemical mutagens and carcinogens (Fig. 3). See CROSSING-OVER (GENETICS); MUTAGENS AND CARCINOGENS; RECOMBINATION (GENETICS).

Since mitotic recombination can be induced at any time during development, and since it provides a method for the genetic labeling of single cells, it is a

useful tool for cell lineage studies. It has been widely employed in *Drosophila* to study gene activity, programmed cell death, cell proliferation, and the development and functioning of different organs.

Gene mutation. In some cells that are heterozygous for mutations, the wild-type alleles may lose their

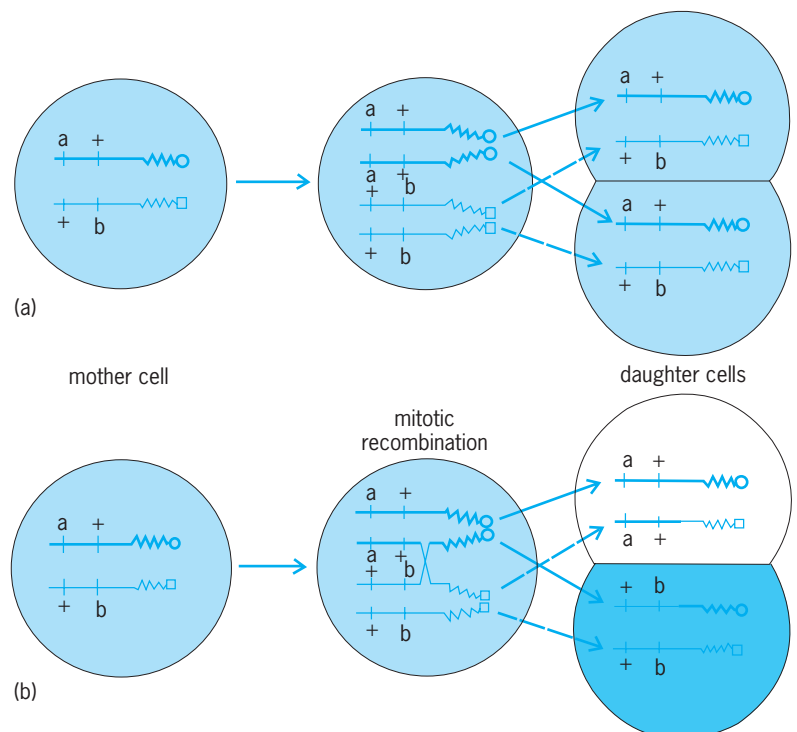


Fig. 2. Mitotic recombination. (a) Mitosis produces daughter cells with genotype identical to that of the mother cell. (b) After mitotic recombination the daughter cells become homozygous for the marker mutations a and b, and the progeny of these cells form mosaic spots.

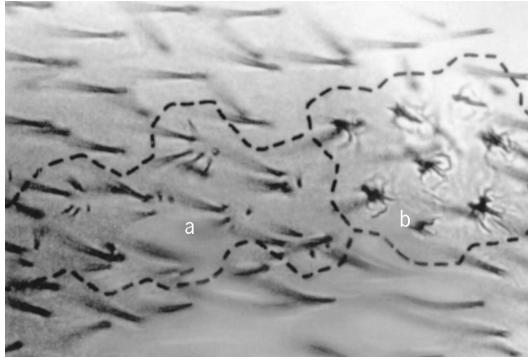


Fig. 3. A twin spot from a wing of *Drosophila melanogaster*, formed after x-ray-induced mitotic recombination. Letters a and b indicate separate clones of cells showing different mutant phenotypes; the background tissue shows the normal phenotype. (Photo by Andrew W. Wayne)

activities as a consequence of a new mutation and thus mosaicism can develop. Mutagens and carcinogens are known to induce such mutations and, indeed, these substances can be screened on the basis of this behavior. Activity of the wild-type allele can be suppressed due to nucleotide substitutions, or to deletions or additions of single nucleotides, and also because certain deoxyribonucleic acid sequences, the so-called transposable controlling elements, may become integrated into the gene. Elements of this sort, discovered in maize by B. McClintock in 1950, produce a variegated phenotype. See TRANSPOSONS.

Functional mosaicism. Although the genetic constitution is the same (XX) in each of the cells of female eutherian mammals, they show mosaicism for the majority of X-linked genes. This functional mosaicism is due to X-chromosome inactivation. Maternally and paternally derived X chromosomes are turned off randomly during embryogenesis, and consequently recessive X-linked genes can be expressed despite their heterozygous condition. An inactive X chromosome forms a condensed mass of chromatin called a Barr body, after M. L. Barr who first described these bodies in 1949. For example, the calico cat is heterozygous for black and orange alleles of an X-linked coat color gene. Due to X-chromosome inactivation, calico cats exhibit an irregular patchwork of black and orange sectors, with each sector representing a clone of cells derived from one cell. The random inactivation of one of the two X chromosomes in all the somatic cells of females ensures that the somatic cells of males and females generate equivalent levels of X-linked gene products.

Another type of mosaicism develops when genes are juxtaposed to blocks of heterochromatin as a consequence of inversion or translocation of chromosome fragments. These genes may become inactive in some cells but not in others, and thus mosaicism develops. This type of mosaicism is called position effect variegation. See GENETICS; SEX-LINKED INHERITANCE.

Janos Szabad

Bibliography. W. Beermann et al. (eds.), *Genetic Mosaics and Cell Differentiation: Results and Prob-*

lems in Cell Differentiation, vol. 9, pp. 1-315, 1978; A. Garcia-Bellido, P. A. Lawrence, and G. Morata, Compartments in animal development, *Sci. Amer.*, 24:102-110, 1979; L. B. Russel (ed.), *Genetic Mosaics and Chimeras in Mammals*, vol. 12 of *Basic Life Sciences*, 1978; C. Stern, *Genetic Mosaics and Other Essays*, 1968; R. A. E. Tilney-Bassett, The inheritance and genetic behaviour of plastids, in J. T. O. Kirk and R. A. E. Tilney-Bassett (eds.), *The Plastids*, pp. 251-524, 1978.

Mosquito

Any member of the family Culicidae in the insect order Diptera. Mosquitoes are holometabolous insects and all larval stages are aquatic. Adults are recognized by their long proboscis for piercing and sucking, and characteristic scaled wing venation (Figs. 1 and 2).

This is a relatively large group of well-known flies with nearly 3000 species in 34 genera reported in the world. There are 13 genera and 167 recognized species of mosquitoes in North America north of Mexico. Almost 75% of these species belong to three genera: *Aedes* (78 species), *Culex* (29 species), and *Anopheles* (16 species).

Biology. Adult females lay their eggs on or near water: the genera *Culex* and *Culiseta* deposit egg rafts on the surface; *Anopheles* lays eggs singly on the surface; and *Aedes* and *Psorophora* lay batches of eggs on substrates near the water. The latter two genera are called floodwater mosquitoes because their eggs require a period of drying for early embryonation to occur and hatch when rains and flooding cover them with water. Although mosquitoes may be found in nearly all aquatic habitats (ponds, streams, salt lakes, tree holes, artificial containers), the majority use still-water habitats.

Most larvae, or wrigglers, feed on algae and organic debris that they filter from the water with their oral brushes, although certain genera (*Toxorhynchites* and *Psorophora*) may be predaceous and feed on other mosquito larvae. Larvae go through three molts and four instars before pupation. Pupae, or tumblers, are active but nonfeeding stages in which metamorphosis to the adult stage occurs. Both larvae and pupae usually breathe through air tubes at the surface of the water.

Adult male mosquitoes are relatively short-lived, and do not suck blood, but feed primarily on nectar and other plant juices. Females also feed on nectar as their primary energy source, but they require a blood meal for egg production in most species. Some mosquito species are very host-specific, blood-feeding only on humans, birds, mammals, or even reptiles and amphibians, although many species will feed on any available host.

Importance. Mosquitoes are of major importance in both human and veterinary medicine. They can cause severe annoyance and blood loss when they

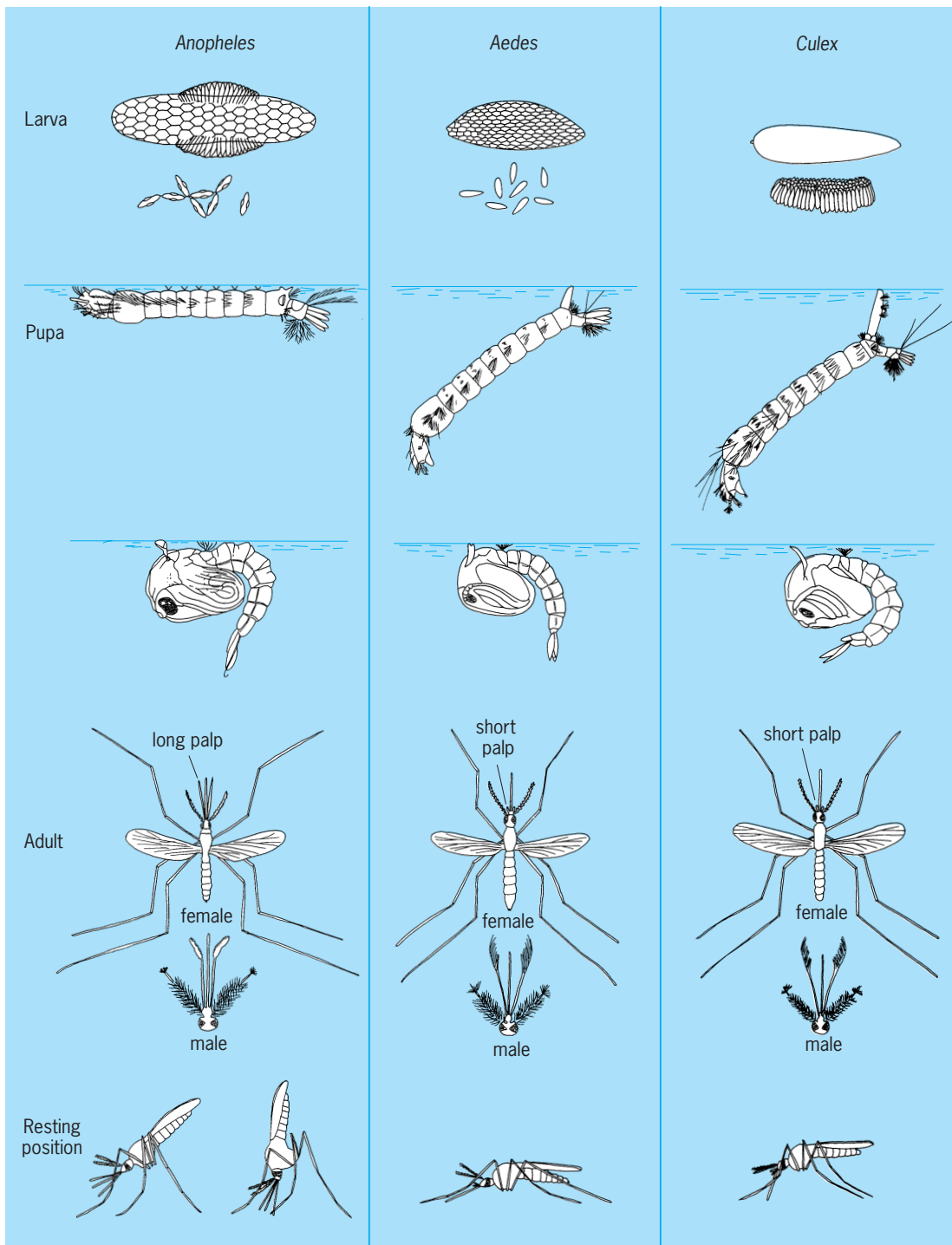


Fig. 1. Morphological and behavioral characteristics of three common genera of mosquitoes. (After *Pictorial Keys to Arthropods, Reptiles, Birds and Mammals of Public Health Significance*, PHS Publ. 1955, June 1969)

occur in dense populations and they act as vectors of three important groups of disease-causing organisms: *Plasmodium*, the protozoan parasite that produces malaria; filarial worms, parasitic nematodes causing elephantiasis in humans and heartworm disease in canines; and arboviruses, which are the causative agents of yellow fever, dengue fever, LaCrosse encephalitis, St. Louis encephalitis, western equine encephalomyelitis, eastern and

Venezuelan equine encephalitis, and several other viral diseases. Human malaria is transmitted exclusively by *Anopheles*, filariasis by *Culex*, *Anopheles*, and *Aedes*, and arboviruses primarily by *Culex* and *Aedes* species. See ARBOVIRAL ENCEPHALITIDES; HEARTWORMS; MALARIA; YELLOW FEVER.

All of these disease agents enter the mosquito when it feeds on an infected host, and within the vector mosquito the agent undergoes development



Fig. 2. The mosquito, a member of the insect order Diptera.

or multiplication. Following complete development, mosquitoes transmit the agent to susceptible hosts in a subsequent blood feeding; either the parasite is injected with the mosquito saliva (malaria and arboviruses), or the parasite breaks out of the proboscis and crawls into the wound produced during blood feeding (filariasis). See ENTOMOLOGY, ECONOMIC; INSECTA; MEDICAL PARASITOL-OGY.

Bruce M. Christensen

Bibliography. A. N. Clements, *Biology of Mosquitoes*, I: *Development, Nutrition, and Reproduction*, 1992; R. F. Darsie, Jr., and R. A. Ward, *Identification and Geographical Distribution of the Mosquitoes of North America, North of Mexico*, American Mosquito Control Ass., 1981; K. L. Knight and A. Stone, *A Catalog of the Mosquitoes of the World (Diptera: Culicidae)*, Entomological Society of America, 1977; M. W. Service, *Mosquito Ecology: Field Sampling Methods*, 2d ed., 1993.

Mössbauer effect

Recoil-free gamma-ray resonance absorption. The Mössbauer effect, also called nuclear gamma resonance fluorescence, has become the basis for a type of spectroscopy which has found wide application in nuclear physics, structural and inorganic chemistry, biological sciences, the study of the solid state, and many related areas of science.

Theory of effect. The fundamental physics of this effect involves the transition (decay) of a nucleus from an excited state of energy E_e to a ground state of energy E_g with the emission of a gamma ray of energy E_γ . If the emitting nucleus is free to recoil, so as to conserve momentum, the emitted gamma ray energy is $E_\gamma = (E_e - E_g) - E_r$, where E_r is the recoil energy of the nucleus. The magnitude of E_r is given classically by the relationship $E_r = E_\gamma^2/2mc^2$, where m is the mass of the recoiling atom and c is

the speed of light. Since E_r is a positive number, the E_γ will always be less than the difference $E_e - E_g$, and if the gamma ray is now absorbed by another nucleus, its energy is insufficient to promote the transition from the nuclear ground state E_g to the excited state E_e .

In 1957 R. L. Mössbauer discovered that if the emitting nucleus is held by strong bonding forces in the lattice of a solid, the whole lattice takes up the recoil energy, and the mass in the recoil energy equation given above becomes the mass of the whole lattice. Since this mass typically corresponds to that of 10^{10} to 10^{20} atoms, the recoil energy is reduced by a factor of 10^{-10} to 10^{-20} , with the important result that $E_r \sim 0$ so that $E_\gamma = E_e - E_g$; that is, the emitted gamma-ray energy is exactly equal to the difference between the nuclear ground-state energy and the excited-state energy. Consequently, absorption of this gamma ray by a nucleus which is also firmly bound to a solid lattice can result in the "pumping" of the absorber nucleus from the ground state to the excited state. The newly excited nucleus remains, on the average, in its upper energy state for a time given by its mean lifetime τ (a quantity dependent on energy, spin, and parity of the nuclear states involved in the deexcitation process) and then falls back to the ground state by reemission of the gamma ray. An important feature of this reemission process is the fact that it is essentially isotropic; that is, it occurs with equal probability in all directions. See ENERGY LEVEL (QUANTUM MECHANICS); EXCITED STATE; GAMMA RAYS; GROUND STATE.

Energy modulation. Before this phenomenon of resonance fluorescence can be turned into a spectroscopic technique, it is necessary to provide an appropriate energy modulation of the gamma ray emitted in the initial decay process. An estimate of the energy needed to accomplish this can be calculated from a knowledge of the inherent width or sharpness of the excited-state nuclear level. This is given by the Heisenberg uncertainty principle as $\Gamma = \hbar/2\pi\tau$ (\hbar is Planck's constant and τ is the mean lifetime of the excited state). In the case of ^{57}Fe , a nucleus for which resonance fluorescence is especially easy to observe experimentally, $\Gamma = 4.6 \times 10^{-12}$ keV. In order to modulate the emitted gamma-ray energy, which in this case corresponds to 14.4 keV, one can take advantage of the Doppler phenomenon which states that if a radiation source has a velocity relative to an observer of v , its energy will be shifted by an amount equal to $E = (v/c)E_\gamma$. Setting the required Doppler energy equal to the width of the nuclear level and E_γ equal to the nuclear transition energy leads to the equation below. Relative velocities of

$$\begin{aligned} v &= c \frac{\Gamma}{E_\gamma} = 3 \times 10^{10} \text{ cm/s} \times \frac{4.6 \times 10^{-12}}{14.4} \\ &= 0.0096 \text{ cm/s} = 0.0038 \text{ in./s} \end{aligned}$$

this order of magnitude can be used to modulate the gamma ray emitted in a typical Mössbauer transition, that is, to "sweep through" the energy width of the

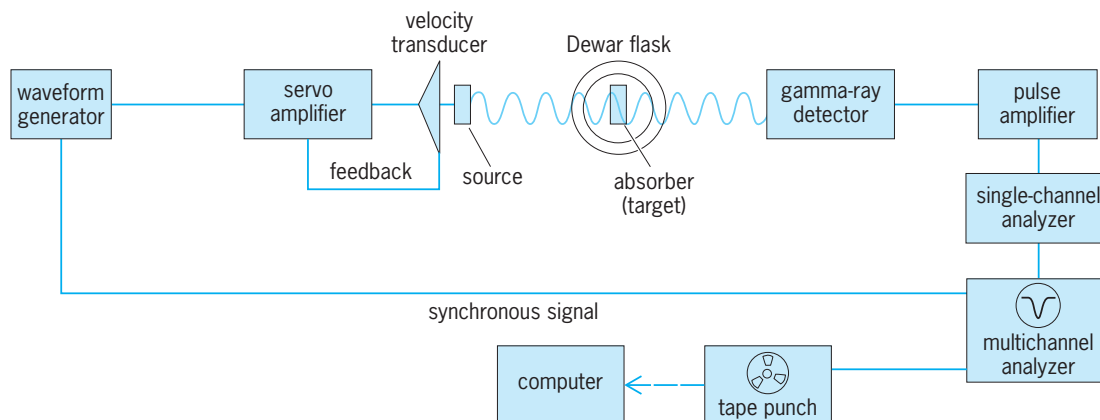


Fig. 1. Experimental arrangement for performing Mössbauer effect spectroscopy. This typical Mössbauer experiment is with ^{57}Fe or ^{119}Sn . (After R. H. Herber, *Mössbauer spectroscopy*, *Sci. Amer.*, 255(4):86–95, October, 1971)

nuclear transition. See DOPPLER EFFECT; RADIOACTIVITY; UNCERTAINTY PRINCIPLE.

Experimental realization. The experimental realization of gamma-ray resonance fluorescence can be achieved with the arrangement illustrated schematically in Fig. 1. In a typical Mössbauer experiment the radioactive source is mounted on a velocity transducer which imparts a smoothly varying motion (relative to the absorber, which is held stationary), up to a maximum of several centimeters per second, to the source of the gamma rays. These gamma rays are incident on the material to be examined (the absorber). Some of the gamma rays (those for which E_γ is exactly equal to $E_e - E_g$) are absorbed and reemitted in all directions, while the remainder of the gamma rays traverse the absorber and are registered in an appropriate detector which causes one or more pulses to be stored in a multichannel analyzer. The electronics are so arranged that the location (address) in the multichannel analyzer, where the transmitted pulses are stored, is synchronized with the magnitude of the relative motion of source and absorber.

A typical display of a Mössbauer spectrum, which is the result of many repetitive scans through the velocity range of the transducer, is shown in Fig. 2. Such a Mössbauer spectrum is characterized by a position δ of the resonance maximum (corresponding to a maximum in the isotropic scattering, and thus a minimum in the intensity of the transmitted radiation), a linewidth Γ , and a resonance effect magnitude ϵ corresponding to the total area A under the resonance curve. See GAMMA-RAY DETECTORS.

In the case of the Mössbauer active nuclides ^{57}Fe and ^{119}Sn , among others, two additional features which are of great interest to chemists and physicists may be experimentally elucidated. One of these is the quadrupole coupling which is observed if the Mössbauer nuclide is located in an environment where the electric charge distribution does not have cubic (that is, tetrahedral or octahedral) symmetry. Such a spectrum is shown in Fig. 3, in which the magnitude of the quadrupole interaction Δ is equal to $e^2qQ/2$, where e is the electron charge, q is the gradient of the electrostatic field at the nucleus, and Q is the

nuclear quadrupole moment. Finally, a Mössbauer spectrum can also give information on the magnitude of the magnetic field H_0 acting on the nucleus through the magnetic hyperfine interaction. This is illustrated in Fig. 4, where only a single resonance line would be observed in the absence of a magnetic interaction. See HYPERFINE STRUCTURE; NUCLEAR MOMENTS.

Moreover, all of these parameters— δ , Δ , Γ , A , and H_0 —are temperature-dependent quantities, and their study over a range of temperatures and conditions can shed a great deal of light on the nature of the environment in which the Mössbauer nuclide is located in the sample under investigation. More than one hundred Mössbauer transitions, involving 43 different elements, have been experimentally observed and reported.

Application. Mössbauer effect experiments have been used to elucidate problems in a very wide range of scientific disciplines, and only a few examples can be cited as representative of the information extracted from such studies.

Nuclear physics and chemistry. One of the narrowest resonance lines which has been observed is that from the 6.8-microsecond, 6.2-keV gamma transition in ^{181}Ta , and detailed Mössbauer effect measurements using a source of 140-day ^{181}W have shown that the magnetic moment of the spin- $9/2$ excited state

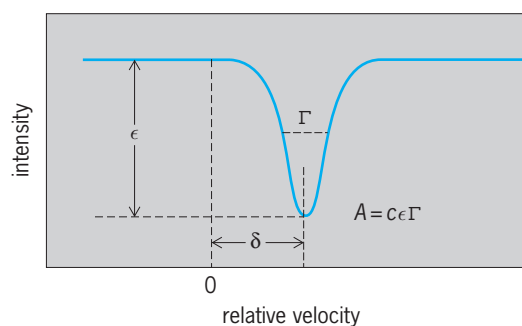


Fig. 2. Mössbauer spectrum of an absorber which gives an upsplit resonance line. The spectrum is characterized by a position δ , a linewidth Γ , and an area A related to the effect magnitude ϵ .

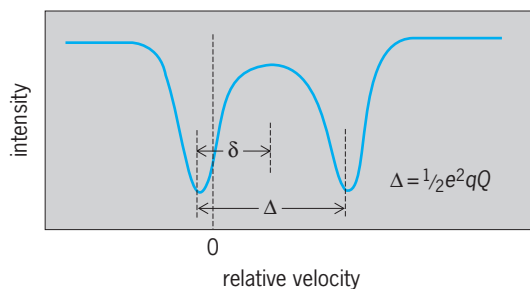


Fig. 3. Mössbauer spectrum of an absorber (containing for example ^{57}Fe or ^{119}Sn) which shows quadrupole splitting Δ .

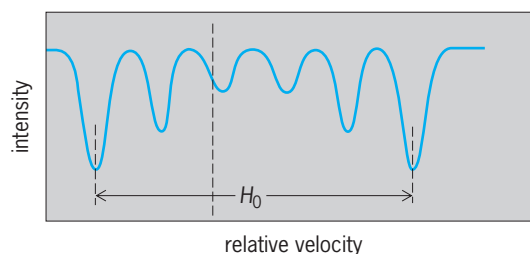


Fig. 4. Mössbauer spectrum of metallic iron showing the splitting of the resonance line by the internal magnetic field ($H_0 = 33 \text{ T} = 330 \text{ kG}$ at room temperature).

in ^{181}Ta is $+5.35 \pm 0.09$ nanometers and the nuclear quadrupole moment of this state is $+4.4 \pm 0.05) \times 10^{-24} \text{ cm}^2$. Such data are of considerable use to nuclear physicists in refining models which describe the fundamental interaction forces in the nucleus. Similarly, the 93.26-keV resonance in ^{67}Zn has been used to determine the magnetic moment of the 12 state (spin = $1/2$, negative parity) in this nuclide and leads to the conclusion that the $1/2^-$ and $5/2^-$ states can be considered minus-quasiparticle states. Mössbauer spectroscopy has also been a potent technique for the study of the electromagnetic moments of nuclei, in particular the magnetic dipole moment and the (mean-square) charge radius of nuclear states. The nuclides ^{191}Ir and ^{193}Ir are typical of those which have been used for detailed studies of nuclear parameters. Such nuclear information is frequently difficult to obtain by non-Mössbauer-effect methods. See NUCLEAR STRUCTURE.

Recoilless gamma-ray resonance experiments have been able to provide detailed information concerning excited-state lifetimes involved in the nuclear decay process. The lifetime values for the nuclides ^{119}Sn , ^{107}Au , and ^{73}Ge , among others, are largely based on Mössbauer effect measurements. See NUCLEAR ISOMERISM.

Recoilless gamma fluorescence spectroscopy has also been used to study the chemical consequences of nuclear decay, and the lifetimes of the Mössbauer transition (typically about 10^{-8} s) provide a convenient time scale to distinguish rapid electronic relaxation processes (typically 10^{-12} to 10^{-14} s) from atomic translation processes (typically slower than 10^{-6} s), and thus study the chemical fate of an atom which results from the decay of a radioactive parent nuclide.

Solid-state physics. Mössbauer effect spectroscopy has made significant contributions to the study of problems in solid-state physics, especially of the nature of the magnetic interactions in iron-containing alloys and the dependence of the magnetic field on composition, temperature, pressure, and other parameters which are of importance in metallurgical processes; solid-state-device fabrication; the structural use of metals and alloys; and numerous related problems of great practical importance. See FERROMAGNETISM.

Combining Mössbauer effect spectroscopy with vibrational spectroscopic studies has led to a clearer understanding of the nature of inter- and intramolecular forces and the relationship of these forces to the properties of polymeric materials. See INTERMOLECULAR FORCES; POLYMER.

It has also been possible, using this technique, to study the effect of high pressure and isotropic compressibility on the chemical properties of materials, especially in the case of experiments with ^{57}Fe , ^{181}Ta , and ^{119}Sn . Such studies have led to the design of high-pressure processes in preparative metallurgy and materials science. See HIGH-PRESSURE CHEMISTRY.

A technique which has also found particular application in metallurgy and the study of catalysts is the use of conversion electron spectroscopy to detect the Mössbauer effect. The advantage of this non-destructive technique is that only the surface layers of the material being investigated are probed by the gamma rays. Thus, in conjunction with the standard transmission type of experimental arrangement shown in Fig. 1, it is possible to differentiate between the atoms on the surface of a material and those in the interior. Such studies frequently can lead to a better understanding of the chemical modification of a surface when solid materials are exposed to a reactive environment as in corrosion and in the poisoning of catalysts. See SURFACE PHYSICS.

At the extremely low end of the temperature scale, Mössbauer effect spectroscopy has been useful in examining the nature of those materials which become superconductive at sufficiently low temperatures and the relationship between chemical composition and structure on the one hand and the superconductive transition on the other, as in the dichalcogen layer compounds $\text{TaS}_2\text{-Sn}$ and $\text{TaS}_2\text{-Sn}_{1/3}$ (both studied using the nuclide ^{119}Sn). Nb_3Sn , a material widely used in the construction of superconducting magnets, has been subjected to detailed Mössbauer effect investigations. See SUPERCONDUCTIVITY.

Structural chemistry. The two Mössbauer nuclides most widely exploited by chemists are ^{57}Fe and ^{119}Sn , although a growing body of data resulting from experiments with ^{129}I , ^{99}Ru , ^{121}Sb , and others has been reported. The position of the resonance maximum δ , also called the isomer or chemical shift, can be related to the systematics of the electron configuration of the atom, and extensive isomer shift correlations for iron- and tin-containing compounds have been tabulated. In particular, the isomer shift of tin compounds is readily related to the oxidation state, since

it has been observed that all stannous (Sn^{2+}) isomer shifts are larger than that observed for metallic tin ($\beta\text{-Sn}$), while those for stannic compounds (Sn^{4+}) are smaller than this value. This observation allows an assignment of oxidation state to be made on the basis of the isomer shift parameter, as in the two-dimensional layer compound SnTa_3S_6 in which the tin atom is clearly identified as a stannous ion, contrary to expectations based on theory. See MOLECULAR ISOMERISM; OXIDATION-REDUCTION.

The use of Mössbauer spectroscopy to identify the charge state of a given chemical species has also made major contributions to the understanding of the phenomena of valence fluctuations, valence instabilities, and mixed valencies in solids. Such studies are of particular importance in rare-earth chemistry, and the use of the nuclides ^{149}Sm , ^{152}Sm , ^{153}Eu , and ^{151}Eu has been particularly fruitful. See VALENCE.

Similarly, the isomer shifts reported for a number of ruthenium compounds, which have been studied using the 89.36-keV resonance in ^{99}Ru , can be correlated systematically with the number of $4d$ electrons involved in the bonding of the metal atom to its nearest-neighbor ligands. Such Mössbauer effect studies have led to a clearer understanding of the nature of "ruthenium red," a trimeric ammonia ruthenium oxide, and a number of other compounds of this relatively rare transition-metal homolog of iron.

In the field of organometallic chemistry, Mössbauer effect spectroscopy has served to clarify the structure of a number of compounds of iron and tin which are of considerable synthetic and industrial importance, including $\text{Fe}_3(\text{CO})_{12}$, $[(\pi\text{C}_5\text{H}_5)_2\text{Fe}(\text{CO})_2]_2$, $[(\text{C}_4\text{H}_9)_3\text{Sn}]_2\text{SO}_4$, and the organotin thioglycolates which are used as stabilizers in the plastics industry. See ORGANOMETALLIC COMPOUND.

Biological science. Many molecules of biological including hemoproteins, iron-sulfur proteins, and iron storage and transport proteins, offer an ideal system in which Mössbauer effect spectroscopy can be used to elucidate the structure and bonding properties of the metal atom in complex systems. The first measurements on such molecules were reported in 1961, and a very large number of iron-containing systems have been studied since then. Paramagnetic iron compounds can be studied at temperatures below the magnetic ordering point (Néel temperature), and it is thus possible, by means of Mössbauer effect spectroscopy, to determine the sign and magnitude of the magnetic field acting on an iron atom in a complex biological material with molecular weights ranging up to 50,000 or more. See PROTEIN.

Success has also been achieved in understanding the magnetic basis by which certain microorganisms can orient themselves in the Earth's magnetic field and thus move in the direction of sediments which contain the nutrients essential to their survival. This magnetic orientation, which is of opposite polarity for organisms living in the Northern and the Southern hemispheres of the Earth, is achieved by the incorporation in the microorganism of particles (mag-

netosomes) which consist of Fe_3O_4 encased in an organic envelope and which function as a compass in the Earth's magnetic flux lines. The new insights in understanding this phenomenon were achieved in part through the application of Mössbauer spectroscopy involving the isotope ^{57}Fe . See MAGNETIC RECEPTION (BIOLOGY).

It is also possible to study antiferromagnetically coupled iron atoms in biological molecules by carrying out the Mössbauer effect measurements in an external magnetic field over a range of temperatures. Typical of such a study is that of oxidized and reduced putidaredoxin, an iron-sulfur protein (molecular weight = 12,500) which acts as a one-electron transfer enzyme. The Mössbauer experiments on this material clearly showed that in the oxidized material the two iron atoms in the molecule occupy chemically equivalent sites. On one electron reduction, one iron atom remains ferric (Fe^{3+}), while the other becomes ferrous (Fe^{2+}), and the two atoms couple antiferromagnetically to give an electronic ground state of $S = 1/2$. Such detailed knowledge of the chemical behavior of the iron atoms in this molecule can elucidate the action of biological catalysts (enzymes) on the molecular level. See ANTIFERROMAGNETISM.

Related fields. Mössbauer effect studies have also played a role in studies in many related fields of science, including archeology, geology, engineering studies, theoretical (relativity) physics, chemical kinetics, and biology. The samples of surface material returned from the Moon by the United States Apollo program have been carefully scrutinized by Mössbauer techniques, as have core samples extracted from deep-drilling experiments on the Earth's outer layer. The geographical distribution of ancient Greek pottery has been traced by making use of characteristic Mössbauer effect data, and the pigments used in painting and decorating have been similarly investigated using this technique.

Rolfe H. Herber

Bibliography. T. E. Cranshaw et al., *Mössbauer Spectroscopy and Its Applications*, 1986; R. S. Drago, *Physical Methods for Chemists*, 2d ed., 1992; D. P. Dickson and F. J. Berry (eds.), *Mössbauer Spectroscopy*, 1987; Yu. M. Kagan and I. S. Lynbutin (eds.), *Applications of the Mössbauer Effect*, 5 vols., 1985; G. J. Long (ed.), *Industrial Applications of the Mössbauer Effect*, 1987; G. J. Long (ed.), *Mössbauer Spectroscopy Applied to Inorganic Chemistry*, 3 vols., 1984, 1987, 1989; G. J. Long and F. Grandjean (eds.), *Mössbauer Spectroscopy Applied to Magnetism and Materials Science*, vol. 1, 1993; S. Mitra, *Applied Mössbauer Spectroscopy: Theory and Practice for Geochemists and Archeologists*, 1993.

Motion

If the position of a material system as measured by a particular observer changes with respect to time, that system is said to be in motion with respect to

the observer. Absolute motion, then, has no significance, and only relative motion may be defined; what one observer measures to be at rest, another observer in a different frame of reference may regard as being in motion. *See* FRAME OF REFERENCE; RELATIVE MOTION.

The time derivatives of the various coordinates used to specify the system may be used to prescribe the motion at any instant of time. How the motion develops in subsequent instants is then determined by the laws of motion. In classical dynamics it is supposed that in principle the motion and configuration of the system may be specified to an arbitrary precision, although in quantum mechanics it is recognized that the measurement of the one disturbs the other.

For a system of f degrees of freedom, the motion may be represented by a point in an f -dimensional velocity space, the coordinates of which are the time rates of change of the coordinates that describe the configuration of the system. For a system under no forces that is described by rectangular cartesian coordinates, these time derivatives are constants. For a single particle, this result is the first of Newton's laws of motion, namely, that a particle remains at rest or in a state of uniform motion unless acted upon by an external force.

The most general theory of motion that has yet been developed is quantum field theory, which combines both quantum mechanics and relativity theory, as well as the experimentally observed fact that elementary particles can be created and annihilated. *See* DEGREE OF FREEDOM (MECHANICS); DYNAMICS; EULER'S EQUATIONS OF MOTION; HAMILTON'S EQUATIONS OF MOTION; HARMONIC MOTION; KINEMATICS; LAGRANGE'S EQUATIONS; NEWTON'S LAWS OF MOTION; OSCILLATION; PERIODIC MOTION; QUANTUM FIELD THEORY; QUANTUM MECHANICS; RECTILINEAR MOTION; RELATIVITY; ROTATIONAL MOTION.

Herbert C. Corben; Bernard Goodman

Motivation

The intentions, desires, goals, and needs that determine human and animal behavior. An inquiry is made into a person's motives in order to explain that person's actions. The answer may have important consequences.

Arousal. Different roles have been assigned to motivational factors in the causation of behavior. Historically, some have defined motivation as a nonspecific energizing of all behavior. Others define it as recruiting and directing behavior, selecting which of many possible actions the organism will perform. Whether there are different states corresponding to different motives or a single state of being motivated is open to question. The likely answer is that both aspects exist. More specific determinants of action may be superimposed on a dimension of activation or arousal that affects a variety of actions nonselectively (**Fig. 1**). The situation determines what the animal does; arousal level affects the vigor, prompt-

ness, or persistence with which the animal does it. Many have suggested that behavioral arousal may reflect activity in the reticular formation and related structures in the brainstem, which may alert to prime higher information-processing systems. Finally, there may be an optimal, intermediate level of arousal. One may be too lethargic for effective action if arousal is too low, too frenzied and frantic if it is too high.

Homeostasis and drive. The notion of arousal is not incompatible with the possibility of separate systems for different motives. A specific "hunger drive" could activate feeding or food-seeking activities, with this effect superimposed upon a more global background of greater or lesser arousal. *See* HUNGER.

Early drive theorists saw motivated behavior as adjunct to physiological mechanisms of homeostasis, that is, the mechanisms by which the body regulates internal variables such as temperature, blood sugar level, and the volume and concentration of body fluids. Faced with dehydration, a land-dwelling animal may slow down its rate of water loss by excreting a more concentrated urine (physiological), but is also becomes disposed to seek and drink water (behavioral). Depleted of energy reserves, an animal may reduce its metabolic rate and so waste fewer calories (physiological). It also becomes disposed to seek and ingest food (behavioral). Thus, motivated behavior forms part of a negative-feedback loop, an arrangement characteristic of regulatory systems.

The physiologist Walter Cannon, the learning theorist Clark Hull, and the clinician Sigmund Freud each proposed variations on homeostasis. As a general theory of motivation, however, the homeostatic model faces difficulties. First, not all "basic biological drives" work this way. For example, there is no evidence of any physiological imbalance that builds up in the absence of sexual activity and corrected by it. In particular, the level of sex hormones in the blood is not reduced by sexual activity and may even be increased by it.

Second, motivated behavior can be influenced by external as well as internal factors. The smell of broiling steak may stimulate hunger, and words and pictures are used to promote sexual arousal. In the rat, provision of specially attractive food can trigger a bout of feeding in the short term, and even produce dietary obesity in the long term; and a male rat's flagging sexual interest may be revived if a new partner is offered. Since these external influences are not coupled with the animal's internal state, they can lead to behavior that does not promote homeostasis and may even threaten it, as in dietary obesity.

To add to the complexity, internal and external factors are not independent and additive; rather they interact with each other. A sodium-deficient rat (internal) becomes hyperresponsive to the taste of salt (external) even if the ingested salt does not relieve the deficiency. The urine of a female guinea pig contains odorous substances (external) that evoke intense interest in intact males but not in castrated males (internal). In such cases, internal influences

affect behavior by setting the animal's responsiveness to certain external signals. The interaction occurs in the opposite direction as well: external signals can affect internal state. Thus, in ring doves the mere sight of a male's courtship display (external) can affect ovarian secretions in the female (internal).

Third, especially in humans, vigorous and persistent goal-directed behavior can occur in the absence of any physiological need. For example, one may strive for success at a task or in one's profession, or to be a good provider, a good student, or a good homemaker. The attempt to explain these motives as "derived from" homeostatic ones has lost its appeal. See HOMEOSTASIS.

The brain. These multiple influences must be registered and translated into behavior by the nervous system. It has been suggested that various neuron clusters in the hypothalamus could be command centers, responsible for the excitation or inhibition of the various drive states.

The theory is no longer tenable in its original form. Expression of any action involves systems, not centers, in which multiple levels of the brain cooperate. It remains true, however, that motivational sequences can be initiated from within the hypothalamus. Thus microinjections of norepinephrine into the hypothalamus can trigger hunger in otherwise satiated rats. In rodents (but not in humans), female sexual behavior is promptly and totally abolished by spaying; but it can be restored by implants of tiny amounts of hormones into the hypothalamus. Finally, the more general idea that motivated behavior reflects a balance of excitatory and inhibitory influences is widely accepted.

Cognition. There is another reason for the homeostatic drive concept's loss of popularity. Even relatively simple motives can be influenced by much more than the existing internal and external situation. They respond to potential or expected factors, as registered by cognitive apparatus.

Drives, anticipations, and representations. Even relatively simple motives such as hunger and thirst are responsive to cognitive factors. A rat may learn how much water it is going to need, and adjust its intake to anticipated rather than present requirements. It can learn which of two flavored fluids is going to supply a large caloric load, and drink less of that fluid in anticipation of this. In humans, an image or fantasy of an attractive conspecific can produce the experience, as well the physical signs, of sexual excitement. Such behaviors imply that an animal is capable of representing anticipated or potential states of affairs, as well as perceiving existing ones. The forward-looking or goal-directed nature of behavior may depend on such capabilities. See THIRST AND SODIUM APPETITE.

Representations and goal-directed behavior. To a hungry rat, food becomes a goal. The rat will make various responses, including arbitrary learned ones or operants, that lead to contact with food. A rat can be trained to press a lever for food reinforcement, but it could just as easily be trained to run down an alley or

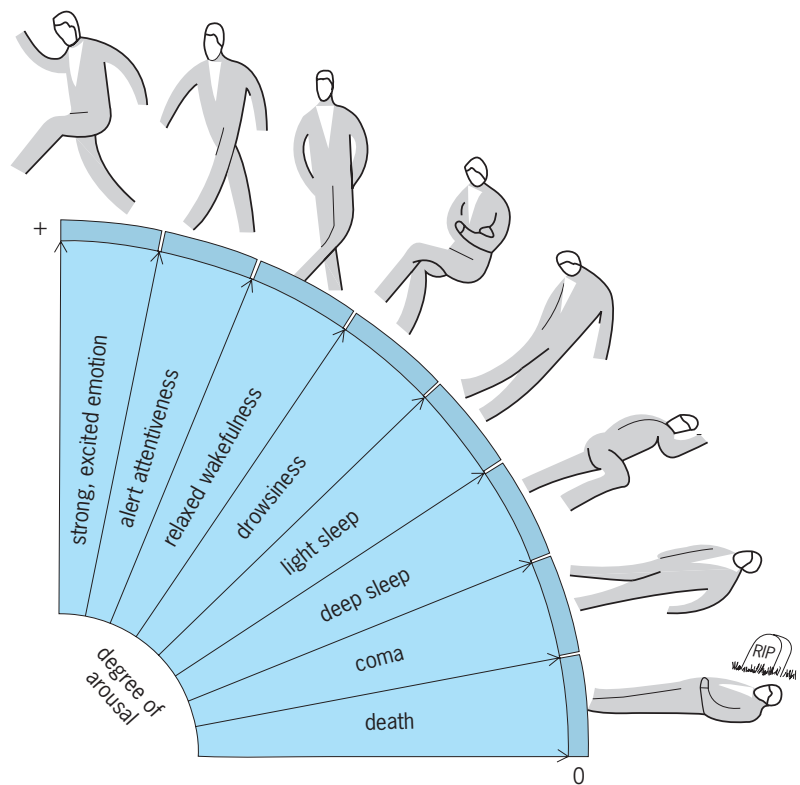


Fig. 1. Diagram of the continuum of arousal.

dance around the cage or do whatever else is necessary (within its capabilities) to attain its goal. It is this flexibility of goal-directed behavior that justifies the concept of motivation. If an animal will do whatever is necessary to obtain food, it must want food. Internal factors then may act by setting the goal status of environmental commodities: the effect of hunger is to make food a goal.

There is a question as to how behavior can be guided by a state or event (goal attainment) that does not yet exist. Modern approaches to this question lean heavily on cognitive concepts. Mammals, birds, and even some insects can represent to themselves a nonexistent state of affairs. They can represent what a goal object is (search images): a chimpanzee may show behavioral signs of surprise if a different food is substituted for the usual one. They can represent where it is (cognitive maps): a digger wasp remembers the location of its nest relative to arbitrary landmarks, and will fly to the wrong place if the landmarks are moved.

If this idea is generalized, motivated behavior can be thought of as guided by a feedback control system with a set point (Fig. 2). A set point, such as a thermostat setting, establishes a goal state which the control system seeks to bring about. Behavior is controlled, not by present external or internal stimuli alone, but by a comparison between the existing state of affairs (registered by the sensory system) and a desired state of affairs, that is, the set point or goal, registered or specified within the brain. The animal then acts to reduce the difference between the existing and the desired state of affairs.

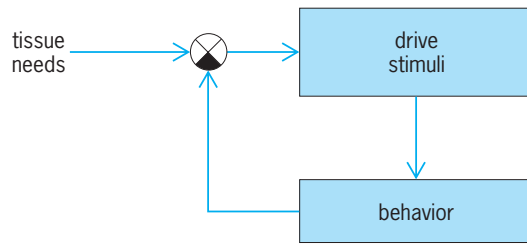


Fig. 2. Motivation as part of a negative-feedback system. Behavior is motivated by drives, which it acts to reduce or remove.

Complex representations in humans. This way of looking at motivation helps bridge the gap between simple motives in animals and complex ones in humans. If to be motivated is to do whatever is necessary to bring about an imagined state of affairs, then human motives can literally be as complex, and be projected as far into the future, as human imaginations permit. Individuals may seek such abstract goals as equitable return for effort, or personal power, or achievement. See COGNITION.

Instinctive behavior. Another approach to motivation comes from the work of ethologists, that is, students of behavior in its natural environment. Historically, ethologists tended to emphasize the species-typical, or instinctive, repertoires of their subjects. These were often described as stereotyped action patterns, responsive not to their consequences (like operants) but to their antecedents: sign or releasing stimuli.

However, old distinctions have broken down, and ethology has formed links with cognitive psychology. The broken-wing display of the piping plover provides an example. If a predator approaches a nest with eggs, the parent bird may behave as if injured (hence easy prey) and thus lead the intruder away from the nest. This action pattern is characteristic of the species and unlearned in its gross topography; yet the bird monitors the intruder's behavior and modulates the display accordingly. It may approach more closely and intensify the display if the intruder is not at first diverted from its path. Thus a species-typical action pattern can be used in ways suggestive of purpose and goal direction: the bird modifies it as necessary to promote the goal of diverting the intruder. See ETHOLOGY.

Emotion. Motivation and emotion are closely related. Indeed, it has been argued that emotions are the true motivators and that other factors internal, situational, and cognitive take hold of behavior by way of the emotions they evoke. In the simplest case, pleasure and displeasure have been recognized for centuries as having motivational force. If a nondeprived rat eats substantially when offered palatable food, this may simply mean that the food tastes good. In more complex cases, the role of cognitive operations, such as how an individual feels about an event, as well as what is done about it, can depend heavily on how an individual thinks about it.

A further suggestion is that different persons may develop different explanatory styles by which to make causal attributions. A person may habitually

explain unfortunate events as resulting from his or her own (internal, stable, uncontrollable) inadequacies, whereas good events may be explained as resulting from luck or chance. Such a pessimistic and self-denigrating explanatory style may play a role in generating the emotional and behavioral symptoms of depression.

Culture. The culture in which an individual is raised has a powerful effect on how the individual behaves. It has been argued that culture teaches its members what to believe are the consequences of a specific action (cognitive), and how the individuals should feel about those consequences or about the actions themselves (emotional/motivational). Douglas G. Mook

Bibliography. R. D'Andrade and C. Strauss (eds.), *Human Motivation and Cultural Models*, 1993; D. G. Mook, *Motivation: The Organization of Action*, 2d ed., 1997; D. Pfaff (ed.), *Physiological Mechanisms of Motivation*, 1982; C. A. Ristau (ed.), *Cognitive Ethology*, 1991; B. Weiner, *Human Motivation*, 1989.

Motor

A machine that converts electrical into mechanical energy. Motors that develop rotational mechanical motion are most common, but linear motors are also used. A rotary motor delivers mechanical power by means of a rotating shaft extending from one or both ends of its enclosure. Such a motor is shown in **Fig. 1**, cut away to show the internal parts. The shaft is attached internally to the rotor. Shaft bearings permit the rotor to turn freely. The rotor is mounted coaxially with the stationary part, or stator, of the motor. The small space between the rotor and stator is called the air gap, even though fluids other than air may fill this gap in certain applications. **Figure 2** illustrates the relationship between the rotor, stator, air gap, and shaft.

Linear motors may have a stationary part that forms an elongated track, along which the movable part carries loads. Alternatively, the stator may be the shorter of the two members, and the movable part may act as a continuous conveyor. Linear motors operate on the same physical principles as rotary alternating-current motors.

In a motor, practically all of the electromechanical energy conversion takes place in the air gap. Although a few electrostatic motors have been built, they require such prohibitively high voltages that all commercial motors employ magnetic fields as the energy link between the electrical input and the mechanical output. The air-gap magnetic field is set up by current-carrying windings located in the rotor or the stator, or by a combination of windings and permanent magnets. The magnetic field exerts forces between the rotor and stator to produce the mechanical shaft torque; at the same time, in accord with Faraday's law, the magnetic field induces voltages in the windings. The voltage induced in the winding connected to the electrical energy source is often called

a countervoltage because it is in opposition to the source voltage. By its magnitude and, in the case of alternating-current (ac) motors, its phase angle, the countervoltage controls the flow of current into the motor's electrical terminals and hence the electrical power input. The physical phenomena underlying motor operation are such that the power input is adjusted automatically to meet the requirements of the mechanical load on the shaft. See ELECTROMAGNETIC INDUCTION; MAGNET; MAGNETISM; WINDINGS IN ELECTRIC MACHINERY.

Construction. Both the rotor and stator have a cylindrical core of ferromagnetic material, usually steel. The parts of the core that are subjected to alternating magnetic flux are built up of thin sheet steel laminations that are electrically insulated from each other to impede the flow of eddy currents, which would otherwise greatly reduce motor efficiency. The windings consist of coils of insulated copper or aluminum wire or, in some cases, heavy, rigid insulated conductors. The coils may be placed around pole pieces, called salient poles, projecting into the air gap from one of the cores, or they may be embedded in radial slots cut into the core surface facing the air gap. In a slotted core, the core material remaining between the slots is in the form of teeth, which should not be confused with magnetic poles. A disassembled universal motor, in Fig. 3, illustrates several of these features. See EDDY CURRENT.

Direct-current (dc) motors usually have salient poles on the stator and slotted rotors. Polyphase ac synchronous motors may have salient poles on the rotor and slotted stators. Rotors and stators are both slotted in induction motors. Permanent magnets may be inserted into salient pole pieces, or they may be cemented to the core surface to form the salient poles.

The windings and permanent magnets produce magnetic poles on the rotor and stator surfaces facing each other across the air gap. If a motor is to develop torque, the number of rotor poles must equal the number of stator poles, and this number must be even because the poles on either member must alternate in polarity (north, south, north, south) circularly around the air gap. If there are four magnetic poles on the rotor, there are also four on the stator, and the motor is said to be a four-pole motor. A high-speed motor has fewer poles than a low-speed motor.

Classification of rotary motors. Motors are classified according to power rating as subfractional horsepower, fractional horsepower, or integral horsepower. They are also classified by application as general purpose, definite purpose, and special purpose. In the United States, definitions of these terms and detailed standards for dimensions and operating characteristics are set by the National Electrical Manufacturers Association and the Small Motors Manufacturers Association. The table classifies electric motors in common use according to operating principles; it does not include control motors.

Alternating-current motors. Electric motors operating directly from polyphase (usually three-phase) lines

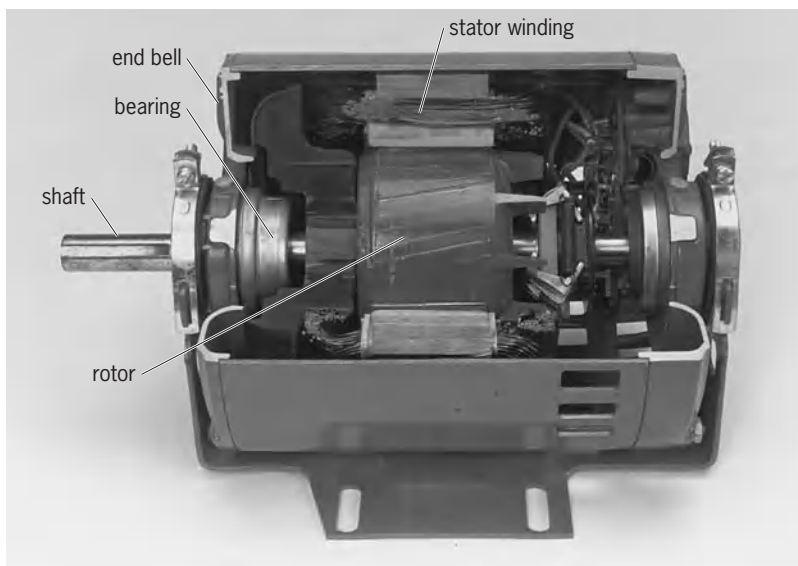


Fig. 1. Cutaway view of a single-phase induction motor. (Emerson Motor Division)

include synchronous motors and induction motors. The speed of synchronous motors is determined precisely by the supply frequency, the speed in revolutions per minute (rpm) being given by $120f/p$, where

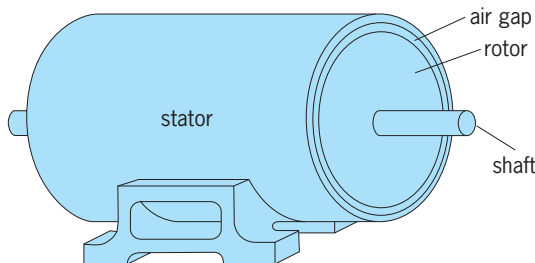


Fig. 2. Elements of an electric motor. (After G. McPherson and R. D. Laramore, *An Introduction to Electrical Machines and Transformers*, 2d ed., John Wiley and Sons, 1989)

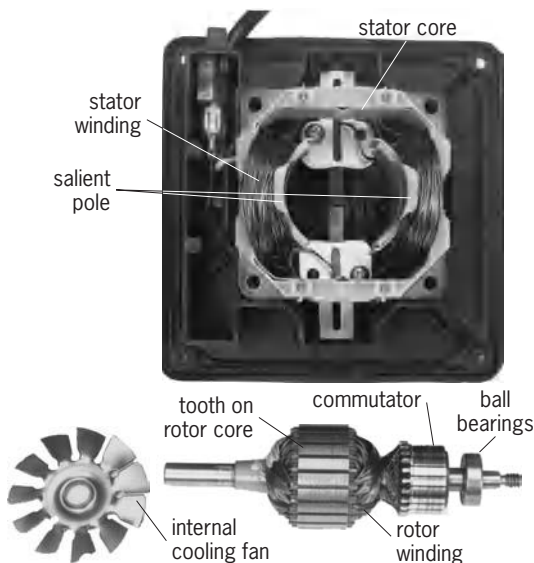


Fig. 3. Disassembled universal motor. (Emerson Motor Division)

| Varieties of electric motors in common use | | |
|---|---|---|
| Alternating-current motors | | |
| Polyphase | Single phase | Direct-current motors |
| Synchronous Induction Wound rotor Squirrel cage Single cage Low resistance (design A) High resistance (design D) Deep bar (design B) Multicage (design C) | Synchronous Hysteresis Reluctance Universal (ac or dc) ←————→ Induction Shaded pole Stepped pole Split phase Capacitor start Two-value capacitor Permanent-split (PSC) capacitor | PMDC (permanent magnet poles) Wound poles Shunt Series Compound |

f is the frequency in hertz and p is the number of poles of the motor. Polyphase induction motors operate at speeds slightly less than those calculated by this formula. See SLIP (ELECTRICITY).

Single-phase induction motors have no inherent starting torque and are classified according to the means used to get them started. These classifications include resistance-split-phase (usually simply called split-phase) motors, capacitor-start motors, permanent-split capacitor (PSC) motors, two-value capacitor motors, shaded-pole motors, and stepped-pole motors. Single-phase synchronous motors include reluctance motors and hysteresis motors. See ALTERNATING-CURRENT MOTOR; HYSTERESIS MOTOR; INDUCTION MOTOR; RELUCTANCE MOTOR; SYNCHRONOUS MOTOR.

Direct-current motors. Direct-current motors are classified according to the means used to magnetize (excite) their salient field poles. Those employing permanent magnets for this purpose are called PMDC machines. Those having wound poles are called shunt, series, or compound motors, depending on whether their field-pole windings are connected in parallel (shunt = parallel) or in series with the armature, or whether each pole has two windings, one involved in shunt connection and the other in series. A brushless dc motor is a polyphase synchronous motor connected to a dc source through an electronic inverter that is controlled by rotor position sensors in such a way that the system behaves exactly as a shunt dc motor. See DIRECT-CURRENT MOTOR.

Universal motors. Universal motors are series dc motors so designed as to make them capable of operating also on single-phase alternating current. They achieve speeds above those possible with synchronous and induction motors, and these high speeds produce high horsepower ratings for very compact motors. See UNIVERSAL MOTOR.

Control motors. Motors to be used for precise control of motions are designed to have very low inertia. They include stepper motors, permanent magnet motors, two-phase induction motors, and brushless dc motors. See SERVOMECHANISM; STEPPING MOTOR.

Torque development. The windings in motors are largely shielded from the air-gap magnetic field by

the steel pole pieces and core teeth, and very little force is exerted by the field on the conductors themselves. Motor torque is best calculated by application of the principle of virtual work to the field energy in the air gap. Under the conditions in the air gap, this principle states that torque is produced if a small rotation θ of the rotor and its magnetic poles results in a change in the energy stored in the air-gap field. If W_{fld} is the energy of the air-gap field, the torque τ is given by Eq. (1). The torque is in newton-meters

$$\tau = \frac{\partial W_{fld}}{\partial \theta} \tag{1}$$

(N · m) [or pound-feet (lbf · ft)] if W_{fld} is in joules (J) [or foot-pounds (ft · lbf)] and θ in mechanical radians. At least two torque-producing phenomena are evident from Eq. (1). First, torque results from the attraction of rotor and stator poles of unlike polarity, because their alignment would increase the flux per pole and thus increase the field energy. Second, when salient poles are involved, a differential rotation may change the field energy by changing the magnetic reluctance to the flow of flux across the gap. The second phenomenon is the basis for the operation of reluctance and stepping motors and is called reluctance torque. See ROTATIONAL MOTION; TORQUE; VIRTUAL WORK PRINCIPLE.

Energy considerations allow formulas to be developed to determine the torque for different motor configurations. Typical is that for a motor with a cylindrical air gap having a diameter D and a length L , and a magnetic flux density that varies sinusoidally with angle around the gap. The formula is written in terms of the specific electric loading (SEL) and the specific magnetic loading (SML). The specific electric loading in this case is the number of conductors of the stator winding per meter of distance measured around the inside surface of the stator, multiplied by the root-mean-square (rms) winding current. The specific magnetic loading is the rms value of the air-gap magnetic flux density. The torque formula (2) applies directly to polyphase induction and

$$\tau = \frac{\pi}{2} D^2 L (SEL)(SML) k_w \sin \delta_{SR} \tag{2}$$

cylindrical-rotor synchronous motors, where δ_{SR} is the angular displacement between the poles of magnetomotive force developed by the stator winding and the poles of the air-gap flux field, and k_w is a winding factor. The factor D^2L shows that the torque capability of a motor is proportional to its volume. The torque is independent of the number of poles and hence independent of the rated speed. Thus, for a fixed volume, increasing the number of poles in a motor reduces its horsepower rating because the rated speed is reduced with no change in torque. A similar result is available for dc motors. See ELECTRIC ROTATING MACHINERY.

George McPherson, Jr.

Speed Control of AC Motors

The polyphase synchronous motor is a constant-speed (synchronous), variable-torque, doubly excited machine. The stator is supplied with polyphase alternating current of a specific frequency (line frequency), and the rotor field is supplied with direct current. The rotor speed of the synchronous motor is a direct function of the number of stator and rotor field poles and the frequency of the alternating current applied to the stator. Since the number of rotor poles of the polyphase synchronous motor is not easily modified, the change-of-frequency method is the only way to control the synchronous speed of the motor.

The polyphase induction motor is also a doubly excited machine, whose stator is supplied with polyphase alternating current of line frequency and whose rotor is supplied with induced alternating current whose frequency depends on the rotor slip. The speed of the polyphase induction motor is asynchronous, and can be varied by one or more of the following methods: (1) by changing the applied frequency to the stator (the same method as used for the synchronous motor); (2) by controlling the rotor slip by means of rotor resistance control (used for wound-rotor induction motors); (3) by changing the number of poles of both stator and rotor; and (4) by "external voltage" control, obtained by conductively or inductively introducing applied voltages of the proper frequency to the rotor.

A number of electromechanical or purely mechanical methods also provide speed control of alternating-current motors. One is referred to as the Rossman drive, in which an induction motor stator is mounted on trunnion bearings and driven with an auxiliary motor, providing the desired change in slip between stator and rotor. Polyphase induction and synchronous motors having essentially constant speed characteristics at rated voltage are also assembled in packaged drive units employing gears, cylindrical and conical pulleys, and even hydraulic pumps to produce a variable speed output. In addition to reversal of rotation, some of these units employ magnetic slip clutches and solenoids to control the various mechanical and hydraulic arrangements through which a relatively smooth control of speed is achieved.

The principal method of speed control used for fractional-horsepower single-phase induction-type, shaded-pole reluctance, series, and universal motors is the method of primary line-voltage control. It involves a reduction in line voltage applied to the stator winding (of the induction type) or to the armature of series and universal motors. In the former, this produces a reduction of torque and increase in rotor slip. In the latter, it is simply a means of controlling speed by armature voltage control or field flux control or both. The reduction in line voltage is usually accomplished by one of six methods: auto-transformer control, series reactance control, tapped main-winding control, saturable reactor (or magnetic amplifier) control, silicon controlled rectifier feedback control, and variable-speed drives. See AUTO-TRANSFORMER; SATURABLE REACTOR.

Electronic control techniques. The development of power electronic switching devices such as the thyristor, or silicon controlled rectifier (SCR), created unlimited possibilities for control of virtually all types of motors (single-phase, direct-current, and polyphase). Silicon controlled rectifiers in sizes up to 1600 A (rms) with voltage ratings up to 1600 V are available. See RECTIFIER; SEMICONDUCTOR RECTIFIER.

SCR control of series motors. Fractional-horse-power universal, alternating-current, and direct-current series motors may be speed-controlled from a single-phase 110- or 220-V supply using the half-wave circuit involving a diode D_1 and silicon controlled rectifier (Fig. 4a). The trigger point of the silicon controlled rectifier is adjusted via the potentiometer R_v , which phase-shifts the gate turn-on voltage of the silicon controlled rectifier whenever its anode A is positive with respect to the cathode K . During the negative half-cycle of input voltage, the silicon controlled rectifier conduction is off and the gating pulse is blocked by the diode D_1 . When the positive half-cycle is initiated once again, the capacitor C charges to provide the required gating pulse at the time preset by the time constant R_vC .

Replacing the diode D_1 by a DIAC and the silicon controlled rectifier by a TRIAC converts the circuit from half-wave to full-wave operation and control (Fig. 4b). Compared with Fig. 4a, this circuit provides almost twice the torque and improved speed regulation. Neither circuit, however, is capable of automatic maintenance of desired speed because of the inherently poor speed regulation of series-type alternating-current, universal, or direct-current motors, with application or variation of motor loading.

Automatic speed regulation is achieved by using the half-wave feedback circuit shown in Fig. 4c. This circuit requires, however, that both field leads (f_1, f_2) and both armature leads (A_1, A_2) are separately brought out for connection as shown in Fig. 4c. The desired speed is set by the potentiometer R_v , in terms of reference voltage V_d across the diode D_1 . The actual motor speed is sensed in terms of the armature voltage V_a . Diode D_1 conducts only when the reference voltage V_d (desired speed) exceeds the

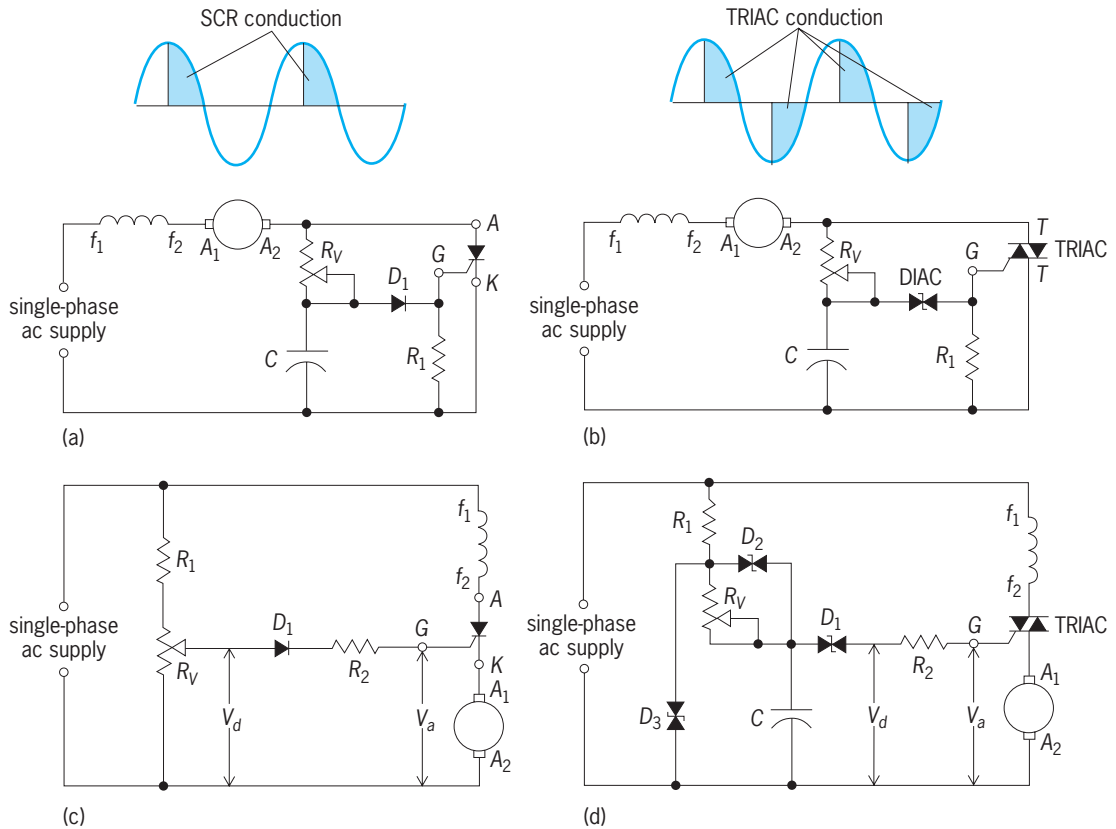


Fig. 4. Single-phase, universal or dc motor control using half- and full-wave (ac) controls, with and without voltage feedback. (a) Half-wave circuit. (b) Full-wave circuit. (c) Half-wave circuit with voltage feedback. (d) Full-wave circuit with voltage feedback. (After I. L. Kosow, *Control of Electric Machines*, 1973)

voltage V_a to restore the motor speed to its desired setting. Thus, when the motor load is increased and the speed drops, the silicon controlled rectifier is gated earlier in the cycle because the diode D_1 conducts whenever V_d exceeds V_a .

As with the previous half-wave circuit, the feedback circuit may be converted to full-wave operation by replacing the diode D_1 with a DIAC and the silicon controlled rectifier with a TRIAC (Fig. 4d). This circuit also is shown with additional minor modifications to improve the regulation of speed at lower speeds: the addition of the DIACs D_2 and D_3 and the capacitor C . The advent of high-power TRIACs and silicon controlled rectifiers has also extended the use of the circuit shown in Fig. 4d to larger, integral-horse-power alternating-current series motors.

When direct-current motors are operated by using the electronic techniques shown in Fig. 4, it is customary to derate motors proportionately because of the heating effect created by alternating-current components in both half- and full-wave waveforms. There is no reversal of direction in either the alternating-current or direct-current series motor under full-wave operation because both the armature and field currents have been reversed.

SCR control of polyphase ac motors. There are fundamentally only three types of polyphase alternating-current motors: synchronous motors (SMs), squirrel-cage induction motors (SCIMs), and wound-rotor induction motors (WRIMs). All three employ identical

stator constructions. As a result, larger polyphase motors (up to 10,000 hp or 7500 kW) are controlled by silicon controlled rectifier packages which employ some so-called universal method of speed control. In these methods, both the frequency and stator voltages are varied (in the same proportion) to maintain constant polyphase stator flux densities and avoid operating the motor in the saturation region, thus avoiding overheating.

Two major classes of solid-state adjustable voltage age/frequency drives have emerged, namely the cycloconverter (an ac/ac package) and the rectifier-inverter (an ac-dc-ac package). Both packages convert three-phase fixed-frequency alternating current to three-phase variable voltage/variable frequency.

The half-wave cycloconverter (Fig. 5a) is capable of supplying from 0 to 20 Hz to an alternating-current polyphase motor. It uses six silicon controlled rectifiers per phase but is incapable of either phase-sequence reversal or frequencies above 20 Hz. The full-wave converter (Fig. 5b) uses twice the number of silicon-controlled rectifiers (12 per phase) and possesses advantages of wider frequency variation (from +30 Hz down to 0 and up to -30 Hz), potentialities for dynamic braking, and capability of power regeneration.

The symbol for a polyphase squirrel-cage induction motor driven by a cycloconverter is shown in Fig. 5c, and a synchronous motor in Fig. 5d. The symbols imply frequency/voltage control of identical

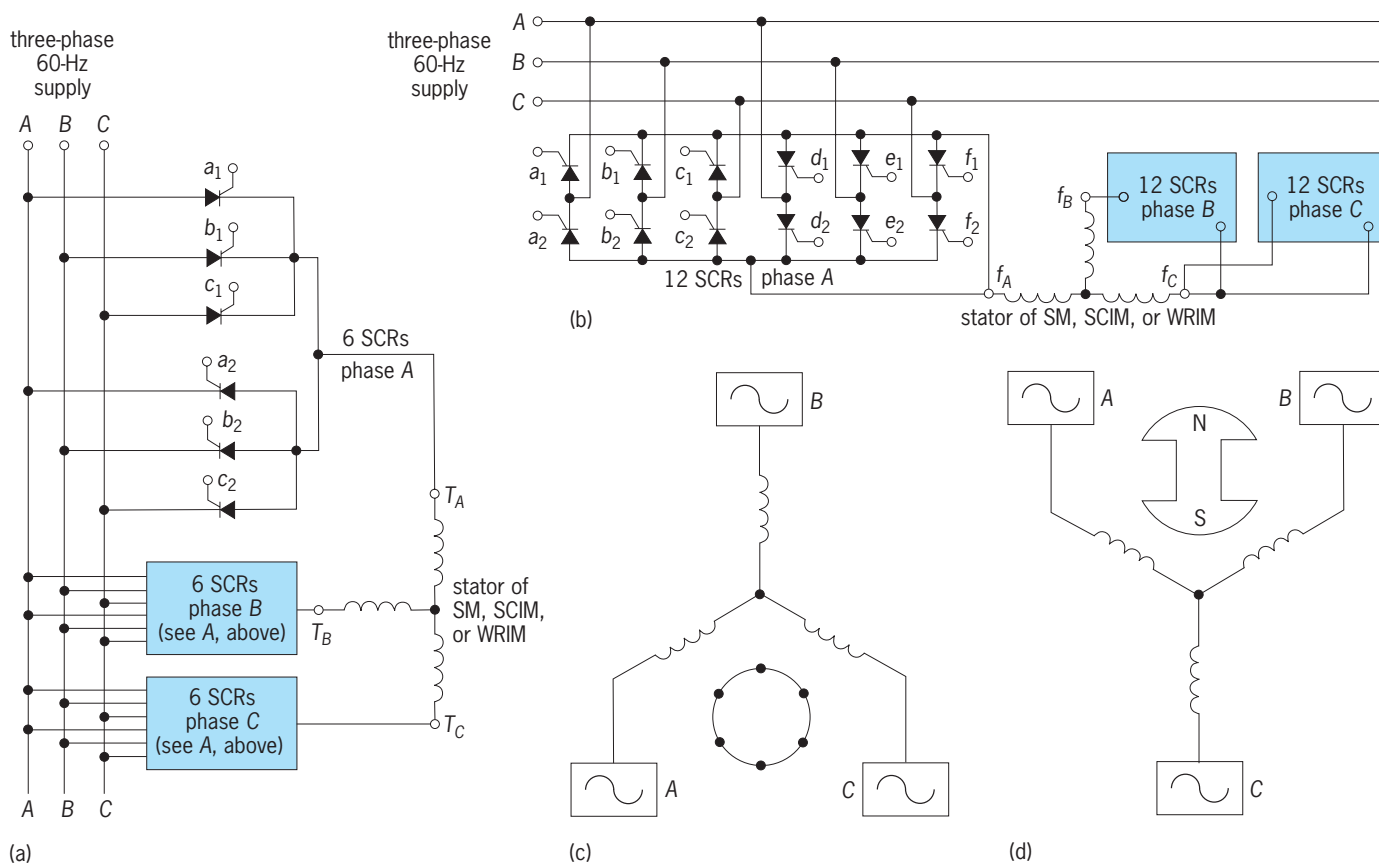


Fig. 5. Half-wave and full-wave cycloconverters with graphical symbols for simplified representation. (a) Half-wave cycloconverter (exclusive of triggering of silicon controlled rectifier gates). (b) Full-wave cycloconverter (exclusive of silicon controlled rectifier gating). (c) Symbol for squirrel-cage induction motor driven by a cycloconverter. (d) Symbol for synchronous motor driven by a cycloconverter. (After I. L. Kosow, *Control of Electric Machines*, 1973)

stators, and the nature of the motor is determined solely by its rotor.

The second class of solid-state drive package is the rectifier-inverter, which converts fixed-frequency polyphase alternating current to direct current (variable voltage). The direct current is then inverted by a three-phase inverter (using a minimum of 12 silicon controlled rectifiers commercially) to produce variable-voltage, variable-frequency three-phase alternating current for application to the motor stator. The block diagram of a solid-state rectifier-inverter package is shown in Fig. 6a. The three-phase 60-Hz supply is first rectified to produce variable direct current (Fig. 6b) by appropriate phase shifts of silicon controlled rectifier gates. The variable direct current is then applied to the direct-current bus of Fig. 6a. Inversion is accomplished by appropriate phase shift of gates of the 12 silicon controlled rectifiers to produce phase and line voltages displaced, respectively, by 120°. This three-phase output voltage is applied to a transformer whose secondary is applied to the stator of the motor (Fig. 6c). The rectifier-inverter is capable of power regeneration through the feedback loop shown in Fig. 6a.

Brush-shifting motors. By use of a regulating winding, commutator, and a brush-shifting device, the speed of a polyphase induction motor can be controlled similarly to that of a direct-current shunt

motor. Such motors are used for knitting and spinning machines, paper mills, and other industrial services that require controlled variable-speed drive. The primary winding on the rotor is supplied from the line through slip rings. The stator windings are the secondary windings, and the third winding, also in the rotor, is an adjusting winding provided with a commutator. Voltages collected from the commutator are fed into the secondary circuit. Three brushes are mounted 120 electrical degrees apart on a movable yoke, and three other brushes are similarly mounted on a separate movable yoke. Each set of brushes can be moved as a group. Thus both the spacing between sets of brushes and the angular position of the brushes are adjustable. Brush spacing determines the magnitude of the voltage applied to the secondary. When brush sets are so adjusted that pairs of brushes are in contact with the same commutator segment, the secondary is short-circuited and no voltage is supplied. Under these conditions the motor behaves as an ordinary induction motor. The speed can be reduced by separating the brushes so that secondary current produces a negative torque. The machine can be operated above synchronism by interchanging the position of the brushes, so the voltage collected is in a direction to produce a positive torque. The motor can be reversed by reversing two of the leads supplying the primary.

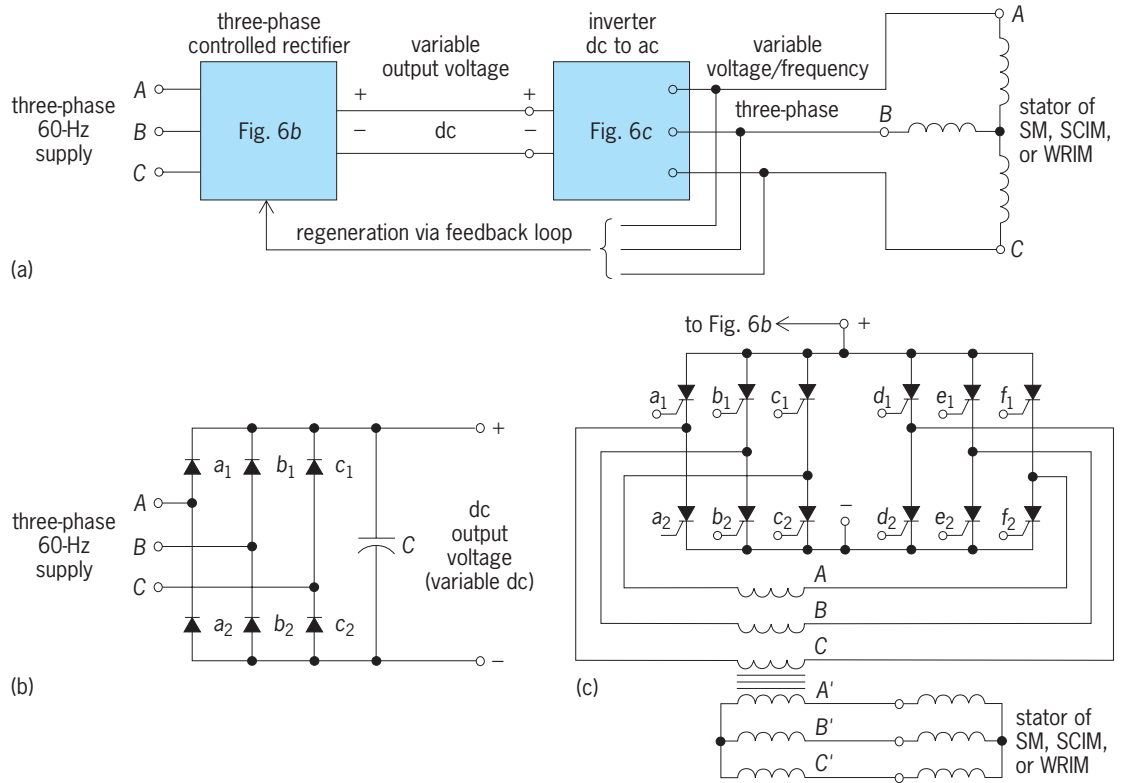


Fig. 6. Rectifier-inverter circuitry for variable voltage/frequency control of three-phase motors. (a) Block diagram of rectifier-inverter. (b) Three-phase full-wave controlled rectifier. (c) Basic inverter dc to three-phase ac for phase-shift voltage control. (After I. L. Kosow, *Control of Electric Machines*, 1973)

Closed-loop control techniques. As mentioned above, slip control of the speed of a wound-rotor induction motor can be achieved by manually controlling an external resistance in series with the rotor winding. This method of speed control has the disad-

vantages of poor speed regulation at any given resistance setting, reduced motor efficiency due to power lost in the external resistors, and the requirement of a human operator to manually adjust the motor to a desired speed. This method of speed control falls

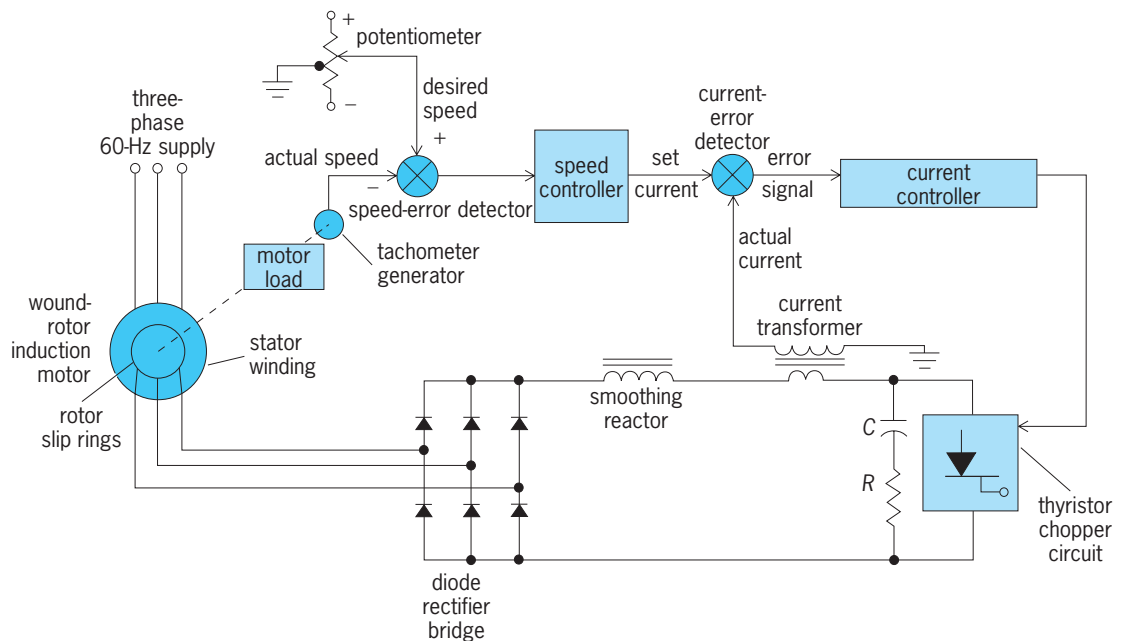


Fig. 7. Speed control of wound-rotor induction motor by static variation of external rotor resistance, using a high-frequency thyristor-controlled chopper circuit.

under the general category of open-loop control.

Thyristor-controlled chopper circuit. A modern method of closed-loop control (Fig. 7) uses a high-frequency thyristor-controlled chopper circuit that permits the external rotor resistance to be varied steplessly by using static semiconductor elements. The feedback circuit shown provides a relatively inexpensive variable-speed drive with good dynamic response, a wider range of speed control, and improved speed regulation. See CONTROL SYSTEMS.

The desired speed in Fig. 7 is set by means of a potentiometer. The speed-error detector senses the difference between the actual rotor speed and the desired (set) speed, which in turn provides the "set speed" signal to a speed controller. The latter provides a "set current" signal to a second error detector that controls the thyristor chopper and rotor slip power. The thyristor chopper, which parallels the RC branch, is switched (on and off) at a frequency of approximately 1 kHz. The ratio of on/off time controls the effective rotor resistance and consequently the motor speed by modification of the torque-speed characteristic.

Cascade drive. The circuit of Fig. 7 uses a single closed-loop method of automatic control of a wound-rotor induction motor. A somewhat more sophisticated method of automatic closed-loop control is the so-called cascade drive system (Fig. 8), which uses two closed loops. The inner control loop uses a variation in the thyristor firing angle to control and adjust the rotor current in the cascade circuit. The rotor current actually determines the motor torque and speed, in part. The outer or speed loop senses the rotor current (on the alternating-current side of the thyristor inverter) and uses it as feedback to set the desired motor speed via a speed-error detector and current-error detector, respectively. In effect, the dual closed-loop system permits the speed-error signal to produce the necessary motor torque and rotor current to reduce the error between the desired (set) speed and the actual rotor speed, regardless of changes in motor load. This method of control provides speed regulation of less than

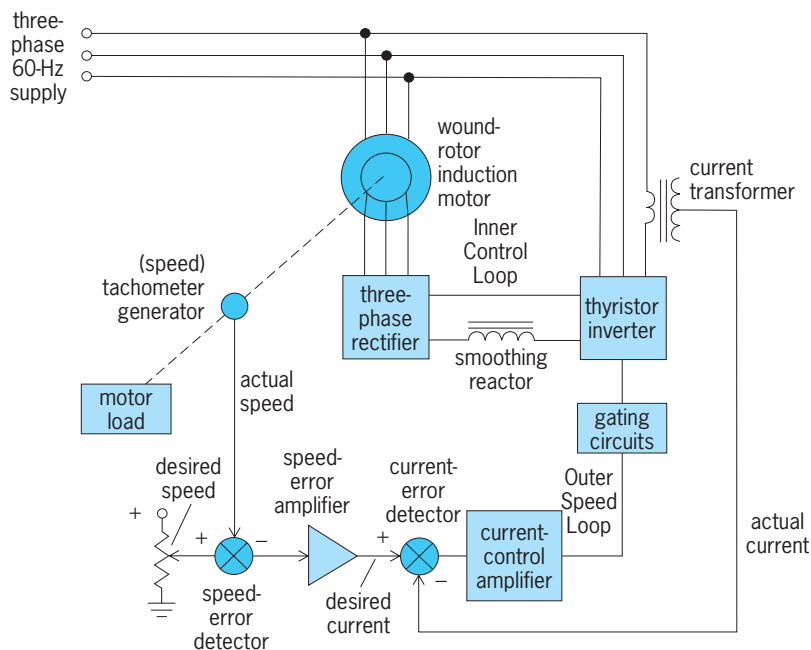


Fig. 8. Closed-loop control of a wound-rotor induction motor by a static cascade drive system.

0.1% from no-load to full load at any desired speed setting.

Current-source inverter. The basic voltage-source inverter (VSI) of Fig. 6c is modified in Fig. 9 to provide a current-source inverter (CSI). Unlike the voltage-source inverter, which uses a wide variety of switching devices and commutation arrangements, the current-source inverter uses the conventional and preferred arrangement shown in Fig. 9. The name current-source inverter stems from the fact that the current to the motor stator is automatically and sequentially commutated from one full-wave thyristor to the next, after each simple gate signal to the oncoming thyristor. The capacitors are selected in conjunction with the single inductor to ensure good commutation. Unlike the voltage-source

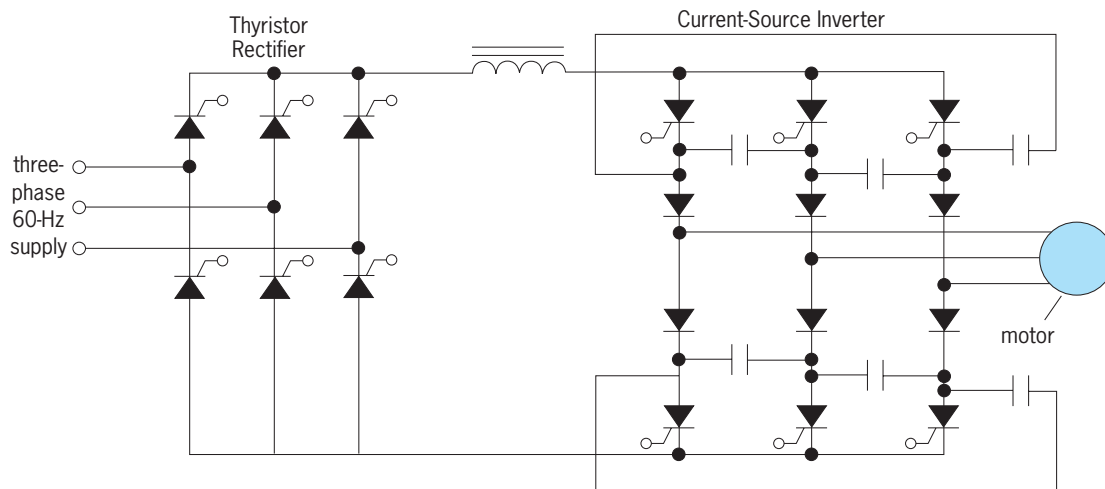


Fig. 9. Full-wave thyristor rectifier and autosequentially commutated current source inverter.

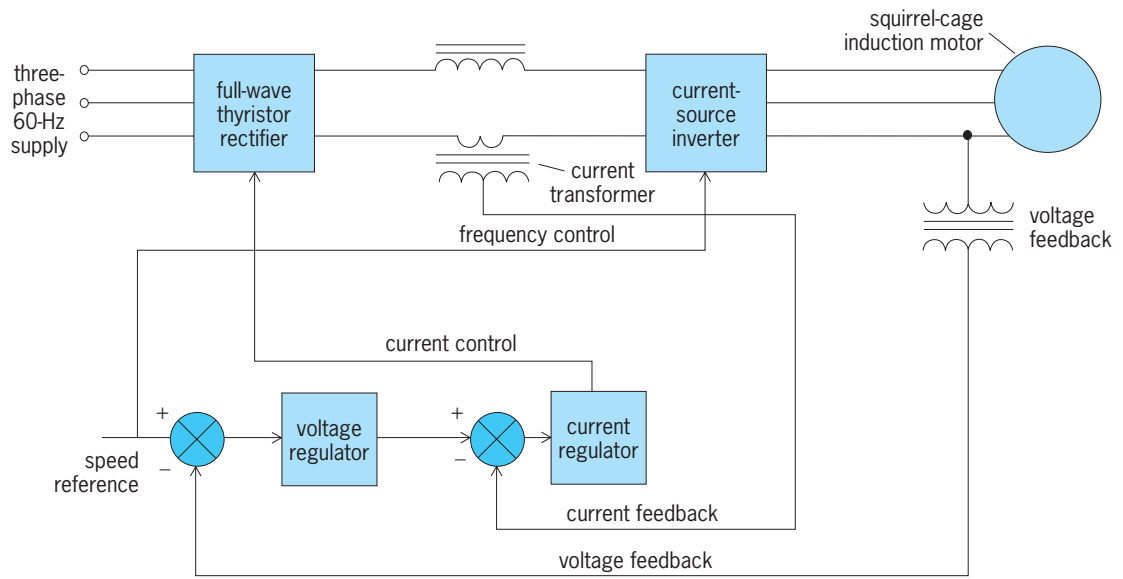


Fig. 10. Closed-loop control of a squirrel-cage induction motor using a current-source inverter with a current-regulating loop and a voltage-regulating loop.

inverter shown in Fig. 6c, which provides open-loop control of the stator voltage source, the current-source inverter requires feedback control to permit its use as an adjustable speed drive for a squirrel-cage induction motor or wound-rotor induction motor.

The simplest application of a current-source inverter, providing both current and voltage feedback, is shown in Fig. 10. The outer voltage feedback loop senses the stator terminal voltage as an indication of motor speed. The error signal to the voltage regulator serves as a reference to the current-regulator error detector. The current regulator controls the direct-current supply of the thyristor phase-controlled rectifier.

In effect, the circuit of Fig. 10 converts the current-source inverter to an equivalent feedback-controlled voltage source. The advantages of the current-source inverter over conventional pulse-width modulator drives or voltage-source inverter drives are found in three areas: motor harmonics, regeneration, and fault currents. In the current-source inverter, unlike other drives mentioned above, the inverter determines the harmonic currents, which are always proportional to the motor line current at the fundamental frequency. Regeneration is much simpler with the current-source inverter than other drives because only the link voltage reverses in regeneration, rather than the link current, eliminating the need for a dual or other bidirectional input converter. The large link inductor tends to oppose any rapid current changes, making fault protection much simpler than in voltage-source inversion or pulse-width modulation systems, which have large link capacitors.

Irving L. Kosow

Speed Control of DC Motors

Compared to an alternating-current induction motor, the direct-current motor is more expensive and complex. However, the direct-current motor has the ad-

vantage of having the capability to develop a wide variety of torque-speed characteristics. Direct-current motors find applications in electric traction and in industrial drives requiring large speed variations in both directions of rotation, or requiring a precise variation of speed over a narrow range. In other words, the direct-current motor is highly controllable.

Speed equation. The direct-current motor has two accessible electric circuits, the field circuit and the armature circuit (Fig. 11), which may be used to control the motor speed and other characteristics. The motor steady-state speed n is given by Eq. (3),

$$n = \frac{V_t - I_a(R_a + R_{ax})}{\phi(I_f)} \quad (3)$$

where V_t is the terminal voltage across the armature, I_a is the armature current, R_a is the armature winding resistance, R_{ax} is the external resistance in the armature circuit, and ϕ is some function of the field current I_f . If the magnetic circuit of the motor is not saturated, Eq. (3) may be written as Eq. (4), where k is a constant.

$$n = \frac{V_t - I_a(R_a + R_{ax})}{kI_f} \quad (4)$$

Armature and field control. From Eqs. (3) and (4), it follows that the motor speed may be controlled by varying the terminal voltage V_t , varying the field current I_f by changing the field-circuit external resistance R_{fx} , and varying the armature-circuit external resistance R_{ax} . The method of varying the terminal voltage is known as the armature control, and the method of varying the field-circuit external resistance is called the field control. The method of varying the armature-circuit external resistance is rather inefficient (or lossy) and is suitable for very small motors. From Fig. 11, it is clear that the field current I_f is limited by the field-circuit resistance, $R_f + R_{fx}$. Thus,

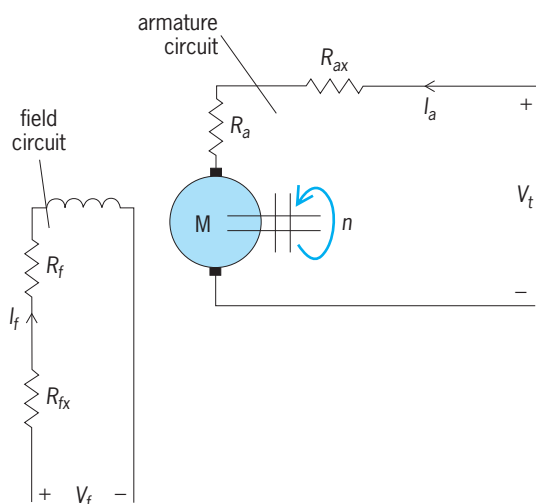


Fig. 11. Field and armature circuits of a dc motor. M = motor; V_t = terminal voltage across the armature; I_a = armature current; R_a = armature winding resistance; R_{ax} = external resistance in armature circuit; V_f = voltage across the field; I_f = field current; R_f = field winding resistance; R_{fx} = external resistance in field circuit; n = speed.

I_f cannot exceed a value determined by R_f given by Eq. (5), where V_f is the voltage across the field.

$$(I_f)_{\max} = \frac{V_f}{R_f} \quad (5)$$

Of course, I_f can be made to decrease below $(I_f)_{\max}$ by increasing the external resistance R_{fx} inserted in the field circuit. Because the speed of the motor depends inversely on the field current, as governed by Eq. (3), for a given voltage V_f , field control can be used only to increase the motor speed beyond the minimum determined by $(I_f)_{\max}$. From Eq. (3), however, it follows that armature control, achieved by varying V_t , may be used to either increase or decrease the motor speed from a given reference. Prior to the advent of power semiconductors, armature control was used in special cases.

Ward-Leonard system. The methods of armature control and field control have been combined to obtain the Ward-Leonard system. In this system a generator driven by an auxiliary motor supplies the motor to be controlled. By varying the generator field-circuit resistance and the motor field-circuit resistance, this system gives a wide range of speed control in both directions of rotation of the motor. The Ward-Leonard system has found applications in mine hoists and rolling mills to drive and reverse large and varying loads. The major drawback of the system is the capital cost of a set of three machines of almost equal ratings, and the space occupied by the motor-generator set. See MOTOR-GENERATOR SET.

Solid-state control. Resistance control, either in the field or armature of a direct-current motor, results in poor efficiency. High-power solid-state controllers offer the most practical, reliable, and efficient method of motor control. The most commonly used solid-state devices in motor control are power

transistors and thyristors or silicon controlled rectifiers. The principal differences between the two are that the transistor requires a continuous driving signal during conduction, whereas the thyristor requires only a pulse to initiate conduction; and the transistor switches off when the driving signal is removed, but the thyristor turns off when the load current is reduced to zero. Application of a reverse-polarity voltage shortens the recovery time to full reverse-blocking capability. Utilizing power semiconductors, choppers and converters are the most common electronic controllers employed in direct-current drives. See TRANSISTOR.

Choppers. In principle, a chopper is an on-off switch connecting the load to and disconnecting it from the battery (or direct-current source), thus producing a chopped voltage across the load. Symbolically, a chopper as a switch is represented in Fig. 12a, and a basic chopper circuit is shown in Fig. 12b. In this circuit, when the thyristor does not conduct, the load current flows through the freewheeling diode D , which provides a path for the armature current when the silicon controlled rectifier is not conducting. From Fig. 12a it is clear that the average voltage across the load, V_0 , is given by Eqs. (6), where V is

$$V_0 = \frac{t_{\text{on}}}{t_{\text{on}} + t_{\text{off}}} \quad V = \frac{t_{\text{on}}}{T} \quad V = \alpha V \quad (6)$$

the battery voltage, the various times are shown in Fig. 12a, T is known as the chopping period, and $\alpha = t_{\text{on}}/T$ is called the duty cycle. Thus the voltage across the load varies with the duty cycle.

There are three ways in which the chopper output voltage can be varied, and these are illustrated in Fig. 12c-e. The first method, in which the chopping frequency is kept constant and the pulse width (or on-time t_{on}) is varied (Fig. 12c), is known as pulse-width modulation. In the second method, called frequency modulation, either t_{on} or t_{off} is fixed, and the chopping period is variable (Fig. 12d). The preceding two methods can be combined to obtain pulse-width and frequency modulation (Fig. 12e), which is used in current limit control. In a method involving frequency modulation, the frequency must not be decreased to a value that may cause a pulsating effect or a discontinuous armature current; and the frequency should not be increased to such a high value as to result in excessive switching losses. The switching frequency of most choppers for electric drives range from 100 to 1000 pulses per second. The drawback of a high-frequency chopper is that current interruption is accompanied by high-frequency noise. See MODULATION.

A simplified form of a chopper circuit is shown in Fig. 12f, where the chopper is shown to supply a direct-current motor (having the field winding in series with the armature). The circuit is a pulse-width modulation circuit, where t_{on} and t_{off} determine the average voltage across the motor, as given by Eq. (6). As mentioned above, the silicon controlled rectifier cannot turn itself off once it begins to conduct. Thus, turning the silicon controlled rectifier off requires a commutating circuit that

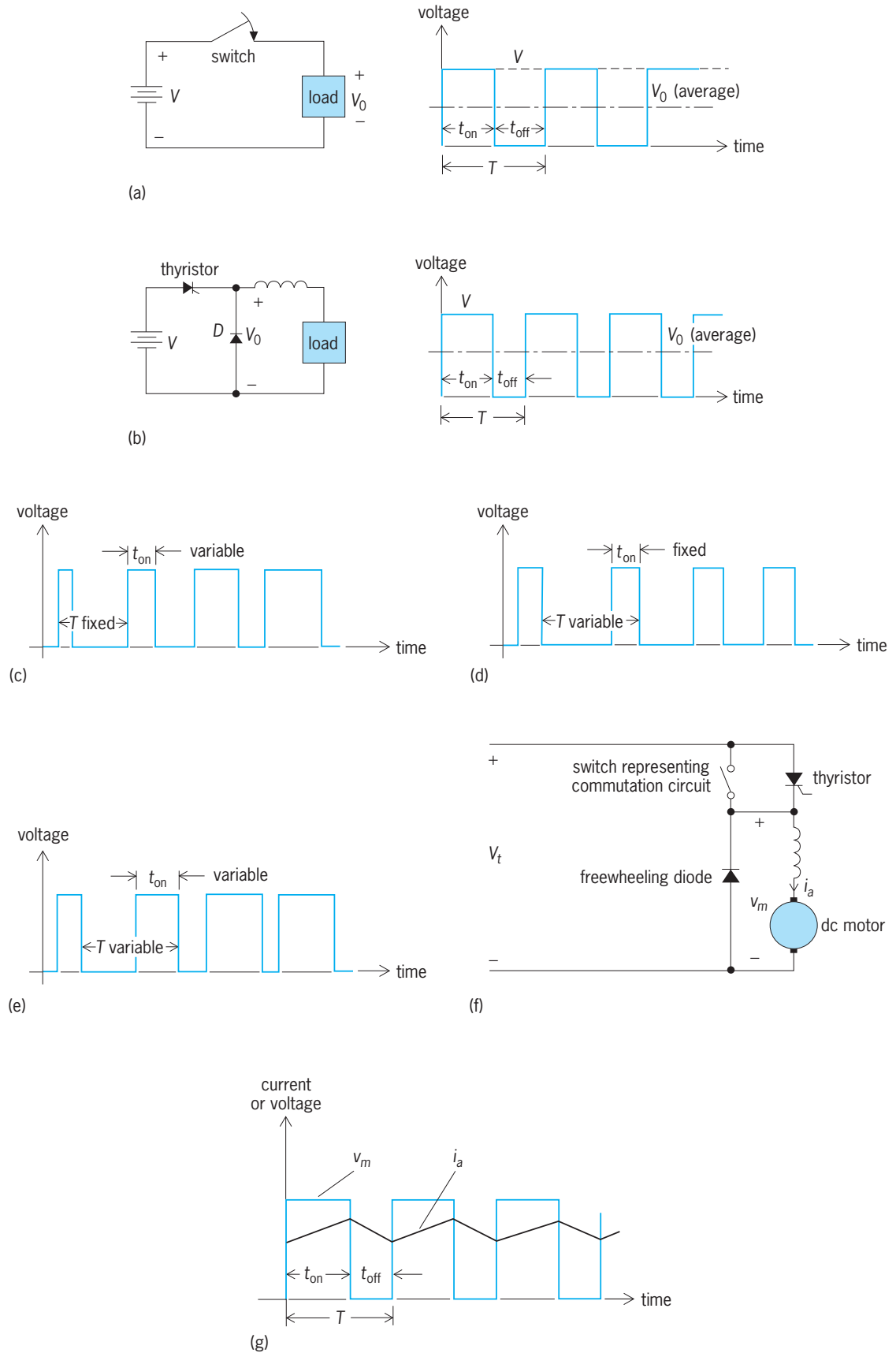


Fig. 12. Chopper control of dc motors. (a) Symbolic representation of a chopper. (b) Basic chopper circuit. (c) Constant-frequency, variable-pulse-width waveform. (d) Variable-frequency, constant-pulse-width waveform. (e) Variable-frequency, variable-pulse-width waveform. (f) A dc motor driven by a chopper. (g) Motor current and voltage waveforms. V_t = applied voltage; v_m = motor voltage; i_a = armature current. Lowercase letters denote instantaneous quantities.

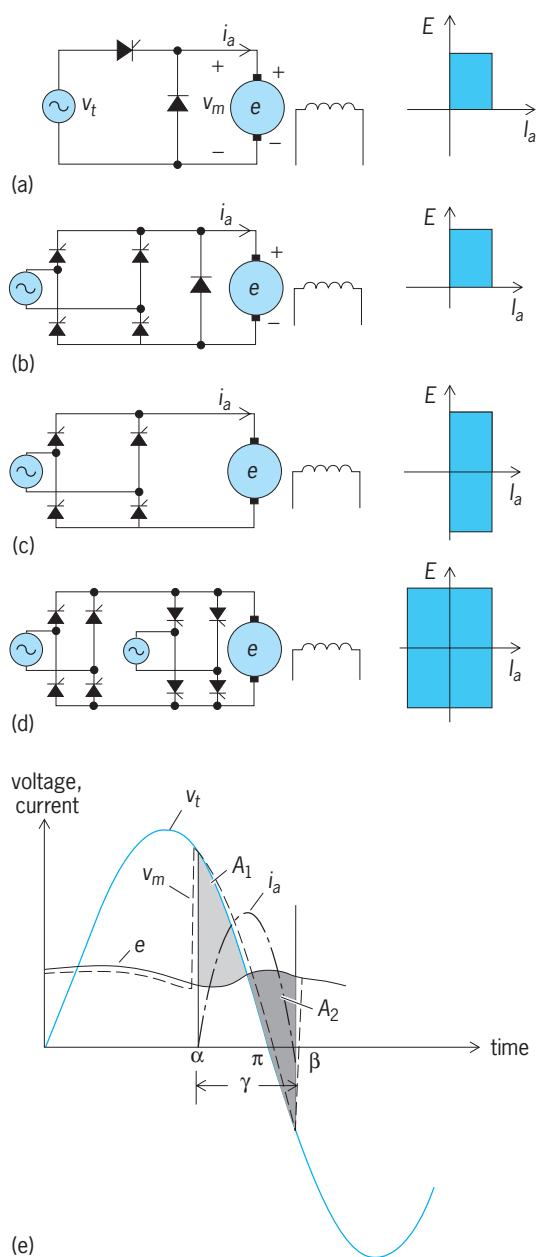


Fig. 13. Single-phase converters and their quadrant operations. (a) Half-wave converter. (b) Semiconverter. (c) Full converter, (d) Dual converter. (e) Waveforms in a dc motor controlled by a single-phase half-wave converter. E = voltage at dc terminals; i_a = current at dc terminals; v_m = motor voltage; v_t = applied voltage; i_a = armature current; e = motor back electromotive force (emf). Lowercase letters denote instantaneous quantities.

impresses a negative voltage on the silicon controlled rectifier for a very short time (of the order of microseconds). For simplicity, the circuitry for commutation, which is often quite involved, is denoted by a switch in Fig. 12f. The motor current and voltage waveforms are shown in Fig. 12g. The silicon controlled rectifier is turned on by a gating signal at $t = 0$. The armature current i_a builds up, and the motor starts and picks up speed. After the time t_{on} , the silicon controlled rectifier is turned off, and remains off for a period t_{off} . During this period the armature current continues to drop through the freewheel-

ing diode circuit. Again, at the end of t_{off} the silicon controlled rectifier is turned on, and the on-off cycle continues. The chopper thus acts as a variable-voltage source.

Converters. A converter changes an alternating-current input voltage to a controllable direct-current voltage. Converters have an advantage over choppers in that natural or line commutation is possible in converters; therefore, no complex commutation circuitry is required. Controlled converters use silicon controlled rectifiers and operate either on single-phase or three-phase alternating current. The four types of phase-controlled converters commonly used for direct-current motor control are half-wave converters, semiconverters, full converters, and dual converters. Each of these could be either a single-phase or a three-phase converter. Half-wave converters are used for motors of ratings up to 0.5 hp (0.4 kW). Semiconverters are one-quadrant converters in that the polarities of voltage and current at the direct-current terminals do not reverse. In a full converter, the polarity of the voltage can reverse, but the current is unidirectional. In this sense, a full converter is a two-quadrant converter. Dual converters are four-quadrant converters. The four types of single-phase converters and their respective quadrant operations are shown in Fig. 13a-d.

The simplest converter, the single-phase half-wave converter, will be used to demonstrate the operation of a converter-fed direct-current motor. The waveforms of motor speed, current, and voltage are given in Fig. 13e, if the armature resistance is neglected. By controlling the firing angle of the silicon controlled rectifier, the voltage across the motor armature, and hence its speed, is controlled. In Fig. 13e, α is the firing angle for the silicon controlled rectifier. However, the armature current i_a does not begin to flow immediately after the silicon controlled rectifier is turned on. The armature current begins to flow only when the line voltage v_t becomes greater than the motor voltage v_m . The current continues to flow for the period γ shown in Fig. 13e. This period is also known as the conduction angle, and is determined by the equality of the shaded areas A_1 and A_2 . The area A_1 corresponds to the energy stored in the inductance L of the motor circuit while the armature current is building up. This stored energy is returned to the source during the period that the armature current decreases and ultimately becomes zero. The silicon controlled rectifier then blocks until it is turned on again. The motor speed is controlled by the firing angle α . See CONVERTER. Syed A. Nasar

Bibliography. B. J. Chalmers (ed.), *Electric Motor Handbook*, 1988; S. B. Dewan, G. R. Slemon, and A. Straughen, *Power Semiconductor Drives*, 1984; D. G. Fink and H. W. Beaty (eds.), *Standard Handbook for Electrical Engineers*, 14th ed., 2000; A. E. Fitzgerald, C. Kingsley, and S. D. Umans, *Electric Machinery*, 6th ed., 2003; I. L. Kosow, *Electric Machinery and Transformers*, 2d ed., 1991; G. McPherson and R. D. Laramore, *An Introduction to Electrical Machines and Transformers*, 2d ed., 1990; S. A. Nasar, *Electric Machines and Power*

Systems, I: Electric Machines, 1995; S. A. Nasar and I. Boldea, *Linear Electric Motors*, 1987; National Electrical Manufacturers Association, *Motors and Generators*, Publ. MG1-2003, 2003; P. C. Sen, *Thyristor DC Drives*, 1981, reprint 1991; H. A. Toliyat and G. B. Kliman (eds.), *Handbook of Electric Motors*, 2d ed., 2004.

Motor-generator set

A motor and one or more generators, with their shafts mechanically coupled, used to convert the voltage or frequency of an available power source to another desired frequency or voltage. The motor of the set is selected so that it operates from the available power supply; the generators are designed to provide the desired output voltage or frequency. Motor-generator sets are also employed to provide special control features for the output voltage.

The principal advantage of a motor-generator set over other conversion systems is the flexibility offered by the use of separate machines for each function. Assemblies of standard machines may often be employed with a minimum of additional engineering required. Since energy conversion is employed twice, electrical to mechanical and back to electrical, the efficiency of this system is lower than that in most other conversion methods. In a two-unit set the efficiency is the product of the efficiencies of the motor and of the generator.

Motor-generator sets are used for a variety of purposes, such as providing a precisely regulated dc current for a welding application, a high-frequency ac power for an induction-heating application, or a continuously and rapidly adjustable dc voltage to the armature of a dc motor employed in a position control system. See GENERATOR; MOTOR. Arthur R. Eckels

Motor systems

Those portions of nervous systems that regulate and control the contractile activity of muscle and the secretory activity of glands. Muscles and glands are the two types of organ by which an organism reacts to its environment; together they constitute the machinery of behavior. Cardiac muscle and some smooth muscle and glandular structures can function independently of the nervous system but in a poorly coordinated fashion. Skeletal muscle activity, however, is entirely dependent on neural control. Destruction of the nerves supplying skeletal muscles results in paralysis, or inability to move. The somatic motor system includes those regions of the central nervous system involved in controlling the contraction of skeletal muscles in a manner appropriate to environmental conditions and internal states. This article discusses skeletal muscle innervation and contraction; the motor unit; and neural centers for motor control, particularly the motor cortex. See GLAND; MUSCLE; MUSCULAR SYSTEM.

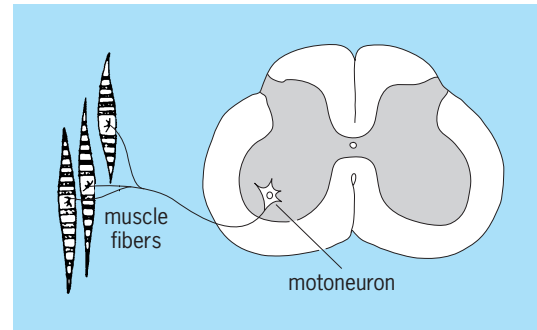


Fig. 1. Diagrammatic representation of a single motor unit. A motoneuron in the ventral horn of the spinal cord innervates multiple muscle fibers (typically more than the three shown).

Skeletal Muscle

A skeletal or striated muscle consists of a bundle of individual contractile elements called muscle fibers. The fibers are held together in the muscle by connective tissue, and their ends are attached via tendons to movable bones; thus, muscle shortening during active contraction results in movement. The nerve supply to skeletal muscles of the limbs and trunk is derived from large nerve cells called motoneurons, whose cell bodies are located in the ventral horn of the spinal cord (**Fig. 1**). Muscles of the face and head are innervated by motoneurons in the brainstem. The axons of the motoneurons traverse the ventral spinal roots (or the appropriate cranial nerve roots) and reach the muscles via peripheral nerve trunks. In the muscle, the axon of every motoneuron divides repeatedly into many terminal branches, each of which innervates a single muscle fiber. The region of innervation, called the neuromuscular junction, or motor end plate, is a secure synaptic contact between the motoneuron terminal and the muscle fiber membrane (**Fig. 2**).

Skeletal muscle contraction. A fixed sequence of events leads from motoneuron activity to contraction of skeletal muscle. An action potential initiated in a motoneuron propagates over the motoneuron axon into its many terminals. At each motor end plate, the action potential causes release of a neuromuscular transmitter agent, which depolarizes the postsynaptic membrane sufficiently to initiate an action potential in the muscle fiber membrane. This muscle fiber action potential propagates along the fiber at a speed of about 5 m/s and triggers the contractile process. See BIOPOTENTIALS AND IONIC CURRENTS.

The neuromuscular transmitter agent is acetylcholine, which is synthesized by the nerve terminals and is contained in tiny vesicles seen in electron micrographs. When the terminal is depolarized, some vesicles rupture at the surface, releasing acetylcholine into the synaptic gap. The acetylcholine diffuses across the gap and combines with receptors in the postsynaptic membrane. This combination opens channels in the membrane, and they conduct ions across the membrane, resulting in its depolarization. The presynaptic action potential liberates enough acetylcholine to depolarize the

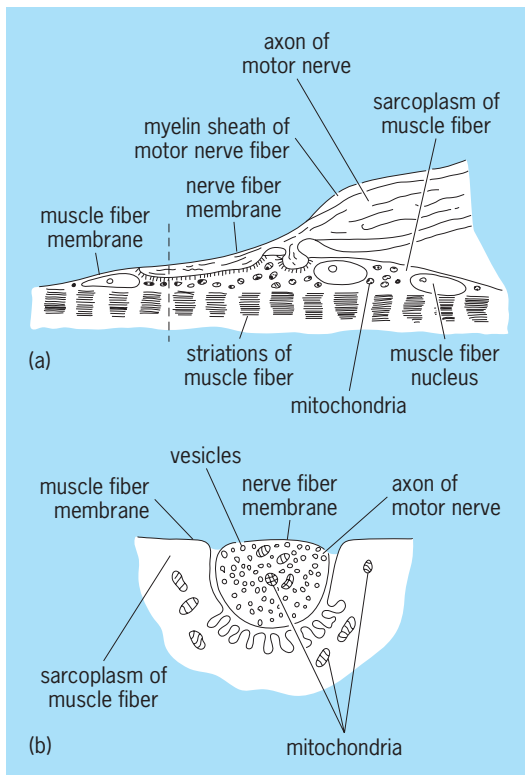


Fig. 2. Diagrams of the neuromuscular junction, or end plate. (a) Section cut parallel to long axis of muscle fiber, as seen by light microscopy. (b) Section cut perpendicular to long axis of muscle fiber (as through broken line in a), as seen by electron microscopy.

muscle membrane past the threshold for initiating an action potential. The receptors in the specialized muscle membrane of the end plate make it about 1000 times more sensitive to acetylcholine than nonjunctional regions of the muscle membrane. See ACETYLCHOLINE; SYNAPTIC TRANSMISSION.

The contractile elements in a muscle fiber consist of two proteins, actin and myosin, present in a ratio of about 1:3. The combination, called actomyosin, forms long molecular chains. In the presence of adenosine triphosphate (ATP), actomyosin filaments contract. Each muscle fiber contains up to 10 million such myofilaments of actomyosin. Electron microscopic studies of muscle show that the actin and myosin elements lie parallel to one another and that during contraction the interdigitated elements slide past each other, thus shortening the muscle. See ADENOSINE TRIPHOSPHATE (ATP); MUSCLE PROTEINS.

Summation of contraction. The mechanical tension generated by a single contraction of a muscle fiber is much more prolonged than the electrical action potential that initiates it. The muscle fiber action potential lasts only a few milliseconds, but the rise and fall of the resulting twitch tension can take over 100 ms (Fig. 3). The muscle action potential, like the action potential of neurons, is an all-or-none phenomenon; its amplitude cannot be modified by changing the strength of the stimulus eliciting it. The contractile mechanism, however, is graded in strength, and the

tension produced by a single twitch contraction is less than the maximal possible tension of the fiber. Because the contractile process greatly outlasts the duration of the muscle action potential, the tension elicited by two action potentials briefly separated in time is greater than that produced by a single twitch. The twitch contraction produced by a second action potential sums with that persisting from the first. Several muscle action potentials initiated in rapid succession produce temporal summation of the successive contractile responses. However, the increment of tension of each contraction diminishes as the net tension increases, until a maximum tension, called the tetanic tension, is reached. In some muscles the tetanic tension is nearly four times greater than the twitch tension. The rate of discharge necessary to produce maximal tetanic tension varies in different muscles, from 30 discharges per second for the slow soleus muscle to 350 discharges per second for the rapidly contracting internal rectus muscle of the eye.

Motor units. Since synaptic transmission at the neuromuscular junction is so secure, an action potential in the motoneuron will produce contraction of every muscle fiber that it contacts. For this reason, the motoneuron and all the fibers it innervates form a functional unit, called the motor unit. The number of muscle fibers in a single motor unit may be as small as 6 (for intrinsic eye muscles) or over 700 (for motor units of large limb muscles). For a given muscle the average number of muscle fibers per motor unit can be determined by dividing the total number of fibers in the muscle by the number of motoneurons in its pool. The typical innervation ratio for limb muscles is on the order of 120–150. In general, muscles involved in delicate rapid movements have fewer muscle fibers per motor unit than large muscles concerned with gross movements.

Physiological properties. Even within a given muscle, different motor units have significantly different physiological properties; some motor units are specialized for steady, prolonged contraction, while others are used for brief periods of high tension. The properties of motor units that underlie their functional specialization include the biochemical makeup of their muscle fibers, the speed and magnitude of their twitch tension, their response to repetitive activation, the size of their motoneuron, and physiological constraints on their relative recruitment. Detailed studies on hindlimb muscles of the cat indicate that these properties tend to vary systematically together. The spectrum of motor unit properties may be understood in terms of two contrasting types (Fig. 3). On one end of the functional spectrum are numerous “slow” type-S motor units, so called because their twitch tension takes longer to reach a peak (80–100 ms) and to decay than the more rapid twitch tension of “fast” type-F motor units at the opposite end of the spectrum. The maximum twitch tension of slow units (for example, 1 gram for gastrocnemius muscle) is smaller than the peak tension of fast units (about 20 g). When stimulated repetitively for prolonged periods of time, the tetanic

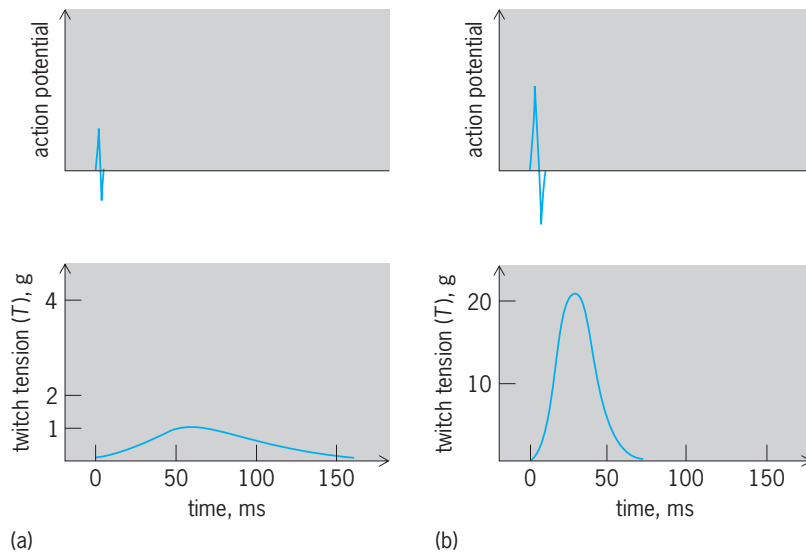


Fig. 3. Electromyographic action potentials and mechanical twitch tensions of two motor units. Type S (slow twitch) on left and type F (fast twitch) on right.

tension of the type-S units is sustained for much longer—on the order of hours—than the tetanic tension of the larger type-F units, which fatigue within minutes. These differences in mechanical performance are related to different metabolic properties of their muscle fibers. As a result, slow motor units are best at providing prolonged periods of tension, whereas fast units are specialized for brief bursts of intense activity.

Motoneuron recruitment. These functional differences are further correlated with differences in the recruitment of their motoneurons. The type-S motor units are generally recruited at lower thresholds in a variety of motor responses than the type-F units. The motoneurons supplying slow motor units tend to have smaller soma size than the large motoneurons of the type-F motor units. In most cases, the motoneurons of a pool are activated in a systematic sequence, in order of increasing size; this recruitment order is the same whether the motoneurons are activated by descending commands from higher centers or by input from peripheral muscle nerves. The fact that properties like twitch tension and recruitment order are related to the motoneuron size has been called the size principle. For example, in isometric contraction of the first dorsal interosseous muscle of the hand, the first motor units to be activated contribute twitch tensions of 0.1 g; as the net muscle force increases, ever larger motor units become active; near the maximum levels of force, the largest units are finally recruited, with a twitch tension of 10 g. Thus, the larger units are recruited at higher levels of net force and also contribute larger twitch tensions. In fact, the size of the motor units' twitch tension is directly proportional to their recruitment level over a wide range of tensions.

Muscle tension. Thus, during voluntary muscle contractions, progressively greater muscle tension is produced by two mechanisms: additional motor units are recruited into activity, in order of small to large; and the firing rates of active motor units increase

with net force. In some muscles, the rate of motor unit discharge may increase from 8 to 30 action potentials per second, as the muscle contraction increases from light to maximal effort. At the lower firing rates, the successive twitch tensions of each motor unit produce a series of separate twitches. Since the different motor units of a muscle typically discharge asynchronously, the net tension in the muscle nevertheless varies smoothly. At higher frequencies of motor unit discharge, the twitch tensions of individual motor units begin to summate, producing still further increases in tension. Thus, increased temporal summation of twitches in individual units as well as increased spatial summation of twitches in different units contribute to increased tension. Moreover, the contribution of the large units recruited at high tension is proportionately greater.

Components of Skeletal Motor System

Motoneurons are activated by nerve impulses arriving through many different neural pathways. Some of their neural input originates in peripheral receptor organs located in the muscles themselves, or in receptors in skin or joints. Many muscle receptors discharge in proportion to muscle length or tension; such receptors have relatively potent connections to motoneurons, either direct monosynaptic connections or relays via one or more interneurons. Similarly, stimulation of skin and joints, particularly painful stimulation, can strongly affect motoneurons. Such simple segmental pathways constitute the basis for spinal reflexes.

The other major source of input to motoneurons arises from supraspinal centers. **Figure 4** shows the main nervous system centers involved in controlling the input to motoneurons. Evidence that these regions play a role in control of motor activity comes from three types of observations: damage of cells in these areas by experimental or clinical lesions produces motor deficits; electrical stimulation in these regions evokes motor responses or interferes with

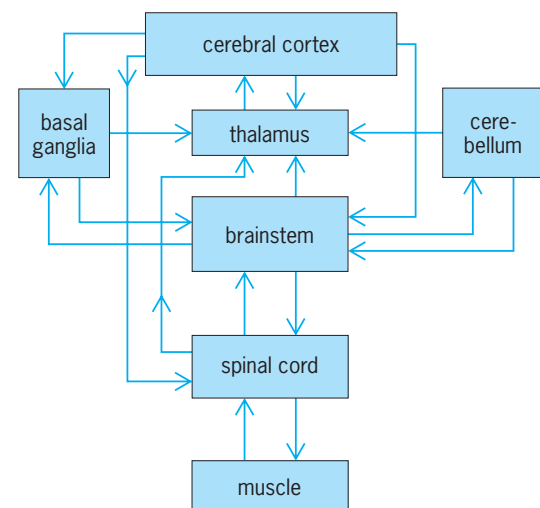


Fig. 4. Schematic diagram of the major components of the vertebrate motor systems. Arrows indicate the main neural connections between regions.

ongoing movements; and activity of cells recorded in these areas in moving animals is clearly related to their motor activity. Figure 4 indicates the major interconnections between cells in these regions; as one might expect, they function together during voluntary movements. The representation of the primate motor system in Fig. 4 is greatly simplified to provide an overview of its major components.

Segmental circuits. At the spinal level, the motoneurons and muscles have a close reciprocal connection. Afferent connections from receptors in the muscles return sensory feedback to the same motoneurons which contract the muscle. Connections to motoneurons of synergist and antagonist muscles are sufficiently potent and appropriately arranged to subserve a variety of reflexes. In animals with all higher centers removed, these segmental circuits may function by themselves to produce simple reflex responses. Under normal conditions, however, the activity of segmental circuits is largely controlled by supraspinal centers. Descending tracts arise from two major supraspinal centers: the cerebral cortex and the brainstem.

Brainstem. The brainstem, which includes medulla and pons, is a major and complex integrating center which combines signals descending from other higher centers, as well as afferent input arising from peripheral receptors. The descending output from brainstem neurons affects motor and sensory cells in the spinal cord. Brainstem centers considerably extend the motor capacity of an animal beyond the stereotyped reflex reactions mediated by the spinal cord. Even without any higher centers, animals with a brainstem can perform integrated activities such as standing, walking, and making appropriate postural adjustments. In contrast to segmental reflexes, these motor responses involve coordination of muscles over the whole body. In decerebrate cats, electrical stimulation of certain brainstem regions evokes a form of locomotion; the rate of walking is influenced by proprioceptive feedback from the limbs. Another major motor function of the brainstem is postural control, exerted via the vestibular nuclei. Sensory receptors in the vestibule of the ear, signaling the orientation of the head in space and changes in head position, have potent influence on the vestibular nuclei, which in turn effectively regulate limb and trunk muscles to respond in a manner that maintains postural stability.

Besides neurons controlling limb muscles, the brainstem also contains a number of important neural centers involved in regulating eye movements. These include motoneurons of the eye muscles and various types of interneurons that mediate the effects of vestibular and visual input on eye movement. For example, one important pathway mediates the vestibular ocular reflex; it transforms vestibular signals, initiated by head movements, to generate compensatory eye movements, with the result that eye position can remain fixed in space. Visual input can generate two different types of eye movements, each mediated by its own neural circuitry: rapid saccadic eye movements, which move

the eye as quickly as possible from one visual target to another; and smooth-pursuit eye movements, which follow a slowly moving target. In contrast to limb muscles, which must deal with changing loads, the eye muscles control a fixed and predictable load. This, plus the need for maintaining accurate vision, has led to a highly developed oculomotor system, residing largely in the brainstem.

Cerebellum. Another important coordinating center in the motor system is the cerebellum, an intricately organized network of cells closely interconnected with the brainstem. The cerebellum receives a massive inflow of sensory signals from peripheral receptors in muscles, tendons, joints, and skin, as well as from visual, auditory, and vestibular receptors. Higher centers, particularly the cerebral cortex, also provide extensive input to the cerebellum via pontine brainstem relays. The integration of this massive amount of neural input in the cerebellum somehow serves to smooth out the intended movements and coordinate the activity of muscles. Without the cerebellum, voluntary movements become erratic, and the animal has difficulty accurately terminating and initiating responses. The output of the cerebellum affects primarily brainstem nuclei, but it also provides important signals to the cerebral cortex.

The function of the cerebellum has been likened to a servocontrol mechanism which detects and corrects errors during the course of a movement. Following injury to the cerebellum, movements are ataxic; that is, they suffer from errors in rate, range, force, and direction. Starting, stopping, and changing the direction of movement are especially disturbed; in reaching for an object the hand often misses the target. Voluntary movement is characterized by tremor which increases as the movement progresses. All these disturbances appear to be due to defects, not in generating movement, but in detecting errors and correcting deviations. Thus, if in reaching for an object the hand strays from the correct path, correction is begun too late and proceeds too far, so that the hand overshoots the direct path. A series of such overcorrections results in the oscillating tremor typical of cerebellar injury.

Injury to those portions of the cerebellum which receive afferent input from the vestibular nuclei produces marked disturbances of equilibrium and gait, so that the victim, although showing little uncoordination of movement while lying in bed, is unable to maintain balance and to coordinate limb muscles in walking.

Basal ganglia. At another level of motor system are the basal ganglia. These massive subcortical nuclei receive descending input connections from all parts of the cerebral cortex. Their output projections send recurrent information to cerebral cortex via thalamus, and their other major output is to brainstem cells. Their substantial size suggests that the basal ganglia perform an important motor function, but the precise nature of the function remains unclear. Pathological lesions of the basal ganglia result in two types of motor disturbances, either decreased

motor activity or generation of involuntary movements; such lesions involve no sensory deficits, nor do they impair mental capacities.

Decreased motor activity caused by disruption of basal ganglia, called negative signs, include akinesia (a disinclination to use the affected part of the body), delays and slowing of limb movements, absence of normal postural adjustments necessary for stable standing, and deficits in facial expression. In other cases, basal ganglia disorders produce excessive involuntary motor activity; such positive signs of basal ganglia pathology include muscle rigidity (resistance to passive stretch caused by tonic motor activity), rhythmic muscle tremor, characteristic of parkinsonism, and uncontrollable movements of the hands or feet (alletosis) or the entire limb (ballismus). These motor symptoms can sometimes be alleviated by surgical or chemical treatments affecting the basal ganglia. The fact that both the positive and negative signs of basal ganglia pathology represent a disorder in generation of movements suggests that basal ganglia play a role in the initiation of movements.

Cerebral cortex. At the highest level of the nervous system is the cerebral cortex, which exerts control over the entire motor system. The cerebral cortex performs two kinds of motor function: certain motor areas exert relatively direct control over segmental motoneurons, via a direct corticospinal pathway, the pyramidal tract, and also through extrapyramidal connections via supraspinal motor centers. The second function, performed in various cortical association areas, involves the programming of movements appropriate in the context of sensory information, and the initiation of voluntary movements on the basis of central states. Cortical language areas, for example, contain the circuitry essential to generate the intricate motor patterns of speech. Limb movements to targets in extrapersonal space appear to be programmed in parietal association cortex. Such cortical areas involved in motor programming exert their effects via corticocortical connections to the motor cortex, and by descending connections to subcortical centers, principally basal ganglia and brainstem.

As indicated in Fig. 4, the motor centers are all heavily interconnected, so none really functions in isolation. In fact, some of these connections are so massive that they may form functional loops, acting as subsystems within the motor system. For example, most regions of the cerebral cortex have close reciprocal interconnections with underlying thalamic nuclei, and the corticothalamic system may be considered to form a functional unit. Another example is extensive connection from cerebral cortex to pontine regions of the brainstem, controlling cells that project to the cerebellum, which in turn projects back via the thalamus to cerebral cortex. Such functional loops are at least as important in understanding motor coordination as the individual centers themselves. With this overview in mind, the role of motor cortex in controlling movement will now be considered. The anatomical and physiological information available on the organization of motor cortex also

exemplifies the information required to understand other motor centers. See BRAIN.

Motor Cortex

The cerebral cortex is an extensive network of nerve cells covering the forebrain. The axons of some cortical cells pass into the underlying white matter and descend through the brainstem to affect the motoneurons of the brainstem or spinal cord. Areas of cortex designated as motor cortical areas have many such corticofugal cells. Electrical stimulation of these motor areas elicits movement, and their excision produces either paralysis or paresis (weakness). Activity of cells recorded in these regions is closely related to limb movements.

Motor areas. Systematic stimulation of the exposed cerebral cortex with electrical pulses reveals several prominent motor areas.

Precentral area. In primates the major one is the precentral motor area, which lies anterior to the central fissure of Rolando. In the monkey it extends over the precentral gyrus (Fig. 5); in humans it is largely buried in the anterior wall of the central fissure. Repetitive electrical stimulation within this region can elicit discrete movement, such as flexion of the thumb or a retraction of the lip. The evoked movement is always on the side of the body contralateral (opposite) to the stimulated hemisphere, since the descending pathways cross the midline before reaching the spinal cord. Moreover, the particular muscles activated depend on the stimulated site within the precentral motor area. Beginning at the midline and proceeding laterally along the precentral gyrus, one finds an orderly array of motor points for leg, trunk, arm, and face musculature. The topographical relations may be represented by displaying the body superimposed along the precentral gyrus as shown for the monkey in Fig. 5. In addition to this mediolateral representation of lower to upper

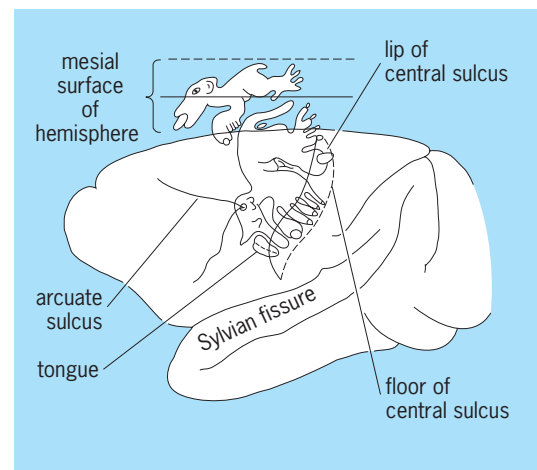


Fig. 5. Brain of monkey, with figurines showing motor representation of body in precentral motor cortex (lower figurine) and supplementary motor area (upper figurine). Parts of the body are drawn superimposed on the cortical regions that regulate their movement. The representation of toes and fingers buried in the precentral bank is shown with the mesial surface of the hemisphere folded out.

body musculature, there is an anterior-posterior organization, with the musculature of the apical portions of the body (fingers, toes, lower lip, and tip of tongue) located posteriorly, mostly in the depths of the central fissure, and the musculature of trunk and back represented more anteriorly, on the free surface of the precentral gyrus. A striking feature of motor cortex organization is the disproportionate area devoted to different muscles. The muscles of fingers, toes, lips, and tongue, which are involved in delicate, precise movements, have relatively large cortical areas devoted to their control. Comparatively smaller cortical areas are devoted to trunk musculature, which explains the distortion of the figurines in Fig. 5.

Precise stimulation of precentral motor cortex sites can sometimes evoke contraction of individual muscles. However, while the threshold response to stimulation may be activation of a single muscle, in most cases such stimuli also evoke subthreshold effects in motoneurons of other muscles. This is due to the fact that the stimuli affect numerous cortical cells projecting to different motoneurons. In the primate, forelimb motoneurons receive direct input from a colony of corticomotoneuronal cells, whose cell bodies are distributed over several square millimeters of cortex. Careful analysis indicates that the cortical colonies converging on motoneurons of different muscles may overlap extensively. In fact, single corticomotoneuronal cells may send divergent terminal projections to motoneurons of several different muscles. This means that the activity of such cells would affect a group of muscles.

Besides having motor output affecting muscles, the cortical cells also receive sensory input from receptors in muscles and skin. The topographic map over precentral cortex in Fig. 5 represents not only the parts of the body that move when the cortex is stimulated, but also the regions of the body from which the cortical cells can be activated. In other words, the sensory map and motor map are superimposed in register: cortical cells receive sensory input from the same regions to which they send motor output. Such input-output relations indicate that normally the cells would be involved in a functional loop interconnecting the cortical site with a peripheral locus.

Premotor and supplementary areas. The precentral motor area is the most elaborate motor cortical area with the most direct output connections to motoneurons. In addition, other motor cortical areas have also been demonstrated. One of these, called the premotor area, is located anterior and adjacent to the primary motor cortex. Cells in the premotor cortex are activated when the monkey prepares to make a movement, suggesting a role in motor programming. Another complete motor representation, the supplementary motor area, lies on the medial surface of the hemisphere (Fig. 5). The threshold for evoking movements by electrical stimulation is higher for supplementary motor area than for precentral motor area; such responses often involve the musculature of both sides of the body.

Postcentral cortex. A fourth significant motor representation is in postcentral cortex. Its topographical organization would be represented by a figurine which is the mirror image of that shown on the precentral cortex in Fig. 5. Like the precentral map, the postcentral map also represents both the sensory input from peripheral receptors as well as motor output to the relevant regions. The precentral and postcentral maps are heavily interconnected by cells making corticocortical connections. Nevertheless, some motor output from postcentral cortex is independent of its connections to precentral cells, since movements can be evoked by stimulating postcentral regions after motor cortex has been removed. The threshold for evoking motor effects from postcentral cortex is higher than from precentral cortex.

Eye and neck movements. Stimulation of the cortex in the region anterior to the arcuate fissure elicits movements of neck and eyes. The typical response consists of conjugate movement of the eyes and turning of the head to the side opposite the stimulated hemisphere. For this reason the area enclosed between the two limbs of the arcuate fissure is often called the frontal eye field. Since the major ascending pathways leading into the cortex cross the midline before reaching the cortex, stimuli applied to the right side of the body generate nerve impulses which reach the left hemisphere. Such impulses activating the motor eye fields on the left would cause movement of the head and eyes toward the right, that is, toward the stimulated side of the body. Such orientational movements may be part of the motor component of attention.

Effects of ablations. Removal of motor cortex within and just anterior to the central fissure causes a partial loss of voluntary movement in the contralateral muscles represented in the ablated cortex. Lesions confined to the leg area cause paralysis of the leg, leaving the arm and face musculature unaffected; thus the topographical organization of motor cortex revealed by stimulation is confirmed by ablation experiments. The paralyzed extremity is flaccid; that is, it hangs limply and displays no resistance to passive flexion or extension. Paralysis of voluntary movement is not permanent, however; the duration and severity of the deficit vary with species, being more severe in the human, ape, and monkey than in lower mammals such as dog and cat. This species difference reflects the increased dominance of the cerebral cortex in phylogenetically advanced forms. Following ablation of the arm motor cortex in monkeys and chimpanzees, movement at the shoulder reappears within a few days. Later, elbow and wrist movements reappear, and finally (8–12 weeks in chimpanzees) crude movements of fingers are restored. However, finger movements never regain the preoperative delicacy; thumb-finger approximation such as occurs in grooming remains permanently defective and awkward. This reflects the critical role of cortex in control of the finger muscles, which have an extensive cortical representation.

Following isolated unilateral ablation of the frontal eye fields in monkeys, the animal's head and eyes

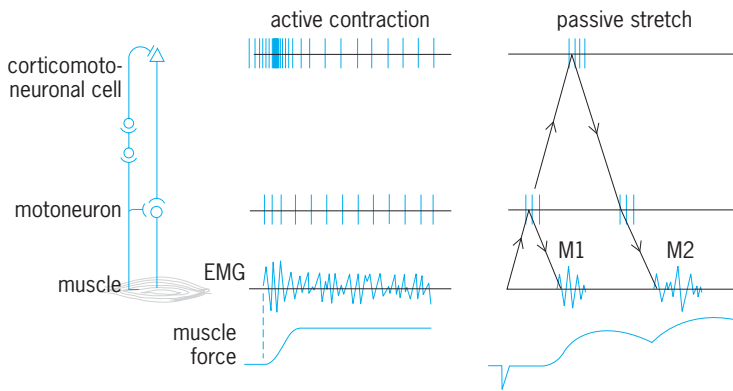


Fig. 6. Response of corticomotoneuronal cell and target muscle during active muscle contraction and in response to passive stretch of muscle. Diagram at left shows neural pathways: input from muscle stretch receptors affects motoneuron and corticomotoneuronal cell; the corticomotoneuronal cell also receives centrally originating input during active movement, and projects to motoneuron. Passive muscle stretch evokes impulses conducted to motoneuron via segmental reflex, producing M1 response, and to corticomotoneuronal cell, contributing to M2 response via transcortical loop.

turn toward the side of the lesion; there is transient paralysis of conjugate deviation of the eyes toward the opposite side. In walking, the animal tends to turn toward the side of the lesion; that is, it tends to follow the eyes and head, so that locomotion is circular. Such circling gait may persist after paralysis of the eye muscles has disappeared.

Activity of cortical neurons. The activity of single neurons recorded in monkeys trained to make specific movements further confirms the role of cerebral cortex cells in initiating and controlling movement. The location of motor cortex neurons which discharge in relation to active movement of particular parts of the body is again represented by the topographic map in Fig. 5. When the monkey initiates a simple hand movement—for example, pressing a key when a light is illuminated—many cells in the precentral hand region become active well before the onset of agonist muscle activity; many neurons continue to fire as long as the related limb muscles are active.

In monkeys making wrist movements requiring different degrees of active force, and independent wrist displacement, experimenters found that motor cortex cell activity was more closely related to the active force of a movement than the resultant limb position. **Figure 6** illustrates the typical response pattern of a corticomotoneuronal cell during an active wrist movement. Such corticomotoneuronal cells facilitate the activity of one or more target muscles. At onset of movement, the cell discharges a burst of activity beginning well before onset of activity in its target muscles. During the static hold period most corticomotoneuronal cells fire tonically at a steady rate; this tonic firing rate was found to increase in proportion to the amount of active force exerted. Thus, the firing pattern of a corticomotoneuronal cell shows a contribution to maintaining steady force during the hold period, as well as an additional burst related to initiating a change in force.

Besides firing during active movement, corticomotoneuronal cells also respond to passive limb movements, typically passive joint movements that stretch

their target muscles. Sensory input from stretch receptors in the same muscles which the corticomotoneuronal cell facilitates would make functional sense, since it would help compensate for load changes. A sudden increase in load during an active movement, for example, would stretch the agonist muscles, and would activate corticomotoneuronal cells whose output in turn would facilitate activity of those muscles, tending to overcome the increased load. Such a transcortical load compensation reflex is entirely analogous to the segmental stretch reflex mediated by direct input to motoneurons from stretch receptors.

Figure 6 also illustrates the responses evoked by a sudden stretch of a muscle. The first, short-latency electromyogram (EMG) response (M1) is mediated by the segmental stretch reflex—the direct connection from stretch receptors to motoneurons. The second EMG response (M2) is mediated by long-loop reflexes, including those through motor cortex. In contrast to the spinal circuit, the gain of the transcortical loop could be changed by the state of higher centers. Indeed, the response of cortical cells to peripheral stimulation is enhanced when the animal is prepared to make a movement in which the cells are involved. Corticomotoneuronal cells with monosynaptic connections to motoneurons are an important component of the pyramidal system, which controls spinal cells directly.

Pyramidal tract. One of the most important corticofugal pathways is the pyramidal tract, or corticospinal tract, which originates from cortical cells and runs without interruption to the spinal cord. The human medullary pyramid contains about 1 million fibers, mostly of small diameter. Almost half of the pyramidal tract originates from neurons in the precentral motor areas. Some of these are the giant pyramidal cells of Betz, large conical cells which histologically distinguish precentral motor cortex. In humans, the motor area of each hemisphere contains about 34,000 Betz cells, enough to account for about 2% of the pyramidal fibers, presumably those of large (11–20-micrometer) diameter. The supplementary motor area also contributes some fibers to the pyramidal tract. The rest of the pyramidal tract originates in the postcentral gyrus, posterior to the central fissure.

The axons of the pyramidal tract leave their cells of origin in cortical layer 5 to enter the subcortical white matter and then pass through the internal capsule, the cerebral peduncles, and the pons (**Fig. 7**). Some fibers terminate in the brainstem on motoneurons of cranial nerve nuclei. Below the pontomedullary junction the tract continues on the ventral surface of the medulla as a recognizable band of fibers called the pyramid. At the caudal border of the medulla most of the fibers cross to the opposite side of the neuraxis to descend in the dorsolateral column of the spinal cord. Some pyramidal fibers do not decussate but continue either in the ventral white funiculi or in the lateral column of the same side. The spinal portion of the corticospinal tract extends throughout the full extent of the spinal cord

but progressively diminishes in size as it passes from cervical to lumbar segments. Terminations are particularly numerous in the cervical and lumbar enlargements, which contain the motoneurons supplying the musculature of upper and lower limbs.

The role of the pyramidal tract in movement in monkeys and chimpanzees has been studied by sectioning the medullary pyramid and observing the resulting changes in motor performance. The most prominent defect following unilateral pyramidotomy is contralateral paresis involving the musculature from the neck down. The affliction is more severe in chimpanzees than in monkeys; in the former, even stereotyped movements of progression are impaired but not abolished. In neither animal is paralysis so grave as to render the affected parts useless, but there is severe impairment of voluntary movement and loss of such fine movements as opposition of thumb and index finger in grooming and individual movements of the fingers in manual exploration. This type of deficit has been observed to persist up to 4 years after operation, and thus may be considered permanent. Associated with the paresis is flaccidity like that seen following precentral cortical ablations; the extremities are limp and offer no resistance to movement at the joints.

Extrapyramidal systems. Isolated destruction of the corticospinal tract does not cause complete loss of voluntary movement. Moreover, electrical stimulation of the motor cortical areas after pyramidotomy still elicits muscular contraction. Therefore, the cortex must give rise to motor pathways other than the pyramidal tract. These other pathways are often referred to collectively as cortically originating

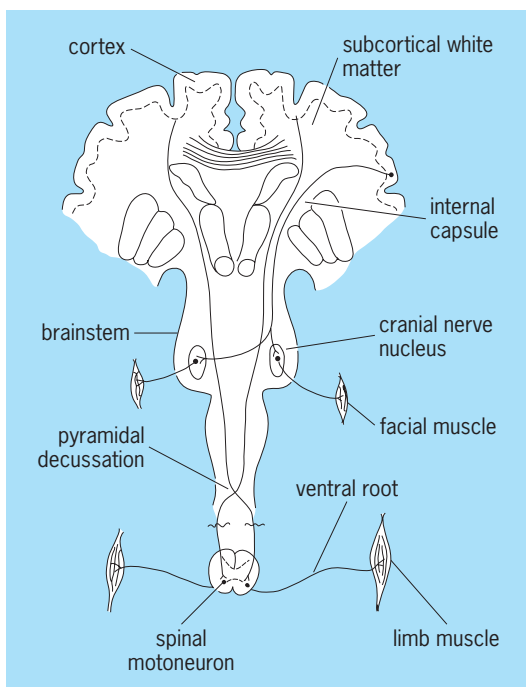


Fig. 7. Diagram of pyramidal system, showing origin and course of corticospinal (pyramidal) tract. In a similar manner, corticobulbar fibers affect cranial motoneurons of facial muscles.

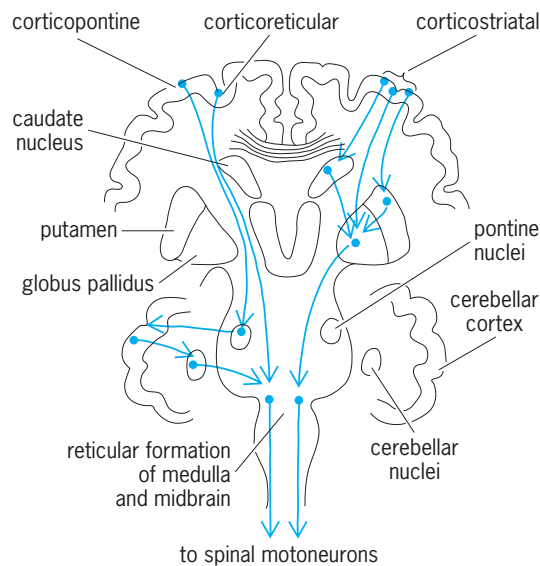


Fig. 8. Diagram of extrapyramidal systems, showing major descending connections from cerebral cortex.

extrapyramidal pathways. Although extrapyramidal projections probably arise from nearly all portions of the cortex, the precentral motor area appears to give rise to particularly large contributions. **Figure 8** illustrates the anatomic relation of some of the better-known extrapyramidal systems.

Corticostriatal and corticopallidal systems. These two systems originate from rostral precentral cortex and project to the caudate nucleus and the putamen, which are portions of the basal ganglia. From the basal ganglia, impulses are relayed to the brainstem, which in turn affects the spinal levels:

Corticoreticular systems. These systems originate from the cortex around the central fissure, especially from the motor area. The axons terminate in the region of the pons and medulla on the diffusely organized neurons constituting the brainstem reticulum. The projection is bilateral and poorly, if at all, organized somatotopically. Impulses are presumably relayed to the spinal cord via the reticulospinal tracts which traverse the ventral and lateral white columns of the cord. See RETICULAR FORMATION.

Corticopontine systems. These systems arise from the cortex of each of the four lobes of the brain, although the contribution from the rostral precentral gyrus is most prominent. A significant contribution also comes from the supplementary motor cortex on the mesial surface of the hemispheres. The fibers terminate in the pontine nuclei, which in turn project to the cerebellum. From the cerebellum, projection systems feed impulses into the medullary reticulum, which in turn projects through the reticulospinal tracts to the spinal cord as described above. See BRAIN; NERVOUS SYSTEM (VERTEBRATE).
Eberhard E. Fetz

Bibliography. E. R. Kandel, J. H. Schwartz, and T. M. Jessel, *Principles of Natural Science*, 4th ed., 2000; V. B. Mountcastle, *Medical Physiology*, 2 vols., 1980; H. D. Patton et al., *Textbook of Physiology*, 1989; R. Porter and R. Lemon (eds.), *Corticospinal Function and Voluntary Movement*, 1993.

Motorcycle

A two- or three-wheeled self-propelled motor vehicle. Power is usually supplied by an internal combustion engine of one or more cylinders with either air or liquid cooling, although electric-powered motorcycles have been built. Transmissions commonly have four or five forward speeds; motorcycles with automatic transmissions have been commercially available, but have not achieved great popularity. Some three-wheeled motorcycles are equipped with a reverse gear. Power is transmitted from the transmission to the rear wheel by either a roller chain, a belt, or a shaft drive. *See* INTERNAL COMBUSTION ENGINE.

Motorcycles are classified by engine displacement and by type of use. Vehicles powered by engines of 50 cm³ or lower displacement are classified as mopeds; vehicles powered by engines with displacements of 700 cm³ and above are classified as heavyweights. Vehicles between these displacements are referred to as lightweight or medium-weight motorcycles depending on where they are in the range. Motorcycles are generally used for recreation or transportation on paved roads. Some light- and medium-weight motorcycles are specifically designed for off-road use.

Recreational motorcycles are further classified as sports or touring vehicles depending on their primary intended use. A typical two-wheeled modern touring motorcycle (see *illus.*) is designed to carry a driver and passenger and is equipped with saddlebags and a tour pack for carrying clothing and supplies, as well as a fairing and windshield for protection from the elements. It is not unusual to find a touring motorcycle also equipped with a stereo radio, a cassette player, a citizens' band radio, and other accessories. As the illustration shows, a two-wheeled motorcycle generally has its engine mounted in a tubular frame and has a rear wheel on a swing arm which can rotate relative to the frame in a vertical plane but which is restrained by one or two coil springs and hydraulic shock absorbers. The front wheel is generally mounted in a fork which contains two telescopically operating spring hydraulic damper systems. *See* SHOCK ABSORBER.

A motorcycle engine is started either by kicking a foot-operated cranking mechanism or by activating



Touring motorcycle. (Harley-Davidson Motor Co., Inc.)

an electric starter to rotate the crankshaft. Operating controls are standardized so that the throttle is activated by twisting the grip on the right handlebar, the clutch lever is located on the left handlebar, the gear-shift lever is activated by a foot pedal on the left side of the motorcycle, and the front and rear brakes are operated by hand and foot levers respectively on the right side of the motorcycle. Brakes are of either the disk or drum type, with disk brakes predominating on high-performance motorcycles. A two-wheeled motorcycle is steered by turning the handlebars and leaning in the direction of the turn. The steering of three-wheeled motorcycles or two-wheeled motorcycles equipped with a sidecar requires different techniques. *See* AUTOMOTIVE BRAKE.

J. L. Bleustein

Bibliography. J. Bennett, *The Complete Motorcycle Book*, 2d ed., 1999; W. N. Crouse and D. L. Anglin, *Motorcycle Mechanics*, 1982; Society of Automotive Engineers, *Motorcycle Dynamics and Rider Control*, Spec. Publ. 428, February 1978.

Mountain

A feature of the Earth's surface that rises high above its base and has generally steep slopes and a relatively small summit area. Commonly the features designated as mountains have local heights measurable in thousands of feet or many hundreds of meters, lesser features of the same type being called hills, but there are many exceptions. *See* HILL AND MOUNTAIN TERRAIN.

Mountains rarely occur as isolated individuals. Instead they are usually found in roughly circular groups or massifs, such as the Olympic Mountains of northwestern Washington and the Harz Mountains of northern Germany, or in elongated ranges, like the Sierra Nevada of California, the Bighorn Range of Wyoming, or the Sierra de Guadarrama of central Spain. An array of linked ranges and groups, such as the Rocky Mountains, the Alps, or the Himalayas, is a mountain system. North America, South America, and Eurasia possess extensive cordilleran belts, within which the bulk of their higher mountains occur. *See* CORDILLERAN BELT; MASSIF; MOUNTAIN SYSTEMS.

As a rule, mountains represent portions of the Earth's crust that have been raised above their surroundings by upwarping, folding, or buckling, and have been deeply carved by streams or glaciers into their present surface form. Some individual peaks and massifs have been constructed upon the surface by outpourings of lava or eruptions of volcanic ash. *See* VOLCANO.

Edwin H. Hammond

Bibliography. D. R. Butler, S. J. Walsh, and G. P. Malanson (eds.), *Mountain Geomorphology: Integrating Earth Systems*, 2003; J. Gerrard, *Mountain Environments: An Examination of the Physical Geography of Mountains*, 1990; P. N. Owens and O. Slaymaker (eds.), *Mountain Geomorphology*, 2004; M. Price, *Mountains: Geology, Natural History, & Ecosystems*, 2002.

Mountain meteorology

The effects of mountains on the atmosphere, ranging over all scales of motion, including very small (such as turbulence), local (for instance, cloud formations over individual peaks or ridges), and global (such as the monsoons of Asia and North America).

Lee waves. The most readily perceived effects of a mountain, or even of a hill, are related to the blocking of air flow. When there is sufficient wind, the air either goes around the obstacle or over it, causing waves in the flow similar to those in a river washing over a boulder. Since ascending air cools by adiabatic expansion, the saturation point of water vapor may be reached in such waves as they form over an obstacle, and a cloud then forms in the ascending branch of the wave motion. Such a cloud dissipates in the descending branch where adiabatic warming takes place, with a sometimes spectacular display of lenticular cloud formations (**Fig. 1**). See CLOUD; CLOUD PHYSICS.



Fig. 1. Lenticular cloud, southwest of Boulder, Colorado, October 24, 1980, 17:50 MST. (Ron Holle, NOAA/ERL, Boulder)

The shapes and amplitudes of these lee waves (they form over and to the lee of mountains) depend not only on the thermal stability and on the vertical wind shear in the overlying atmosphere but also on the shape of the underlying terrain. Such waves have been modeled successfully by towing obstacles through a tank in which salt solutions of different concentrations are stacked upon each other, mimicking various degrees of atmospheric stability (**Fig. 2**). They also have been modeled numerically in computers. See SIMULATION; WAVE (PHYSICS).

Mountain torque. On a grander scale, mountain ranges, such as the Sierras of North and South America, place an obstacle in the path of the westerly

winds (that is, winds from the west), which generally prevail in middle latitudes. Such a blockage tends to generate a high-pressure region upwind from the mountains (this may be viewed as air piling up as it prepares to jump the hurdle), and a low-pressure area downwind. Thus, there is a stronger push against the mountains on the high-pressure western side than on the low-pressure eastern side. The net effect between these stronger and weaker pushes would be to move the mountains toward the east. However, since the mountains are firmly attached to the Earth, an extremely slight acceleration of the Earth's rotation is a result. While minute, this acceleration can be measured by atomic clocks as a variation in the length of day. Since the strength of the westerlies causing this effect varies from day to day, there will be slight variations in the length of day smaller than milliseconds. This variation is known as the mountain torque effect. As this torque speeds up the rotation of the Earth, it will slow down the rotation of the atmosphere, because the total rotational momentum of the earth-atmosphere-ocean system is conserved as the system spins in space. Thus, slowing down the atmospheric flow is precisely what a mountain obstacle does. See TORQUE.

Planetary waves. Less subtle than mountain torque effects are the large-scale meanders that develop in the global flow patterns, once they have been perturbed, mainly by the North and South American Andes and by the Plateau of Tibet and its Himalayan mountain ranges. All other mountain ranges on Earth are dwarfed by these gigantic, crustal upheavals. These meanders in the large-scale flow are known as planetary waves. They appear prominently in the pressure patterns of hemispheric or global weather maps. They can also be seen from space as they manifest themselves in large-scale swirls of cloud formations (**Fig. 3**). See WEATHER MAP.

Diurnal heating effects. There are important thermal effects attached to mountains in addition to their serving as large obstacles thrown into the path of the atmosphere, a moving fluid. As satellite measurements have shown, surface temperatures of the Earth do not change much with terrain elevation. They depend mostly on the exposure angle to the Sun, the vegetation cover, and the type of soil. Surface temperatures vary dramatically, however, between snow-free and snow-covered ground, mostly because of changes in the reflectivity, or albedo. Thus, the snow-free ground surface temperatures during the day in the Rocky Mountains or in Tibet are not much different from those in the adjacent plains. At night, however, the thinner and drier mountain

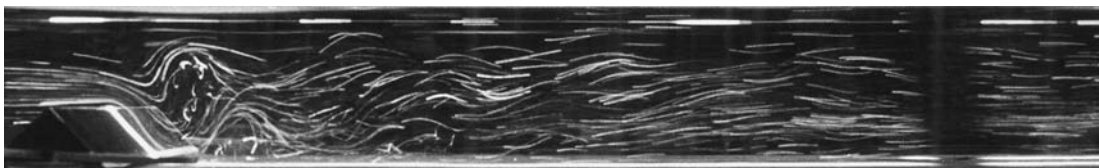


Fig. 2. Lee waves produced by towing an obstacle through a Plexiglas channel filled with stratified salt solutions of various densities, mimicking a stable atmosphere. (D. L. Boyer and L. Tao, University of Wyoming)



Fig. 3. Infrared image taken by the *GOES West* satellite on May 4, 1979, at 2:45 GMT. The large swirls of clouds in middle latitudes of both hemispheres indicate the position of planetary waves in the jet stream, and of cyclones. (NASA)

atmosphere lets more infrared radiation escape to space than does the thicker and moister atmosphere over the plains. Therefore, the ground surface in the mountains cools off more readily than in the plains, causing greater diurnal variations in surface temperatures of elevated terrain. See ALBEDO; TERRESTRIAL RADIATION.

The free atmosphere at the same height or pressure levels corresponding to the elevations of adjacent mountain ranges remains cooler during the day, as it is farther away from the surface heat source; and it is warmer during the night, since air does not emit radiation as readily as does the solid surface of the ground. Therefore, during the day there will be

a gradient of temperature from the warm mountain surface to the cooler atmosphere at the same altitude over the plains. That gradient reverses at night. As a consequence, during the day a so-called heat low tends to develop over mountain ridges; and it reverses to anticyclonic, high-pressure conditions at night. Since air tends to flow from high to low pressure, especially if the horizontal space dimensions and the time scales involved do not allow the deflecting force of the Earth's rotation (the Coriolis force) to become fully effective, there is flow toward the mountains (valley breezes) during the day and away from the mountains (mountain breezes) at night. See CORIOLIS ACCELERATION; CYCLONE.

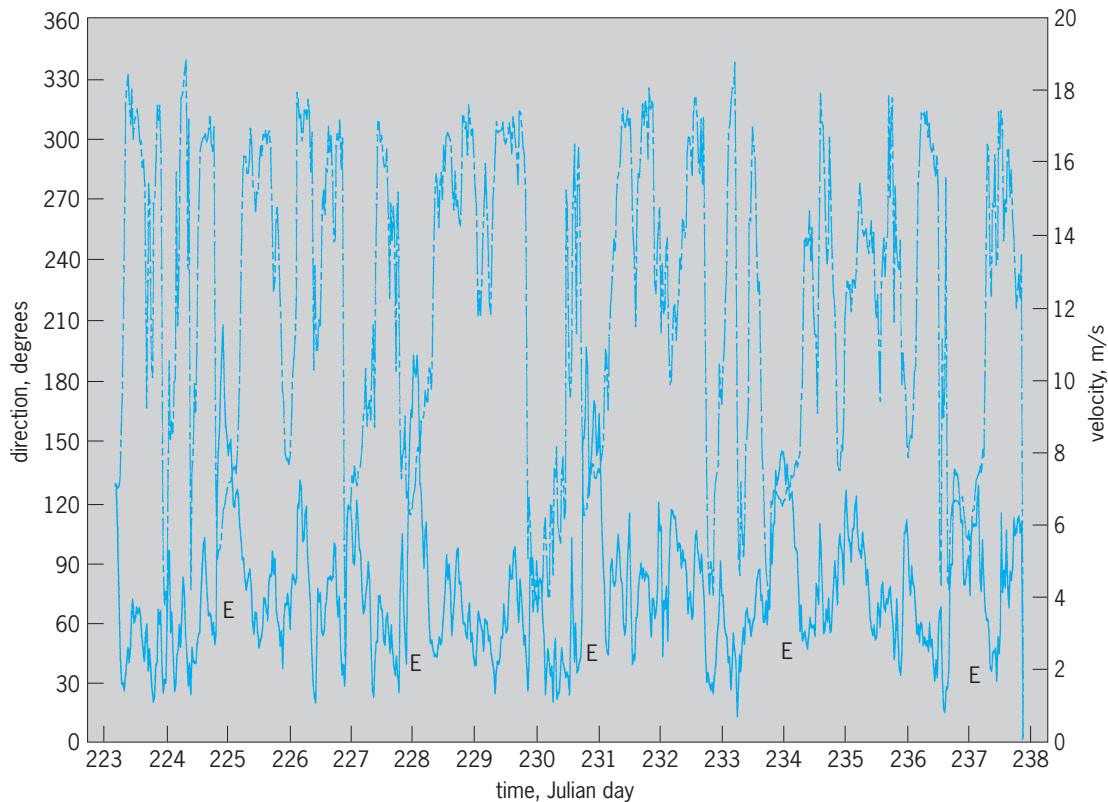


Fig. 4. Wind direction and speed measurements at the peak of Mount Werner, Colorado, during August 1984. During the day, winds tend to be from the northwest; in the evening, they tend to shift to the southeast, at times resulting in rather strong events (E) that might endanger small aircraft flying close to the peak. The layer in which these strong winds are contained typically extends to about 0.6 mi (1 km) above mountain peak elevation. 1 m = 3.3 ft.

Plateau circulation system. A circulation system similar to that generated by diurnal heating effects but on a grander scale develops if large plateaus, such as the Tibetan Plateau or the Rocky Mountains, are involved in the diurnal heating and cooling cycle. The resulting plateau circulation system over North America extends as far as the Mississippi Valley during summer, when this diurnal cycle is best developed. As the air flows toward the mountains, thunderstorms tend to develop over the major ranges and peaks, with a frequency maximum during the early afternoon. With the reversal of the temperature and pressure gradients at night, these thunderstorms tend to collapse, but new ones form at the leading edge of the air mass now flowing away from the mountains. Perhaps it is a result of this circulation that the Great Plains west of the Mississippi show a nocturnal thunderstorm maximum, as late as past midnight. See AIR MASS; THUNDERSTORM.

Because the more intense daytime heating of the mountains, as compared to the free atmosphere over the plains, is most prominent in summer, the tendency of air flowing toward the large plateaus and ascending along their periphery also favors that season. Thus the large monsoonal weather systems of Asia and also of North America are caused, at least in part, by the heating effects of the adjacent, large and high plateaus.

Latent heat effects. There is an additional mountain effect associated with the summer thunderstorm and

monsoon activities. The ground surface temperature rises rapidly in the morning, especially during summer, producing the temperature and pressure gradients described above, with the resulting wind systems. As air converges toward the mountains and ascends along their slopes, clouds soon form that shade the ground from further insolation and prevent further heating of the soil. However, once cumulus cloud formation has started, especially during the summer thunderstorm season, the latent heat of condensation of water vapor makes the interior of the clouds warmer than their environment, causing them to grow until they develop into full-fledged thunderstorm systems. As heavy rain starts falling from these systems in the late afternoon, its evaporation contributes to the rapid cooling of the subcloud layer, causing a rapid reversal from low-pressure conditions with inflow toward the mountains to high-pressure conditions with flow away from the mountains (Fig. 4). See AIR PRESSURE; INSOLATION; PRECIPITATION (METEOROLOGY).

Monsoons. It appears that the latent heat cycle—warming by condensation and cooling by evaporation—assumes a dominant part in the formation of the diurnal plateau circulation cycle. A similar concept holds for the monsoons of Asia and North America on a seasonal, rather than diurnal, basis. For instance, the Tibetan Plateau heats in spring; it attracts inflow into a heat-generated low-pressure system near the Earth's surface,

compensated by outflow aloft from an anticyclone that takes up summer residence over the plateau. The steadiness and vigor of this pressure distribution and its attendant monsoonal circulation system, however, is maintained in large part, by the release of latent heat in the torrential rains that fall over India and southwest China whenever the monsoon is at its peak.

It is well known that the major monsoon circulations interact with the global circulation, shaped in part by sea-surface temperature anomalies in the equatorial Pacific. This is why, in certain years, the monsoon rains tend to fail. The various aspects of mountain meteorology, therefore, have to be viewed within the larger picture. There is a continuous interaction between the weather effects on all space and time scales generated by the mountains and the weather patterns that prevail elsewhere on the Earth. Thus, what is occurring over Tibet has an effect on weather over the Pacific and North America, and vice versa; all weather effects are part of a large, complicated system that is only beginning to be elucidated. *See* ATMOSPHERIC GENERAL CIRCULATION; METEOROLOGY; MONSOON METEOROLOGY; WIND.

Elmar R. Reiter

Bibliography. R. G. Barry, *Mountain Weather and Climate*, 2d ed., 1992; E. R. Reiter (ed.), GARP-ALPEx: The Alpine Experiment, *Meteorol. Atm. Phys.*, 36:1-296, 1987; E. R. Reiter et al., ROMPEX-85, The Rocky Mountain Peaks Experiment of 1985, *Bull. Amer. Meteorol. Soc.*, 68:321-328, 1987; E. R. Reiter et al., Tibet revisited: TIPMEX-86, *Bull. Amer. Meteorol. Soc.*, 68:607-615, 1987; E. R. Reiter and M.-C. Tang, Plateau effects on diurnal circulation patterns, *Mon. Weath. Rev.*, 112:638-651, 1984.

Mountain systems

Long, broad, linear to arcuate belts in the Earth's crust where extreme mechanical deformation and thermal activity have been (or are being) concentrated.

Geological properties. Mountain systems in the general sense occur both on continents and in ocean basins, but the geological properties of the systems in continental as opposed to oceanic settings are distinctly different. The mechanical strain in classical, continental mountain systems is expressed in the presence of major folds, faults, and intensive fracturing and cleavage. Thermal effects are in the form of vast volcanic outpourings, intruded bodies of igneous magma, and metamorphism. Uplift and deformation in young mountain systems are conspicuously displayed in the physiographic forms whose bold topographic relief and beauty are so awesome. Where mountain building is presently taking place, the dynamics are partly expressed in warping of the land surface (as in Palmdale Bulge, California) and significant shallow or deep earthquake activity. Locations of ancient mountain systems in continental regions now beveled flat by erosion are clearly disclosed by the presence of highly deformed, intruded, and metamorphosed rocks. Excellent examples of

such mountain systems are the Appalachians of eastern North America, the Caledonian system of Norway and the British Isles, the Hercynian of England, France, and Germany, and the Urals of Russia. Additionally, the Precambrian shield provinces of the world contain the roots of ancient mountain systems. *See* VOLCANOLOGY.

Evolution. The generally accepted view among scientists today is that mountain systems evolve through the movements and interference of rigid plates of lithosphere which compose the Earth's crust and upper mantle. The movements and interactions likewise predetermine the history and nature of sediment accumulations which eventually become bound up in the mountain systems. Some plates consist exclusively of oceanic crust (Pacific, Cocos, Nazca, Caribbean, and Philippine), and a few of continental crust (such as Arabia), but the largest are generally some combination of oceanic and continental crust (America, Eurasia, Africa, India, Antarctica, and Somalia). The relation of the present distribution and movements of these plates to active mountain belts and to sedimentation patterns provides the key to understanding the complicated evolution of mountain systems in the past. *See* LITHOSPHERE.

Oceanic systems. Two basic classes of oceanic mountain systems exist. A world-encircling oceanic rift mountain system has been built along the extensional tectonic boundary between plates diverging at rates of 0.8-2.4 in./year (2-6 cm/year) from the mid-oceanic ridges. This rift mountain system, exposed to partial view in Iceland, is about 36,000 mi (60,000 km) long, averages 3200 mi (2000 km) in breadth, and displays a topographic relief of 6600 ft (2000 m). Rocks along the mid-oceanic rift mountain system are mainly faulted basaltic volcanics. The ridge crest of the mountain system, marking the locus of generation of oceanic crust, is offset repeatedly by transform faults. *See* MARINE GEOLOGY.

The second type, island arc mountain systems, occurs in oceanic basins where the crust dives downward at trench sites, thus underthrusting adjacent oceanic crust. The process results in imbricate faulting, folding, gravity sliding, metamorphism, and partial melting of the crust to produce magmas and volcanic piles. Rates of descent are on the order of 2-4.7 in./year (5-12 cm/year). Such island arc mountain building is now taking place in the western Pacific from the Aleutian arc southward through the belt of archipelagos which includes the Japan, Mariana, Philippine, and New Hebrides arcs. The topographic relief of such island arc mountain systems is staggering, in some places exceeding 33,000 ft (10,000 m). All traces of ancient rift and island arc mountain systems which once existed in the ocean basins have been completely destroyed by the dynamics of sea-floor spreading and the coordinated descent (subduction) of oceanic crust into the Earth's mantle. In effect, these mountains, and the rocks of which they were composed, have been recycled. *See* OCEANIC ISLANDS.

Classical systems. The classical, conspicuous mountain systems of the Earth occur at the

continent/ocean interface, for this is the site where plate convergence has led to major sedimentation, subduction of oceanic crust under continents, collision of island arc mountain systems with continents, and head-on collision of continents. Less commonly, the continent-ocean interface may be the site of transform faulting (as in the San Andreas Fault). The South American Cordillera or Andean mountain system has resulted from stresses generated during underthrusting of South America by oceanic crust of the Pacific plate. Melting of the subducted oceanic plate has led to outpourings of volcanic materials which constitute the volcanic edifice of the Andes. The Alpine, Apennine, Carpathian, and Himalayan systems define an east-west-trending Mediterranean system which has resulted from convergence and collision of the African and Indian plates against Eurasia. The ancient Tethys Sea was closed (repeatedly) by the northerly relative movement of Africa, thrusting oceanic crust and sediment onto the European platform. The leading edge of the Indian plate crashed at a rate of 6 in./year (15 cm/year) into Eurasia, producing the spectacular Himalayan system. See CORDILLERAN BELT; FAULT AND FAULT STRUCTURES.

Mountain systems then result from mechanical and thermal responses to sometimes complicated and oftentimes superimposed (and perhaps fortuitous) plate movements. The Cordillera of western North America is one of the best examples of a mountain system of long and complex history. Its unusual evolution is partly expressed in the number and diversity of its components (Coast Ranges, Sierra Nevada, Basin and Range, Colorado and Columbia plateaus, Rocky Mountains, and Rio Grande Rift). The geology of the Cordillera mountain system is the product of superimposed rift, compressional, and extensional movements, related mainly to the interaction of the American plate with oceanic plates to the west. See OROGENY; PLATE TECTONICS.

George H. Davis

Bibliography. J. B. Bird (ed.), *Plate Tectonics*, rev. ed., 1980; J. Erickson, *Plate Tectonics: Unraveling the Mysteries of the Earth*, 1992, reprint 2001; X. LePichon et al., *Plate Tectonics*, 1976; A. Miyashiro et al., *Orogeny*, 1982; R. B. Singh, *Dynamics of Mountain Geo-Systems*, 1992; S. P. Sychanthavong (ed.), *Crustal Evolution and Orogeny*, 1992; J. T. Wilson, A new class of faults and their bearing on continental drift, *Nature*, 207:343-347, 1965.

Mouth

The oral or buccal cavity and its related structures. The oral cavity forms in the embryo from an in-pocketing of the skin, the stomodeum; it is thus lined by ectoderm and is not, properly speaking, part of the digestive tract. Functionally, however, the mouth forms the first portion of both the digestive and respiratory systems. Various special structures are found in, or associated with, the mouths of most vertebrates. See DIGESTIVE SYSTEM; RESPIRATORY SYSTEM.

Teeth may be present to help grasp or grind food.

In most vertebrates they are relatively simple cones but in some, especially mammals, they are of diverse shapes. Teeth may be present only along the margins of each jaw, as in mammals, or they may be more widely distributed throughout the mouth; they may be set into sockets or fused to the skeleton of the jaws in various ways. See DENTITION; TOOTH.

Various glands, salivary glands in a broad sense, are associated with the mouth. These are of infrequent occurrence in fish but are found in most tetrapods. Humans have three pairs of salivary glands: the parotid, submaxillary, and sublingual. In forms such as some snakes salivary glands may produce a poison used to subdue prey.

Other structures also vary greatly. Most tetrapods have a mobile tongue attached to the floor of the mouth, but few fish do. The structure of the roof of the mouth, or palate, is quite different in different groups. Cyclostomes, the lampreys and hagfishes, have especially highly modified oral cavities with "teeth" and a "tongue" quite unlike those of other vertebrates. See PALATE; TONGUE.

In mammals, including humans, the margins of the lips mark the junction between the outer skin and the inner mucous lining of the oral cavity. The mucosa of the mouth forms the lining and the gums surrounding the teeth and covers the surface of the tongue. The roof of the mammalian mouth consists of the hard palate and, behind this, the soft palate which merges into the oropharynx. The lateral walls consist of the distensible cheeks, and the floor is formed principally by the tongue and the soft tissues that lie between the two sides of the lower jaw, or mandible.

The posterior limit of the oral cavity of mammals is marked by the fauces, an aperture which leads to the pharynx. On either side of the fauces are two muscular arches covered by mucosa, the glossopalatine and pharyngopalatine arches; between them lie masses of lymphoid tissue, the tonsils. Suspended from the posterior portion of the soft palate is the soft retractable uvula. See MOUTH DISORDERS; TONSIL.

Thomas S. Parsons

Mouth disorders

The mouth, or oral cavity, which comprises the lips, tongue, teeth, gums, and related structures, is subject to a large number of disease processes. Periodontitis, an inflammatory disease of the tissues supporting and surrounding the teeth, and dental decay are the most common diseases; together they account for almost all tooth loss. Other diseases of the mouth can be classified as cysts; diseases of the salivary glands; keratotic, inflammatory, ulcerative, and proliferative lesions; oral infections; and an unusual form of rapidly destructive periodontitis observed in some patients with acquired immune deficiency syndrome (AIDS).

Cysts. A cyst is a cavity lined with epithelium and filled or partly filled with fluid. Cysts of the mouth are of odontogenic or nonodontogenic origin. The periapical cyst is the most common odontogenic type

and forms at the apex of the tooth root. Its formation is an attempt by the body to wall off necrotic material emanating from the root of an infected tooth. A follicular cyst arises from the epithelium of the enamel organ of an unerupted tooth and is seen most often around third molars (wisdom teeth). Cysts of non-odontogenic origin develop from epithelium which has been sequestered in bone suture lines during embryonic development, and are named according to their location. The most commonly occurring are nasoalveolar, nasopalatine, globulomaxillary, and incisive canal cysts. Cysts gradually enlarge and may lead to extensive bone destruction. They are treated by surgical excision.

Neoplasms. The oral cavity is made up of many varieties of tissue, each of which can give rise to neoplasms. Thus, several types of malignant and benign neoplasms can be found in and around the mouth. The benign tumors most often seen are fibromas, papillomas, and hemangiomas. Fibromas are tumors of collagenous connective tissue which appear clinically as smooth, elevated areas with broad bases. A papilloma is made up of a thin, connective tissue stalk covered with a thick layer of epithelium. Clinically, this lesion exhibits a verrucous (warty) appearance. Hemangiomas are benign neoplasms of newly formed blood vessels.

Oral squamous-cell carcinoma (epidermoid carcinoma) accounts for over 90% of all malignancies of the mouth and about 5% of all malignancies of the entire body. Its most common site is the lower lip, and it is more common in males than in females. Lesions of the lip are associated with exposure to sunlight, and those in the oral cavity with tobacco use and consumption of alcoholic beverages. Even though leukemia is not a primary tumor of the oral cavity, it can be included among the important mouth disorders because swelling and inflammation of the gingiva may be the first clinical signs of the presence of the disease. *See ONCOLOGY.*

Salivary gland diseases. The major salivary glands are the parotids, submaxillaries, and sublinguals. Mumps is a viral infection of the secretory cells of these glands and occurs most often in one or both parotid glands. The salivary glands can become infected by retrograde passage of bacteria through the secretory duct, causing acute or chronic sialadenitis. Salivary ducts can also become blocked by calculi (stones), resulting in the accumulation of secretions and swelling. This lesion is called a ranula. The most frequent neoplasm is a mixed tumor. Immunosuppression, irradiation, and administration of cytotoxic drugs tend to decrease salivary flow and to result in enhanced caries and periodontal disease and in painful inflammatory lesions of the oral mucosa. *See ORAL GLANDS.*

White and keratotic lesions. Many disorders of the mouth appear clinically as white or hyperkeratotic lesions. Lichen planus is a dermatologic disease which also occurs in the mouth, on the tongue (keratosis linguae), or on the lips (keratosis labialis) as smooth, lacy networks of white lines or, less commonly, as white patches. The patches may become ulcerative.

The cause of the disease is not known. Leukoplakia is the most important white lesion of the mouth since it is considered to be premalignant. It appears clinically as granular white patches on the oral mucosa. Only a small percentage of such lesions progress to carcinoma. These lesions exhibit an accumulation of excessive keratin and an alteration in the pattern of maturation of the epithelium. An inflammatory infiltrate develops in the underlying connective tissue. Traumatic injury and burns can also appear clinically as white lesions.

Oral infections. Gingivitis and periodontitis, as well as dental caries, are a result of oral bacterial infections. Mycotic infection by *Candida albicans* is recognized clinically by the presence of red, inflamed mucosa bearing colonies of white fungi with the appearance of milk curds. The disease is called candidiasis or thrush, and it occurs most commonly in children, in adults after extensive antibiotic therapy, and in some patients with acquired immune deficiency syndrome (AIDS).

Infection of the mouth and lips by the virus herpes simplex type I is possibly the most common human viral infection, next to the common cold. Initially, the lesions appear as small blebs or blisters which rupture to form open ulcers. The ulcers are extremely painful and can involve the entire mouth. The disease is self-limiting and usually disappears within 1–2 weeks. *See HERPES.*

Ulcers of the mouth can also be caused by blood dyscrasias, lichen planus, and several other dermatologic diseases.

Health factors. Generalized systemic disease may play a large role in the development of mouth disorders. Vitamin deficiency can result in glossitis (inflammation of the tongue); and diabetes mellitus, stress, and abnormalities in blood leukocytes may predispose the individual to development of gingivitis and periodontitis. *See ACQUIRED IMMUNE DEFICIENCY SYNDROME (AIDS); DIABETES; PERIODONTAL DISEASE; VITAMIN.* Roy C. Page

Bibliography. J. A. Regezi and J. J. Sciubba, *Oral Pathology*, 3d ed., 1999; W. G. Shafer, M. G. Hine, and B. M. Levy, *A Textbook of Oral Pathology*, 4th ed., 1983.

Moving-target indication

A method of presenting pulse-radar echoes in a manner that discriminates in favor of moving targets and suppresses stationary objects. Moving-target indication (MTI) is almost a necessity when moving targets are being sought over a region from which the ground clutter echoes are very strong. The most common presentation of the output of a radar with MTI is a plan position indicator (PPI) display. The moving targets appear as bright echoes, while ground clutter is suppressed.

Principles of operation. MTI is based upon the use of the Doppler effect; that is, the carrier frequency of the echo from a target moving toward or away from the radar shifts by an amount proportional to

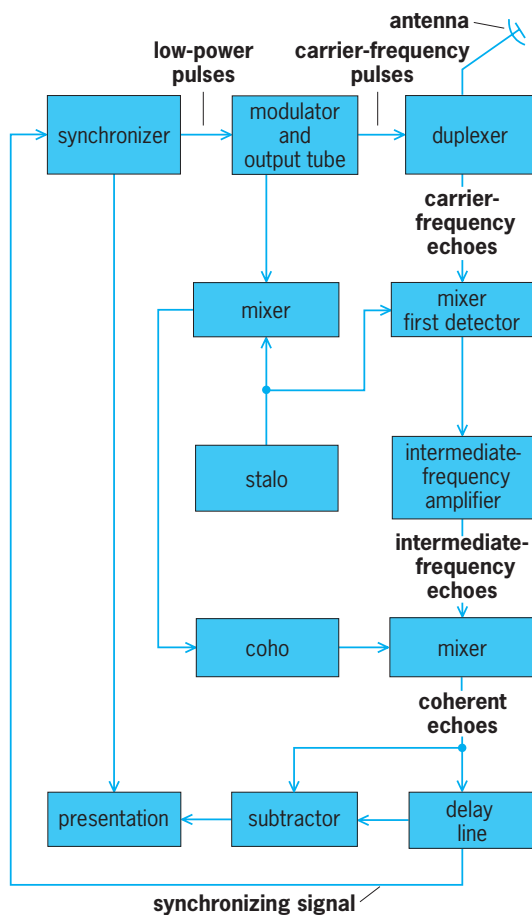


Diagram of radar system with moving-target indication.

the product of radial velocity and transmitted frequency (see *illus.*). A stable oscillator in the receiver is synchronized with the transmitter, providing a continuous reference of the transmitter frequency and phase. Echoes are heterodyned with the reference oscillator. Stationary objects supply echoes having a carrier phase shift which is constant from pulse to pulse, because the time for the signal to travel to a stationary object and back is always the same. Therefore the heterodyner output for stationary echoes does not change in phase from pulse to pulse. However, phase shift in echoes from moving targets changes from pulse to pulse because of the change in signal propagation time. See DOPPLER RADAR.

To utilize the lack of pulse-to-pulse phase difference to discriminate against clutter echoes, the heterodyner output is fed to a delay line which stores the signal for a period of time exactly equal to the period between pulses. Then the output of the delay line is subtracted from the freshly produced output of the heterodyner. If the two outputs are identical, as will be the case for clutter, the difference is zero. But this difference is not zero for moving targets because of their pulse-to-pulse phase shift. The difference between consecutive echoes is presented as the MTI output. The delay line usually employed utilizes acoustic propagation within fused quartz.

A variety of internal arrangements are in use for assuring that the reference oscillator is synchronized

with the transmitter output. The reference oscillator may be at the carrier frequency but usually operates at an intermediate frequency, because better stability can be obtained in a low-frequency oscillator. This arrangement requires that the local oscillator used for conversion from the carrier frequency to the intermediate frequency be extremely stable, and the term *stalo* (stable local oscillator) is used to denote circuits that satisfy this requirement. The reference oscillator is called the *coho* (coherent oscillator) to denote that it remains coherent (fixed phase relation) with the transmitter output.

Measures of merit. The figure of merit of an MTI system, clutter suppression factor, is defined as the ratio of root-mean-square (rms) clutter amplitude before MTI to rms clutter fluctuation after MTI. It is the amount by which the clutter is suppressed. It depends in turn upon several other factors, some of which are system parameters and thus controllable, and some of which are functions of terrain and weather.

Another measure of merit for MTI radar is the signal-to-clutter ratio (SCR). However, it has been found that increasing the SCR simultaneously reduces the signal-to-noise ratio (SNR), a measure of merit of the sensitivity of the radar. Targets in a clutter-free environment are less detectable with MTI than without it because of reduced SNR. A parallel bank of Doppler filters can provide a simultaneous improvement in both SCR and SNR. See SIGNAL-TO-NOISE RATIO.

Use of charge-transfer devices. Charge-transfer devices can be utilized to implement MTI with excellent capability for canceling clutter. A charge-transfer device is a capacitorlike semiconductor device of the bucket-brigade variety. As used in MTI, samples of the returned radar signal are stored and processed as discrete quantities of electrical charge. This implementation has become known as discrete-signal moving-target indication (DSMTI). Metal oxide semiconductor field-effect transistors are used as charge-transfer devices to provide the delay that permits subtracting a radar return delayed by one pulse-repetition interval (PRI) from the currently incoming radar return.

Adaptive MTI. Radar systems designers have long appreciated that the best target range and velocity resolution are obtained by using recurrent pulses whose duration is short compared with the PRI. However, this practice gives rise to ambiguities in determining the range and velocity of the target. In fact, it causes the radar to be blind to targets moving at certain velocities when MTI is in use. A way to avoid blind velocities is to implement a random staggering of the radar's pulse-repetition frequency (PRF). Fast changes in the PRF combined with filtering systems and integration methods that adapt automatically to these changes permit ambiguities in determining the range and velocity of the target to be resolved in a single look at the target. Clutter filters switched in synchronism with the PRF also reduce undesirable clutter sidebands even when the PRF is switched at a high rate, and the irregular pulse spacing

alleviates the problem of blind velocities. The clutter filter can be a notch-type filter that has its rejection band centered about the carrier frequency. See ELECTRIC FILTER.

This adaptive MTI (AMTI) proved to be useful in achieving acceptable detection of targets in the presence of natural precipitation static and chaff dispersed by the enemy as a defensive measure against radar. See ELECTRONIC WARFARE; RADAR.

John M. Carroll

Bibliography. J. L. Eaves and E. K. Reedy, *Principles of Modern Radar*, 1987; M. I. Skolnik, *Introduction to Radar Systems*, 2d ed., 1986; M. I. Skolnik, *Radar Handbook*, 2d ed., 1990.

Mucilage

A naturally occurring, high-molecular-weight (ranging 200,000 and up), organic plant product of unknown detailed structure. The term is loosely used, often interchangeably with the term gum. Chemically, mucilage is closely allied to gums and pectins but differs in certain physical properties. Although gums swell in water to form sticky, colloidal dispersions and pectins gelatinize in water, mucilages form slippery, aqueous colloidal dispersions which are optically active and can be hydrolyzed and fermented. Mucilages are not pathological products but are formed in normal plant growth within the plant by mucilage-secreting hairs, sacs, and canals, but they are not found on the surface as exudates as a result of bacterial or fungal action after mechanical injury, as are gums. Mucilages occur in nearly all classes of plants in various parts of the plant, usually in relatively small percentages, and are not infrequently associated with other substances, such as tannins. The most common sources are the root, bark, and seed, but they are also found in the flower, leaf, and cell wall. Any biological functions within the plant are unknown, but they may be considered to aid in water storage, decrease diffusion in aquatic plants, aid in seed dispersal and germination, and act as a membrane thickener and food reserve. Mucilages are commonly identified by physical properties, most recently by infrared spectroscopy.

Most mucilages are considered to be cellulosic polysaccharides containing the same group of sugars as gums and pectins, and commonly occur as salts of acids (acid type). The cations are chiefly of calcium, magnesium, potassium, and sodium. The most common acids are uronic acids, the most important being galacturonic acid, although other acid residues including sulfonic acids are known. Some are known to exist with the acid function absent (neutral type) and consist only of simple sugars connected by a glycosidic linkage. Hydrolysis of mucilages yields pentoses and hexoses, the most common being arabinose, galactose, glucose, mannose, rhamnose, and xylose.

The chief industrial sources of mucilages are Icelandic and Irish moss, linseed, locust bean, slippery

elm bark, and quince seed. They are obtained by milling if they occur in the endosperm, but more commonly by extraction with water or dilute sodium carbonate solution when they are found outside the seed coating. They are purified by precipitation with alcohol or salt solutions from the aqueous solution and marketed as powders. Synthetic mucilages can be formed by decomposition reactions, as in the hydrolysis of starch to form dextrans which are mixed with gums to form adhesives. Mucilages find applications in cosmetics (hand lotions and hair sets), medicinals (laxatives and diuretics), pharmaceuticals (emulsifying agents and materials for prevention of precipitation by colloidal suspension), and industry (thickening agents, emulsion stabilizers, binders, sizing of silks, textile printing, and paper manufacture). See ADHESIVE; GUM; PECTIN.

Elbert H. Hadley

Bibliography. A. J. Kinloch, *Adhesion and Adhesives: Science and Technology*, 1987; A. A. Lawrence, *Natural Gums for Edible Purposes*, 1977; W. Pigman (ed.), *The Carbohydrates*, 2d ed., vols. 1A-2B, 1970-1978.

Muffler

A device used to attenuate sound while also allowing fluid (usually gas) to flow through it; also known as silencer in British usage. Mufflers are extensively used to reduce the intake and exhaust noise from pumps, fans, compressors, and internal combustion engines. Although active noise control techniques are emerging, most mufflers continue to use passive silencing methods. Passive mufflers are categorized as reactive or dissipative based on their primary method of attenuation. Reactive mufflers reflect sound back toward the noise source, and dissipative mufflers use porous materials to absorb the sound (Fig. 1).

Muffler attenuation as a function of frequency is often characterized by the transmission loss (L_{TL}) and insertion loss (L_{IL}), in units of decibels (dB). Transmission loss is defined as $10 \log_{10} (W_{in}/W_{out})$, where W_{in} and W_{out} represent the sound power entering and exiting the muffler. Insertion loss is defined as the difference in sound pressure level at a fixed location caused by inserting the muffler in the system. Transmission loss is convenient for describing the behavior of the muffler itself, while insertion loss gives a better indication of muffler attenuation in a particular system. For some dissipative duct mufflers, acoustic performance is described by the drop in sound pressure level per unit length. See DECIBEL; SOUND PRESSURE.

In addition to attenuation, several other criteria need to be considered in the design of these elements. The pressure drop due to flow restrictions usually needs to be minimized, since flow losses waste energy and will increase with flow rate. Other design considerations include size, durability, the magnitude of self-noise or flow-generated noise, and the initial, maintenance, and replacement costs.

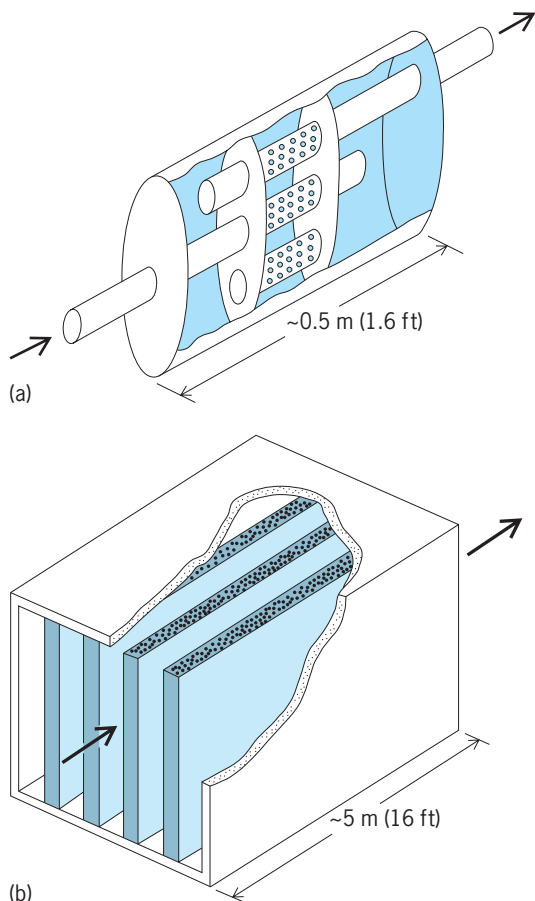


Fig. 1. Mufflers. (a) Reactive muffler for automobile exhaust system. (b) Dissipative muffler for industrial application.

Often the design criteria conflict: The attenuation generally increases with size and flow restriction, and this increase typically competes with the need to minimize these latter parameters.

Reactive mufflers. Reactive mufflers reflect acoustic waves at locations where a duct expands, contracts, or branches. Often a combination of reactive elements such as expansion chambers, resonators, and flow reversals is used (Fig. 1a). Reactive mufflers can be designed to provide better low-frequency attenuation than a dissipative muffler of similar size. Also, reactive mufflers can be used in harsh environments that dissipative or active mufflers might not withstand.

In most cases, reactive mufflers are best suited for low-to-moderate frequencies, where acoustic wavelengths are larger than any cross dimension of the muffler. At these frequencies, mufflers can exhibit resonance and/or broadband attenuation behavior. A resonance suggests a small region of high attenuation centered around a frequency. Resonance behavior is exhibited by the Helmholtz (Fig. 2a) and side-branch (Fig. 2b) resonators. Reactive mufflers with broadband behavior, such as a simple expansion chamber, may exhibit lower maximum attenuation than a resonator, but attenuate over a wider frequency range (Fig. 2c). The side-branch resonator and the

expansion chamber demonstrate another characteristic of many reactive muffler elements where the attenuation is cyclic with respect to frequency. Locations for attenuation maxima and minima are related to ratios between muffler lengths and the acoustic wavelength. See RESONANCE (ACOUSTICS AND MECHANICS).

Dissipative mufflers. Dissipative mufflers use absorptive materials that dissipate the acoustic energy into heat. A variety of porous media can be used for absorption, with fibrous materials such as fiberglass being common. The dissipative material can be incorporated in a number of ways. For example, it

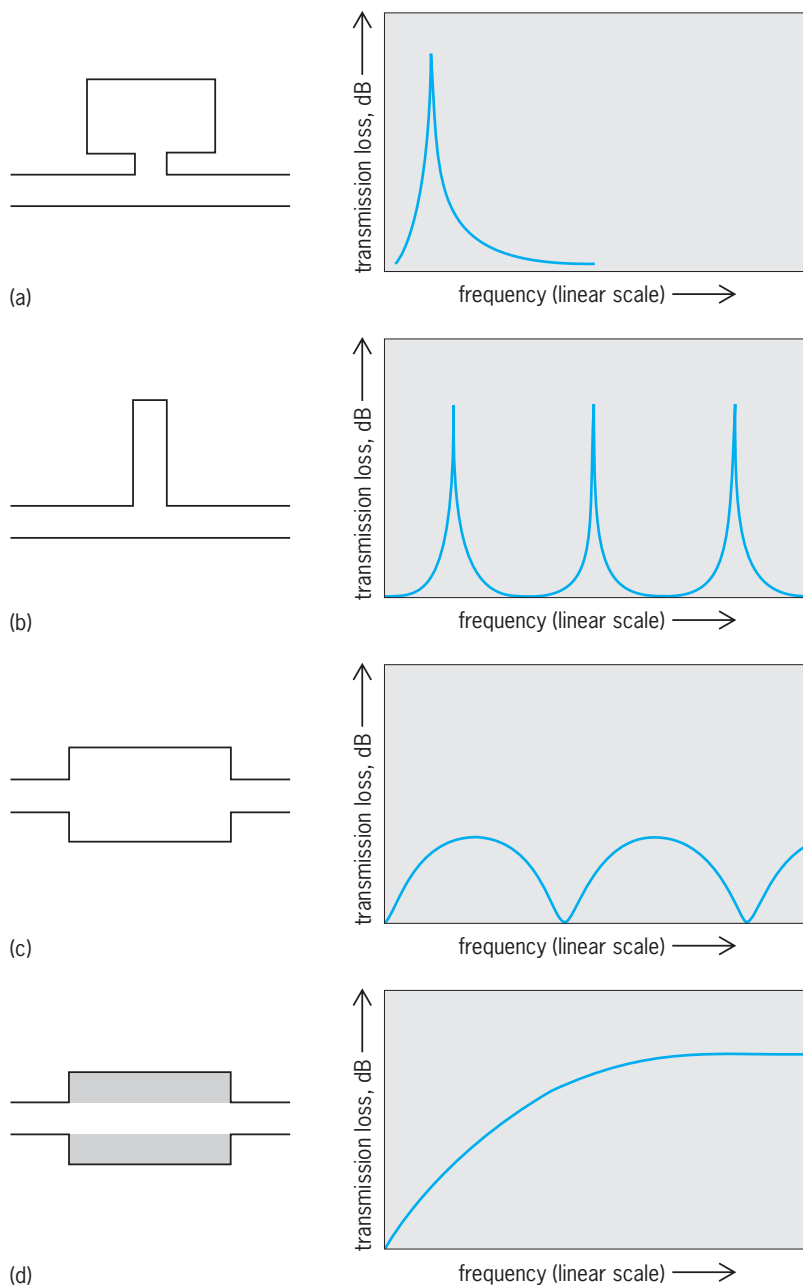


Fig. 2. Simple muffler elements, and their representative transmission losses. (a) Helmholtz resonator. (b) Side-branch resonator. (c) Expansion chamber. (d) Expansion chamber with dissipative material.

can fill the outer cavity of an expansion chamber (Fig. 2*d*), or it can be used as linings for the duct walls and baffles that separate a duct into smaller channels (Fig. 1*b*). The linings and baffles can be flat, contoured, constructed from layers of different materials, or mixed and matched for a particular application. Absorptive materials may face challenges due to harsh conditions such as high temperatures and potential clogging from particulate-laden flows. See SOUND ABSORPTION.

Dissipative mufflers are best suited for moderate-to-high frequencies, since absorption is less effective at low frequencies. At frequencies where the absorptive materials are effective, the attenuation is broadband, and the passbands exhibited by reactive mufflers are reduced or eliminated (Fig. 2*d*). Compared to reactive mufflers of similar size, dissipative mufflers can have higher attenuation (except at resonances for the reactive muffler) and lower pressure drop. At higher frequencies, where the acoustic wavelength is smaller than the duct width, the attenuation of a dissipative muffler may decrease considerably.

In practice, no muffler is entirely dissipative or reactive, as some amount of both types of behavior is inevitable. Many mufflers combine reactive and dissipative components to capitalize on their respective strengths. Reactive elements are used for low-frequency attenuation, and dissipative elements are used to absorb moderate- and high-frequency noise.

Active mufflers. Active mufflers attenuate unwanted noise by adding sound to counteract it. In a simplified example, a loudspeaker is used to generate sound waves that are an exact mirror image of the incoming noise. The disturbances add algebraically, resulting in a cancellation of the unwanted sound. An active muffler consists of sensors (such as microphones), a controller, and actuators (such as loudspeakers). The controller unit processes the signals from the sensor, and computes an appropriate signal for the actuator. Numerous control systems and strategies exist, and are under continuous development.

Active mufflers are best suited for low frequencies where the sound field is relatively simple. The effectiveness of active mufflers has been demonstrated for a number of situations, but several challenges remain, which are the topic of ongoing research. There is a need for rugged sensors and actuators that can withstand high temperatures and harsh environments. Also, high-intensity disturbances at low frequencies require large-displacement, high-power actuators. See ACTIVE SOUND CONTROL.

Design tools. Many predictive tools are available to assist in the design of mufflers. For simple geometries and low frequencies, analytical wave methods are often adequate. For higher frequencies and more complex mufflers, computer models can be used effectively. Currently, however, experimental testing of numerous prototypes is often required to finalize a muffler design. Several factors that can affect muffler behavior are difficult to fully describe and predict. These include mean flow effects, flow-generated

noise, intense sound levels, and acoustic propagation through porous interfaces and media. Research is continuing in these and other areas to more fully understand and more accurately predict the muffler behavior. See ACOUSTIC NOISE. Ahmet Selamet

Bibliography. L. L. Beranek and I. L. Vér, *Noise and Vibration Control Engineering*, Wiley, 1992; D. D. Davis, Jr., Acoustical filters and mufflers, in C. M. Harris (ed.), *Handbook of Noise Control*, McGraw-Hill, 1957; M. L. Munjal, *Acoustics of Ducts and Mufflers*, Wiley, 1987.

Mugiliformes

An order in the class Actinopterygii, subdivision Euteleostei, superorder Acanthopterygii. The Mugiliformes comprise 17 genera and about 72 species, all in the single family Mugilidae, the mullets (also known as grey mullets). Past classifications considered mullets as primitive perciforms; however, they are presently deemed subperciforms and are the only order in Mugilomorpha, one of three series making up the Acanthopterygii. See ACTINOPTERYGII; PERCIFORMES.

Morphology. The following characters distinguish the family: The head and body sectors are more or less round in cross section and the tail sector is moderately compressed; the dorsal fins are well separated, the first has four weak spines, and the second 8 to 10 soft rays; the pectoral fins are high on the body; the pelvic fins are subabdominal, each with one spine and five rays; the anal fin has two or three spines and 7 to 11 soft rays; an adipose eyelid is usually present; the mouth is small and terminal; teeth are small or absent; gill rakers are long; the stomach is usually muscular (gizzardlike), and the intestine is long and coiled; scales of adults are usually ctenoid (or cycloid in juveniles); and the lateral line is absent or very faint. Adult mullet species range from 15 cm (6 in.) to 1.2 m (4 ft) in maximum length, but most species are about 30 to 70 cm (11.8 to 27.5 in.) long (see **illustration**).

Habitat. Mulletts inhabit coastal marine and brackish-water environments in all tropical and temperate seas, as well as freshwater. Some marine species enter freshwater only temporarily and may ascend great distances up rivers, such as the striped mullet (*Mugil cephalus*), a common and economically important species that occurs practically worldwide in warm waters. Adult mountain mullet (*Agonostomus monticola*), a western Atlantic species ranging from North Carolina to Venezuela, including the West Indies, is confined to freshwater streams, but its juveniles occur in offshore coastal waters. Probably the only species that does not spend part of its life in a marine environment is the Abu mullet (*Lisa abu*) of Turkey, Iraq, Pakistan, and southeastern Asia. Most species form schools, some quite large. Their primary food is detritus, plant material, and microorganisms that are filtered from muddy substrates by a pharyngobranchial organ,



Largescale mullet. (Photo courtesy of John E. Randall)

macerated by the muscular stomach, and absorbed by the long coiled intestine. Herbert Boschung

Bibliography. I. J. Harrison, Mugilidae, pp. 1071–1085 in K. E. Carpenter (ed.), *The Living Marine Resources of the Western Central Atlantic: FAO Species Identification Guide for Fishery Purposes*, vol. 2, FAO, Rome, 2003 (dated 2002); I. J. Harrison and G. J. Howes, The pharyngobranchial organ of mugilid fishes: Its structure, variability, ontogeny, possible function and taxonomic utility, *Bull. Brit. Mus. Nat. Hist. (Zool.)*, 57(2):111–132, 1991; J. S. Nelson, *Fishes of the World*, 4th ed., Wiley, New York, 2006; M. L. J. Stiassny, What are grey mullets?, *Bull. Mar. Sci.*, 52(1):197–219, 1993.

Mulberry

A genus, *Morus*, of trees characterized by milky sap and simple, often lobed, alternate leaves. White mulberry (*M. alba*) was introduced into the United States from China during the nineteenth century as a source of food for silkworms. The silkworm project was unsuccessful, but the trees remained. This species has smooth shiny leaves that are usually lobed (Fig. 1), and a very irregular habit of branching. The fruit is white and insipid. Red mulberry (*M. rubra*) attains a height of 60 ft (18 m) and grows in the eastern half of the United States and in southern Ontario.



Fig. 1. White mulberry (*Morus alba*).

Its leaves are rough above and soft and pubescent (hairy) beneath, and sometimes have two or three lobes (Fig. 2). The fruit is about 1 in. (2.5 cm) long,

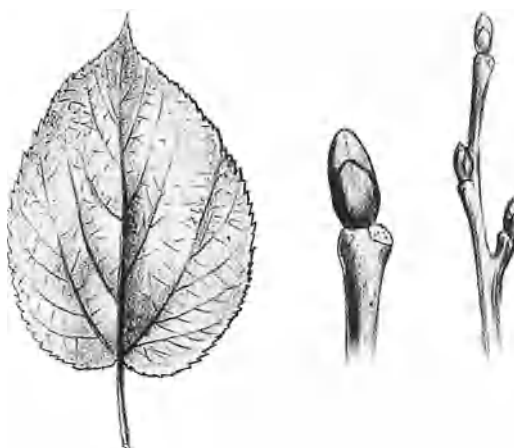


Fig. 2. Red mulberry (*Morus rubra*).

dark purple or black, and sweet, and is used as food by both domestic and wild animals. The wood is used for fence posts, furniture, interior finish, agricultural implements, and barrels. See FOREST AND FORESTRY; TREE.

Arthur H. Graves; Kenneth P. Davis

Mule

The mule is a hybrid sired by a male ass (*Equus asinus*) out of a female horse (*E. caballus*). The opposite cross, very seldom made, produces the hinny (a hybrid between a stallion and a female ass). The mule and the hinny are usually sterile, but there have been two authenticated cases of mules producing living progeny. Male mules, often called horse mules, are almost always castrated to make them more tractable as work animals.

Jacks and jennets are the males and females, respectively, of the domestic ass family. In the United States, most jacks have been used to sire mules, whereas almost all jennets have been kept solely to produce more jackstock. Small asses, about the size of Shetland ponies, are commonly called donkeys or burros. They are used as pack animals in a few

mountainous areas of the Southwest, and occasionally as children's pets in other regions. *See* HORSE PRODUCTION.

American jacks, sometimes called Mammoth jacks, were developed from a blending of several European strains of jackstock, principally the Catalonian, Maltese, Majorcan, Andalusian, Italian, and Poitou. Large jacks may measure over 15 hands (60 in. or 1.5 m) and weigh 1100 lb (500 kg). Compared to the horse, jacks have shorter hair on the tail and mane, much longer ears, smaller and deeper hoofs, and no callosities (hard or thickened areas) on the inside of the hind legs. The loud, harsh voice of jacks is called a bray. The mule also shows these characteristics, but usually is more full bodied and more smoothly constructed than the jack.

A jack to be used for mule production should be raised with horses after weaning and not permitted to see jennets and mules until he has been broken to breed mares. Otherwise, the jack may be very slow or even unwilling to breed a mare. Heavy draft mares bred to jacks produce big draft mules. Mares of finer quality produce lighter, finer-boned mules of the kind that were formerly used by many tobacco growers and cotton farmers in the South.

Mules are noted for their endurance, sure footedness, and ability to stand hard work in hot weather. They can safely be self-fed in lots or corrals, whereas horses cannot. Usually steady and free from nervous excitability, mules can be handled by inexperienced or careless farm labor. *See* AGRICULTURAL SCIENCE (ANIMAL). John M. Kays

Multiaccess computer

A computer system in which computational and data resources are made available simultaneously to a number of users. Users access the system through terminal devices, normally on an interactive or conversational basis. A multiaccess computer system may consist of only a single central processor connected directly to a number of terminals (that is, a star configuration), or it may consist of a number of processing systems which are distributed and interconnected with each other as well as with the user terminals.

The primary purpose of multiaccess computer systems is to share resources. The resources being shared may be simply the data-processing capabilities of the central processor, or they may be the programs and the databases they utilize. The earliest examples of the first mode of sharing are the general-purpose, time-sharing, computational services. Examples of the latter mode are airlines reservation systems in which it is essential that all ticket agents have immediate access to current information. Both classes of systems are still popular, with the proportion and importance of distributed application systems continuing to increase in areas such as corporate management and operations. *See* DATABASE MANAGEMENT SYSTEM.

There are a number of economic factors supporting the growth of multiaccess systems. Primary among these are improvements in hardware perfor-

mance and economics, with the latter being the principal factor. Digital-computer hardware continues to exhibit rapidly decreasing costs while its operational capabilities increase. At the same time most of the other cost factors involved with corporate operations such as labor, travel, and communications are increasing, and the cost of time delays in operations such as inventory management and confirming orders, as well as the cost of investment funding, is becoming significant.

System components. The major hardware components of a multiaccess computer system are terminals or data entry/display devices, communication lines to interconnect the terminals to the central processors, a central processor, and on-line mass storage.

Terminals may be quite simple, providing only the capabilities for entering or displaying data. Also utilized in multiaccess systems are terminals having an appreciable amount of "local intelligence" to support simple operations like editing of the displayed text without requiring the involvement of the central processor. Some terminals provide even more extensive support, such as the local storage of small amounts of text. It is also possible to assemble clusters of terminals to share local logic and storage.

The interconnecting communication lines can be provided by utilizing the common-user telephone system or by obtaining leased, private lines from the telephone company or a specialized carrier. Another important source of data circuits is a value-added network specializing in providing data transmission services. *See* DATA COMMUNICATIONS.

The communications interface for the processor may be provided by an integral hardware component or by a separate device known as a communications controller or front-end (communications) processor. The detailed operations involved in the control of a single communications line are not very great, but servicing a large number of such lines may present an appreciable load to the central processor; thus, separate hardware for communications interfacing is almost essential in a multiaccess computer system.

A central processor suitable for use in a multiaccess system must include the capability to support a large central memory as well as the communications interface mentioned above and the on-line mass storage discussed below.

The desirable characteristics of a mass storage device are: the ability to quickly locate any desired data; the ability to transfer data at a very high speed; the provision of economical storage for a large quantity of data; and a high degree of reliability.

The most common form of mass storage is rotating magnetic disks. Originally, magnetic drums were utilized for this purpose; however, the price, performance, and capacity of disks have resulted in their replacing drums. *See* COMPUTER STORAGE TECHNOLOGY.

System operating requirements. A multiaccess system must include the following functional capabilities: (1) multiline communications capabilities that will support simultaneous conversations with a reasonably large number of remote terminals; (2) concurrent execution of a number of programs

with the ability to quickly switch from executing the program of one user to executing that of another; (3) ability to quickly locate and make available data stored on the mass storage devices while at the same time protecting such data from unauthorized access.

The ability of a system to support a number of simultaneous sessions with remote users is an extension of the capability commonly known as multiprogramming. In order to provide such service, certain hardware and software features should be available in the central processor. Primary among these is the ability to quickly switch from executing one program to another while protecting all programs from interference with one another. These capabilities are normally provided by including a large central memory in the processor, by providing hardware features that support rapid program switching, and by providing high-speed transfers between the central memory and the mass storage unit on which the programs and the data they utilize are stored.

Memory sharing is essential to the efficient operation of a multiaccess system. It permits a number of programs to be simultaneously resident in the central memory so that switching execution between programs involves only changing the contents of the control registers. The programs that are resident in central memory are protected from interfering with one another by a number of techniques. In earlier multiprogramming systems, this was accomplished by assigning contiguous memory space to each program and then checking every access to memory to ensure that a program was accessing only locations in its assigned space. The drawback to this type of memory allocation is that the entire program had to be loaded into memory whenever any portion of it was to be executed. A popular memory management technique is the utilization of paging. The program is broken into a number of fixed-size increments called pages. Similarly, central memory is divided into segments of the same size called page frames. (Typical sizes for pages and page frames are 512 to 4096 bytes.) Under the concept known as demand paging, only those pages that are currently required by the program are loaded into central memory. Page frames may be assigned to a given program in a random checkerboard fashion. Hardware capabilities are provided to automatically manage the assignment of page frames as well as to make them appear to the program as one continuous address space. At the same time, the hardware provides memory access protection.

Software capabilities. The control software component of most interest to an interactive user is the command interpreter. It is only one portion of a larger control program known as the operating system; however, in multiaccess systems it differs greatly from the command interpreter in a system providing batch multiprogramming service on a non-interactive, non-terminal-oriented bases. This routine interacts directly with users, accepting requests for service and translating them into the internal form required by the remainder of the operating system, as well as controlling all interaction with the system.

The operating system must also have the capability for controlling multiprogramming; that is, the concurrent execution of a number of user programs quickly switching from one to another during their execution as well as controlling memory sharing. The capability to page the memory as outlined above can be utilized to provide users with the impression that each has available a memory space much larger than is actually assigned. Such a system is said to provide a virtual memory environment. Similarly, the ability of the operating system to quickly change context from one executing program to another will result in users' receiving the impression that each has an individual processor. This is especially true when considering the large difference between the response and thinking time of a human compared to the computer's processing time. For the interactive users, such a capability results in the impression of having a private virtual machine.

The mass storage device mentioned above as hardware required for a multiaccess system must be supported by an efficient file management system. The latter is responsible for maintaining current information as to the physical location of the data stored on the mass storage device as well as providing a capability for quickly locating those data and controlling their transfer to the central memory for utilization. In addition, the file management system must provide protection of data from unauthorized access. See DIGITAL COMPUTER. Philip H. Enslow, Jr.

Bibliography. V. D. Black, *Computer Networks*, 2d ed., 1993; F. Halsall, *Data Communications, Computer Networks, and Open Systems*, 4th ed., 1996; G. Held and R. Sarch, *Data Communications: A Comprehensive Approach*, 6th ed., 1999; W. Stallings, *Data and Computer Communications*, 6th ed., 1999.

Multidimensional scaling

Any of a number of procedures for the fitting of a selected geometric model to a body of multivariate data. In the original form advanced by Warren S. Torgerson, multidimensional scaling dealt with judgments about stimuli, with its major aim being the derivation of psychological dimensions presumed to underlie the judgments, and the determination of scale values for stimuli along the derived dimensions. Classical multidimensional scaling was a significant methodological advance in the characterization of complex stimulus domains. If forced to rely exclusively on unidimensional scaling methods, a researcher could characterize a multidimensional stimulus domain only by first scaling the stimuli on individual dimensions and later assembling the scaled dimensions in a postulated model. Even then, the model would often be incomplete because of insufficient foresight in choice of dimensions, or inability to determine the relations between, and relative importance of, the chosen constituent dimensions.

Advantages. With the advent of multidimensional scaling and related techniques, psychological inquiry benefited in two important ways. First, the data

collection procedures associated with the new techniques did not require prior specification of constituent dimensions before scaling could take place. As a consequence, it was necessary to instruct the subjects in the specific dimensions to be considered in making judgments. Since most multidimensional scaling procedures made use of judgments on the metadimensions of similarity of preference rather than judgments on specific psychological dimensions (such as asking whether something is good or bad), the elicited responses were considered to be less artificially constrained by the investigator, and hence more likely to reflect the true nature of the subject's psychological experience. Second, the multidimensional scaling procedures allowed all relevant stimulus dimensions to be extracted in the same operation and to exhibit their relative strengths and interrelationships within the confines of a single model. This feature thus freed researchers of the requirement of specifying dimensions in advance. These two features, rightly or wrongly, served to promote the idea that multidimensional scaling would allow "recovery" of the cognitive structure underlying the judgments.

Fields other than psychology also found applications for multidimensional scaling. Through the use of metadimensions such as proximity or dominance, analysis could be extended to well-defined surrogates, such as indices of overlap or confusion, measures of net migration, rankings of dominance, or coefficients of interaction. The procedure then allowed psychological scaling methodology to be applied in nonpsychological sciences and permitted results to be interpreted, not in terms of psychological processes, but in terms of processes appropriate for the particular field of application.

Evolution. The evolution of multidimensional scaling from its origins in the classical euclidean model to its present status as a more general model can be chronicled in terms of four somewhat overlapping but distinct generations of methodology.

First generation. In the first generation, that of classical multidimensional scaling, the main models were euclidean in nature, with data having the property of ratio scale required. With scaled dissimilarities considered as distances in euclidean space, or with normed similarities considered as cosines of angles in euclidean space, the classical model used linear algebra to construct basis dimensions through a standard eigenvector decomposition of a scalar product matrix. The resulting axes were orthogonal and were ordered in magnitude according to the square roots of the extracted eigenvalues. Solutions were analytic, as opposed to iterative, and produced the full complement of dimensions simultaneously. Violations of metric axioms in the judged data would sometimes give imaginary dimensions, and nonlinearity of data with respect to the underlying structure often generated extra, spurious dimensions. Data for the classical model were usually aggregated over subjects prior to analysis, since single data sets were not considered adequate for the stringent input data requirements. Little was done to scale subjects individually or to relate one subject's pattern of judgments to that of

another. In addition, coordinate dimensions were initially defined solely in terms of the axes of eigenvectors without any consideration for the psychological plausibility of such artificial dimensions.

Second generation. The second distinguishable phase in multidimensional scaling methodology was initiated by the contributions of Roger N. Shepard and has become known as nonmetric multidimensional scaling. Nonmetric multidimensional scaling departed from the classical approach in several respects. First, the procedures incorporated a nonlinear psychological response function directly into the model rather than treating such distortion as simple error or correcting for it prior to the analysis. Although many plausible families of functions, such as exponential or power curves, could have been assumed, the multidimensional scaling packages emerging during this phase made only the very minimal assumption that the psychological function was simple monotonic with respect to the underlying structure. By imposing only ordinal constraints and by allowing input to be transformed as necessary as long as ordinality was preserved, it was found that the fitting of such ordinally transformed data to a space of low dimensionality was accompanied by such a degree of overdetermination that a spatial configuration could be estimated with essentially ratio-level precision.

A further distinctive attribute introduced with nonmetric multidimensional scaling was the algorithmic basis of the solution procedure. Most such procedures were iterative in nature, since the determination of the configuration and estimation of the function by monotone least-squares regression required a succession of approximations programmed on the computer. The main computational problem was the determination of an appropriate loss function for comparing the original to the transformed data, and the determination of an efficient procedure for minimizing the function. Since there were several possibilities for defining the stress, or loss function, as well as a number of heuristics for producing the minimization algorithm, a variety of computer programs emerged. The most widely circulated versions originated from the University of Michigan, the University of North Carolina, and Bell Laboratories. The aims of the various approaches were alike in the sense of attempting to fit a ratio-level geometric model to ordinal-level data. The use of ordinal data relieved the researcher of the requirement to convert original data into ratio-level distance estimates prior to scaling, and was found to be very attractive in psychological research contexts in which data collection procedures could not hope to meet the stringent ratio-level requirements of classical scaling.

A related feature of the nonmetric multidimensional scaling phase was a pluralism resulting from the proliferation of options. Input observations could be judgments of similarities or dissimilarities, conditional or unconditional, complete or partially missing, and so on. The fitted model could be a distance model or a scalar product model; it could be euclidean or noneuclidean, or it could be symmetric or nonsymmetric. And for the fitting procedure,

decision criteria could differ in the initialization, convergence, or termination phases, in the choice of a loss function, or in the specification of dimensionality, to mention only the most relevant. As a result, the hope of agreeing on standard criteria for a unique solution was nearly abandoned. The abundance of options gave rise to a curiously paradoxical state of affairs wherein nonmetric multidimensional scaling seemed at one and the same time to be very precise by producing ratio-level solutions from ordinal data, yet rather imprecise by allowing a whole range of such solutions to be considered about equally acceptable for a given set of data. While the variants of a solution for a given set of data were often highly similar to one another, there was still no easy way to assess the superiority of any one solution over another. Thus, the proliferation of possibilities encouraged a certain permissiveness or tolerance with respect to solutions that seemed to be at variance with one another.

Third generation. In the third generation, that of individual differences models, scaling capability expanded to three-way analyses with the estimation of weights for each subject on individual dimensions. By introducing an additional set of parameters to the estimation process, the analysis was forced to abandon the gradient method of function minimization of the previous stage in favor of using an alternating least-square procedure. This breakthrough, typified by the work of J. Douglas Carroll, offered answers to two long-standing problems in multidimensional scaling. First, inaccuracies and inconsistencies caused by aggregating over subjects were reduced by giving each subject an individual place in the model. Not only that, but by including all subject parameters as an integral part of a single model, a more useful solution could be obtained to give a group space representing a general trend, plus individual spaces with weighted dimensions. A second, and possibly more remarkable, advance was the “dimensional invariance” that offered an answer to the rotation problem. Thus, in three-way weighted euclidean models, the extracted group dimensions were found not to be arbitrary, as they had been in classical and nonmetric multidimensional scaling, but rather, they were fixed by consistent variation underlying the individual subjects’ judgments. A further benefit arising from the introduction of three-way analyses was that the technical capability of multidimensional scaling was brought closer in line with current developments in the field of numerical analysis, thereby demonstrating that large parameter sets could be estimated without problems of local minima, and suggesting that other multiple parameter sets could be proposed and successfully programmed by using alternating least-squares procedures. Three-way scaling thus opened the way for a development toward greater computational complexity.

Fourth generation. The fourth and current generation of multidimensional scaling is that of complex model development. Work of this nature actually goes back to the early studies proposing the city-block metric as an alternative to the euclidean metric. As the current phase is gaining momentum, its emphasis is not

so much on properties of the input data or on the efficiency of the fitting algorithm, but on the adequacy of the structural model itself for the representation of the kinds of psychological data found in complex stimulus domains. Throughout the development of multidimensional scaling, there has been an acknowledgment that cognitive structures were more than euclidean in nature, and that in certain cases the euclidean spatial model might be totally inappropriate. Early efforts to capture the complexity through noneuclidean spatial modeling were thwarted, however, by lack of computational procedures powerful enough to fit hypothesized structures without being susceptible to degenerate solutions, premature termination of the algorithm at local minima, and other technical problems. In testing two alternative models on the same data, it was often difficult to determine whether one model or the other provided the better fit. But with computation problems and numerical analysis now more under control, the emphasis has turned to development of noneuclidean spatial models and nonspatial geometric models, as well as the exploration of hybrid models that incorporate elements of taxonomic and continuous scaling. To complement previous work on scaling of Minkowski p metrics, research has focused on models for riemannian space, on discrete algebraic models, and on models that are especially appropriate for distinctive feature domains. With increased interest in detailed exploration of cognitive processes, more emphasis has been placed on cluster analysis procedures and on possible means for integrating hierarchical and taxonomic structures within traditional scaling models.

Richard Degerman

Bibliography. J. D. Carroll and P. Arabie, *Multidimensional scaling*. *Annu. Rev. Psychol.*, 31:607-649, 1980; P. E. Green et al., *Multidimensional Scaling: Concepts and Applications*, 1989; W. S. Torgerson, *Theory and Methods of Scaling*, 1958; F. W. Young and R. M. Hamer (eds.), *Multidimensional Scaling: History, Theory and Applications*, 1987.

Multilevel control theory

An approach to the control of large-scale systems based on (1) decomposition of the complex overall control problem into simpler and more easily managed subproblems and (2) coordination of the subproblems so that overall system objectives and constraints are satisfied.

The controllers are organized in a multilevel hierarchical structure according to three basic criteria: functional, plant, and temporal decomposition. Functional and temporal decompositions are often classified as multilayer or vertical structures; plant decomposition is often classified as a multilevel or horizontal structure.

Functional decomposition. In this approach the overall control problem is partitioned into a nested set of generic control functions (**Fig. 1**). The regulatory or direct control function interfaces with the plant to implement the decisions of the optimizing controller inputted in the form of setpoints,

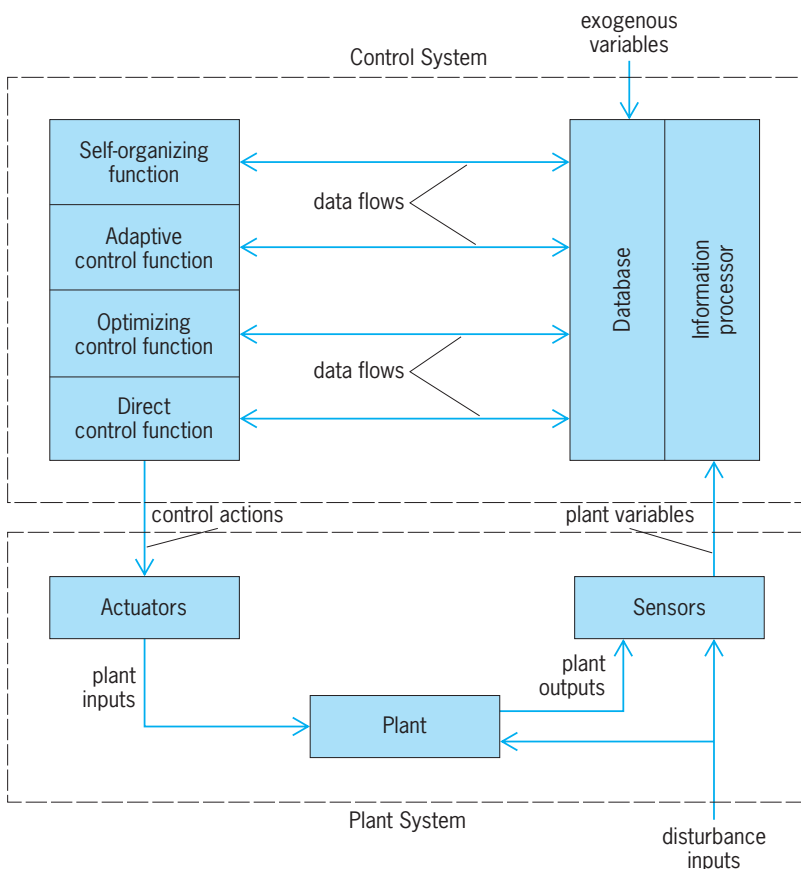


Fig. 1. Functional control hierarchy.

desired trajectories, or targets. The optimizing control function determines the necessary relationships among the variables of the system to achieve an optimal (or perhaps suboptimal but satisfactory) performance based on a given approximate (simplified and aggregated) model of the plant and its environment. The adaptive control function updates parameters of the optimizing control model to achieve a best fit to current plant behavior, and updates parameters of the direct controller to achieve good dynamic response of the closed-loop system. Finally, the self-organizing function is concerned with changing the structure of the control system (for example, nonparametric changes in control, decision-making and information-processing algorithms) in response to changes in system objectives, market conditions, plant characteristics (induced, say, by replacement of obsolete components), major contingency events, and so forth. See ADAPTIVE CONTROL; OPTIMAL CONTROL THEORY.

An alternate formulation of the functional hierarchy is characterized by the following distinct layers of control and information processing: direct control, supervisory control, intraregion coordination, production scheduling and operational management, and management information. This formulation differs from the previous one (Fig. 1) in the explicit identification of management information and decision-making functions (level 4), human-machine interfaces (operator, supervisor, and man-

ager consoles), and couplings with other subsystems (resulting from, say, a horizontal decomposition of the overall system).

Plant decomposition. In this multilevel approach the controlled system (plant) is partitioned into subsystems along lines of weak interaction. In a two-level control hierarchy (Fig. 2), each subsystem has its own (first-level) controller which acts to satisfy local objectives and constraints. A second-level controller (coordinator) influences the actions of the local controllers to compensate for subsystem interactions so that overall objectives and constraints are satisfied. The approach extends readily to a hierarchy with three or more levels.

The multilevel control problem is typically formulated as a constrained optimization problem where the original complex problem is replaced by a number of optimization subproblems of reduced dimensionality. This simplifies the analysis, and it may also result in reduced computational effort.

There are several methods advanced for coordinating the subproblem solutions. In the goal coordination method (also known as the interaction balance or nonfeasible method), Lagrange multipliers enter into the subsystem cost functions as "shadow" prices. These are adjusted by the second-level controller in an iterative procedure which culminates (if the method is applicable) in the satisfaction of the subsystem coupling relationships. In the interaction prediction method (the feasible method), the interaction variables are specified by the second-level controller according to overall optimality conditions. The subproblems are solved to satisfy local optimality conditions constrained by the specified values of the interaction variables. Again, an iterative procedure is implied.

As far as the plant is concerned, only the final result of the iterative procedure is important. However, since the optimum depends on the exogenous inputs to the system (for example, disturbances), and these are assumed to vary with time, the solution process must be repetitive. Thus, in on-line control applications of the multilevel structure:

1. The first-level controllers compensate for local effects of variations of the exogenous inputs; that is, they maintain local performance close to the optimum while ensuring that local constraints are not violated.

2. The second-level controller modifies the criteria or the constraints, or both, for the first-level controllers in response to changing requirements on the system, so that actions of the local controllers are consistent with the overall objectives of the system. In effect, the second-level controller compensates for the mean effects of variations in the interaction variables.

3. The control signals generated by the first-level controllers represent computed optimizing control conditions to be implemented on the plant. In general, these are implemented via direct controllers with the control signals as setpoints.

Temporal decomposition. In this approach, the control or decision-making problem is partitioned into

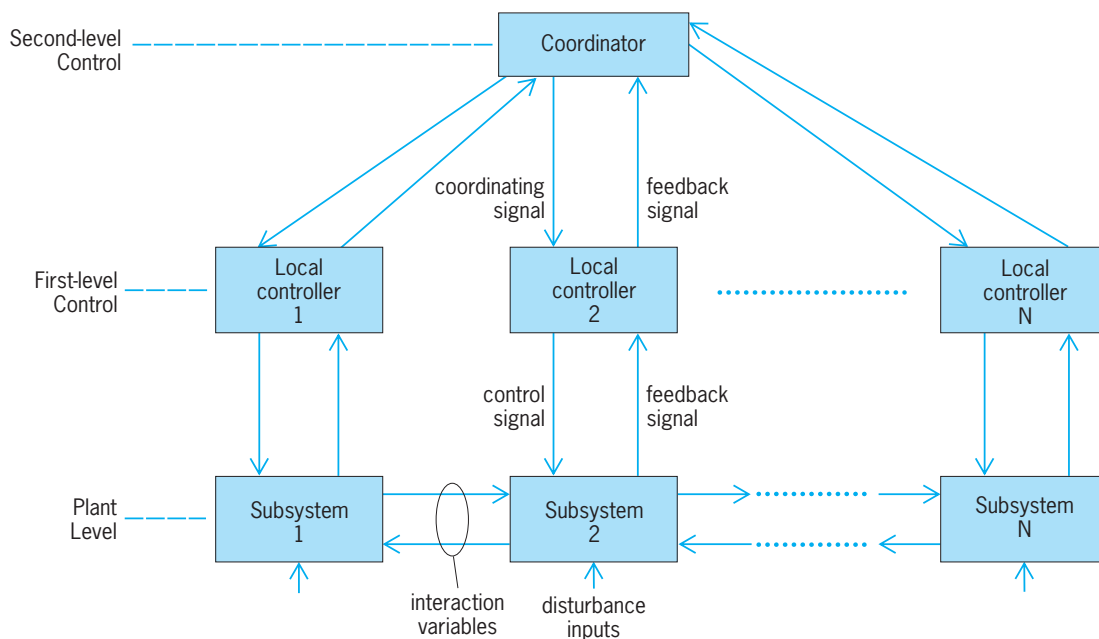


Fig. 2. Two-level control hierarchy.

subproblems based on the different time scales relevant to the associated action functions. These time scales reflect such factors as the response time of the plant, the bandwidth characteristics of the disturbance inputs, and trade-off considerations relating the benefit of control action to its cost. At each layer of the hierarchy, decisions and control actions determined at higher layers (corresponding to lower-frequency events) are treated as constants; disturbances and actions associated with lower layers are treated as noise to be represented by their respective mean values. The temporal hierarchy may embrace a broad spectrum of control and decision-making activities, including process control, production control, scheduling, and planning functions. Time scales may range from seconds to years. Other characteristics of the temporal hierarchy relate to the form of the model, the degree of uncertainty involved in the decision making, the level of aggregation, the information flow requirements, and so forth. In a typical production system hierarchy, three levels of decision making are defined based on time scale: medium-term production planning at the upper level; short-term production planning at the middle level; and scheduling of machines and processes at the lower level. *See* PROCESS CONTROL; PRODUCTION ENGINEERING.

Nontraditional control technologies. Multilevel structures may be used to integrate different knowledge-based and nontraditional control technologies (such as fuzzy control, neural-net control, and expert systems) within a given systems control application. The choice of technology may be based on time-scale requirements, nature of the knowledge base, or characteristics of the control problem. Thus, in one application, fuzzy control is applied at the first level, rule-based control at the second level, and algorithmic-based decision-making

at the third. In another application, direct control is carried out by conventional linear controllers under the supervision of a heuristic rule-based controller at the second level. In another structure, a heuristic approach based on fuzzy set theory is applied to the lower three layers of a production planning and scheduling hierarchy organized according to time scale. The fuzzy decision system helps to reduce the effort at the knowledge engineering and design stage, simplify the computer implementation, provide an effective means of handling uncertainties inherent in the problem formulation, and improve coordination between the human operator and the computer control system. *See* EXPERT SYSTEMS; FUZZY SETS AND SYSTEMS; NEURAL NETWORK.

One feature of expert-systems methodology is that it permits the human-machine interface to be shifted to higher layers or levels of the control hierarchy. Therefore, the ability to incorporate heuristics, judgmental, and generally nonquantifiable factors into the decision-making problem means that more of the tasks requiring human intervention may be embedded into the computer control structure.

In the traditional formulation of the hierarchical control/decision-making problem, each subsystem is characterized by a single objective function. However, this simplifying assumption often leads to unsatisfactory or unrealistic results because of multiple, conflicting, or nonquantifiable objectives. In computer-integrated manufacturing systems, for example, the objectives of maximizing product quality, productivity, profit, and production flexibility are often in conflict. The problem is handled by decomposing it into a series of independent multiobjective subproblems arranged in a hierarchy with vertical and horizontal interdependencies. The individual subproblems are solved by standard multiobjective methodologies, in terms of constraints imposed by

the supramal unit (local manager) and feedbacks provided by lower-level units. An interactive approach is applied to handle the vertical interactions, for example, adjusting upper and lower bounds on available resources to achieve local feasible and efficient solutions.

Applications. Hierarchical control has become accepted technology in many industries. Even when the hierarchical structure is not explicitly identified, it is often implicitly embedded in the design of the system, providing the conceptual framework for integrated (or plant-wide) systems control. Motivating considerations are improved productivity, operating efficiency, product quality, or other economic-based objectives. Following is a sampling of industrial applications. The major factor contributing to the economic feasibility of these applications has been the availability of modern, highly sophisticated, distributed microprocessor-based control systems. Hierarchical control has been applied to various nonindustrial systems, such as robotics and free-way systems.

Steel and process industries. Some of the most advanced applications of multilayer and multilevel control concepts may be found in the steel industry. Tasks are assigned to a hierarchy of computers according to production unit, time scale, and the nature of the task with respect to computational and informational requirements. Similar applications may be found in chemical, oil, paper, and other large-scale process plants.

Electric power system. The system decomposes naturally with the following levels explicitly identified.

First level: Generating units (steam generator, turbine, electric generator).

Second level: Set of generating units within a power station.

Third level: Set of power stations belonging to a utility company.

Fourth level: Group of utilities interchanging power.

In addition, a temporal hierarchy may interleave the multilevel structure with successive layers (at increasing time scales) associated with economic dispatch, unit commitment, production scheduling, and maintenance scheduling. *See* ELECTRIC POWER SYSTEMS.

Batch processing. Special features of hierarchical control applied to batch process systems (for example, specialty chemicals and pharmaceuticals) include dynamic optimization and superposition of such discrete decision-making functions as scheduling, time allocation, and production planning.

Multimode operations. An occasional requirement of many production systems is operation in different processing modes or stages, with each mode characterized by a distinct set of objectives, constraints, and process relationships. Typical examples are startup and shutdown sequences for large or critical processing units (for example, distillation columns and turbine-generator units), and restorative control of an electric power system to normal operation following a contingency event (such as a generator outage or line fault). In these applications, multilevel con-

trol structures provide effective means of controlling the system within each successive mode, as well as coordinating the transitions from one mode to the next.

Robotics. The control problem associated with robotics lends itself naturally to a multilayer decomposition. Typically, first-layer servocontrollers have responsibility for positional and velocity control of the motion components of the robot. The second level coordinates the actions of the servocontrollers so that the robot achieves the specified task or trajectory. The task specifications are determined at the third level so as to optimize overall performance, satisfying constraints (such as avoiding obstacles), and considering interactions with other robots. Finally, at the fourth level, adaptive modifications are incorporated to respond to changing objectives, environmental conditions, and so forth. *See* ROBOTICS.

Manufacturing systems. These systems are characterized by discrete event models and logic-based controls. Computer-integrated manufacturing (CIM) is the conceptual basis for integrating the information flows and decision-making associated with product design, production planning and scheduling, and plant operations. As in the case of integrated control of process-type systems, a hierarchical control architecture forms a natural basis for implementing the integration. In a typical multilevel scheme, individual machines and processes are controlled at level 1; at level 2, related or sequential machines or processes are coordinated to achieve optimum sequencing, maximum throughput, and efficient use of resources; finally, at level 3, the full system is brought together by using resource planning, maintenance planning, and other systems tools.

A natural extension of the hierarchical approach to computer-integrated manufacturing is to automate the entire process from product specification to manufacture by integrating computer-aided design (CAD), computer-aided process planning (CAPP), and computer-aided manufacturing (CAM). This capability reduces workforce requirements and the time required to implement product changes, a valuable asset in job shop production and in achieving rapid response to changing market demand. This approach can be extended to batch processing systems (as in pharmaceutical production). An important step in this direction is the development of a formal approach to the integration problem, that is, the development of generalized formal models, procedures, and protocols for implementing the integration for different classes of manufacturing and production systems. *See* COMPUTER-AIDED DESIGN AND MANUFACTURING; COMPUTER-INTEGRATED MANUFACTURING.

Air-traffic management systems. A multiscaled hierarchical approach has been considered for the next generation of air-traffic management systems (ATM). The proposed structure consists of a horizontal dimension which relates to geographical air space partitions (or control sectors) and a vertical dimension associated with space-time scale decomposition of the decision-making function. The horizontal dimension provides for coordination of the interfaces

between air space partitions with respect to individual users and ground-based control. The vertical hierarchy comprises, at the first level, air-traffic control (ATC), which is concerned with control of individual flights and their interfaces; at the second level, air space traffic flow management, which sets conditions and constraints for air-traffic control; and at the third level, air space management, which oversees the overall air space structure and sets rules assuring successful operation of the lower decentralized levels of control. Increasing uncertainties at each higher level (corresponding to increasing time horizons) are handled by aggregation of the information provided by the lower levels. See AIR-TRAFFIC CONTROL; CONTROL SYSTEMS; DISTRIBUTED SYSTEMS (CONTROL SYSTEMS).

Irving Lefkowitz

Bibliography. S. B. Gershwin, Hierarchical flow control: A framework for scheduling and planning discrete events in manufacturing systems, *Proc. IEEE*, 77:195–208, 1989; Y. Y. Haimes et al., *Hierarchical-Multiobjective Analysis of Large-Scale Systems*, 1990; I. Lefkowitz, Integrated control of industrial systems, *Phil. Trans. Roy. Soc. London*, A287:443–465, 1977; S. N. Salthe, *Evolving Hierarchical Systems*, Columbia University Press, New York, 1985; S. G. Tzafestas and J. K. Pal (eds.), *Real Time Microcomputer Control of Industrial Processes*, Kluwer Academic, Dordrecht, 1990; T. J. Williams, *Analysis and Design of Hierarchical Control Systems*, 1985.

Multimedia technology

Computer-based, interactive applications having multiple media elements, including text, graphics, animations, video, and sound. Multimedia technology refers to both the hardware and software used to create and run such systems.

Applications. The earliest multimedia programs, in the 1970s, were arcade and home video games. Their popularity, combined with the introduction of personal computers in the late 1970s and early 1980s, led to more sophisticated and colorful computer-based games. The development of new hardware then made possible CD-ROM (compact disc-read only memory) technology for storage and distribution of programs requiring large amounts of data, such as encyclopedias and other reference works. New technology also permitted the proliferation of graphics, sound files, and the earliest virtual reality programs. The 1990s brought the Internet and the World Wide Web (WWW) into popular use, and this broadened the audience for multimedia applications of all types.

By the beginning of the twenty-first century, the favorite and most widely used multimedia systems still were those designed for entertainment and games. However, multimedia technology is also used to create extensive reference works, such as encyclopedias, dictionaries, and databases. Businesses use multimedia for briefings, advertising, and sales presentations. Teachers and corporate trainers employ multimedia to design lecture aids or tutorial programs.

The mode of delivery for each application depends

| Application | Most common delivery mode |
|---|---------------------------|
| Briefings, presentations, and advertising | Disk, WWW |
| Reference and databases | CD-ROM, DVD, WWW |
| Museums, education, and training | Kiosk, Disk, CD-ROM |
| Entertainment and games | CD-ROM, DVD, WWW |

on the amount of information that must be stored, the privacy desired, and the potential expertise of the users. Applications that require large amounts of data are usually distributed on CD-ROMs, while personal presentations might be made directly from a computer using an attached projector. Advertising and some training materials are often placed on the WWW for easy public access. Museums make use of multimedia kiosks with touch screens and earphones. A listing of the most common delivery modes used by each kind of application is given in the **table**. See COMPACT DISK; INTERNET; VIDEO GAMES; WORLD WIDE WEB.

Hardware. Multimedia products may be created and run on the commonly used computer environments, such as Microsoft Windows[®] and Apple Macintosh.[®] Multimedia system users may employ a variety of input devices in addition to the keyboard and mouse, such as joysticks and trackballs. Touch screens provide both input and display capabilities and are often the choice when potentially large numbers of novices may use the system. Other display devices include high-resolution monitors and computer projectors. Generally the abundance of graphics and video in multimedia applications requires the highest resolution and deepest color capacity possible in display devices.

Input devices for the creation of multimedia applications include graphics tablets, which are pressure-sensitive surfaces for drawing with special pens; digital cameras, which take pictures electronically; and scanners, which convert existing pictures and graphics into digital form. Other hardware devices, such as a video card and video digitizing board, are required both to create and to play digital video elements.

The hardware for incorporating sound elements into multimedia systems includes microphones, voice-recognition systems, sound chips within the computer, and speakers, which come in a wide variety of forms with varying capabilities and quality. See SPEECH RECOGNITION.

When all the elements of multimedia are united in a single application, the result is a large file requiring a high-capacity storage medium. Large, mass-produced products are usually packaged using CD-ROM or DVD (digital video disc). CD-ROM players and recorders became commonplace in the early 1990s and enabled the delivery of high-capacity products. DVDs, introduced in the middle 1990s, are capable of even larger storage capacity and full-motion video combined with high-quality audio.

Multimedia applications can also be presented using the WWW. The advantage is that users are

not required to have CD-ROM or DVD players on their computers. Moreover, content can be updated frequently. The disadvantage is that the large multimedia applications result in long download time and uneven quality, especially for users with low-bandwidth connections. As the technology surrounding the WWW improves, this kind of delivery should become more common. *See* COMPUTER GRAPHICS; COMPUTER PERIPHERAL DEVICES; COMPUTER STORAGE TECHNOLOGY; ELECTRONIC DISPLAY.

Component development software. Specific software is needed to develop each component of a multimedia program, such as text, graphics and images, animation, digital audio, and digital video.

Text. Text helps the user navigate through the multimedia system; explains pages, screens, or pictures; and communicates the information or content of the program. Designers vary the effect of text by using different fonts, sizes, and colors available in word processing programs. The weight or thickness of each letter greatly affects readability. Text can also be encoded on WWW pages. Many software programs are available to create unusual lettering that is shadowed, warped, curved, or fuzzy. The basic concept in text design is that it should be clear and readable over the background. It should contribute to the intended communication, not detract from it.

Graphics and images. In multimedia presentations the images should be clear and concise without distracting lines and backgrounds. Complex designs should be used only if they are central to the theme of the system. There are two kinds of image files: vector images and bitmaps. Vector images define the picture using mathematical properties of coordinates, length, or size. They result in very small files, but the pictures are not photo-realistic. Bitmap files represent the image as a series of dots or pixels and store information about the color of each pixel. Bitmap images produce high-resolution, clear images, but the files are quite large. Compression is often used to reduce the file size. Some popular formats for compressed files are Graphics Interchange Format (GIF) and Joint Photographic Experts Group (JPEG and JPG). Software is available that allows the designer to draw vector images, create the bitmaps, scan designs or pictures created elsewhere, and edit or crop any digital image. *See* DATA COMPRESSION.

Animation. Images can be moved across the screen or displayed quickly in sequence to create animation. Many editing programs are available to help the user design computer animation files. In fact, most software packages that allow the editing of GIF files also feature the ability to create animated GIF files in a compressed format. Some software allows the user to specify the beginning and ending point for the animation, and then the program fills in the animation required in the middle. Single images can be warped, stretched, or manipulated in any way, and multiple images can be edited, transformed, or morphed, changing from one image to another.

Digital sound. The addition of sound to a multimedia program adds dramatically to its appeal. Computer generated sound has progressed from the

early “beep” to realistic playing of symphonic music. Sound files usually come in one of two formats: WAV and MIDI. Digital audio files (WAV) are similar to bitmap images in that the sound is cut into pieces (“sampled”); and attributes, such as frequency and amplitude of the analog sound wave, are stored about individual samples. When more samples are stored, the file becomes bigger and the resulting sound is more realistic when played. Fewer samples result in smaller files but lower-quality sound. Musical Instrument Digital Interface (MIDI) files are used for the recording and playing of musical instruments on the computer. MIDI files contain header sections with information about the arrangement and timing of the music, and one or more tracks are used to combine the sounds of several instruments. MIDI files are smaller than WAV files, but are more difficult to create, are more dependent upon the hardware for quality, and are not appropriate for spoken dialog. WAV files, however, can be used for any sound and have a more consistent quality when played, but are difficult to edit. *See* MUSICAL INSTRUMENTS; VOICE RESPONSE.

Digital video. Digital video combines the technologies of graphics, animation, and sound and is usually the most challenging part of a multimedia system. Video cameras capture the analog signal of color, synchronization, and sound. Video capture cards then convert the video into digital format that can be played back on the computer.

Newer video cameras are available to record directly into digital form. Also called camcorders, these cameras are capable of capturing both high-resolution digital video and still photos. It is possible to record and transfer the video or still images directly to a computer, TV, or VCR. Software can then be employed to copy or edit the video, or to add special effects.

The quality of video is measured by the frame rate (number of frames per second). A higher frame rate yields better-quality video but larger files. Usually the designer attempts to use the smallest acceptable frame rate to keep the file size as small as possible but maintain adequate quality. The file size is also affected by the frame size, or the window size of the video as it is displayed. Full-screen video of television quality and sound results in unmanageably large file sizes. Therefore most digital video makes use of compression to reduce the data stored in the video file. There are two kinds of compression/decompression programs: lossy and lossless. Lossless compression results in quality equivalent to the original during the playback, but the files are larger. Lossy compression algorithms result in a loss of quality between the original video and the video that has been compressed and decompressed for playback. However, lossy algorithms are usually acceptable for digital video sections of multimedia systems. The two main file standards are JPEG and MPEG. JPEG was developed for use with still images but is useful also for video. Moving Picture Experts Group (MPEG) files are widely used for video for WWW and DVD applications.

Graphical user interface. Central to the multimedia system is the graphical user interface. Elements such as buttons, checkboxes, and text boxes allow the system to be interactive. Usually these graphical elements are created and combined with text and/or images for communication purposes. These features also allow users to have more control of their experience with the multimedia application. *See* HUMAN-COMPUTER INTERACTION.

System development software. Once all the individual elements have been prepared, the entire system must be designed and assembled. The most obvious way to create the overall program is to write the entire program using a programming language. The programmer has complete control of the finished product. Many modern programming language environments allow easy creation of a graphical user interface and enable the programmer to include both sound and images with a minimum of effort. However, this kind of programming usually requires considerable skill and experience.

The easiest way to assemble a multimedia system is to use presentation software. Business people with no programming experience can create multimedia presentations featuring text, animation, and sound in a short period of time. However, this kind of system development limits the finished product to the features provided by the particular software.

Authoring software provides an alternative to both these platforms. The developer needs to learn to use the software, which takes much less time than learning computer programming, but still retains control over the system being designed. Some authoring software enables the user to piece together a program by dragging icons into a flowchart. Other software allows the user to specify what happens at each point in the system time line. Whatever kind of software is used, the finished product is the result of a wide variety of segments woven together into a multimedia system.

Trends. The future of multimedia technology is dependent upon the evolution of the hardware. As storage devices get faster and larger, multimedia systems will be able to expand, and increased use of DVD should result in improved quality. Rising network speeds will increase the possibility of delivering multimedia applications over the WWW. Currently, Virtual Reality Modeling Language (VRML) is used for some WWW applications and may drastically expand the multimedia experience. Virtual reality is becoming more realistic and will stretch the multimedia experience to envelop the user. The one certainty in multimedia technology is that it will continue to change, to be faster, better, and more realistic. *See* VIRTUAL REALITY.

Pauline K. Cushman; Robert A. Kolvoord

Bibliography. E. England and A. Finney, *Managing Multimedia: Project Management for Interactive Media*, 2d ed., Addison-Wesley, 1999; D. Hillman, *Multimedia Technology & Applications*, Delmar, 1998; D. D. Peck, *Pocket Guide to Multimedia*, Delmar, 1999; T. Vaughan, *Multimedia: Making It Work*, 3d ed., Macromedia, 1996.

Multimeter

An instrument designed to measure electrical quantities. A typical multimeter can measure alternating- and direct-current potential differences (voltages), current, and resistance, with several full-scale ranges provided for each quantity. Sometimes referred to as a volt-ohm meter (VOM), it is a logical development of the electrical meter, providing a general-purpose instrument. Many kinds of special-purpose multimeters are manufactured to meet the needs of such specialists as telephone engineers and automobile mechanics testing ignition circuits. *See* AMMETER; CURRENT MEASUREMENT; OHMMETER; RESISTANCE MEASUREMENT; VOLTAGE MEASUREMENT; VOLTMETER.

Multimeters originated when all electrical measuring instruments used analog techniques. They were generally based on a moving-coil indicator, in which a pointer moves across a graduated scale. Accuracy was typically limited to about 2%, although models achieving 0.1% were available. Analog multimeters are still preferred for some applications. For most purposes, digital instruments are now used. In these, the measured value is presented as a row of numbers in a window. Inexpensive hand-held models perform at least as well as a good analog design. High-resolution multimeters have short-term errors as low as 0.1 part per million (ppm) and drift less than 5 ppm in one year. Many digital multimeters can be commanded by and send their indications to computers or control equipment.

Analog multimeters. The core of the meter is the indicating element, chosen to be a compromise between sensitivity and robustness. A d'Arsonval movement requiring 1 mA direct current for full-scale deflection can withstand rough treatment, but draws relatively high currents when measuring voltages. Voltage ranges are provided by a selection of series resistors; their values are proportional to the required full-scale voltage. In the above example, this value is 1 k Ω /V; this constant provides a convenient quality rating for multimeters. A highly sensitive but relatively fragile meter might use a 20- or 10- μ A full-scale deflection movement. A frequent compromise is a 50- μ A movement, giving a 20-k Ω /V rating. Current ranges are achieved by rearranging the connections to the movement so that it appears across the current-measuring terminals, together with a choice of switch-selectable shunts. Alternatively, a series of current range sockets may be provided in conjunction with a so-called universal shunt. *See* SHUNTING.

Resistance measurement requires an internal source of current, usually a dry battery with a terminal voltage V of 1.5–15 V. The battery, movement, and measuring terminals are connected in series, together with an adjustable limiting resistor R_1 (**Fig. 1a**). The limiting resistor is adjusted with the terminals shorted to bring the meter deflection to full scale, which is marked zero on the ohms ranges. The larger the resistance connected between the measuring terminals, the smaller the meter deflection, which is equal to $V/(R_1 + R_x)$. The resulting scale

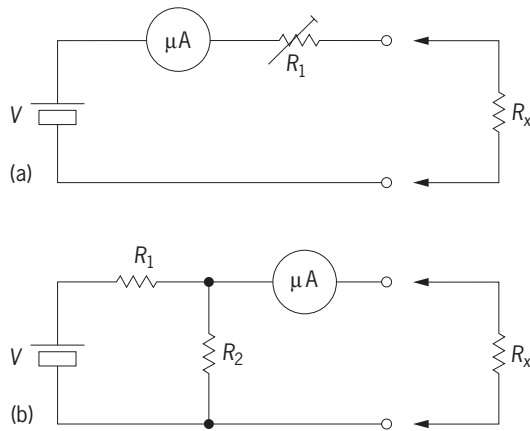


Fig. 1. Circuits for resistance measurements with an analog multimeter over (a) normal-resistance range and (b) low-resistance ranges.

resembles a logarithmic shape, which conveniently allows a wide range of resistances to be measured on a single range. Other ranges can be obtained with the help of a second battery of a higher voltage or a potential divider (Fig. 1*b*).

Digital test multimeters. Digital multimeters intended for test purposes, often described as hand-held models, are closely related to their analog predecessors and are intended for similar use. They are small, operate from batteries, and provide a digital display that is driven by an analog-to-digital converter. Some instruments also include a quasianalog bar display. In a common version of the dual-slope analog-to-digital converter, the analog input signal is integrated for a fixed time, and the integrator is restored to zero by the application of a reference current. The time taken is measured by counting clock cycles, giving the required digital result. At very little extra cost, a microprocessor can be built into the instrument, allowing a variety of features, such as a bar graph display or automatic ranging. See ANALOG-TO-DIGITAL CONVERTER; MICROPROCESSOR.

A power source and active circuits can improve the sensitivity considerably. A typical hand-held unit includes an operational amplifier with an input resistor R_{in} of 10 M Ω to provide all dc ranges (Fig. 2*a*). Different voltage ranges are provided by switching a feedback resistor R_f , by using field-effect transistors to avoid mechanical contacts and simplify automatic ranging. For ac ranges, a lower value of input resistor may improve the frequency response. See OPERATIONAL AMPLIFIER; TRANSISTOR.

Current ranges are provided by measuring the voltage drop across a suitable shunt resistor, using the most sensitive voltage configuration. Restrictions in the power-handling capability of a small shunt limit the current measurement capability of most instruments to 10 A. Resistance ranges are provided by a rearrangement of the input amplifier (Fig. 2*b*). The unknown resistor is used as the feedback element for an operational amplifier which receives a suitably scaled input current from a reference source. The amplifier output is therefore linearly related to the resistor value. A range of reference cur-

rents enables a wide variety of resistor values to be measured.

Digital systems multimeters. Systems digital multimeters have totally different requirements. The additional complexity implies higher power consumption, with supplies drawn from the power line. There is little need for economy, except that any excessive heat developed may make the instrument less stable and reliable.

The resolution, linearity, and stability of several designs enable them to make dc measurements within 1 ppm of the standards against which they have been set. This order of accuracy is normally associated with only the best standards laboratories. Although such multimeters contain the same components as hand-held models, the design details of every part are considerably more refined. Special analog-to-digital conversion techniques combine time-division and charge-balance techniques in such a way that errors due to switching transients and dielectric storage are considerably reduced.

Input isolation. The analog-to-digital converter and input circuits are constructed within a screened enclosure (the guard). Their power is provided by carefully isolated transformers. The timing signals that define the measurement result are transmitted to the grounded output circuits through optoisolators or low-capacitance pulse transformers. By this means, the input circuits are well isolated, and the instrument can be used, with certain restrictions, as if it were not operating from the power line. The actual degree of isolation achieved is a measure of the success of the design. It is normally possible for a dc signal applied to both input leads to influence the display as little as a true input signal 140 dB smaller; that is, the common-mode rejection is 140 dB at dc. As it is much more difficult to obtain low capacitance than high resistance, the comparable figure for ac common-mode rejection is much worse, typically 90 dB at 60 Hz, and degrades linearly with frequency.

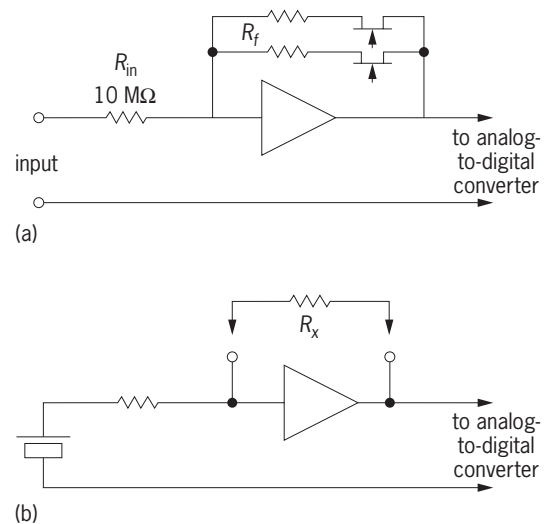


Fig. 2. Digital test multimeter circuits for (a) active input ranging and (b) active measurement of resistance.

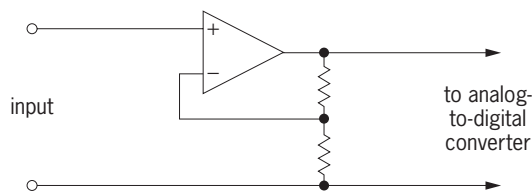


Fig. 3. Circuit for obtaining high input impedance in a digital systems multimeter.

During dc measurements, ac interference is rejected if the integrating time is long or equal to a whole number of periods of the interfering signal. It is also possible to incorporate a filter, resulting in a property defined as normal-mode rejection with a typical value of 70 dB.

Direct-current input circuits. For dc inputs up to about 20 V the signal can be applied to a high-input impedance configuration (Fig. 3). Amplifiers can be constructed with noise levels of the order of tens of nanovolts and input resistances higher than 10 G Ω . The offset current, that is, the steady dc current that flows out of the measurement terminals, should be below 1 picoampere if it is not to be the source of significant errors when measurements are made from source resistances above 10 k Ω . So far, few designs have satisfied this requirement. Resistive attenuators provide higher voltage ranges.

Alternating-current input circuits. For ac inputs, the input arrangement resembles that provided for dc in hand-held meters (Fig. 2a), except that a lower value of resistor is used in conjunction with a far more elaborate amplifier in order to provide better accuracy of gain and a wide bandwidth. An input resistance of 1 M Ω is typical. In addition, an input capacitance of about 150 picofarads results from stray capacitances and guards designed to reduce frequency errors in the input resistor itself. The ac signal is converted into a dc equivalent [invariably the root-mean-square (rms) value] by a thermal or electronic technique. Thermal techniques depend on the heat generated in some device by the ac signal being matched by that generated by a dc signal automatically adjusted by a feedback arrangement. Electronic methods can involve an analog computation using the accurate logarithmic properties of planar transistors or the multiplying properties of configurations in which the transconductance of a gain stage is controlled by the signal. In the most accurate instruments, errors are reduced by carrying out a subsidiary measurement in which a dc equivalent of the ac signal is passed through the same processing path, and using the resulting output to derive corrections for short-term changes in system gain. Sampling methods analogous to those used in digital sampling wattmeters can also be used. See WATTMETER.

Resistance measurements. Resistance measurements are based on principles similar to those of hand-held instruments (Fig. 2b), but refined to improve accuracy by many orders of magnitude. Provision is made for the measurement of lower-value resistors (below 100 k Ω) using four-terminal connections. A further

refinement is to carry out resistance measurement in two stages, first with and then without the test current in order to eliminate errors caused by thermal electromotive forces. The best instruments achieve accuracy of a few parts per million. See THERMOELECTRICITY.

Instrument communication. A vital feature of a systems digital multimeter is its ability to communicate with other instruments. To interconnect instruments from different manufacturers, standard interfaces have been defined, of which two are most commonly found. The IEEE488 bus (also known as the HP-IB and GPIB) carries eight bits of parallel data together with handshaking lines to control data transfer. The RS232 bit serial interface is also widely used, but the control lines and data formats are less well defined, making successful operation a matter of trial and error. USB interfaces are increasingly employed to integrate multimeters under computer control into automated and programmable measurement systems. R. B. D. Knight

Bibliography. W. E. Boyes (ed.), *Instrumentation Reference Book*, 3d ed., 2002; Institute of Electrical and Electronics Engineers, *IEEE Standard for Higher Performance Protocol for the Standard Digital Interface for Programmable Instrumentation*, IEEE Stand. 488.1-2003, 2003; Institute of Electrical and Electronics Engineers, *IEEE Standard Codes, Formats, Protocols and Common Commands for Use with IEEE Std 488.1-1987*, *IEEE Standard Digital Interface for Programmable Instrumentation*, IEEE Stand. 488-2, 1992; International Telegraph and Telephone Consultative Committee (RS232), Stand. V24; B. Kibble et al., *A Guide to Measuring Direct and Alternating Current and Voltage Below 1 MHz*, Institute of Measurement and Control, United Kingdom, 2003.

Multiple cropping

Planting two or more species in the same field in the same year. This general term describes a multiplicity of systems that are as complex as agriculture itself. The first conscious planting by early people consisted of mixtures of cereals, grain legumes, roots, and tubers that were located near permanent dwellings. Perhaps this system of planting was patterned after the natural combinations of trees, shrubs, and other plants from which food was gathered in the wild, or resulted from the unintended growth that originated from seed contained in kitchen refuse that germinated and thrived in early fertile garbage dumps.

Preserved through history to maintain biological, economic, and nutritional diversity, these multiple-species systems still are used by the majority of the world's farmers, especially in developing countries. Where farm size is small and the lack of capital has made it difficult to mechanize and expand, farm families that need a low-risk source of food and income often use multiple cropping. These systems maintain a green and growing crop canopy over the soil

through much of the year, the total season depending on rainfall and temperature. Systems with more than one crop frequently make better use of total sunlight, water, and available nutrients than is possible with a single crop. The family has a more diverse supply of food and more than one source of income, with both spread over much of the year.

Multiple-cropping patterns are described by the number of crops per year and the intensity of crop overlap. Double cropping or triple cropping signifies systems with two or three crops planted sequentially with no overlap in growth cycle. Intercropping indicates that two or more crops are planted at the same time, or at least planted so that significant parts of their growth cycles overlap. Relay cropping describes the planting of a second crop after the first crop has flowered; in this system there still may be some competition for water or nutrients. When a crop is harvested and allowed to regrow from the crowns or root systems, the term ratoon cropping is used. Sugarcane, alfalfa, and sudangrass are commonly produced in this way, while the potential exists for such tropical cereals as sorghum and rice. Mixed cropping, strip cropping, associated cropping, and alternative cropping represent variations of these systems.

Efficient resource use. Natural plant communities are highly diverse, and those containing the largest number of species are found in the tropical lowlands. In these ecosystems, depending on rainfall, there is cover over the soil through all or most of the year. In the absence of some cataclysmic event, these systems may be sustained without human intervention over many centuries. Multiple-cropping systems copy some of the characteristics of these natural ecosystems, especially in the use of resources.

Potato, maize, and climbing beans are intercropped by farmers in the Rio Negro area of Antioquia Department in Colombia (Fig. 1). Potatoes are planted in January in this highland valley at about 7200 ft (2200 m) elevation, and the crop is hilled up in April and maize is planted alongside the potato rows. The potato crop is harvested in May or June, and farmers hill up the maize while digging the potatoes. Climbing beans are planted in July at the base of the maize plants, and bean plants grow up the maize stalks as the season goes on. Both maize and beans are harvested by hand in December and January, residues are returned to the soil, and the cycle begins again.

Figure 1 shows dry-matter accumulation and overlapping growth curves of three crops, potato, maize, and bean, two of which cover the soil during most of the year. These crops intercept sunlight for photosynthesis, break the impact of the pounding rain to reduce soil erosion, and utilize soil moisture and nutrients for growth. Leaf fall and crop residues increase organic matter and contribute to soil fertility. This multiple-species system is self-sustainable over a long period of time, with minimal use of fertilizers and pesticides. A single crop (monoculture), on the other hand, does not use the natural resources available for growth over as much of

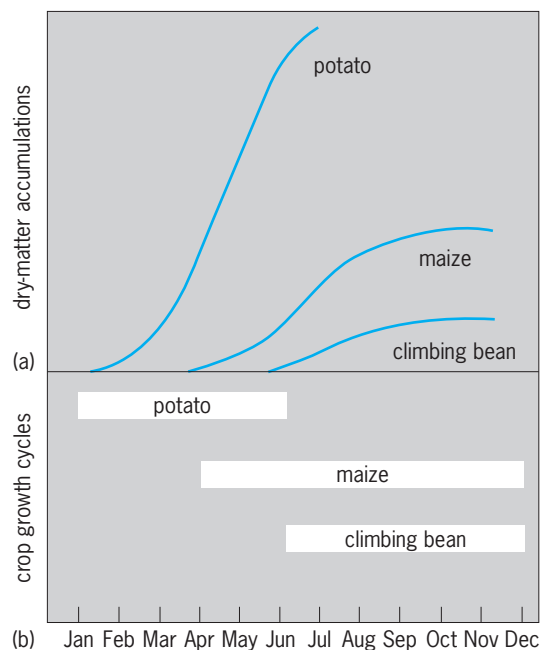


Fig. 1. Graph showing (a) dry-matter accumulation curves and (b) overlapping crop-growth cycles in a multiple-cropping system of potato, maize, and climbing bean, grown in the same field in Rio Negro, Colombia. Each crop growth cycle is indicated from planting date to harvest date.

the total year, nor would a single crop contribute as much residue or provide as much food and income. The multiple-cropping system clearly can take advantage of the available natural resources to sustain good crop growth.

Tropical cropping systems. The potato-maize-bean system described is but one example of the diverse group of crops selected for year-round cropping in the tropics. The crops selected and the intensity of the systems depend on such factors as elevation, rainfall, soil, and also food preferences. In highland areas, maize, sorghum, barley, potato, amaranth, and vegetable crops are often components of these systems. Cycles of these crops overlap, and several species may be found in one field at the same time. Because of lower temperatures at these elevations, crop growth is slow. Traditional maize varieties may require 10 to 13 months to mature, while an improved 7- to 8-month variety or hybrid allows the use of more intensive systems such as the potato-maize-bean combination in Fig. 1.

Lowland tropical systems often are more diverse, with more crops planted together and a shorter growth cycle for each crop. An example from the high-rainfall forest zone in Nigeria is shown in Fig. 2. Raised mounds are established over a period of years, and upland crops such as cassava, yam, sweet potato, and maize are planted on top of the mounds. Rice is planted in the low areas between mounds where water accumulates and helps control weeds. A similar system in Ecuador in the coastal areas that flood each year has rice planted in the lower areas and part of the way up the sides of

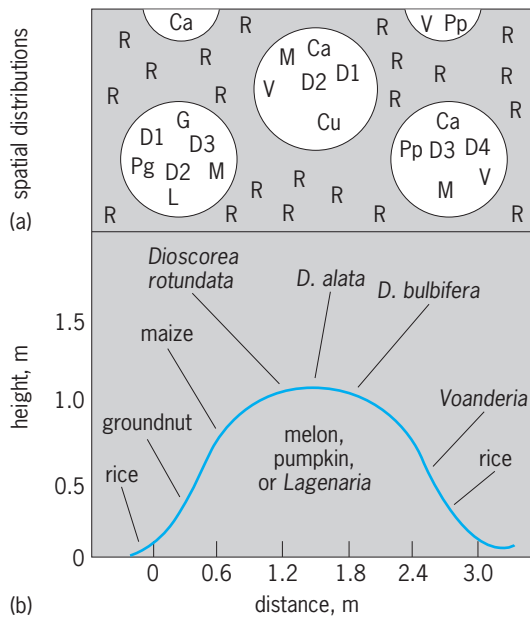


Fig. 2. Diagram showing (a) spatial distributions and (b) heights of various crop species grown on and between raised mounds in Abakaliki, East Central State, Nigeria. M = maize, R = rice, Ca = cassava, Cu = melon, G = groundnut, L = *Lagenaria*, V = *Vigna*, D1 = *Dioscorea rotundata*, D2 = *D. alata*, D3 = *D. bulbifera*, D4 = *D. cayenensis*, Pp = pumpkin, and Pg = pigeon pea. 1 m = 3.3 ft. (After C. A. Francis, *Variety development for multiple cropping system*, *CRC Crit. Rev. Plant Sci.*, 3(2):133–168, 1986)

mounds. Maize is planted on the higher areas and down the slopes to where rice and maize are both planted together. The area floods each year, but the depth of water is not predictable. In years when there is deep water, the rice on the sides of the mounds does well and maize is killed, while in dry years the maize does well and rice does poorly. This is a risk-avoiding strategy used by small farmers. In Asia the *sorjan* system employs the same principle.

Double and triple cropping of rice in southern China began in the twelfth century when short-cycle varieties were selected. In areas with less rain, short-cycle catch crops of mungbean, soybean, or vegetables can be planted after rice to use residual moisture. Cassava is an 8- to 12-month crop that can be established with late rains and survives through the dry season, providing food and income for families in lowland areas. Other crop combinations are sorghum/maize, sorghum/cowpea or peanut, maize/bean, and annual/perennial mixtures. Agroforestry is another dimension of multiple cropping; in this system useful perennial tree crops are interplanted with shorter-cycle crops. There is a nearly infinite series of mixtures that depend on amount and distribution of rainfall, temperature regime, crop adaptation, and food preference in tropical regions. See TERRACING (AGRICULTURE).

Temperate cropping systems. There is less multiple cropping and more monoculture in temperate zones, where mechanization and productivity per unit of labor have shaped agricultural systems. One major exception is pasture systems, such as those in which grasses and legumes are commonly planted together to take advantage of water, nutrients, and light throughout the entire season and to create a more balanced feed source for grazing livestock.

New and less common systems are overseeded or rotated legumes and relay crop patterns of wheat and soybeans (Fig. 3). Overseeding legumes into maize provides ground cover through the fall and spring, produces nitrogen for the next cereal crop, and helps prevent erosion of topsoil by breaking the force of raindrops and contributing surface organic matter to absorb rainfall. This practice is gaining in popularity, with hairy vetch and red clover two of the most common legumes in the system. This sequence is illustrated in Fig. 3.

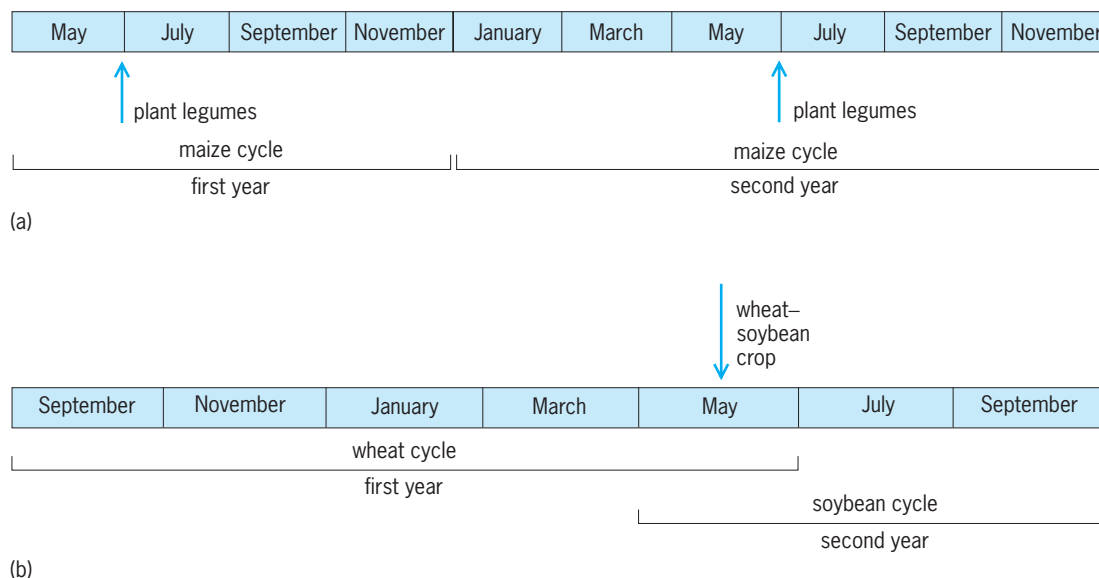


Fig. 3. New multiple-cropping systems for temperate zones in which (a) overseeded legumes in maize and (b) relay-cropped wheat and soybeans have been planted.

Also shown in Fig. 3 is a relay system where soybeans are seeded into growing wheat in May. The wheat is harvested in July, allowing soybeans to grow to their normal maturity in October. If adequate rainfall occurs, two crops can be produced in one year. Other experimental systems are being tested in temperate zones.

Future of multiple cropping. There will be less multiple cropping in the future in some tropical areas as mechanization leads to consolidation of smaller fields in some favored areas. Complex systems will continue to be popular with farmers who seek maximum food production from small areas and those who engage in subsistence farming. Interest by research and extension specialists has stimulated the search for new varieties and for technologies that can improve the productivity of multiple-species systems.

Greater awareness of the potentials of intensive cropping systems in temperate zones will stimulate their adoption by farmers. Pastures that have an appropriate mixture of species can produce more dry matter and reduce risk over a range of different rainfall patterns. Relay systems provide a potential for greater productivity per unit of soil area and divide the fixed costs of production over two crops. Strip crops planted on the contour can help reduce erosion on sloping lands, stabilizing nutrients and capturing rainfall for the next crops. Multiple-cropping systems are likely to remain an important part of agriculture. See AGRICULTURAL SCIENCE (PLANT); AGRICULTURAL SOIL AND CROP PRACTICES; AGRICULTURE; AGRONOMY.

Charles A. Francis

Bibliography. C. A. Francis (ed.), *Multiple Cropping Systems*, 1986; A. A. Gomez and K. A. Gomez, *Multiple Cropping in the Humid Tropics of Asia*, 1983; K. G. Steiner, *Intercropping in Tropical Smallholder Agriculture with Special Reference to West Africa*, 1982.

Multiple proportions, law of

This law states that, when two elements combine together to form more than one compound, the weights of one element that unite with a given weight of the other are in the ratio of small whole numbers. The law can be illustrated by the composition of the five oxides of nitrogen. One gram of nitrogen is combined with 2.85 g of oxygen in nitrogen pentoxide, N_2O_5 ; with 2.28 g in nitrogen dioxide, NO_2 ; with 1.71 g in nitrogen trioxide, N_2O_3 ; with 1.14 g in nitric oxide, NO ; and with 0.57 g in nitrous oxide, N_2O . These numbers are in the simple ratio of 5:4:3:2:1. See DEFINITE COMPOSITION, LAW OF.

Thomas C. Waddington

Multiple sclerosis

A neuromuscular disorder that characteristically involves the destruction of myelin, the insulating material around nerve fibers. The onset of the disease

is unusual in persons under 15 or over 60 years of age, and peak incidence is found in people in their 20s and 30s. Multiple sclerosis affects females more frequently than males by approximately 2:1. Distribution is worldwide, but there is an unusual relationship to latitude, with a much higher incidence at northern latitudes than near the Equator.

Clinical features. The basic mechanism underlying the disease is the destruction of myelin. The effect is analogous to breaking down the insulating material around a telephone wire and exposing the bare wire so that short circuits occur. In multiple sclerosis, only the central nervous system is affected, but both incoming and outgoing processes may be disrupted. Common initial symptoms reflect this underlying disease mechanism. They include blindness in one eye due to disruption of the conduction of the nerve impulse through the optic nerve; weakness of one side of the body due to impairment of the downstream signals from the motor areas of the cerebral cortex through the spinal cord; difficulties with coordination related to problems with cerebellar function; and disturbances in sensation, such as tingling and numbness in an arm or a leg that is related to dysfunction of incoming sensory signals. Multiple sclerosis is a progressive disease, so that over time there is often an accumulation of new symptoms and problems, such as loss of bladder control, difficulty in walking or speaking, and disturbances of mood and cognitive function.

In its classic form, the disease spreads both temporally and anatomically. Temporally, there may be a series of acute attacks, but between attacks a person may recover fully and remain well for some time. Anatomically, areas of disruption of myelin (demyelination) are scattered throughout the nervous system and spinal cord. Thus, symptoms depend upon what part of the nervous system is affected at any given time.

The course of the disease is unpredictable. Some individuals have an acute attack followed by many asymptomatic years. Others have repeated attacks and never quite return to their normal function between attacks. A third group has a chronic progressive form of the disease, usually showing difficulties in walking and use of the hands followed by a subsequent progressive course with no acute exacerbations.

Diagnosis. The diagnosis is based primarily upon the clinical presentation, which makes it very difficult to reach a diagnosis at the time of the first attack. It is the course of the disease that is most helpful in establishing the diagnosis. There are no laboratory tests that define the disease, but abnormalities can aid in the diagnosis. Physiological studies—for example, of the visual, auditory, and sensory systems—can reveal dysfunction of a sensory system that seems clinically normal, indicating that the disease is more widespread than it originally appeared. Areas where myelin has broken down (so-called plaques of demyelination), can be seen by magnetic resonance imaging. However, many people with even a mild form of the disease have the characteristic

lesions of multiple sclerosis on magnetic resonance images.

Lastly, the disturbance of immunological function can be identified by examination of cerebrospinal fluid. There is an increase of immunoglobulins that are synthesized within the nervous system. In addition, the pattern of immunoglobulins shows discrete species of immunoglobulin G (IgG), referred to as oligoclonal bands, which can be detected by electrophoretic studies of cerebrospinal fluid. *See* IMMUNOGLOBULIN.

Etiology. The basic cause of the disease is not known. There clearly are genetic factors in that the incidence of the disease is 20 times higher in first-degree relatives (parents and siblings of an affected individual) than in the general population. If one of a pair of identical twins is affected, the chances of both twins having the disease (the concordance rate) is 28%, whereas in fraternal twins it is 2.5%, which is similar to any group of siblings. Furthermore, certain immunological markers present on lymphocytes are overrepresented in individuals with multiple sclerosis, suggesting a genetic predisposition. However, other factors must be involved, including infection (presumably viral) or possibly an immunological mechanism. The prevailing hypothesis combines these two possible etiological mechanisms to suggest that some type of viral infection occurs early in life to alter the patient's immune system. Thus, the activity and progression of the disease are related to altered immune functions within the central nervous system. *See* HUMAN GENETICS; IMMUNOGENETICS.

Treatment. In keeping with this hypothesis, therapy is aimed at altering the immune status of the affected individual. Therefore, preparations such as prednisone, adrenocorticotrophic hormone, cyclophosphamide, or cyclosporine have been employed in attempts to alter the progression of the disease. There is no panacea for multiple sclerosis, however, and therapy is reserved for those who are having acute attacks or who have clear-cut progression of the disease. There is no therapy to prevent further attacks. *See* NERVOUS SYSTEM DISORDERS.

Guy M. McKhann

Bibliography. R. L. Blaylock, *Multiple Sclerosis*, 1980; C. L. Cazzulo et al. (eds.), *Virology and Immunology in Multiple Sclerosis: Rationale for Therapy*, 1988; S. U. Kim (ed.), *Myelination and Demyelination: Implications for Multiple Sclerosis*, 1989; B. H. Waksman et al., *Research on Multiple Sclerosis*, 3d rev. ed., 1987.

Multiplexing and multiple access

In telecommunications, multiplexing refers to a set of techniques that enable the sharing of the usable electromagnetic spectrum of a telecommunications channel (the channel passband) among multiple users for the transfer of individual information streams. It is assumed that the user information streams join at a common access point to the chan-

nel. The term "multiple access" is usually applied to multiplexing schemes by which multiple users who are geographically dispersed gain access to the shared telecommunications facility or channel. Various methods of multiplexing and multiple access are in common use.

Frequency division. In frequency-division multiplexing (FDM) and frequency-division multiple access (FDMA), the passband of a channel is shared among multiple users by assigning distinct and nonoverlapping sections of the electromagnetic spectrum within the passband to individual users. The information stream from a particular user is encoded into a signal whose energy is confined to the part of the passband assigned to that user. All users may transmit their information simultaneously in time and, although their signals thus interfere in time, the composite signal can be separated into its individual user signals at the receiver by frequency-selective filtering. Care must be exercised so that the signals generated by the users do not overlap in the frequency domain. Thus, the energy of each signal must be strictly limited to its section of the passband, and a guard space between adjacent sections may be required. *See* ELECTRIC FILTER; RADIO SPECTRUM ALLOCATION.

Time division. Time-division multiplexing (TDM) and time-division multiple access (TDMA) permit a user access to the full passband of the channel, but only for a limited time, after which the access right is assigned to another user. Normally the access rights are assigned in a cyclical order to the competing users, with each user periodically gaining access for equal amounts of time. In some applications, however, certain users may be given preference over others by allowing more frequent access or access for longer periods of time. Since in a time-division multiplexing system the user information streams are separated in time, the receiver must have accurate information on the start and end times of each user information stream within a cycle. The process of obtaining such information is known as synchronization.

Statistical multiplexing. For applications in which the user information streams occur intermittently in time, the efficiency of utilization of the channel capacity in time-division multiplexing systems may be very low due to the fact that the access time of a particular user in successive cycles is dedicated, irrespective of whether that user requires the channel. Statistical time-division multiplexing overcomes this problem by assigning time on the channel on a demand basis, which typically increases the number of users who may be accommodated on the same channel. On the negative side, the users may experience delay in accessing the channel during periods when the demand exceeds the supply. Furthermore, since a particular user's information may be transported during different time intervals in successive cycles, a certain amount of identifying information must be transferred either by a separate service channel or along with the user's data, thereby increasing the amount of transmission overhead.

Code division. In code-division multiple access (CDMA), all users are assigned the entire passband of the channel and are permitted to transmit their information streams simultaneously. Thus no separation in frequency or time occurs. To maintain the ability to recover the individual signals at the receiver, at the transmitter each signal has impressed on it a characteristic signature. (For example, it may be coded orthogonally to the other signals.) Correlating the composite received signal against the signature of a particular user has the effect of attenuating the signals from all other users to an acceptable level of interference, and thus allows the extraction of a specific signal to a certain degree of accuracy.

Space division. Space-division multiple access (SDMA) refers to the use of the same portion of the electromagnetic spectrum over two or more spatially distinct transmission paths. In most applications of space-division multiple access, the paths are formed by multibeam antennas, in which each beam is directed toward a different geographic area. The performance of such systems is heavily dependent on the degree to which individual beams can be isolated from each other. See ANTENNA (ELECTROMAGNETISM).

Wavelength division. Transmission systems that employ the optical portion of the electromagnetic spectrum, such as those using fiber-optic cables as the transmission medium, may share the total available passband of the medium by assigning individual information streams to signals of different wavelengths or "colors." In dense wavelength-division multiplexing (WDM) schemes, a typically large number of such wavelengths is available, and consequently the capacity of such systems to carry multiple information streams is large. See OPTICAL COMMUNICATIONS; OPTICAL FIBERS.

Polarization division. Polarization refers to the direction or geometric orientation of the electric field vector of an electromagnetic field. In conventional linear polarization the vector points in either the vertical or horizontal direction in a plane perpendicular to the direction of propagation of the field. In circular polarization the vector rotates in either a clockwise or counterclockwise direction. Polarization-division multiplexing and polarization-division multiple access exploit these differences by assigning electric fields of different polarization to individual channels or users. See POLARIZATION OF WAVES; POLARIZED LIGHT.

Performance of multiplexing systems. The performance of the various multiplexing techniques described above is generally stated in terms of the total information-carrying capacity of the channel and the quality of the signals at the receiver. It is heavily dependent on the attenuation characteristics of the channel passband, the distribution of the energy in the signals representing the data streams, the level and type of interference in the channel, and the multiplexing scheme employed. See ELECTRICAL COMMUNICATIONS.

Hermann J. Helgert

Bibliography. G. Held, *High Speed Transmission Networking*, Wiley, New York, 1999; S. H. Rowe II, *Telecommunications for Managers*, 4th ed., Prentice Hall, 1999; M. Schwartz, *Information Transmission, Modulation, and Noise*, 4th ed., McGraw-Hill, 1990.

Multiplication

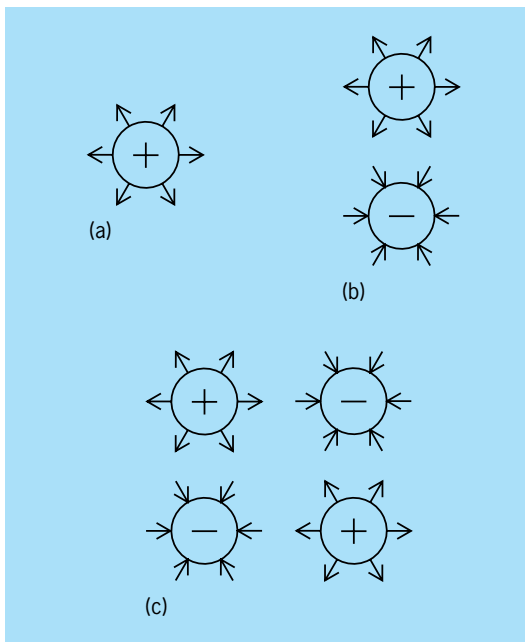
One of the fundamental operations of arithmetic and algebra. The use of the symbol \times , commonly employed in arithmetic to denote multiplication, is attributed to the English mathematician William Oughtred (1574–1660). Because of its resemblance to the letter x , it is rarely used in algebra, where multiplication is frequently denoted by a dot (as in $a \cdot b$) or, most often, merely by juxtaposition of letters (for example, ab). Multiplication of numbers (real or complex) is associative, $a(bc) = (ab)c$; commutative, $ab = ba$; and distributive with respect to addition, $a(b + c) = ab + ac$; but the term has been extended to denote binary operations on many other kinds of objects, and these operations need not possess all the properties of ordinary multiplication listed above (for example, multiplication of matrices is not commutative). Much effort was formerly expended in the computation and design of multiplication tables, and the related endeavor of tables of prime numbers. (A whole number is a prime if it is not exactly divisible by any whole number other than itself and 1.) *Crelle's Journal* (1895) gave a 1000×1000 table, and D. H. Lehmer has given a table of all primes from 1 to 10,006,721. Calculators, well adapted for such work, now are used almost exclusively in such computations. See ADDITION; ALGEBRA; DIVISION; NUMBERING SYSTEMS; NUMBER THEORY; SUBTRACTION.

Leonard M. Blumenthal

Multipole radiation

Standard patterns of radiation distribution about their source. The term radiation applies primarily to the transport of energy by acoustic, elastic, electromagnetic, or gravitational waves, and extends to the transport of atomic or subatomic particles (as represented by quantum-mechanical wave functions). The time dependence of radiation waves is generally analyzed into sinusoidal components with constant frequency ν and wavelength λ (equal to the wave velocity divided by ν), radiation sources being generally smaller than λ . See ELECTROMAGNETIC RADIATION; GRAVITATIONAL RADIATION; QUANTUM MECHANICS; SOUND; WAVE MOTION.

Each multipole pattern reflects the source's geometrical shape (or the shape of a source component). These geometrical features stand out clearly for the static electric potentials generated by fixed charges as shown by the small set of monopole, dipole, and quadrupole charges (see *illus.*), elements of all multipoles being named (in terms of powers of 2)



Static electric potentials generated by fixed multipoles. (a) Monopole ($l = 0$). (b) Dipole ($l = 1$). (c) Quadrupole ($l = 2$).

2^l -poles, with l equal to any nonnegative integer. A monopole ($l = 0$) acoustic wave radiates from a perfectly spherical bubble with oscillating radius; higher multipoles would arise from bubble distortions. So-called transverse waves, elastic or electromagnetic (including light), have only $l \geq 1$ components, gravitational waves only $l \geq 2$. The angular distributions, in azimuth (φ) and colatitude (θ), of 2^l -pole waves have amplitudes distributed in directions (θ, φ) in proportion to the spherical harmonic functions $Y_l^m(\theta, \varphi)$. The index m is a positive or negative integer whose absolute value is equal to less than l . See COORDINATE SYSTEMS; DIPOLE; SPHERICAL HARMONICS.

Symmetries. The amplitude of each multipole wave is proportional to a function $Y_l^m(\theta, \varphi)$, defined with reference to a symmetry axis, for example, the axis of the simple dipole (illus. b). In that case, $l = 1$ and $m = 0$, meaning that the spherical harmonic Y_1^0 , which is proportional to $\cos \theta$, does not depend on the azimuth (that is, the longitude) φ . [The function Y_1^0 pertains to the radiation generated by the simple dipole (see illus.), with oscillating length.] For nonzero values of m , Y_l^m depends sinusoidally on $m\varphi$. Changes of the mutual orientation of the coordinate axis and of the radiation source affect only the value of the index m , leaving the source multipolarity index l fixed. The measurable intensity of radiations is proportional to their squared amplitude, which may be proportional to a squared sum over terms proportional to the Y_l^m functions with different m .

Another symmetry concerns the sign reversal of wave amplitudes upon reflection of the source through its center. The dipole's sign reverses upon this reflection, the monopole and quadrupole do not (see illus.). This reflection reverses the sign

of electric multipoles only for odd values of l ; so-called magnetic multipoles, radiated by current loops with oscillating intensity, reverse their sign only for even values of l . The occurrence of opposite parity classes, represented by $(-1)^l$ or $(-1)^{l+1}$ respectively, extends to all types of multipole radiation. See PARITY (QUANTUM MECHANICS); SYMMETRY LAWS (PHYSICS).

Radiation output. The rate of energy radiated, for example, by a radio antenna, depends on the antenna dimension d , wavelength λ , and multipolarity l , the radiation amplitude being proportional to $(d/\lambda)^l$. For light emission by atoms, the ratio d/λ lies below 10^{-3} , making high-multipole radiations very weak. For gamma-ray emission by nuclei, d is on the order of 10^4 lower than for atoms, but λ may be 10^6 times lower, thus favoring the output of higher-multipole radiations. [The amplitude of magnetic multipole emissions is also proportional to $(v/c)^l$, where v stands for the current carrier's speed, and such emissions thus become important only as v approaches the light velocity, c .] For gravitational radiation, from stellar collisions or collapse sources, both d and λ are huge; terrestrial receiver antennas are, of course, far smaller; this mismatch has prevented its detection thus far. See ANTENNA (ELECTROMAGNETISM); ATOMIC STRUCTURE AND SPECTRA; GAMMA RAYS; RADIOACTIVITY.

Dynamical elements. The multipolarity index l also represents the number of angular momentum quanta \hbar (Planck's constant divided by 2π) radiated together with each energy quantum $h\nu$ (phonon, photon, graviton, and so forth). Detection and measurement of received energy quanta, together with measurement of their detection rate and mapping of their directional distribution, generally serve to diagnose the mechanics of the radiation source, whether atomic, molecular, nuclear, particle-decay, or astrophysical process. Energy and momentum conservation underlie this analysis; so does the conservation of angular momentum which states that the initial angular momentum of the source (\vec{J}_i) equals the vector sum of the final angular momentum of the source (\vec{J}_f) and the angular momentum of the radiation (\vec{L}). The quantitative implications of this vector relation are studied by the branch of quantum theory called angular momentum algebra. The balancing of parity, that is, of each variable's sign reversal (or persistence) under reflection through the source's center, also contributes to the analysis of experimental data. Further, more complex angular-momentum considerations play a role in the analysis of the behavior of spin-carrying particles. See ANGULAR MOMENTUM; CONSERVATION LAWS (PHYSICS); GRAVITON; PHONON; PHOTON; SELECTION RULES (PHYSICS); SPIN (QUANTUM MECHANICS). Ugo Fano

Bibliography. H. A. Bethe and E. E. Salpeter, *Quantum Mechanics of One- and Two-Electron Atoms*, 1957; J. M. Blatt and V. F. Weisskopf, *Theoretical Nuclear Physics*, 2d ed., 1979; R. Courant and D. Hilbert, *Methods of Mathematical Physics*, vol. 1, 1953; J. D. Jackson, *Classical Electrodynamics*, 3d ed., 1998.

Multiprocessing

An organizational technique in which a number of processor units are employed in a single computer system to increase the performance of the system in its application environment above the performance of a single processor. In order to cooperate on a single application or class of applications, the processors share a common resource. Usually this resource is primary memory, and the multiprocessor is called a primary memory multiprocessor. A system in which each processor has a private (local) main memory and shares secondary (global) memory with the others is a secondary memory multiprocessor, sometimes called a multicomputer system because of the looser coupling between processors. The more common multiprocessor systems incorporate only processors of the same type and performance and thus are called homogeneous multiprocessors; however, heterogeneous multiprocessors are also employed. A special case is the attached processor, in which a second processor module is attached to a first processor in a closely coupled fashion so that the first can perform input/output and operating system functions, enabling the attached processor to concentrate on the application workload. *See* COMPUTER STORAGE TECHNOLOGY; OPERATING SYSTEM.

Classification. Multiprocessor systems may be classified into four types: single instruction stream, single data stream (SISD); single instruction stream, multiple data stream (SIMD); multiple instruction stream, single data stream (MISD); and multiple instruction stream, multiple data stream (MIMD). Systems in the MISD category are rarely built. The other three architectures may be distinguished simply by the differences in their respective instruction cycles:

In an SISD architecture there is a single instruction cycle; operands are fetched serially into a single processing unit before execution. Sequential processors fall into this category.

An SIMD architecture also has a single instruction cycle, but multiple sets of operands may be fetched to multiple processing units and may be operated upon simultaneously within a single instruction cycle. Multiple-functional-unit, array, vector, and pipeline processors are in this category. *See* SUPER-COMPUTER.

In an MIMD architecture, several instruction cycles may be active at any given time, each independently fetching instructions and operands into multiple processing units and operating on them in a concurrent fashion. This category includes multiple processor systems in which each processor has its own program control, rather than sharing a single control unit.

MIMD systems can be further classified into throughput-oriented systems, high-availability systems, and response-oriented systems. The goal of throughput-oriented multiprocessing is to obtain high throughput at minimal computing cost (subject to fail-soft equipment redundancy requirements) in a general-purpose computing environment by maximizing the number of independent computing jobs

done in parallel. The techniques employed by multiprocessor operating systems to achieve this goal take advantage of an inherent processing versus input/output balance in the workload to produce balanced, uniform loading of system resources with scheduled response.

High-availability multiprocessing systems are generally interactive, often with never-fail real-time on-line performance requirements. Such application environments are usually centered on a common database and are almost always input/output-limited rather than computer-limited. Tasks are not independent but are often interdependent at the database level. The operating system goal is to maximize the number of cooperating tasks done in parallel. Such systems may also process multiple independent jobs in a background mode. The additional hardware redundancy in a fault-tolerant system over a general-purpose multiprocessor can be considered a trade-off against software complexity and the time required for software check-pointing in a sequential mainframe system. *See* FAULT-TOLERANT SYSTEMS; REAL-TIME SYSTEMS.

The goal of response-oriented multiprocessing (or parallel processing) is to minimize system response time for computational demands. Applications for such systems are naturally computer-intensive, and many such applications can be decomposed into multiple tasks or processes to run concurrently on multiple processors. In the past, successful SIMD and MIMD parallel processors were often special-purpose machines dedicated to a single class of scientific or real-time signal processing applications. The interest in high-performance, low-cost computers able to handle combined numeric, symbolic, and signal processing tasks concurrently, for so-called fifth-generation applications, together with the availability of low-cost very large scale integrated-circuit (VLSI) microprocessors has rekindled interest in response-oriented multiprocessing. *See* CONCURRENT PROCESSING; INTEGRATED CIRCUITS; MICROPROCESSOR.

Throughput-oriented multiprocessing. The performance of a classical shared-memory multiprocessor is limited by the so-called bandwidth of its shared memory (the total data transfer capacity of the memory bus). Access conflicts further reduce effective bandwidth and thus overall system performance. Studies on early multiprocessors showed interesting results on throughput loss as a function of effective memory bandwidth. For example, if a single processor had one unit of throughput, its dual processor had only 10% less throughput than two single processor systems, and a triple processor had 20% less throughput than three individual systems in a multicomputer rather than multiprocessor configuration. This was not a high penalty to pay for fail-soft function in an airline reservation system with 3000 remote agent terminals. Multiprocessors now exhibit similar performance characteristics for up to six processors, that is, up to their effective memory bandwidth performance limit which has been enhanced by a higher degree of memory

interleaving over earlier multiprocessor systems. Operating system software provides the key fail-soft capability in a throughput-oriented multiprocessor. The performance cost of software check-pointing is higher in sequential processors than in a multiprocessor. Since system protective tasks are redundant, the more processors the better, up to the effective bandwidth limits of the system's shared resources.

High-availability multiprocessing. Fault-tolerant multiprocessor systems were a natural development from throughput-oriented multiprocessor systems. While the trade-off that achieves fault tolerance is a hardware one—that is, more hardware units are used in order to achieve greater system availability—the technology employed is primarily a software one. Lower-level hardware redundancy is used in many such systems, but its successful deployment in applications is still a software issue. The basic requisite for a highly available system in most applications is that each major hardware and software component must at least be duplicated. The system requires two or more processors, two paths connecting the processors, and two paths from the processors to the database. The system's disk controllers and communication controllers must be multiported, so that they may be connected to multiple processors. A high-availability database-oriented system requires five essential software ingredients: a network-communication subsystem, a data-communication subsystem, a database manager, a transaction manager, and an operating system. The network communication subsystem supports interprocess communication within a cluster of locally distributed processors. If the highly available system is also a node on a geographically distributed system, the communication subsystem must also support internode communication. *See* DATA COMMUNICATIONS; DATABASE MANAGEMENT SYSTEM.

Response-oriented multiprocessing. The ideal performance characteristic for an N -processor system, on which a given problem could be partitioned into N similar tasks, would be a linear relationship between performance (the rate at which the system can solve a problem, in units of single-processor performance) versus the number of processors. M. Minsky was skeptical of this ideal, conjecturing that for large N the best hope was for $\log_2 N$ performance. In 1967 G. Amdahl suggested Amdahl's (second) law, stating that if a computer has two speeds of operation the slower mode will dominate performance even if the faster mode is infinitely fast. This leads to a $N/\log N$ performance in a multiprocessor performing a single application in multitask mode.

Minsky's conjecture now seems much too pessimistic, and parallel processing performance gains even greater than those predicted by Amdahl's law have become the goal. A sophisticated technique has been developed for the extraction of parallelism from FORTRAN DO-loops that can routinely exceed Amdahl's law for SIMD machines. While 100% efficiency is probably not attainable, many highly tailored applications based on careful manual extraction of paral-

lism have achieved efficiency ratings in the 80–90 range. *See* PROGRAMMING LANGUAGES.

Most computer engineers and architects see a high degree of multiprocessing or parallel processing as essential for achieving performance requirements for future computer systems for scientific computation, for fifth-generation or artificial intelligence (AI) applications, and for dedicated and embedded multiprocessors in control or automation systems. It is also widely accepted that a new parallel application and programming technology must be developed to make effective use of multiprocessors having hundreds or thousands of processors. The promise of high-performance systems made up of low-cost high-volume components (for example, microprocessor chips) rather than high-cost-volume components draws computer architects and applications specialists toward the development of parallel processing technology. *See* ARTIFICIAL INTELLIGENCE; COMPUTER ARCHITECTURE; DIGITAL COMPUTER; EMBEDDED SYSTEMS.

Peter C. Patton

Bibliography. T. J. Fountain and M. J. Shute (eds.), *Multiprocessor Computer Architectures*, 1990; E. Gelenbe, *Multiprocessor Performance*, 1990; N. Suzuki (ed.), *Shared Memory Multiprocessing*, 1992; D. Tabak, *Multiprocessors*, 1989.

Multituberculata

An extinct order in the class Mammalia, subclass Altheria, comprising a major group of early mammals, ranging from Late Jurassic to Late Eocene (about 155 to 35 million years ago). The group is best known from North America, Mongolia, and Europe. Most multituberculates were mouselike in size, although the North American Paleocene *Taeniolabis* was larger, probably closer to the woodchuck (*Marmota monax*) in its proportions. Multituberculates appear primarily to have been terrestrial creatures, but some were probably arboreal (living in trees) and others fossorial (adapted for digging). The Eocene decline and extinction of multituberculates may reflect competition from small placental mammals, especially primates and rodents. *See* ARCHAIC UNGULATE; RODENTIA.

Dentition. Although clearly mammalian and often characterized as rodentlike, multituberculates are unlike rodents or other living mammals in several important features. The dentition is distinctive and, with associated jaw fragments, it is all that is known for most species. In the most primitive multituberculates (traditionally classified in the suborder Plagiauloidea, Late Jurassic to Early Cretaceous), the upper jaw holds three incisors, a canine, and five premolars, and the lower jaw one incisor and four premolars (the lower canine is never present); the second upper incisor and the lower incisor are somewhat enlarged and procumbent. The anterior upper premolars tend to be peglike, but the last upper premolar is larger and more trenchant; the lower premolars are narrow, serrated blades. Each jaw has two molars; these are low-crowned, with their cusps set

in two parallel anteroposterior rows. See DENTITION.

Advanced multituberculates, traditionally classified in the suborders Ptilodontoidea and Taeniolabidoidea (both Late Cretaceous to Eocene), have two upper incisors; the upper canine and the first upper premolar and the anterior lower premolars are absent (a well-developed diastema, a space between two types of teeth, separates incisors from premolars). Molar cusp numbers have increased, and an internal third row of cusps has been added to the upper molars. In ptilodontoids the lower incisor is longer and more slender than in plagiulacoids, and the last (fourth) lower premolar forms an arcuate, serrated blade that sheared against the bladelike most posterior upper premolar; the third lower premolar is a small peglike tooth. In taeniolabidoids the central incisors are much enlarged and in some species have become gnawing teeth with enamel restricted to an anterior band, as in rodents. The most posterior premolars are often reduced and their shearing function lost, while the molars become comparatively enormous, forming a specialized grinding “mill.”

Functional studies using living mammals as analogues suggest that multituberculate incisors were used in grasping, the bladelike premolars in cutting, and the molars in grinding food. Microscopic wear patterns on multituberculate tooth enamel, however, show that jaw motion differed radically from that in other mammals: during chewing, the mandible moved only in the vertical plane, not transversely, and was strongly retracted during the power stroke. These studies also imply that most multituberculates were probably omnivores, although species having gnawing incisors were likely specialized herbivores feeding on tough plant tissues.

Classification. Ongoing research reveals that the traditional classification of multituberculates is inadequate: the relationship between plagiulacoids and ptilodontoids + taeniolabidoids (cimolodontans) is uncertain, as is that between ptilodontoids and taeniolabidoids. The suborder Taeniolabidoidea itself lacks diagnostic characters and likely includes separate lineages having convergent dental resemblances. Unfortunately, ambiguities of this kind will continue to plague multituberculate systematics until better evidence is available in both the anatomic completeness of collected specimens and the phylogenetic continuity between species.

Multituberculate relationships to other mammals are unclear. Although the multituberculate skull and inner ear differ from those in Theria (marsupials + placentals and their antecedents), they fail to identify a special relationship between multituberculates and alternative groups of other, nontherian mammals. For example, much of the posterolateral wall of the multituberculate braincase is formed by an expanded petrosal bone (another bone, the alisphenoid, occupies this space in therians), but this pattern is widespread among nontherians and its phylogenetic meaning unclear. Aspects of the multituberculate postcranial skeleton are therian-like, implying to some workers that multituberculates and therians are sister groups. Unfor-

tunately, these resemblances are probably nonhomologous: in multituberculates, therian-like postcranial features occur in only a few advanced species, but they also evolved independently in triconodontids, primitive Mesozoic mammals that have no special relationship to therians. Dental evidence is of particular importance to the question of multituberculate relationships: although the dentition of plagiulacoids is demonstrably more primitive than that of cimolodontans, it preserves no remnant of passage through a stage in which a therian-like reversed triangle pattern of the postcanine teeth was developed. Consequently, the multituberculate dentition is most parsimoniously interpreted as a modified anteroposterior cusp-in-line dentition, from that in more primitive, nontherian mammals. See ALLOTHERIA; MAMMALIA; MONOTREMATA; THERIA.

Richard C. Fox; Thomas Martin

Bibliography. F. A. Jenkins, Jr. and D. W. Krause, Adaptations for climbing in North American multituberculates (Mammalia), *Science*, 220:712-715, 1983; Z. Kielan-Jaworowska, R. L. Cifelli, and Z.-X. Luo, *Mammals from the Age of Dinosaurs*, Columbia University Press, 2004; Z. Kielan-Jaworowska and J. H. Hurum, Phylogeny and systematics of multituberculate mammals, *Palaeontology*, 44:389-429, 2001; D. W. Krause, Jaw movement, dental function, and diet in the Paleocene multituberculate *Ptilodus*, *Paleobiology*, 8:265-281, 1982; Z.-X. Luo, Z. Kielan-Jaworowska, and R. L. Cifelli, In quest for a phylogeny of Mesozoic mammals, *Acta Palaeontol. Pol.*, 47:1-78, 2002; F. S. Szalay, M. J. Novacek, and M. C. McKenna (eds.), *Mammal Phylogeny*, vol. 1, Springer-Verlag, 1993.

Multivariable control

The control of systems characterized by multiple inputs, which are usually referred to as the controls; or by multiple outputs, which are often the measured variables and the variables to be controlled (**Fig. 1**); or by both multiple inputs and multiple outputs (MIMO). Automobiles, chemical processing and manufacturing plants, aircraft and aerospace vehicles, biological systems, and the national economy are examples of multivariable systems which require and receive some form of regulation or control, be it mathematically contrived or not. Control systems are based on mathematical models of the behavior to be controlled, and the methodologies rely on sound mathematical theories. This article focuses on

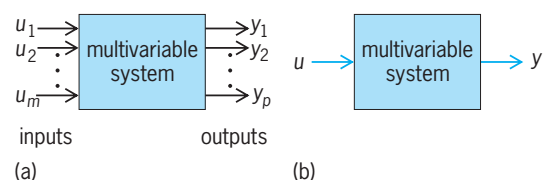


Fig. 1. Two multivariable system representations. (a) Input and output variables shown separately. (b) Input and output variables shown in vector form.

mathematical techniques which are used to design automatic controllers for multivariable systems.

Most of the mathematical control theory which has been developed for multivariable systems has assumed, as a starting point, some form of known linear, time-invariant, continuous, and causal dynamical mathematical model for the system to be controlled. In that model the manipulated inputs and the outputs to be controlled have been identified. Such a mathematical description of the dynamic behavior of variables of interest may be derived directly from first physical principles and then simplified, for example via linearization of a nonlinear description around an operating point.

The most common description is the state-space or state-variable representation of the form of Eqs. (1),

$$\dot{x}(t) = Ax(t) + Bu(t) \quad (1a)$$

$$y(t) = Cx(t) + Du(t) \quad (1b)$$

where $x(t)$ is the n -dimensional state, $u(t)$ is the m -dimensional input, and $y(t)$ is the p -dimensional output. The corresponding $p \times m$ proper, rational, transfer matrix P of the system is given by Eq. (2),

$$P(s) = C(sI - A)^{-1}B + D \quad (2)$$

where s is the Laplace transform variable. All of the concepts that apply to linear systems, such as the notions of stability, controllability, observability, minimal realizations, and poles and zeros, therefore apply here as well. See LINEAR SYSTEM ANALYSIS.

Although the following discussion focuses on linear, time-invariant systems described by Eqs. (1) or (2), some of the multivariable control design methods described below apply to more general systems. For example, the linear quadratic optimal control methodology also applies to time-variant systems, where the parameters—matrices A , B , C , and D in Eq. (1)—depend on time t . Furthermore, all the results discussed below have corresponding results applicable to discrete-time systems described by difference equations instead of differential equations, and transfer matrices expressed in terms of the z -transform variable z instead of the Laplace transform variable s . See OPTIMAL CONTROL (LINEAR SYSTEMS); Z TRANSFORM.

Control objectives. The control objectives to be achieved in the multivariable (MIMO) case include those control objectives which are generally sought in the scalar, that is, single input/single output (SISO) case, and also depend on the specific application being considered. In particular, stability is usually a primary concern in the design of any control system, along with various measures of robustness to uncertainties in the plant model.

While controller failure in the scalar case can lead to catastrophic failure of the overall system, such need not be true in the multivariable case due to the interactions between various input/output pairs. More specifically, in certain applications it may be possible to design a multivariable controller “driven” by all p outputs which performs satisfactorily even

when feedback information from one or more outputs is lost; or the same control objectives may be attained with fewer than m available control inputs. The integrity of such a multivariable controller would be much better than that of its scalar counterpart. See CONTROL SYSTEM STABILITY.

In many control applications, a primary objective is that of tracking, or ensuring that the system output or outputs track a desired input or inputs with little or no steady-state errors. In addition, simultaneous regulation of external disturbances is also often desired; that is, ensuring that any external disturbances affect the plant outputs as little as possible. These performance objectives are to be attained with adequate robustness to uncertainty in the plant and the environment. It is also obvious that simplicity of controller design is a highly desirable design objective in both the scalar and multivariable cases. One final design objective that is unique to the multivariable case is that of minimizing or eliminating the interaction between loops, which is often referred to as decoupling; that is, perfect decoupling would imply a controlled system where each input affects one and only one output or, otherwise stated, a system whose compensated transfer matrix is diagonal and nonsingular.

General considerations. In multivariable control, given a linear, time-invariant, causal system to be controlled described by equations of the type of Eqs. (1) or (2), a linear, time-invariant, causal controller is to be determined that satisfies the control specifications. Here, a causal controller is a controller with proper transfer matrix $C(s)$; $C(s)$ is called proper if $\lim_{s \rightarrow \infty} C(s) < \infty$. Note that more complex controllers may be used, such as nonlinear or switching controllers, but they are considered in practice only when absolutely necessary as they may complicate the control design considerably.

It is useful to consider a general, linear, time-invariant controller that has two degrees of freedom. Two transfer matrices, $C_y(s)$ and $C_r(s)$, may be selected, described in Eq. (3), where U , Y , and R are

$$\begin{aligned} U(s) &= C_y(s)Y(s) + C_r(s)R(s) \\ &= [C_y(s), C_r(s)] \begin{bmatrix} Y(s) \\ U(s) \end{bmatrix}^T \end{aligned} \quad (3)$$

the Laplace transformed variables of the input u , output y , and r , an external reference input; in general Y^T denotes the transpose of Y . The controller $C = [C_y(s), C_r(s)]$ is a rational transfer matrix to be determined. Most of the literature on multivariable control explores methodologies to design the feedback controller $C_y(s)$ to satisfy control specifications such as stability and low sensitivity to parameter variations and disturbances. The controller $C_r(s)$ is also important when some specific response to the reference input r is to be attained. Particular control configurations may impose restrictions on $C_y(s)$ and $C_r(s)$ that may not be selected independently any longer; for example, in the unity or error feedback configuration, $C_r(s) = -C_y(s)$.

Under the control law of Eq. (3), the transfer matrix $T(s)$ [such that $Y(s) = T(s)R(s)$] of the controlled system is given by Eq. (4),

$$T(s) = P(s)M(s) \quad (4)$$

where $M = (I - C_y P)^{-1} C_r(s)$; here, for simplicity, no disturbances are considered ($Y = PU$). Note that $U = MR$; that is, M is the open-loop equivalent transfer matrix between U and R . It can be shown that the plant P (satisfying $Y = PU$) under the control law of Eq. (3) is internally stable, all poles are stable, and no cancellation of unstable poles occurs if and only if (a) the feedback controller $U = C_y Y$ internally stabilizes the system $Y = PU$, and (b) $C_r(s)$ is such that both M and PM are stable, where the rational matrix $M = (I - C_y P)^{-1} C_r$.

In Eq. (4), given a proper and stable T , which determines the desired response of the controlled system, a proper and stable M needs to be determined. Here, M specifies the control action needed, and it is implemented via some combination of feedback and feedforward configuration of proper controllers depending on the problem under consideration. The control input U is almost never implemented via an open-loop controller M . The uncertainties in the plant and environment make it necessary to use a feedback control mechanism. A variety of feedback configurations can be used to implement M , such as linear state feedback with or without state observers or other types of output feedback.

Fundamental limitations. The requirements for stability and causality impose restrictions on attainable control objectives even in the most general case of linear control of Eq. (4); it can be shown that T must contain all the unstable zeros of the plant P , which imposes limitations on the attainable response.

Another fundamental limitation, which is perhaps more widely known, refers to feedback loop properties and is typically expressed in terms of the sensitivity function in the frequency domain via the Bode integral. It states that there is a trade-off in sensitivity; if sensitivity is designed to be low for, say, low frequencies, then unavoidably it will be high at high frequencies, with the situation becoming more restrictive when the system has unstable poles or zeros.

Achievement of objectives. With respect to the achievement of certain or all of the above multivariable design objectives, there are several approaches that may be classified according to the controller type used, whether the approaches are optimal, whether they are time or frequency domain approaches, whether they are driven primarily by industrial applications, or whether they are direct extensions of scalar-case SISO methodologies. Thus, there are state-feedback or output-feedback methodologies that are primarily concerned with pole assignment together with regulation or tracking; methodologies that cast the problems in a cost minimization framework, either in the time domain, such as the linear quadratic regulator problem, or in the frequency domain, such as the H_∞ approach;

methodologies such as the model predictive control approach, which has proven itself in industrial applications; and methodologies that are extensions of classical design techniques. These approaches are described below.

Extensions of classical design techniques. There are multivariable controller design techniques that represent extensions of the well-known and thoroughly tested classical, frequency-response techniques such as the root locus and Nyquist stability criteria. Such techniques are usually based on the ability to first decompose the given multivariable system into essentially independent scalar subsystems that can be compensated individually. One of the foremost procedures, the inverse Nyquist array method, involves the concept of a diagonally dominant transfer matrix description of the system to be controlled. More specifically, the $p \times p$ transfer matrix $P(s) = [p_{ij}(s)]$ is said to be diagonally dominant on some contour Γ of the complex plane if for all $i = 1, 2, \dots, p$, either Eq. (5a) or (5b) holds, where $|p_{ij}(s)|_\Gamma$ represents the

$$|p_{ii}(s)|_\Gamma > \sum_{j \neq i} |p_{ij}(s)|_\Gamma \quad (5a)$$

$$|p_{ii}(s)|_\Gamma > \sum_{j \neq i} |p_{ji}(s)|_\Gamma \quad (5b)$$

magnitude of $p_{ij}(s)$ for any and all s on the contour Γ . Design procedures have been developed for achieving diagonal dominance prior to the employment of the inverse Nyquist array method, which essentially extends the scalar Nyquist stability criteria to the multivariable case.

Another technique that relies on multivariable system decomposition is based on the notion of a dyadic transfer matrix. A $p \times p$ transfer matrix $P(s)$ is said to be dyadic if there exist constant $p \times p$ matrices P_1 and P_2 such that $P(s)$ satisfies Eq. (6). If the dy-

$$P(s) = P_1 \begin{bmatrix} g_1(s) & 0 & \dots & 0 \\ 0 & g_2(s) & \dots & 0 \\ 0 & 0 & \dots & g_p(s) \end{bmatrix} P_2 \quad (6)$$

namical behavior of a system is characterized by a dyadic transfer matrix, it can be decomposed as in Fig. 2. Many systems do exhibit dyadic behavior, and in such cases a forward path controller with transfer matrix given by Eq. (7) results in a unity-feedback closed-loop system with the closed-loop transfer matrix given by Eq. (8).

$$K(s) = P_2^{-1} \begin{bmatrix} k_1(s) & 0 & \dots & 0 \\ 0 & k_2(s) & \dots & 0 \\ 0 & 0 & \dots & k_p(s) \end{bmatrix} P_1^{-1} \quad (7)$$

$$T(s) = P_1 \text{diag} \left\{ \frac{g_i(s) k_i(s)}{1 + g_i(s) k_i(s)} \right\} P_1^{-1} \quad (8)$$

This implies that the dynamical performance of the closed-loop system is governed by the

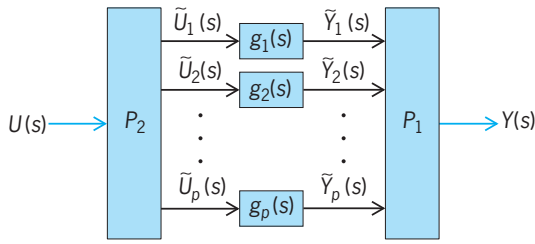


Fig. 2. Decomposition of a dyadic system $P(s)$.

performance of the p scalar feedback systems (9), which can be individually designed by using stan-

$$\frac{g_i(s)k_i(s)}{1 + g_i(s)k_i(s)} \tag{9}$$

dard classical techniques, essentially independent of one another.

Other examples of multivariable extensions of classical design techniques include the characteristic root locus design method and the sequential design method. Virtually all of these classical techniques employ a restrictive form of controller structure in order to facilitate the design.

Control approaches based on state-variable feedback and state estimators. Powerful control design techniques have been developed based on the time-domain, state-variable description of Eq. (1) and the linear state-variable feedback (lsf) control law in Eq. (10), where

$$u(t) = Fx(t) + r(t) \tag{10}$$

$r(t)$ is an m -dimensional external reference input.

Combining Eqs. (1) and (10), it can be seen that the closed-loop dynamics are determined by the eigenvalues of the closed-loop matrix given by Eq. (11).

$$A_F = A + BF \tag{11}$$

The eigenvalues of A_F are exactly the poles of the closed-loop transfer matrix when the system is controllable and observable; the uncontrollable or unobservable eigenvalues cancel in the feedback loop, and in that case the closed-loop poles are only a subset of the eigenvalues of A_F . A celebrated result shows that all the eigenvalues of A_F can be assigned any n desired real or complex-conjugate values if and only if the pair (A, B) is controllable. An important observation is that contrary to the single-input case, given n desired eigenvalues, the feedback matrix F that assigns these eigenvalues to A_F is not unique. This flexibility can be conveniently expressed in terms of the corresponding eigenvectors of A_F where m out of n elements of each eigenvector may be arbitrarily assigned to achieve additional design objectives. These expressions are particularly convenient when the control specifications are expressed in terms of desired eigenvector elements, as is the case in problems such as the flight control of air vehicles.

When the states are not directly measurable (they may be inaccessible or too expensive to measure), they are estimated via an asymptotic state estimator or Luenberger observer, which is driven by the

given system's inputs and measurable outputs. The design of the Luenberger observer is the dual to the state-feedback control problem. The separation principle allows the state feedback and the observer to be designed separately, which greatly simplifies the control design. Optimal control and state estimation methodologies, such as the linear quadratic regulator and the Kálmán filtering approaches, which are based on quadratic cost criteria, may be used to determine optimal controllers and observers. See ESTIMATION THEORY; STOCHASTIC CONTROL THEORY.

Linear state-variable feedback, with possibly an observer L , is combined with input dynamic compensation to achieve a greater diversity of design goals, such as arbitrary eigenvalue or pole placement, decoupling, model matching, output regulation, and zero-error tracking. An input dynamic compensator may be simply seen as a proper series compensator, $K(s)$, placed in the feedforward loop (Fig. 3). Such a control configuration may be seen as a specific implementation instance of the map $M(s)$ of Eq. (4).

Control approaches based on output feedback and frequency domain methods. Linear state-variable feedback-based methods have been used widely in a variety of applications. Alternative methodologies have also been introduced to address stability and performance robustness issues when the control specifications are expressed in the frequency domain.

An important optimal design methodology is the H_∞ optimal control approach, where the cost function is the H_∞ norm of a transfer matrix of interest, typically between some disturbance input and an output. The H_∞ norm provides a measure of the worst possible performance; to illustrate, it can be described as the distance from the origin to the furthest point on the Nyquist plot of the transfer function. In the H_∞ optimal control methodology, a parameterization of all stabilizing controllers is used, and the parameter is selected by minimizing an appropriately selected H_∞ norm, so that the controller not only stabilizes the system but also satisfies additional control objectives expressed in terms of minimizing a norm. The calculations are typically carried out using expressions that involve the constant parameters of some state-variable descriptions. Stability and performance robustness to plant parameter and disturbance uncertainties and the worst case measures in the frequency domain are the important underlying themes of this approach.

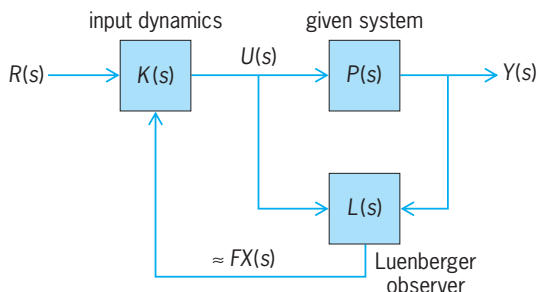


Fig. 3. Linear state variable feedback-input dynamic compensation.

Other frequency domain optimal-control methodologies do exist, for example involving the l_1 norm of appropriate maps, the advantage of the l_1 norm being that solutions may be found efficiently based on linear programming techniques. (The l_1 norm applied to a vector is the sum of the absolute values of all its elements.)

Additional approaches to control design. Optimization theory provides a powerful array of tools that may be used to solve control problems. A relatively recent trend in control design has been the recasting of control problems as linear programming problems or more generally as convex optimization problems, where efficient solution methodologies for large problems exist. Linear matrix inequalities (LMI), which may be solved efficiently by optimization algorithms such as interior point methods, have been used to solve an increasing number of control problems.

An approach that has proven to be very effective in practice, particularly in process control applications, is the model predictive control methodology, which is based on solving constrained optimal control problems on line. In this approach, given the current plant state $x(k)$ in a discrete-time plant description, an open-loop optimal control problem over some future N -step interval is solved subject to current and future constraints on the plant input, state, and output variables. This calculated optimal policy is then applied for only a short interval, one step only, and then the procedure is repeated. (This is a type of receding horizon approach.) The open control law is converted into a closed-loop strategy by using the measured value of $x(k)$ as the current state. This method is robust, deals effectively with constraints, can handle very large problems, and has been applied successfully to nonlinear problems as well.

Panos J. Antsaklis; William A. Wolovich

Relative gain array. The relative gain array is an analytical device well known in process control multivariable applications, also called interaction measure and several similar terms. It is based on the comparison of single-loop control to multivariable control. It is expressed as an array (for all possible input/output pairs) of the ratios of a measure of the single-loop behavior between the input and output variable pair, to a related measure of the behavior of the same input/output pair under some idealization of multivariable control. This is appropriate because most practical multivariable control is achieved through multiple single loops. For a square transfer function matrix T the array $M = [\mu_{ij}]$ is defined by $\mu_{ij} = t_{ij}(s)/\hat{p}_{ij}(s) = t_{ij}(s)\hat{t}_{ij}(s)$, where $\hat{t}_{ij}(s)$ are the elements of $T(s) = T^{-1}(s)$, and \hat{p}_{ij} , equal by definition to $1/\hat{t}_{ji}$, has the interpretation of the transfer function between the paired input and output variables under perfect control of other variables. A number of alternative forms and extensions have been introduced.

The array has a very strong intuitive appeal, particularly in the steady-state form $M(0) = [\mu_{ij}(0)]$: perfect control is an idealization of real control, and the ratio naturally relates single and multiloop areas. $\mu = 1$ if the loop is isolated from the rest of the system. For

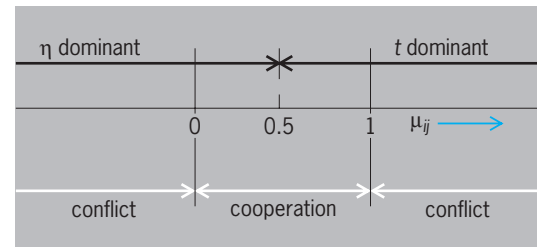


Fig. 4. Classification of ranges of possible values of an element of the relative gain array μ_{ij} .

$\mu < 0$, apparent conflict in behavior should arise. The intuition can be extended by considering \hat{p}_{ij} as being made up of a direct part t_{ij} and an indirect part η_{ij} . The ranges of the values that μ_{ij} can take can be classified in terms of the relative sign and magnitude of μ_{ij} and η_{ij} (Fig. 4): For $0 < \mu_{ij} < 1$, $t_{ij}/\eta_{ij} > 0$; t and η cooperate. For $\mu_{ij} < 0$ or $\mu_{ij} > 1$, $t_{ij}/\eta_{ij} < 0$; t and η conflict. For $\mu_{ij} > 1/2$, $|t_{ij}/\eta_{ij}| > 1$; t dominates. For $\mu_{ij} < 1/2$, $|\eta_{ij}/t_{ij}| > 1$; η dominates. Thus for $\mu < 0$, η dominates and conflicts with t . Intuitively not only is η fighting what would normally be the main path of control, but it overcomes that path. The consequences are indicated as item 2 below.

The intuitive appeal of the array can be formalized in terms of several results, the most important of which are the following:

1. $\mu_{ij}(s)$ is an invariant under scaling and any other single variable compensation or transformation.
2. For normal process and control structures, $\mu_{ij}(0) < 0$ implies that $p_{ij}(s)$ has a right half plane pole or zero, where $p_{ij}(s)$ is the transfer function between the paired input and output variables for any such choice of controls on the rest of the structure. This result can be generalized to relate to consequences in $p_{ij}(s)$ if $\mu_{ij}(s)$ has a right half plane pole or zero.

The obvious use of $\mu(0)$ is to decide on the best pairing of input/output variables for multiloop control using conventional industrial single-loop controllers (for instance, proportional integral or proportional integral derivative controllers). In this case the chosen pairing should generally have the property that the corresponding values of μ_{ij} be positive and preferably be close to 1. The value of μ can usually be related to the expected control behavior and as such can be used to justify more complex controls. Even in transfer function form the measure fails to give information relating to certain issues that are involved in the design of systems, such as detailed dynamic design and disturbance propagation. But experience almost invariably shows that a difficulty in a multivariable control design can be anticipated and diagnosed by use of the array. See CONTROL SYSTEMS; PROCESS CONTROL.

Edgar H. Bristol

Bibliography. P. J. Antsaklis and A. N. Michel, *Linear Systems*, 2006; E. F. Camacho and C. Bordons, *Model Predictive Control*, 2d ed., 2004; J. M. Maciejowski, *Multivariable Feedback Design*, 1989; S. Skogestad and I. Postlethwaite, *Multivariable Feedback Control: Analysis and Design*, 2d ed., 2005; K. Zhou with J. C. Doyle and K. Glover, *Robust and Optimal Control*, 1996.

Multivibrator

A form of electronic circuit that employs positive feedback to cross-couple two devices so that two distinct states are possible, for example, one device ON and the other device OFF, and in which the states of the two devices can be interchanged either by use of external pulses or by internal capacitance coupling. When the circuit is switched between states, transition times are normally very short compared to the ON and OFF periods. Hence, the output waveforms are essentially rectangular in form.

Multivibrators may be classified as bistable, monostable, or astable. A bistable multivibrator, often referred to as a flip-flop, has two possible stable states, each with one device ON and the other OFF, and the states of the two devices can be interchanged only by the application of external pulses. A monostable multivibrator, sometimes referred to as a one-shot, also has two possible states, only one of which is stable. If it is forced to the opposite state by an externally applied trigger, it will recover to the stable state in a period of time usually controlled by a resistance-capacitance (RC) coupling circuit. An astable multivibrator has two possible states, neither of which is stable, and switches between the two states, usually controlled by two RC coupling time constants. The astable circuit is one form of relaxation oscillator, which generates recurrent waveforms at a controllable rate.

Symmetrical bistable multivibrator. In bistable multivibrators, either of the two devices in a completely symmetrical circuit may remain conducting, with the other nonconducting, until the application of an external pulse. Such a multivibrator is said to have two stable states.

JFET circuits. The original form of bistable multivibrator made use of vacuum tubes and was known as the Eccles-Jordan circuit, after its inventors. It was also called a flip-flop or binary circuit because of the

two alternating output voltage levels. The junction field-effect transistor (JFET) circuit (**Fig. 1**) is a solid-state version of the Eccles-Jordan circuit. Its resistance networks between positive and negative supply voltages are such that, with no current flowing to the drain of the first JFET, the voltage at the gate of the second is slightly negative, zero, or limited to, at most, a slightly positive value. The resultant current in the drain circuit of the second JFET causes a voltage drop across the drain load resistor; this drop in turn lowers the voltage at the gate of the first JFET to a sufficiently negative value to continue to reduce the drain current to zero. This condition of the first device OFF and the second ON will be maintained as long as the circuit remains undisturbed. See TRANSISTOR.

If a sharp negative pulse is applied to the gate of the ON transistor, its drain current decreases and its drain voltage rises. A fraction of this rise is applied to the gate of the OFF transistor, causing some drain current to flow. The resultant drop in drain voltage, transferred to the gate of the ON transistor, causes a further rise at its drain. The action is thus one of positive feedback, with nearly instantaneous transfer of conduction from one device to the other. There is one such reversal each time a pulse is applied to the gate of the ON transistor. Normally pulses are applied to both transistors simultaneously so that whichever device is ON will be turned off by the action. The capacitances between the gate of one transistor and the drain of the other play no role other than to improve the high-frequency response of the voltage divider network by compensating for the input capacitances of the transistors and thereby improving the speed of transition.

Bipolar transistor circuits. A bipolar transistor counterpart of the JFET bistable multivibrator with nnp bipolar transistors is shown in **Fig. 2**. The base of the transistor corresponds to the gate, the emitter to the source, and the collector to the drain. Although

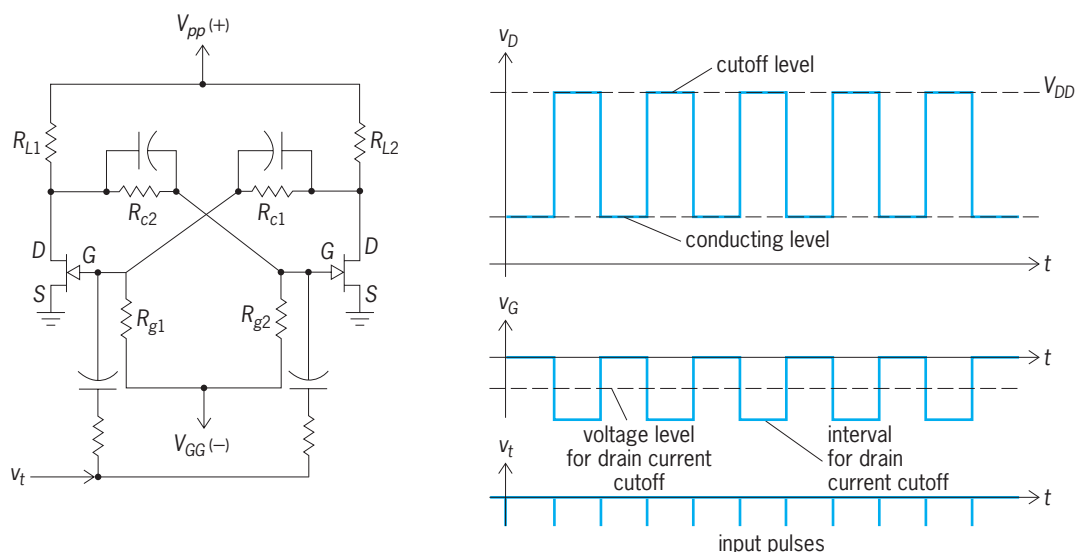


Fig. 1. Bistable multivibrator with triggering, gate, and drain waveforms shown for one transistor.

waveforms are of the same polarity and the action is roughly similar to that of the JFET circuit, there are important differences. The effective resistance of the base-emitter circuit, when it is forward-biased and being used to control collector current, is much lower than the input gate resistance of the JFET when the latter resistance is used to control drain current (a few thousand ohms compared to a few megohms). This fact must be taken into account when the divider networks are designed. If *pnp* transistors are used, all voltage polarities and current directions are reversed.

IGFET circuits. Insulated-gate field-effect transistors (IGFETs) may be used effectively in multivibrators. Use of enhancement-mode insulated-gate field-effect transistors permits direct cross-coupling with considerable simplification (Fig. 3). The circuit shown uses *p*-channel FETs with a negative supply voltage. Similarly, *n*-channel FETs may be used with a positive voltage supply.

Unsymmetrical bistable circuits. Bistable action can be obtained in the emitter- or source-coupled circuit with one of the set of cross-coupling elements removed (Fig. 4). In this case, regenerative feedback necessary for bistable action is obtained by the one remaining common coupling element, leaving one

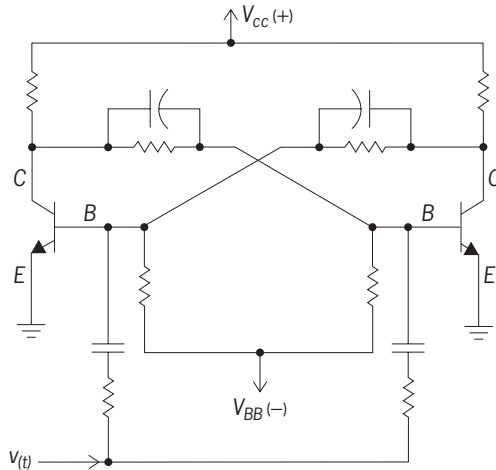


Fig. 2. Bistable multivibrator using *npn* bipolar transistors. Action is similar to vacuum-tube circuit.

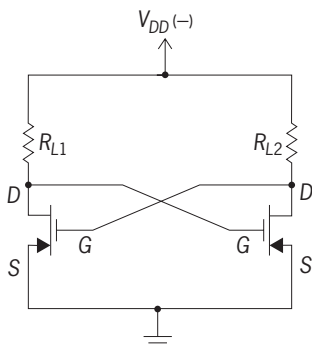


Fig. 3. Direct-coupled, bistable multivibrator which uses *p*-channel enhancement-mode, field-effect transistors, resulting in considerable simplification.

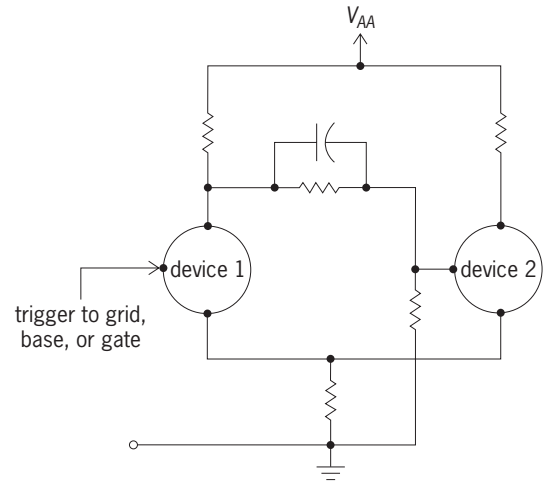
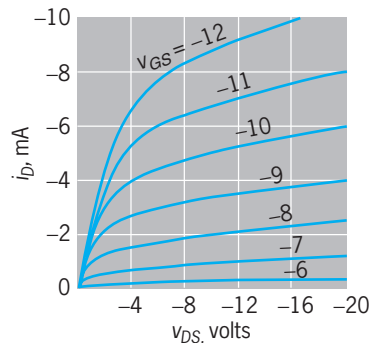


Fig. 4. Unsymmetrical bistable multivibrator.

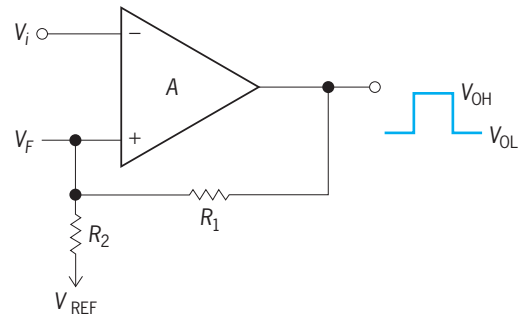


Fig. 5. Operational amplifier comparator used as Schmitt trigger circuit.

emitter or gate free for triggering action. Biases can be adjusted such that device 1 is ON, forcing device 2 to be OFF. In this case, a pulse can be applied to the free input in such a direction as to reverse the states. Alternatively, device 1 may initially be OFF with device 2 ON. Then an opposite polarity pulse is required to reverse states. Such an unsymmetrical bistable circuit, historically referred to as the Schmitt trigger circuit, finds widespread use in many applications.

A Schmitt trigger circuit often employs a high-gain operational amplifier of the type normally used as a comparator in a positive-feedback or regenerative mode (Fig. 5). When $V_i > V_F$, the output is in its low state V_{OL} , with V_F in turn determined by V_{OL} , V_{REF} , R_1 , and R_2 . When $V_i < V_F$, the output will be in its high state V_{OH} , and V_F will be at a correspondingly higher level. Thus if the input V_i switches between two levels, the output will switch between its low and high states. These switching levels are different depending upon whether V_i is increasing or decreasing. The difference in $V_i = V_F$ for the two levels is known as the hysteresis of the circuit. When the input V_i is between the low and high V_F levels, the output can be in either its high or low state depending upon the previous turn-on history; hence the term bistable circuit. See COMPARATOR.

Monostable multivibrator. A monostable or one-shot multivibrator has only one stable state. If one

of the normally active devices is in the conducting state, it remains so until an external pulse is applied to make it nonconducting. The second device is thus made conducting and remains so for a duration dependent upon RC time constants within the circuit itself.

A typical monostable multivibrator is shown in Fig. 6. The input of field-effect transistor (FET) 1 is capacitance-coupled to the output of FET 2. In the absence of external pulses, FET 1 is conducting, with its gate at zero potential, or limited to a slight positive value by saturation. The resultant drain current limits the drain voltage to a value that makes the gate of FET 2 sufficiently negative to keep FET 2 cut off. If a negative pulse is applied to the gate of FET 1, this FET is cut off and FET 2 conducts. The circuit action is similar to that of the bistable circuits except that the voltage drop at the drain of FET 2 is transferred to the gate of FET 1 through the capacitance C_c . This change, or transition, cannot be maintained indefinitely because the current flowing through C_c and R_{G1} causes a decrease in voltage across C_c and rise in voltage at the gate of FET 1, as shown. The initial drop is of the same magnitude as that at the drain of FET 2 at the time the trigger is applied. The ensuing rise is exponential in form and (if R_{G1} is much greater than R_{L2}) is given by Eq. (1). When

$$V_{G1} = (V_{GG} - V_{DD}) \exp\left(\frac{-t}{R_{G1}C_c}\right) + V_{DD} \quad (1)$$

the rising voltage reaches the level V_{c1} , drain current again flows in FET 1. Positive feedback quickly causes FET 1 to become fully conducting and limited by gate current saturation. The duration of the nonconducting interval for FET 1 time (T_1 in Fig. 6) is found by solving Eq. (1) for time $t = T_1$ and for $V_{G1} = V_{c1}$. The result is Eq. (2). The circuit will re-

$$T_1 = -R_{G1}C_c \ln \frac{V_{c1} - V_{DD}}{V_{GG} - V_{DD}} \quad (2)$$

main in this state until another initiating trigger is applied.

Monostable multivibrators are available commercially in integrated chip form, replacing the discrete circuits described here. See INTEGRATED CIRCUITS.

Astable multivibrator. The astable multivibrator has capacitance coupling between both of the active devices and therefore has no permanently stable state. Each of the two devices functions in a manner similar to that of the capacitance-coupled half of the monostable multivibrator, as shown in Fig. 7. It will therefore generate a periodic rectangular waveform at the output with a period equal to the sum of the OFF periods of the two devices. The duration of each of the two periods is governed by an equation of the form of Eq. (2), with the appropriate values for each of the two parts of the circuit used. The transistor astable multivibrator similarly functions as the combination of two transistor monostable sections coupled together.

Astable multivibrators, although normally free-running, can be synchronized with input pulses re-

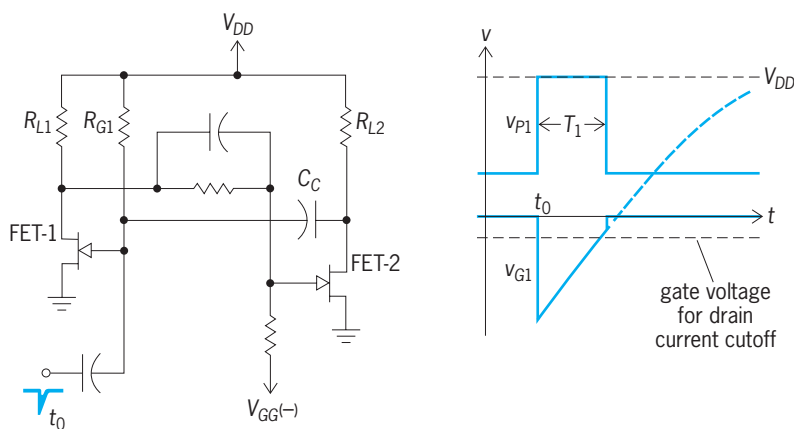


Fig. 6. Basic monostable multivibrator using junction field-effect transistors.

current at a rate slightly faster than the natural recurrence rate of the device itself. This is illustrated in Fig. 8, which shows the relation between the internal waveform and the applied synchronizing pulses. If the synchronizing pulses are of sufficient amplitude, they will bring the internal waveform to the conduction level at an earlier than normal time and will thereby determine the recurrence rate.

Triggering of multivibrators. The period of the bistable, the monostable, and the synchronized

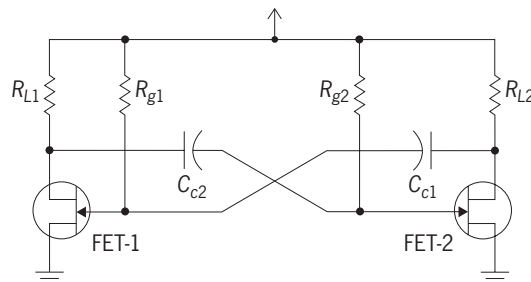


Fig. 7. Astable or free-running multivibrator.

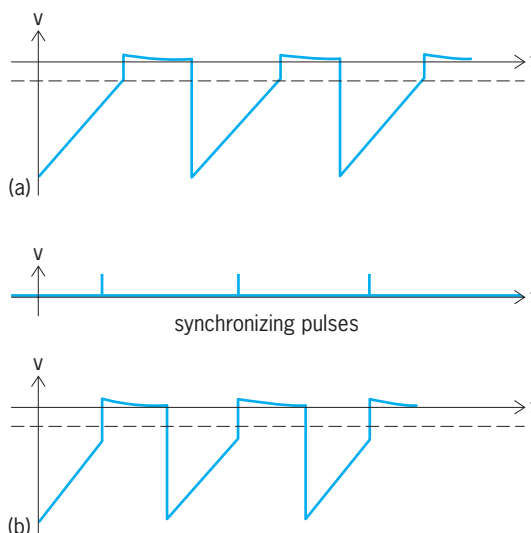


Fig. 8. Comparison of the free-running and synchronized multivibrator waveforms. (a) Free-running waveform. (b) Synchronized waveform.

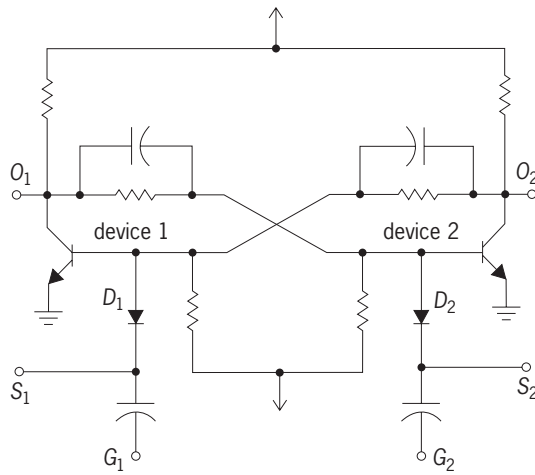


Fig. 9. Multivibrator triggering techniques.

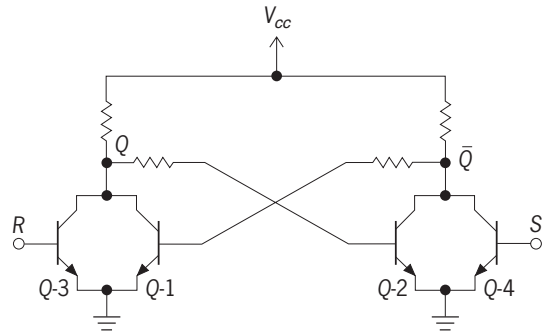
astable multivibrator is controlled by pulses (triggers) from an external source. These triggers may be applied to the circuit in various fashions. The initiating trigger should be sufficiently wide for the circuit to respond (as limited by its high-frequency response) before the pulse is over, but not so wide as to interfere with normal action of the multivibrator once the transition has taken place. The trigger should be coupled to the multivibrator through a small capacitance so that loading by the trigger source is negligible. Usually a faster transition can be achieved if triggers turn off a normally ON device. Triggers are usually applied to the appropriate input, but they may also be applied to the outputs and reach the inputs through the coupling networks. In some cases any coupling between trigger source and multivibrator is objectionable, and an isolating amplifier is used with its drain or collector and that of the multivibrator connected together. Auxiliary diodes are frequently used to provide trigger isolation and to improve triggering stability. See TRIGGER CIRCUIT.

Also, particularly where bistable circuits are used in computer logic systems, it is necessary to reset initial states so that the desired device is ON before a subsequent set of triggers is applied. Such a multivibrator provides extra available terminals for triggering purposes, as illustrated in Fig. 9.

In modern integrated-circuit digital technology, capacitance coupling of triggering pulses is rarely used. Rather, they are applied through a set of logic circuits to effect the ON-OFF periods.

Logic gate multivibrators. Multivibrators may be formed by using two cross-coupled logic gates, with the unused input terminals used for triggering purposes. The bistable forms of such circuits are usually referred to as flip-flops with the term multivibrator rarely used. For example, the circuit of Fig. 10, which is like that of Fig. 2, may be thought of as the cross connection of two, two-input RTL (resistor-transistor-logic) gates. The two extra inputs are available for set-reset triggering functions. The inputs and outputs are at standard logic levels (0) and (1). The output levels after inputs at R and S are

removed depend upon the input gate combinations. This is illustrated by the accompanying truth table, where Q_n and \bar{Q}_n are the output states before the input gate signals are applied, and Q_{n+1} and \bar{Q}_{n+1} are the outputs after signals at R and S are removed. The outputs are independent of the prior states unless R and S are both zero, in which case they remain unchanged. Also if both inputs are logic (1) the output is indeterminate, since if they both go to (0) at the same time the outputs may return to either bistable

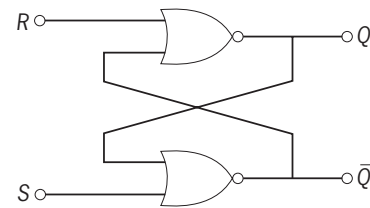


(a)

| Q_n | \bar{Q}_n | R | S | Q_{n+1} | \bar{Q}_{n+1} |
|-------|-------------|---|---|-----------|-----------------|
| 1 | 0 | 0 | 0 | 1 | 0 |
| 0 | 1 | 0 | 0 | 0 | 1 |
| 1 | 0 | 1 | 0 | 0 | 1 |
| 0 | 1 | 0 | 1 | 1 | 0 |
| 0 | 1 | 1 | 0 | 0 | 1 |
| 1 | 0 | 0 | 1 | 1 | 0 |
| 1 | 0 | 1 | 1 | ? | ? |
| 0 | 1 | 1 | 1 | ? | ? |

(b)

Fig. 10. Cross-coupled RTL gates as R-S flip-flop. (a) Circuit. (b) Truth table.



(a)

| R | S | Q_{n+1} | \bar{Q}_{n+1} |
|---|---|-----------|-----------------|
| 0 | 0 | Q_n | \bar{Q}_n |
| 1 | 0 | 0 | 1 |
| 0 | 1 | 1 | 0 |
| 1 | 1 | not used | |

(b)

Fig. 11. NOR gate R-S flip-flop. (a) Symbolic representation. (b) Abridged truth table.

condition. This combination is generally indicated as “not used.” The RTL flip-flop is representative of two cross-coupled NOR gates shown symbolically in Fig. 11 with an abridged truth table leaving out the unnecessary initial state columns. The circuit function illustrated by these two circuits is generally referred to as an *R-S* flip-flop. Such flip-flops identified by the same truth table can be constructed by using cross-coupled NAND gates with inverting stages added at the inputs.

If the *R-S* flip-flop inputs are preceded by AND gates [Fig. 12, where R' or S' can be (1) only if both inputs to the AND gate are (1)] with a precisely timed “clock” pulse applied during the existence of the *R* and *S* functions, the output state of the flip-flop is set at the time that the clock pulse is applied. This is called a clocked *R-S* flip-flop and has the same truth table as the unlocked version.

If a clocked *R-S* flip-flop uses a three-input NAND gate at the inputs with a feedback path consisting of

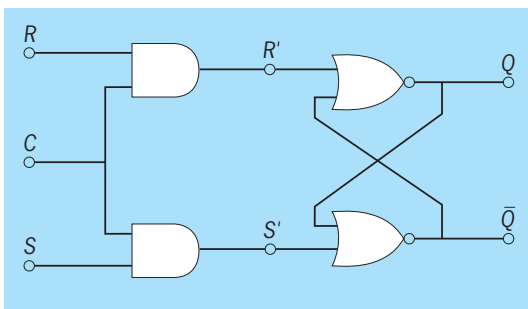
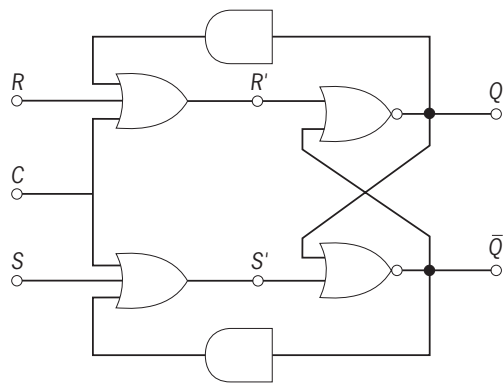


Fig. 12. Clocked *R-S* flip-flop.

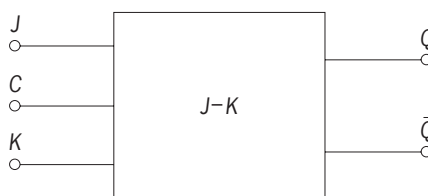


(a)

| <i>R</i> | <i>S</i> | Q_{n+1} | \bar{Q}_{n+1} |
|----------|----------|-------------|-----------------|
| 0 | 0 | Q_n | \bar{Q}_n |
| 1 | 0 | 0 | 1 |
| 0 | 1 | 1 | 0 |
| 1 | 1 | \bar{Q}_n | Q_n |

(b)

Fig. 13. Pulse-triggered *R-S* flip-flop. (a) Circuit. (b) Truth table.



(a)

| <i>J</i> | <i>K</i> | Q_{n+1} | \bar{Q}_{n+1} |
|----------|----------|-------------|-----------------|
| 0 | 0 | Q_n | \bar{Q}_n |
| 0 | 1 | 0 | 1 |
| 1 | 0 | 1 | 0 |
| 1 | 1 | \bar{Q}_n | Q_n |

(b)

Fig. 14. *J-K* flip-flop. (a) Symbolic representation. (b) Truth table with standard output designation.

a single-gate noninverting delay (Fig. 13), the inputs are “steered” in such a way that the not-used state is removed as indicated by the accompanying truth table. This works only if the clock pulse is very narrow so that it is removed before the feedback action is completed. Such a circuit is often referred to as a pulse-triggered *R-S* flip-flop.

Additional circuits which make the operation of the flip-flop independent of pulse width result in a circuit known as a *J-K* flip-flop, shown symbolically in Fig. 14 with a truth table which identifies *J* and *K* inputs in a standard manner.

Sometimes problems exist with *J-K* flip-flops for slow rising pulses at inputs *J* and *K*. Additional circuits are sometimes used whereby one flip-flop called a master drives another flip-flop called a slave with feedback connections so that relative independence of clock pulse width or rise and fall times is achieved. Such a circuit is called a master-slave *J-K* flip-flop. The truth table is that of the ordinary *J-K* flip-flop.

Bistable, monostable, and astable circuits are all commercially available in integrated-circuit chip form as an alternative to the discrete logic-gate building-block approach. See LOGIC CIRCUITS.

Glenn M. Glasford

Bibliography. G. M. Glasford, *Analog Electronic Circuits*, 1986; G. M. Glasford, *Digital Electronic Circuits*, 1988; J. Millman and A. E. Grabel, *Microelectronics*, 2d ed., 1987; A. S. Sedra and K. C. Smith, *Microelectronic Circuits*, 4th ed., 1997.

Mumps

An acute contagious viral disease, characterized chiefly by enlargement of the parotid glands (parotitis). The virus, 175–200 nanometers in diameter, is a member of the *Paramyxovirus* genus in the

family *Paramyxoviridae*. It will grow in monkeys, newborn hamsters, embryonated eggs, and tissue cultures. See ANIMAL VIRUS.

Besides fever, the chief signs and symptoms are the direct mechanical effect of swelling of glands or organs where the virus localizes. One or both parotids may swell rapidly, producing severe pain when the mouth is opened. In orchitis, the testicle is inflamed but is enclosed by an inelastic membrane and cannot swell; pressure necrosis produces atrophy, and if both testicles are affected, sterility may result. The ovary may enlarge, without sequelae. Meningoencephalitis and thyroiditis are seen, but rarely. Cases without parotitis occur; 30–40% of cases are subclinical.

Diagnosis is by isolation of virus from acute saliva or cerebrospinal fluid in chick embryos or monkey kidney tissue cultures; or by enzyme immunoassay, complement-fixing responses, or neutralizing antibody responses. See COMPLEMENT-FIXATION TEST; EMBRYONATED EGG CULTURE; IMMUNOASSAY; TISSUE CULTURE.

Mumps is endemic throughout the world, in all seasons. Epidemics occur under crowded conditions, particularly among 5–15-year-olds, but also among military troops. Humans are the only known reservoir. In cities, a high proportion of the population is immune by the age of 15 years.

Gamma globulin prepared from mumps convalescent serum, given as soon as parotitis is observed, can decrease the incidence of orchitis. An attenuated live virus vaccine, produced by repeated passage through eggs, can induce immunity without parotitis. The vaccine is recommended for children over 1 year of age and for adolescents and adults who have not had mumps parotitis. It is contraindicated in pregnancy and in persons who are allergic to egg protein or neomycin. A single dose of the vaccine given subcutaneously induces detectable antibodies in 95% of vaccinated individuals, and antibody persists for at least 10 years.

Mumps vaccine is available in monovalent form (mumps only) or in combinations with rubella (MR) or with measles and rubella (MMR) live vaccines. Combination live virus vaccines produce antibodies to each of the viruses in about 95% of vaccinated individuals.

In 1967, the year mumps vaccine was licensed, there were about 200,000 mumps cases (and 900 patients with encephalitis) in the United States. In 1985, after 18 years of vaccine use, the number of mumps cases was less than 3000, with fewer than 20 cases of encephalitis. However, mumps cases increased somewhat during the 1986–1988 period, particularly among persons over 14 years old, who in 1987 accounted for 38% of the cases (compared to 8% in the prevaccine period). A number of the outbreaks during this period occurred among college students.

The principal strategy to control mumps in the United States is to achieve and maintain high immunization levels, primarily among infants and young children. Trivalent measles-mumps-rubella (MMR)

vaccine is the preferred formulation. All persons thought to be susceptible should be vaccinated, including those without documentation of physician-diagnosed mumps, immunization with live mumps virus vaccine at 12 months of age or older, or laboratory evidence of immunity.

Ensuring immunity for adolescents and young adults is especially important in view of the shift in risk of the disease to these age groups. This trend is probably attributable to the relative underimmunization of individuals born between 1967 and 1977. The shift in risk to older persons is limited to states without comprehensive mumps immunization school laws; this provides further evidence that the relative resurgence of mumps in the United States is not due to vaccine failure but to a failure to vaccinate. See VACCINATION. Joseph L. Melnick

Bibliography. R. B. Belshe, *Human Virology*, 2d ed., 1990; A. S. Evans (ed.), *Viral Infections of Humans: Epidemiology and Control*, 4th ed., 1994.

Muonium

An exotic atom, mu or (μ^+e^-), formed when a positively charged muon (μ^+) and an electron are bound by their mutual electrical attraction. It is a light, unstable isotope of hydrogen, with a muon replacing the proton. Muonium has a mass 0.11 times that of a hydrogen atom due to the lighter mass of the muon, and a mean lifetime of 2.2 microseconds, determined by the spontaneous decay of the muon ($\mu^+ \rightarrow e^+ \nu_e \bar{\nu}_\mu$).

Muonium is formed when beams of μ^+ produced in particle accelerators are stopped in certain non-metallic targets. The μ^+ beams are generally spin-polarized; that is, the average spin angular momentum of the muons in the beam points in a definite spatial direction. The muon spin retains this spatial orientation after picking up an electron to form muonium in the reaction $\mu^+ + X \rightarrow \text{Mu} + X^+$, where X represents an atom of the target material. Since the positron in μ^+ decay is emitted preferentially along the muon spin direction, the muonium polarization can be monitored very simply by measuring the spatial distribution of the decay positron. This technique of polarization measurement allows extremely small amounts of muonium to be detected and studied. See SPIN (QUANTUM MECHANICS).

Muonium was first observed by V. W. Hughes and coworkers in 1960. The characteristic Larmor precession of the muonium polarization in a magnetic field when positive muons were stopped in a target of argon gas indicated the formation of muonium. The observed Larmor frequency of 14 GHz/tesla is a unique signature of the triplet ($F = 1$) bound state of an electron and a muon. See LARMOR PRECESSION.

Since muonium is a system consisting only of leptons, it serves as an ideal testing ground for the theory of quantum electrodynamics (QED), which describes the electromagnetic interaction between particles. Indeed, experimental measurement of the hyperfine structure interval of the ground-state

muonium levels have shown that the quantum electrodynamic theory of this system is accurate to the level of one part per million. Such measurements also provide the best available values for the muon mass and magnetic moment. *See* LEPTON; QUANTUM ELECTRODYNAMICS.

Muonium chemistry and muonium spin rotation (MSR) are subfields which concentrate on study of muonium in matter. Chemical reaction rates for a wide variety of muonium reactions have been measured and compared with hydrogen rates. Formation, spin precession, and depolarization of muonium in gases, semiconductors, and insulators have been studied in some detail. Such experiments seek to understand the chemical and physical behavior of a light hydrogen isotope in matter, and to use the extremely sensitive muon polarization measurement technique to probe the structure of materials. *See* POSITRONIUM.

Patrick O. Egan

Bibliography. H. J. Ache (ed.), *Positronium and Muonium Chemistry*, 1979; V. Hughes and C. S. Wu (eds.), *Muon Physics*, 3 vols., 1979; A. Schenck, *Muon Spin Rotation Spectroscopy*, 1985.

Muscle

The tissue in the body in which cellular contractility has become most apparent. Almost all forms of protoplasm exhibit some degree of contractility, but in muscle fibers specialization has led to the preeminence of this property.

Anatomy

In vertebrates three major types of muscle are recognized: smooth, cardiac, and skeletal.

Smooth muscle. Smooth muscle, also designated visceral and sometimes involuntary, is the simplest type. These muscles consist of elongated fusiform cells which contain a central oval nucleus. The length of such fibers varies greatly, from a few micrometers up to 0.02 in. (0.5 mm). These fibers contract relatively slowly and have the ability to maintain contraction for a long time. Smooth muscle forms the major contractile elements of the viscera, especially those of the respiratory and digestive tracts, and the blood vessels. Smooth muscle fibers in the skin regulate heat loss from the body. Those in the walls of various ducts and tubes in the body act to move the contents to their destinations, as in the biliary system, ureters, and reproductive tubes.

Smooth muscle is usually arranged in sheets or layers, commonly oriented in different directions. The major physiological properties of these muscles are their intrinsic ability to contract spontaneously and their dual regulation by the autonomic nerves of the sympathetic and parasympathetic systems.

Cardiac muscle. Cardiac muscle has many properties in common with smooth muscle; for example, it is innervated by the autonomic system and retains the ability to contract spontaneously. Presumably, cardiac muscle evolved as a specialized type from the general smooth muscle of the circulatory vessels.

Its rhythmic contraction begins early in embryonic development and continues until death. Variations in the rate of contraction are induced by autonomic regulation and by many other local and systemic factors. *See* AUTONOMIC NERVOUS SYSTEM.

The cardiac fiber, like smooth muscle, has a central nucleus, but the cell is elongated and not symmetrical. It is a syncytium, a multinuclear cell or a multicellular structure without cell walls. Histologically, cardiac muscle has cross-striations very similar to those of skeletal muscle, and dense transverse bands, the intercalated disks, which occur at short intervals.

The heart contains its own specialized system for initiation and spread of contraction in a wavelike form over the myocardium. This conducting system, which is composed of the sinoauricular and atrioventricular nodes and intervening bundles of special tissue, transmits the primary impulses. It is modified cardiac muscle, sometimes called the neuromuscular system to indicate its dual characteristics. *See* HEART (VERTEBRATE).

Skeletal muscle. Skeletal muscle is also called striated, somatic, and voluntary muscle, depending on whether the description is based on the appearance, the location, or the innervation. The individual cells or fibers are distinct from one another and vary greatly in length from over 6 in. (15 cm) to less than 0.04 in. (1 mm). These fibers do not ordinarily branch, and they are surrounded by a complex membrane, the sarcolemma. Within each fiber are many nuclei; thus it is actually a syncytium formed by the fusion of many precursor cells.

The transverse striations of skeletal muscle form a characteristic pattern of light and dark bands within which are narrower bands. These bands are dependent upon the arrangement of the two sets of sliding filaments and the connections between them.

A number of different types of vertebrate skeletal fibers are known, including twitch and tonus, red and white, fast twitch and slow twitch, large- and small-diametered, and so forth. Details of the correlations between these and other properties of fiber types are far from satisfactorily known.

Morphology. When organized into muscles, skeletal muscle fibers are arranged in an orderly fashion with the axis of the fibers orienting roughly between the two points of muscle attachment. Surrounding the individual muscle fibers is a connective tissue layer, the endomysium. Groups of fibers form bundles, or fasciculi, which are ensheathed with collagenous fibered connective tissue, the perimysium. This perimysium is continuous throughout the muscles and with the epimysium surrounding the entire muscle. These sheaths of connective tissue grade rather abruptly into the dense collagenous structure of tendons at the ends of the muscle.

The size of a muscle—the number and length of fibers—is directly related to the function of that muscle. For a muscle of a fixed volume, the number of fibers is inversely proportional to the length of the fibers. Hence, as the number of fibers increases with the force development of the muscle, their length

decreases with a reduction in the ability of the muscle to shorten. Conversely, as the fibers increase in length, allowing the muscle to shorten over a greater distance, its force decreases with the reduction in the number of fibers. Pinnate muscles have the fibers arranged obliquely to the central tendon of the muscle, permitting a large number of shorter fibers to be packed into a conveniently shaped muscle. These muscles can develop great force but can shorten through a very restricted distance.

Skeletal muscles tend to be distributed about a movable structure (bone) so as to produce antagonistic actions, allowing the force of each muscle to relengthen the opposing muscle. In most parts of the body, movement of the bones is achieved through the synergistic contraction of groups of muscles rather than by the action of individual muscles.

The innervation of skeletal muscles by motor fibers of the peripheral nervous system presents a quite constant pattern. This pattern appears early in embryonic life when the motor nerves grow into the muscle tissue of each body segment. During the ontogeny of limbs and specialized features of the trunk, neck, and head, a great deal of rearrangement occurs, but the early relationships are maintained so long as the muscle persists as a postembryonic structure. *See MUSCULAR SYSTEM.* Walter Bock

Biophysics

Muscle is a biological system for generating mechanical force and requires the expenditure of free energy. A solution of adenosine triphosphate (ATP) has substantially more free energy than its breakdown products adenosine diphosphate (ADP) and inorganic phosphate (P_i). The loss in free energy when ATP and water are converted into ADP and P_i provides the required expenditure of energy for muscle. Thus ATP and water may be considered the fuel for operating the muscle machine. *See ADENOSINE TRIPHOSPHATE (ATP).*

This energy expenditure may or may not permit the muscle to do useful work. If the force generated is applied to an immovable object, such as a weight

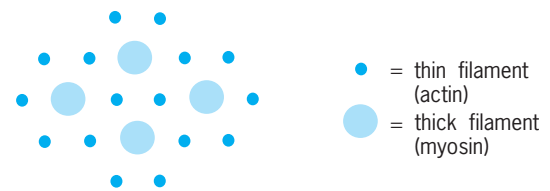


Fig. 2. Cross section of a muscle fibril.

too heavy to lift, the result is merely the generation of static tension in the muscle. The muscle is then said to be contracting isometrically—it is prevented by the heavy weight from changing length. If the force generated exceeds the weight, then the muscle, by shortening, will move the weight. Such a muscle is contracting isotonicly. An animal normally employs both types of contraction; isometric contraction occurs in applying pressure or in holding the skeleton against gravitational forces, while isotonic contractions result in moving the skeleton.

All three types of muscle operate with the same machinery. The following discussion focuses on research on the contractile apparatus of vertebrate skeletal (striated) muscle.

Microstructure. A striated muscle is a cylindrical bundle of fibers; each fiber is, in turn, a cylindrical bundle of fibrils (myofibrils). The myofibrils are segmented into sarcomeres, each a diminutive cylinder about $1\ \mu\text{m}$ in diameter and $2.5\ \mu\text{m}$ long. The sarcomeres of adjacent myofibrils are in register, thus giving the entire fiber a banded (or striated) appearance.

At either end the sarcomeres are bounded by end plates called Z membranes. The central sectors of sarcomeres are highly birefringent and are known as A (for anisotropic) bands, while the sections to either side of the Z members are practically nonbirefringent and are known as I (for isotropic) bands. Both the A and I bands result from the existence of two independent arrays of filaments which interpenetrate extensively in the A band (Fig. 1): these filament arrays are hexagonal in cross section (Fig. 2). Thin filaments are attached to each Z membrane and extend inward into the sarcomere but are not continuous across it. Thick filaments exist in the central sector. When the fibril is at rest length, the thick and thin filaments interpenetrate, but the degree to which they do so varies with the length of the fibril. The relatively clear central zone separating the two edges of the thin filament arrays is called the H zone (Fig. 1). The Z membrane, which is common to two sarcomeres, is composed of a special weave. From an end-on view it is a cubic lattice, but in longitudinal section it shows thin filaments in a hexagonal array.

Mechanistic model. It is possible to discuss the operation of muscle in mechanistic terms. In this sense a muscle cell has three recognizable machines: the sarcoplasmic reticulum, the calcium sensor, and the contractile apparatus.

Sarcoplasmic reticulum. The sarcolemma (cell membrane) of a single muscle fiber is perforated in a regular manner by tubular invaginations (sarcotubules). In contact with these tubules are elongated flat sacs

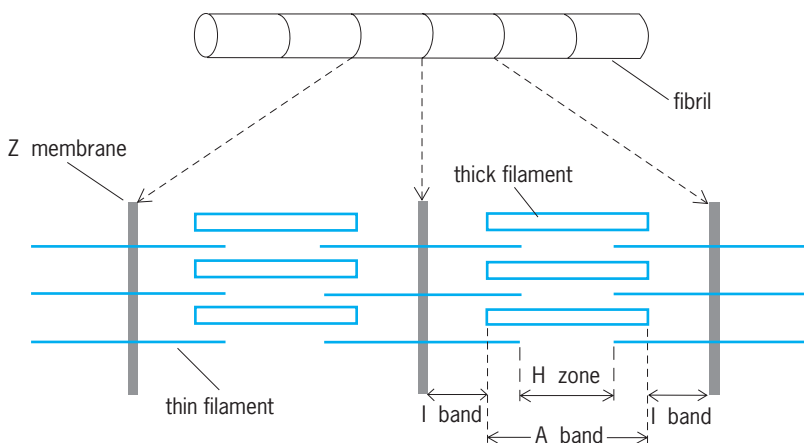


Fig. 1. Schematic representation of two adjacent sarcomeres showing banding patterns as revealed by electron microscopy.

or cisternae which surround subbundles of fibrils like a blanket. Collectively these cisternae are known as the sarcoplasmic reticulum (Fig. 3). Both the sarcolemma and the walls of the cisternae are bilayered. There is no direct opening between the tubules and the cisternae; however, at the points of contact there are four distinct layers (Fig. 3b). Communication between tubules and cisternae is therefore not by passage of a transmitter substance but perhaps by a field effect. Like bilayered membranes of other cells, the membranous sarcoplasmic reticulum contains essential phospholipid and structural protein, as well as certain enzyme proteins. See CELL MEMBRANES.

In response to nerve signals the sarcoplasmic reticulum can explosively release calcium ions into the contractile apparatus. Contractile impulses are initiated at the motor end plates, which are the points of contact of the nerve endings with the muscle cell. The nerve endings release the chemical acetylcholine, and the muscle cell then initiates its own stimulus which is conducted along the sarcolemma to the tubules. The tubules extend inward to surround each myofibril at the Z membrane. From here the electrical signal is somehow transferred to the adjacent anastomosing sheath of cisternae surrounding each myofibril. Sodium and potassium ions are intimately involved in the transmission of the impulses or stimuli. As the stimulus wave travels along the sarcolemma, sodium ions move into the membrane and potassium ions move outward. This shift of cations is not matched by a corresponding shift of anions, and as a result of the change of electrical charge on either side of the membrane an electrical potential change is created which propagates the stimulus wave. As a result of the nerve impulse a transition occurs in the cisternae membranes, causing them to become more permeable and thereby creating a very leaky pump. Because of the ion gradient that exists, calcium ions diffuse out of the cisternae very rapidly. This explains the rapid contractile response to a nerve impulse, since it is the calcium ion (Ca^{2+}) concentration which controls the contractile mechanism at the molecular level. The sarcoplasmic reticulum stores large amounts of Ca^{2+} in the lumen of the cisternae. See ACETYLCHOLINE; BIOPOTENTIALS AND IONIC CURRENTS.

After contraction, the electrical effect stops, the impermeability of the cisternae walls is restored, and calcium ion is pumped back into the cisternae, which brings about relaxation. Since calcium must move counter to the electrical gradient, this phase therefore requires work, is associated with ATP hydrolysis, and is another example of active transport.

Calcium sensor. This device is attached to the contractile apparatus and "reads" the ambient Ca^{2+} concentration. When adequate Ca^{2+} is present, it removes blocks and permits the apparatus to work. In the absence of Ca^{2+} it restores these blocks and prevents the apparatus from contracting. This sensor device is described in more detail below.

Contractile apparatus. This machine is constructed from two proteins and runs on energy provided by magnesium ions (Mg^{2+}) and ATP. Muscle fibers

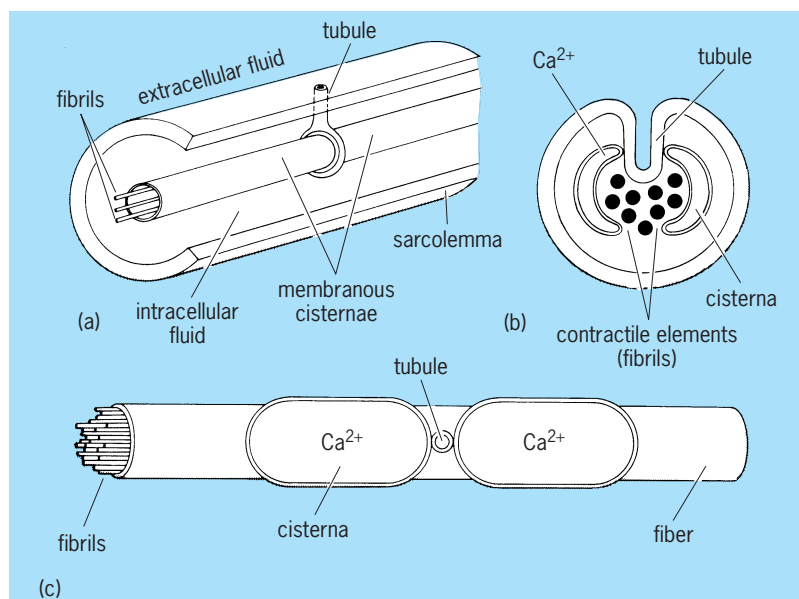


Fig. 3. Schematic representations of the sarcoplasmic and sarcotubular systems. (a) Cutaway of a muscle fiber. (b) Cross section of a fiber. (c) Fiber viewed from above (sarcolemma removed) showing triad formation of tubule and cisternae.

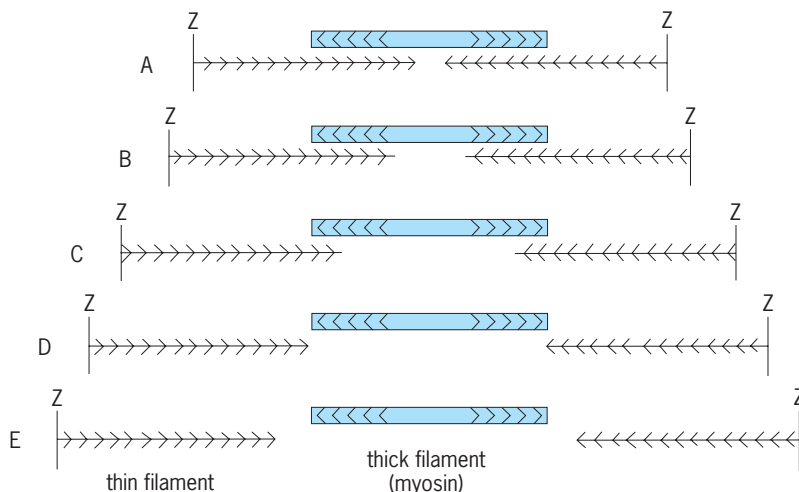


Fig. 4. A sarcomere changing length (Z-Z distance). The thick and thin filaments, though remaining at constant length, translate relative to one another. The chevrons indicate the polarity and location of the molecules constituting the filaments. The number of myosin molecules next to actin differs at different extensions (A, B, C, D, E). Not shown in this schema is the fact that sarcomeres contract at constant volume, so that shortening is accompanied by separation of the filaments in the transverse directions.

shorten or lengthen because their sarcomeres shorten or lengthen, that is, because the filaments within the sarcomeres translate (slide) relative to one another. The self-propelled, active shortening of fibers is always accompanied by the hydrolysis reaction, $\text{ATP} + \text{H}_2\text{O} \rightarrow \text{ADP} + \text{P}_i$. Fibers lengthen or get stretched only when external forces are applied to them. An emerging generalization seems to be that all biological movement (such as all muscular, amoeboid, and ciliary movement) is associated with the translation of filaments and with the hydrolysis of a triphosphate. However, the array arrangement, the proteins constituting the filaments, and the particular phosphate may differ from system to system.

The thin and thick filaments are principally composed of the proteins actin and myosin, respectively (Fig. 4). Both filaments are polymeric and polar, and each is a serial arrangement of “building blocks” or monomers. One “end” of a monomer is different from the other, and in the serial arrangement of a filament the monomers are assembled “head to tail,” so that the filament itself has a direction. The actin filaments can be thought to originate at the Z disks of the sarcomere and to point toward a central midplate. Each thick filament can be thought of as consisting of two coaxial myosin filaments; both originate at the midplane, and each points to one Z disk. The midplate is thus a plane of symmetry for the sarcomere. The directions assigned to the filaments can be taken as the directions of relative motion when a sarcomere shortens.

In both actin and myosin filaments the serial arrangement of monomers is helical. In cross section these actin and myosin helices form a well-ordered lattice. The net result of this geometry is that parallel to the fiber axis some actins of a particular actin filament are closest to some myosins of an adjacent myosin filament (the remainder of the actins are closest to other myosin filaments, and the remainder of the myosins are closest to other actin filaments). The ratio of actins to myosins need not be unity (it is in fact 5) because the pitches of the two types of helices need not be the same. See MUSCLE PROTEINS.

Actin monomer. This monomer is a dumbbell-shaped, single-chain globular protein of about 42,000 daltons and a maximum dimension of about 5 nanometers. Actin in monomeric form is known as G-actin, while in the two-stranded helical polymer form it is known as F-actin. Under physiological conditions F-actin is insoluble. In F-actin each monomer is held to the monomer before and after it on the same strand, as well as to two monomers of the other strand. These four interactin bonds are not covalent, but are quite strong. Nevertheless, the F-actin filament is rather flexible, implying that the monomers themselves may be deformable under physiological conditions. Along the filament axis the F-actin double helix repeats every 13 monomers, or every 36 nm.

Myosin monomer. This monomer is a very long (about 150 nm) and very heavy (about 480,000 daltons) multichain molecule resembling a Y with a knob or head at the end of each arm (Fig. 5). The main structure of the Y consists of two heavy chains whose C-terminal (carboxyl end) regions are intertwined to form the stem, and whose N-terminal (amino end) regions unravel to form the arms and heads. Attached to each head is a set of two light chains, one of which (the regulatory light chain) contains a high-affinity Ca^{2+} -binding site and a reactive serine that can be phosphorylated. The amino acid composition of the light chains is often different among different muscle cells (for example, fast skeletal muscle, smooth muscle, and embryonic muscle). This circumstance gives rise to many myosin isozymes.

By selective proteolysis it is possible to cut away and isolate just the heads of myosin. Such a head is

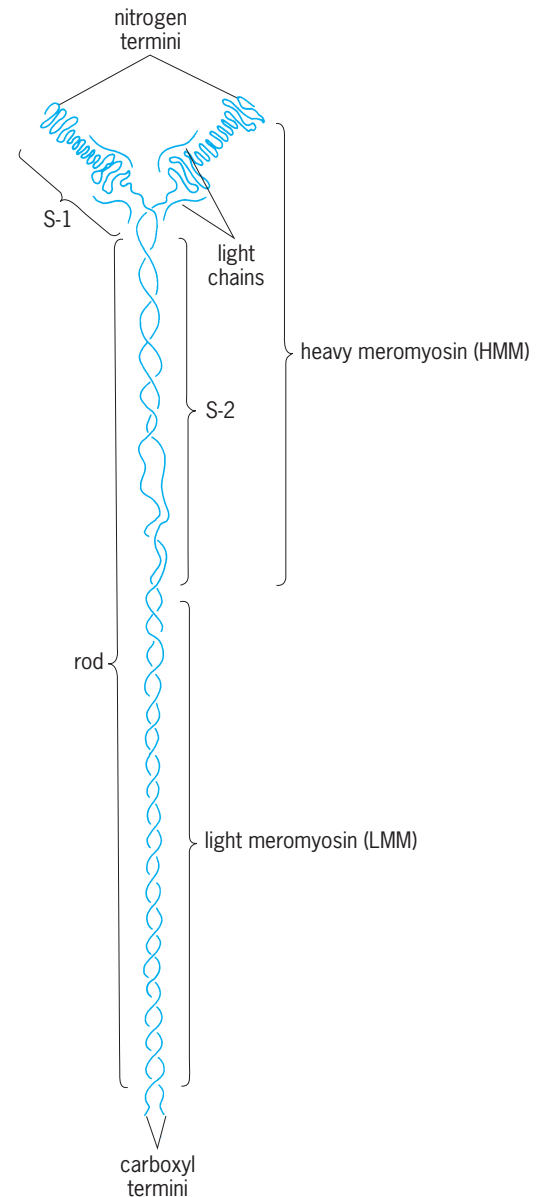


Fig. 5. Diagram of myosin, which is about 150 nm long. The various segments are obtainable after suitable proteolysis; light chains separate out in denaturing solvents. It is currently speculated that S-1 itself contains three domains.

known as a myosin subfragment 1 or S-1, while the remaining structure is known as a myosin rod. Proteolysis can again be used to cut this rod and generate two more segments; an N-terminal myosin subfragment 2 or S-2, and a C-terminal light meromyosin (LMM). When light meromyosin is cut away from the myosin molecule, the two-headed structure remaining is known as heavy meromyosin (HMM). The places at which the proteases cut are regions of flexibility in the original structure. Specifically, the heads can rotate about the arms, and there is a region of flexibility about 40 nm from the head-rod junction. Light meromyosin and any structure containing light meromyosin are insoluble in intracellular fluid, but all other segments are readily soluble. Myosin molecules are assembled into a thick filament so that there is a central core of strongly

interacting, insoluble light meromyosin segments with radially protruding soluble heavy meromyosin portions. Thus the central core is devoid of radiating heavy meromyosin and the S-1 moieties are situated adjacent to the actin filaments (Fig. 6). In the electron microscope the heavy meromyosin moieties appear as projections issuing from the thick filaments which occasionally touch adjacent actin filaments. These projections have been called cross-bridges. A cross-bridge is mainly an S-1 moiety, perhaps with some contribution from the S-2.

Fueling reaction. It is widely accepted that the hydrolysis of ATP (or its analogs) is the fueling reaction for a great variety of biological machines, including muscle. There are, however, four aspects of this reaction that are especially important in the present context. First, the participants, in this reaction ATP, ADP, and P_i , are highly ionic. Each interacts strongly with both H^+ and Mg^{2+} . This means that the free energy obtainable from the reaction depends on these ionic concentrations. Second, the reaction does not proceed at an appreciable rate in the absence of catalysis, or even in the presence of the catalyst myosin. It proceeds rapidly only when both actin and myosin are present in the catalytic form, actomyosin. These circumstances ensure that fuel is degraded only when work is being performed. Third, enzymatic catalysis always proceeds in steps, so the free energy is dissipated in steps. The algebraic sum of these stepwise decrements must equal the total free energy of the ATP hydrolysis, but the pattern of the steps is characteristic of the particular enzyme. Fourth, the steady-state concentration of MgATP in muscle is around 3 millimolar. The MgATP would be quickly exhausted in heavy exertion were it not that the breakdown products, ADP and P_i , are swiftly synthesized back into ATP by the creatine kinase system (at the expense of creatine phosphate), and more slowly but more extensively synthesized by the glycolytic system (at the expense of stored glycogen) and by mitochondria (at the expense of oxygen and foodstuffs).

Physiology. Under experimental conditions (when the membrane of a muscle fiber has been removed), the chemical composition of the fluid bathing the contractile apparatus can be controlled. It is then possible to correlate unambiguously the composition and the physiological state of a muscle fiber. The results are as follows: (1) When the medium is devoid of ATP, as in exhaustion or death, the fiber is stiff and inextensible, and is said to be in rigor. This physiological state arises because in the absence of ATP the erstwhile high affinity of the S-1 moieties for the actins literally bonds all the cross-bridges to adjacent actins. (2) When the medium contains Mg^{2+} and ATP but no Ca^{2+} , the fiber is flaccid and unresistingly extensible; it is then said to be relaxed. This state comes about because the calcium sensor is preventing any interaction between the cross-bridges and the adjacent actins. (3) When the muscle medium contains Mg^{2+} , ATP, and Ca^{2+} , the fiber is active. If it is prevented from shortening (isometric), it is developing tension and consuming ATP at a low rate, and if it

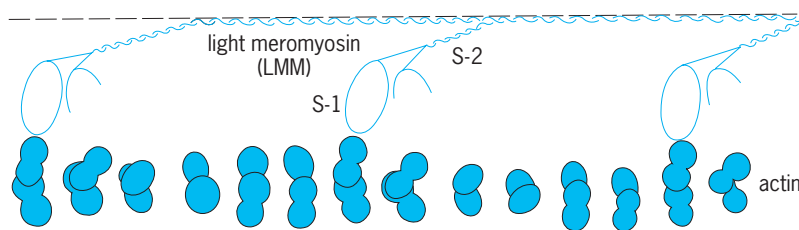


Fig. 6. Diagram of the molecules in a plane of nearest actin-myosin couples. The second S-1 of each myosin head is shown in limited fashion; each would be nearest to a separate actin filament. There is approximately one myosin head to every five actins.

is not restrained (isotonic), it is shortening while at the same time consuming ATP at a rate dependent on the speed of shortening. It is thought that under these circumstances, force interactions tending to produce the filament interdigitation are going on between the cross-bridges and the adjacent actins.

Calcium sensor. The calcium sensor serves as a regulator of muscle action. Along the two grooves of the actin helix in every thin filament, firmly attached to actin, are deployed two long chains of a protein called tropomyosin. Periodically along each tropomyosin strand are clumps of globular proteins called troponins (T, I, and C). There is reason to think that such clumps are attached to both tropomyosin and actin. Troponin C contains several Ca^{2+} binding sites, two of which have a very strong affinity for the Ca^{2+} and strongly prefer it to Mg^{2+} . These sites fit the sensor role.

Experimentation with certain of the Ca^{2+} -binding proteins led to what seemed a well-documented theory of Ca^{2+} regulation in vertebrate skeletal muscle. In relaxation, along its helical course, tropomyosin is bound to actin monomers so that it masks all their S-1 binding sites, thus blocking all contractile activity. As the Ca^{2+} concentration in the bulk interfibrillar solution rises (this Ca^{2+} concentration is itself regulated by the sarcoplasmic reticulum), it binds to the troponins which then tilt so as to pull the masks off the actins, thus permitting contraction. There is reason to think that, while elements of this theory remain correct for vertebrate skeletal muscle, other tissues may have other operative components in the Ca^{2+} sensing system. It has been found, for example, that in mollusks, Ca^{2+} sensing is performed solely by the light chains, tropomyosin/troponin being unnecessary. There is mounting evidence that something of the same myosin-linked mechanism operates in vertebrate smooth muscle. Even in vertebrate skeletal muscle the discovery that the light chains must be intact for regulation to occur suggests that these elements work in parallel with tropomyosin/troponin. Evidence has also appeared that not only physical separation of the myosin and actin but also some other means of myosin ATPase inhibition bring about relaxation. Finally, whether and how the Ca^{2+} -binding site of actin has a role in the actin-activation of ATPase remains to be discovered. Perhaps if it does, it too will have a role in regulation. At the present time the field of Ca^{2+} regulation has a surfeit of proposed mechanisms.

Length-tension relationship. The explanation of the active state discussed above arises from a classical physiological experiment known as the length-tension relationship. When the sarcomere is at rest length, every cross-bridge in a plane of nearest actins and myosins has access to actin. This continues to be so for small extensions of the fiber and its sarcomeres. But with increasing stretch a length is reached at which this ceases to be so. For stretches beyond this critical length the number of cross-bridges with adjacent actin decreases in proportion to stretch. At high extensions a second critical length is reached beyond which no cross-bridges have adjacent actins. In the experiment, the fiber, in the relaxed state, is drawn out to the desired length and clamped, then ambient condition (3) above is imposed to activate it (in the case of a membrated fiber, electrical excitation accomplishes the same condition). Two quantities are measured at each fixed length: the static tension developed against the length clamp and the ATPase activity. These two quantities, along with the number of cross-bridges opposite to actins, are then plotted against extension (Fig. 7). The graphs are surprisingly similar: flat at first, then falling linearly, finally flat at zero. These results strongly suggest that the ATPase-driven force generators are the cross-bridge-actin couples.

The previous experiment shows that the total force generated by the fiber is simply the sum of the unitary forces generated by a fixed number of cross-bridge-actin couples, but because of clamping, no shortening was permitted. When shortening occurs, the number of couples that can interact can still be kept approximately constant by limiting the experiment to small (around 10%) contractions, but even so a new phenomenon enters, since time changes the relative position of a particular cross-bridge and a particular actin. In this investigation the fiber is allowed

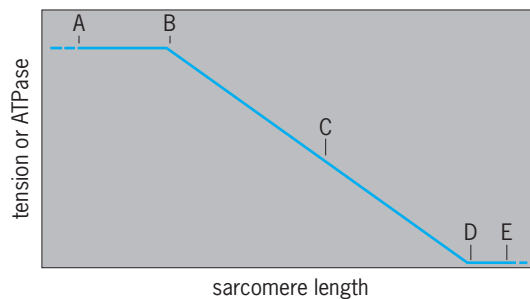


Fig. 7. Sketch of how either tension- or actin-activated ATPase depends on sarcomere length. Letters correspond to sarcomere length in Fig. 4. The correspondence is the basis for thinking that the force-generating unit is a myosin S-1 and its associated actin.

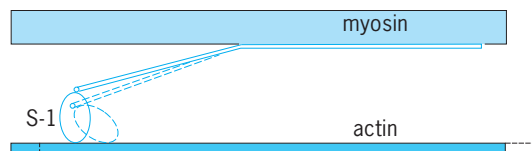


Fig. 8. Diagram suggesting that shortening would result if actin and an S-1 moiety could bind at two different angles.

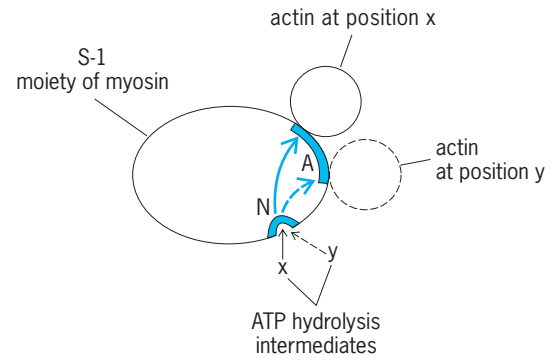


Fig. 9. The hypothesized mechanism of the transformation of chemical energy into mechanical energy in muscle. N is the enzymatic site that catalyzes the hydrolysis of ATP; A is the site at which actin binds; x and y are successive intermediates in ATP hydrolysis. When x is occupying N, the actin binds at position x on the A site; when y is occupying N, the A site binds actin at position y. During catalysis, when x is chemically converted into y, actin is caused to move from one binding position to another. This movement is forced by the chemical conversion, so the product of the distance moved and the force constitutes mechanical work.

to shorten against a constant external force (f) that it can overcome, such as a light weight (force = mass \times gravitational acceleration), and the velocity of shortening (v) is measured. The empirical result is the characteristic equation of isotonic contraction: $(f + a)v = b(f_0 - f)$, where f = force (load); v = velocity; f_0 = maximum load; and a, b = constants. The arguments are now more complicated, but this behavior too can be rationalized by assumptions as to how the force interaction between a couple depends on time and distance.

Energy transformation. The structural-functional interpretation of the length-tension diagram (above) suggests that a myosin S-1 piece (attached to the thick filament) reacting with an actin monomer (attached to the thin filament) is the unitary engine causing the interfilament slide (Fig. 8). The S-1 piece binds actin through its A site and catalyzes ATPase at its N site (Fig. 9). The intermediates in the breakdown of ATP appear sequentially on the N site: first x, then y, then z, and so on. When one ATP is totally degraded, then a fresh ATP initiates the next sequence (x, y, z, and so on). Enzymatic activity is cyclic. Because of the way that S-1 is built, it seems that to each intermediate there corresponds a specific way of holding actin at the A site, namely, in position x, in position y, and so on. Thus, corresponding to endlessly repeated ATPase cycles there are endlessly repeated cycles in the manner of holding actin, that is, mechanical cycles. Each mechanical cycle constitutes a force impulse. When the enzymatic and mechanical activity of all the S-1 pieces acting in concert is summed, a steady force is maintained at the cost of steady ATP hydrolysis. Ultimately, therefore, the energy transduction in muscle is carried out by the molecular structures of the S-1 pieces of myosin.

Patricia Rainford

Bibliography. C. Emerson et al., *Molecular Biology of Muscle Development*, 1986; A. Engle and C. Franzini-Armstrong, *Myology*, 2 vols., 2d ed., 1994; G. H. Pollack, *Muscles and Molecules: Uncovering*

the *Principles of Biological Motion*, 1990; D. J. Schneck, *The Mechanics of Muscle*, 1992; R. J. Stone and J. A. Stone, *Atlas of the Skeletal Muscles*, 3d ed., 1999.

Muscle proteins

Specialized proteins in muscle cells that are the building blocks of the structures constituting the moving and regulatory machinery of muscle. The moving machinery comprises myofilaments that are discernible by electron microscopy. These myofilaments are of two kinds, myosin and actin, and their regular arrangement within the cell gives the striated pattern to skeletal muscle fibers (Fig. 1). It is recognized that the sliding of the two sets of filaments relative to each other is the molecular basis of muscle contraction. To understand the ultimate mechanism that causes the movement of these filaments relative to each other, it is necessary to consider the features of the individual molecules that make up these filaments. Practically all nonmuscle cells, although lacking the filaments of muscle, contain proteins similar to those found in muscle; these proteins are likely to be involved in cell motility and in determining properties of cell membranes. See CELL MEMBRANES.

The banding pattern of muscle fibers produced by the regular arrangement of myosin and actin is further enhanced by the Z membranes, or disks, which appear as dark bands when the myofibril is viewed longitudinally and which separate adjacent sarcomeres. There is now considerable evidence for elastic connections occurring between the Z disks and interacting with the myosin filaments. The main constituent of these filaments is titin, a protein with a mass of about 3×10^6 daltons; it is also known as connectin. These connections represent individual titin molecules spanning the distance between Z lines and M lines. The length of these molecules varies with the length of the sarcomere. The cDNA encoding large stretches of titin reveal 11 domain superrepeats of titin spanning the 43-nanometer repeat of the thick filament helix. Each superrepeat contains two types of motif; one resembles subunits of fibronectin, a protein involved in establishing links between cells and the extracellular matrix, the other motif being related to the C-2 immunoglobulins. Another giant protein, nebulin (mass 8×10^5 daltons), is also involved in maintaining continuity between Z bands and M filaments forming presumably elastic connections with actin filaments. Nebulin also contains repeating domains, each consisting of 35 amino acid residues and each constituting an actin-binding domain. It has been suggested that nebulin plays a role in determining the length of actin filaments. These proteins may play a role in maintaining the orderly arrangements of thick and thin filaments.

The main component of the Z disk is the protein α -actinin. Other proteins, including desmin, are also involved. Smooth muscle also contains thick and thin filaments that slide past each other as the muscle contracts. In these muscles, however, the filaments

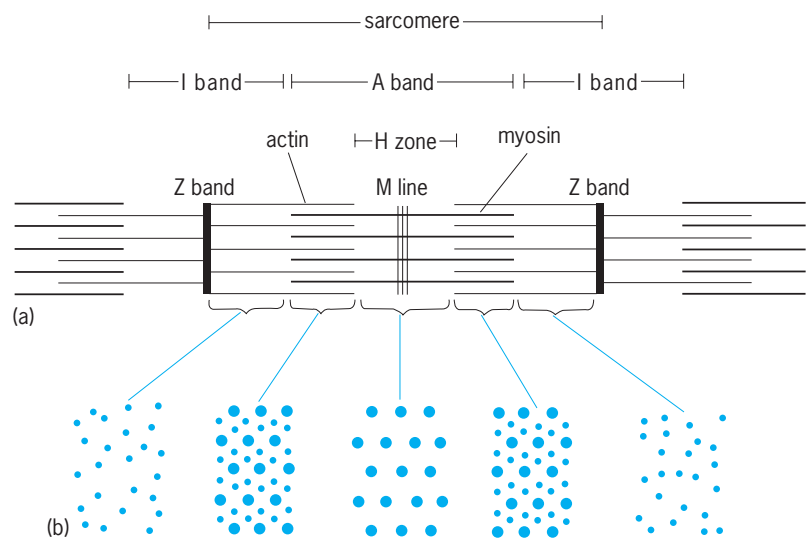


Fig. 1. Origin of striations in muscle. (a) Banding pattern of a sarcomere, showing the arrangement of myosin and actin molecules. (b) Cross-sectional appearance of sarcomere at different points along its length. (After R. Craig, *The structure of the contractile filaments*, in A. G. Engle and B. Q. Banker, eds., *Myology*, McGraw-Hill, 1986)

are not in register, and so sarcomeres and Z disks do not exist and striation patterns cannot be seen. Thin filaments are attached to irregularly disposed structures, the so-called dense bodies.

Both chief constituents of the muscle filaments can exist as single molecules (monomeric form) or in self-assembled filamentous structures (polymeric form). When isolated under the usual conditions, both myosin and actin are obtained as monomers.

Myosin. Molecules of myosin, amounting to about 60% of the total muscle protein, are arranged in filaments occupying the central zone of each segment (sarcomere) of the fibril, the A band. Myosin is an elongated molecule whose overall length is about 150 nm (Fig. 2). In contrast to many other fibrous molecules, myosin is not uniform throughout its length. It is made up of two intertwined heavy

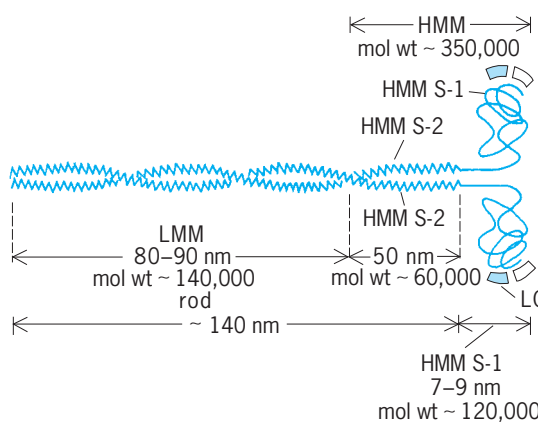


Fig. 2. Schematic representation of the structure of the myosin molecule. Hinge regions postulated in the mechanism of contraction are at the junction of HMM S-1 and HMM S-2 and of HMM S-2 and LMM (light meromyosin). It should be noted that HMM S-1 has one chief polypeptide chain, while the other fragments have two. The scheme suggests the presence of two different subunits in each HMM S-1. (After J. N. Walton, ed., *Disorders of Voluntary Muscle*, 3d ed., Longman, 1974)

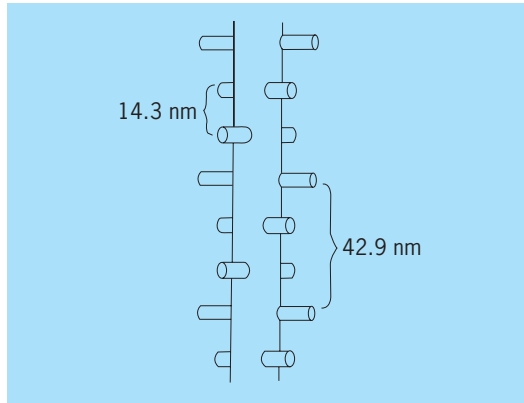


Fig. 3. Schematic diagram of the arrangement of cross bridges in a vertebrate thick filament, showing a three-stranded helix. The peg-shaped bridges are purely schematic and should be interpreted to show only the location and not the shape of the cross bridges. (After A. T. Bull et al., eds., *Companion to Biochemistry: Selected Topics for Further Study*, Longman, 1974)

peptide chains (molecular weight of about 2×10^5) whose ends form two separate globular structures. The intertwined portion has a high α -helical content and forms a rigid rod. Attached to each head is a set of two light chains, one of which can be phosphorylated. A light chain of the latter type is involved in the regulation of smooth muscle and molluscan muscle, hence the name regulatory light chain. The amino acid sequence of both light and heavy chains is different among different muscle cells, such as those of fast skeletal muscle, smooth muscle, cardiac muscle, and embryonic muscle. That gives rise to many myosin forms.

The diversity of myosin is due to the existence of a family of genes coding for the heavy and light chains. These genes show slight variation between species. Within a given species, the expression of these genes shows developmental control as well as tissue specificity in muscles of different function, and it is also susceptible to change depending on the activity of the muscle.

Myosin functions as an ATP-cleaving enzyme, adenosine triphosphatase (ATPase). Each myosin head has an ATP binding site as well as an actin-binding site. The ATPase activity of myosin itself is low in media that contain magnesium ion, but if actin is added to myosin at low ionic strength in the presence of magnesium ion, considerable activation of ATPase takes place. This actin-activated ATPase reac-

tion, which furnishes energy of contraction and the interaction of myosin with actin, is an essential element of the contraction mechanism. See ADENOSINE TRIPHOSPHATE (ATP).

Light meromyosin shows a tendency to form regular aggregates, which is thought to be involved in the process leading to the assembly of myosin molecules into a thick filament. Thus there is a central core of strongly interacting light meromyosin segments with radially protruding soluble heavy meromyosin portions (Fig. 3). Under the electron microscope, the heavy meromyosin moieties appear as projections issuing from the thick filaments that occasionally touch adjacent actin filaments. These projections have been called cross bridges; a cross bridge is an S-1 moiety attached by an S-2 to the core of the thick filament. In vertebrate muscle there are three cross bridges, separated by 120° angles on planes axially separated by 14.3 nm. The cross-bridge triplets are rotated by 40° on successive planes. The central portions of the thick filaments are devoid of cross bridges.

Actin. In the monomeric form, actin is a bilobular-shaped, nearly globular unit with a diameter of about 6 nm, a height of about 4 nm, and a molecular mass of 4.2×10^4 daltons. Actin, as isolated in the monomeric form, contains a tightly bound atom of magnesium and one bound ATP molecule in a cleft between the two lobes. The monomeric units are arranged in a helical fashion in the thin filament. Since the actin units are not bound by covalent bonds, the helical arrangement can be regarded either as one involving a single helix or one describable as the intertwining of two helices. The change from the globular to the fibrous form can be brought about in the test tube by adding salts and is accompanied by the hydrolysis of the bound ATP to bound adenosine diphosphate (ADP). The transformation of ATP to ADP taking place during polymerization is not involved in muscle contraction. This reaction presumably occurs when actin filaments are laid down in the course of development, growth, or regeneration. See ADENOSINE DIPHOSPHATE (ADP).

Activation by calcium. The interaction of actin and myosin in striated muscle is regulated by the cellular calcium level through a protein complex attached at regular intervals to the thin filament. Two protein complexes associated with the actin filaments are tropomyosin and troponin. Tropomyosin is an α -helical protein. A stretch of actin contains 13 monomers in contact with a pair of tropomyosin molecules arranged in a diametrically opposite fashion. Each tropomyosin molecule carries a troponin complex of three proteins, troponin-C, -I, and -T (Fig. 4). Troponin-T is the subunit most likely to be involved in association with tropomyosin. Troponin-I is known for its ability to inhibit the interaction of actin and myosin. Troponin-C acts as a calcium sensor to detect the myoplasmic calcium concentration, which in resting muscle is kept low by the calcium pump of the sarcoplasmic reticulum, an intracellular membrane system associated with the T tubule, which in turn is an invagination of the cell

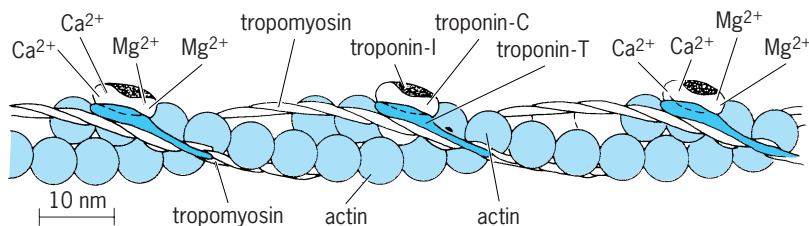


Fig. 4. Schematic representation of regulated thin filament. (After R. L. Moss, J. D. Allen, and M. L. Greaser, *Effects of partial extraction of troponin complex upon the tension-pCa relation in rabbit skeletal muscle: Further evidence that tension development involves cooperative effects within the thin filament*, *J. Gen. Physiol.*, 87:761-774, 1986)

membrane. Other components include a ryanodine receptor, triadine, and a voltage sensor protein participating in the process of calcium ion release. When the calcium level is high enough, the contractile machinery is turned on. As the stimulus ceases, the calcium level drops and the machinery is again turned off. Changes in the protein structure of troponin-C that occur upon binding of calcium ions lead to changes in the other components of the thin filament and finally to activation of the actin-myosin cross-bridge interaction.

The mechanism by which the calcium ion-induced changes in the thin filament lead to activation of muscle contraction is still not fully understood. Earlier views regarded the attachment of myosin heads to actin as the regulated step. In contrast, later biochemical studies of myosin ATPase activated by regulated actin in the test tube suggest that activation involves a shift among states of myosin attached to actin, from one of weak binding to one of strong binding. Time-resolved x-ray diffraction on live muscle shows that the first structural changes in the thin filaments may involve a movement of tropomyosin, but the link between the movement and the changes in actin-myosin interaction require further clarification.

In striated muscles of invertebrates, the regulatory machinery is different. In mollusks, for example, sensing of calcium ion is performed solely by myosin, with the light chains playing a large role and tropomyosin troponin being unnecessary. A myosin-linked mechanism also operates in vertebrate smooth muscle, where phosphorylation triggered by the increase in calcium ion is involved. In smooth muscle, the thin filaments do not contain troponin. Regulation involves the phosphorylation of myosin, the reaction being catalyzed by an enzyme (myosin light-chain kinase), that is activated by the combination of calcium ions released on stimulation with calmodulin, a protein present in all cells. See CALCIUM METABOLISM; MUSCLE.

The structure of dystrophin and the associated complex of proteins containing sugarlike molecules (glycoproteins) attached to the cell membrane has been clarified. The absence of dystrophin and associated proteins leads to Duchenne muscular dystrophy. See MUSCULAR DYSTROPHY. John Gergely

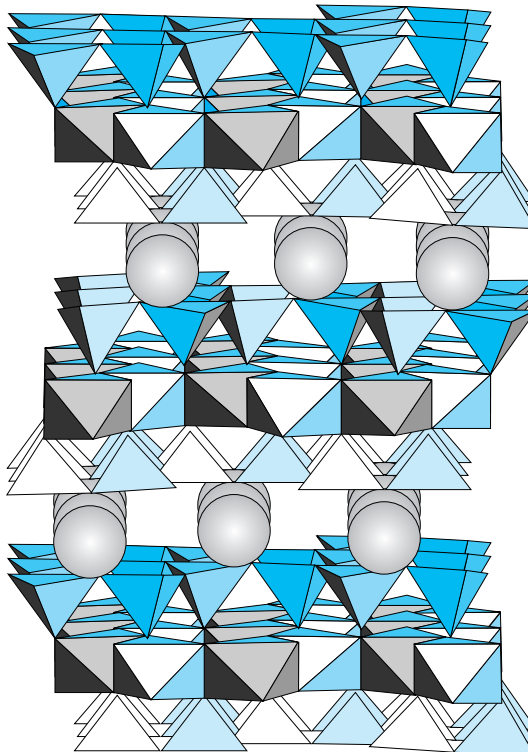
Bibliography. A. G. Engel and B. Q. Banker, *Myology*, 2d ed., 1994; Z. Grabarek, T. Tao, and J. Gergely, Molecular mechanism of troponin-C function, *J. Muscle Res. Cell Motil.*, 13:383-393, 1992; K. Holmes and W. Kabsch, Muscle proteins: Actin, *Curr. Opin. Struc. Biol.*, 1:120, 1991; K. Matsumura and K. P. Campbell, Dystrophin-glycoprotein complex: Its role in the molecular pathogenesis of muscular dystrophies, *Muscle and Nerve*, 17:2, 1994; I. Rayment et al., Structure of the actin-myosin complex and its implications for muscle contraction, *Science*, 261:58, 1993; I. Rayment et al., The three-dimensional structure of myosin subfragment 1: A molecular motor, *Science*, 261:50, 1993; J. Trinick, Understanding the functions of titin and nebulin, *FEBS Lett.* (Federation of European Biochemical So-

cieties), 307:44-48, 1992; Lord Walton (ed.), *Disorders of Voluntary Muscle*, 6th ed., 1994.

Muscovite

A mineral of the mica group with an ideal composition of $KAl_2(AlSi_3)O_{10}(OH)_2$. Sometimes it is referred to as a white mica or potash mica.

Structure. Muscovite is a dioctahedral mica, where two of the three possible octahedral cation sites are occupied by aluminum atoms (Al; referred to as M2 sites) and the other site (M1) is vacant. The M2 octahedra are formed with six oxygen atoms (O) or hydroxyl groups surrounding the aluminum atoms; these octahedra are linked laterally by the sharing of edges to form an octahedral sheet (see *illus.*). Tetrahedra are located on either side of the octahedral sheet. Each tetrahedron comprises silicon (Si) and aluminum (T site) bonded to four oxygen atoms, with generally 25% random substitution of silicon by aluminum at this site. A two-dimensional network of sixfold rings (ditrigonal) is formed by corner sharing of tetrahedra to produce a tetrahedral sheet. The single octahedral sheet sandwiched between the two tetrahedral sheets forms a 2:1 layer. Potassium (K)



A polyhedral model illustrating the muscovite-2M₁ atomic structure. The top and bottom layers are identical, but the middle layer differs; note that the octahedral faces are oriented differently. The structural unit includes the center layer and either the top or bottom layer. The corners of the tetrahedra represent oxygen atoms, whereas octahedral corners represent oxygen or hydroxyl groups. The octahedra contain aluminum ions, whereas the tetrahedra contain aluminum or silicon. The polyhedral interiors, the aluminum and (aluminum/silicon) ions, are not shown. Large spheres are interlayer potassium ions, which serve to connect the layers. Hydrogen atoms are not shown.

ions (A site) are located in the interlayer region, where this large cation locks together ditrigonal rings from adjacent tetrahedral sheets.

The stable form of muscovite is a two-layer ($2M_1$) structure, with overall monoclinic symmetry in space group $C2/c$. Regular monoclinic one-layer ($1M$) and disordered forms ($1M_d$) are less abundant, and a trigonal three-layer ($3T$) and a two-layer form ($2M_2$) different from $2M_1$ are rare. An apparent stability sequence exists with $1M_d$ transforming to $1M$ and then to $2M_1$ as either temperature or annealing time is increased. K-deficient forms are generally $1M$ in structure. See CRYSTAL STRUCTURE.

Chemistry. The A sites in muscovite may contain K, sodium (Na), barium (Ba), cesium (Cs), and rubidium (Rb); the M2 sites may have minor amounts of magnesium (Mg), ferrous iron [Fe(II)], ferric iron [Fe(III)], lithium (Li), manganese (Mn), titanium (Ti), and chromium (Cr), in addition to aluminum; the T sites contain only silicon and aluminum, and minor amounts of fluorine (F) can replace the hydroxyl group. In the older literature, the term fuchsite was used for the green variety of muscovite, which contains chromium. Solid solutions between other micas, either dioctahedral or trioctahedral, are quite limited. Related minerals include illite, which is similar to muscovite but has more silicon, magnesium, and water (H_2O) and less potassium; and phengite, which contains significant amounts of divalent cations in the M sites and more silicon than muscovite. See ILLITE; LEPIDOLITE; SOLID SOLUTION.

Properties. The layerlike qualities of the structure produce a perfect basal (001) cleavage. Specific gravity is 2.76–2.88, hardness on the Mohs scale is 2–2.5, and luster is vitreous to pearly. Thin sheets are flexible and may be colorless, with books (thick crystals) translucent, yellow, brown, reddish, or green. Cation exchange capacity is low because the interlayer cation fully compensates the residual charge on the 2:1 layer. Upon heating in air to temperatures above dehydroxylation [the actual temperatures of dehydroxylation are kinetically controlled, about 800°C (1470°F) for larger crystals, and 500°C (930°F) for fine-grained material], muscovite transforms topotactically to a dehydroxylate structure where the M sites become five-coordinated. See COORDINATION CHEMISTRY; HARDNESS SCALES.

Occurrence. Muscovite occurs commonly in all the major rock types, in igneous rocks (granites, pegmatites, and hydrothermal alteration products), in metamorphic rocks (slates, phyllites, schists and gneisses), and in sedimentary rocks (sandstones and other clastic rocks). See ROCK.

Uses. As larger flakes, muscovite is used as an electrical insulator, both for its dielectric properties and for its resistance to heat. Ground muscovite is used for fireproofing, as an additive to paint to provide a sheen and for durability, as a filler, and for many other applications. See MICA; SILICATE MINERALS.

Stephen Guggenheim

Bibliography. S. W. Bailey (ed.), *Micas, Reviews in Mineralogy*, vol. 13, 1984.

Muscular dystrophy

A group of muscle diseases that are hereditary and characterized by progressive muscle weakness and wasting.

Etiology. The muscular dystrophies are primary diseases of the muscle cells characterized by progressive degeneration and replacement by fibrous tissue, resulting in progressive muscle weakness. In some types of muscular dystrophies, the disease appears to be restricted to the skeletal muscles alone (facioscapulohumeral muscular dystrophy, limb-girdle muscular dystrophy), and in others skeletal-muscle involvement is a part of a more generalized process, with abnormalities in other organ systems as well (Duchenne's muscular dystrophy, myotonic dystrophy). These features as well as the differing patterns of inheritance (see **table**) indicate that the various muscular dystrophies are different diseases with different genetic and biochemical abnormalities underlying them.

The gene for the Duchenne and the Becker muscular dystrophy has been identified and localized to band Xp21 on the short arm of the X chromosome. This gene produces the muscle protein dystrophin, absent in Duchenne's dystrophy and qualitatively altered in Becker's dystrophy. Although the gene for myotonic dystrophy has not been identified, it has been found to be closely linked to the secretor locus and has been mapped to the proximal long arm of chromosome 19. See HUMAN GENETICS; MUSCLE PROTEINS.

Pathology. The muscle shows varying degrees of necrosis of individual muscle fibers with phagocytosis, as well as regeneration of some fibers. In addition, in many fibers there is an internal migration of the muscle cell nuclei. The individual muscle fibers become rounded and show considerable variability in their diameter in a random fashion. As the disease progresses, the necrotic muscle fibers are replaced by fibrous connective tissue and fat. The changes in the muscle are much more marked in the more rapidly progressive forms of muscular dystrophy (for example, Duchenne's dystrophy), and may be minimal in the slowly progressive ones (for example, myotonic dystrophy).

As a reflection of muscle degeneration, enzymes originating in skeletal muscle can often be detected in the serum in increased amounts.

Diagnosis. The diagnosis of the various types of muscular dystrophies still remains largely descriptive, and is based on the sex of the patient, the age of onset, the muscle groups initially affected, other muscular and nonmuscular abnormalities, and the pattern of inheritance. Since no specific laboratory test is available for most muscular dystrophies, other than the X-linked types, the number of different types varies depending on the authority consulted. The most common forms are the X-linked recessive Duchenne's muscular dystrophy and the dominantly inherited myotonic dystrophy. Other, less common forms include the milder X-linked Becker's muscular dystrophy, which is allelically related to Duchenne's;

| Characterization of the muscular dystrophies | | | | |
|--|----------------------|-------------------------------|--|-----------------------|
| Type | Age | Involved muscles | Other features | Inheritance |
| Pseudohypertrophic Duchenne | Less than 4 years | Pelvic girdle | Pseudohypertrophy, mental retardation, abnormal cardiogram | X-linked |
| Becker | 8–10 years | Pelvic girdle | Pseudohypertrophy, abnormal cardiogram | X-linked |
| Facioscapulohumeral | Adolescence | Face, shoulder, upper arm | — | Dominant |
| Myotonic | Young adult | Face, neck, hands | Myotonia, cataract, gonadal atrophy, balding | Dominant |
| Limb girdle | Adolescence or older | Shoulder and pelvic girdle | — | Recessive |
| Ocular | Any age | Extraocular muscles | Very variable | Dominant or recessive |

dominantly inherited facioscapulohumeral muscular dystrophy; and the recessively inherited limb-girdle muscular dystrophy. The ocular myopathies appear to include several different disease entities, all of which have as a common feature the progressive paralysis of extraocular muscles.

Duchenne's muscular dystrophy. This is the most rapidly progressive form of muscular dystrophy. It affects boys before the age of 4, and is characterized initially by progressive weakness of the hip muscles with difficulty in rising from the floor or chair and in climbing stairs. This is accompanied by enlargement of the calf muscles, which are infiltrated by fat and fibrous tissue (pseudohypertrophy). Weakness of the muscles of the upper arms and shoulder muscles follows. By 10 to 12 years of age, most of the affected children are restricted to a wheelchair, and the majority die in the late teens or early twenties. In the majority of cases, their mothers, who are carriers of this gene, usually have few or no symptoms, but often have elevated serum creatine kinase levels. Although their daughters do not develop this disease, each has a 50% chance of being a carrier of this gene, with a 50% possibility of transmitting this disease to each of her male offspring. In the remaining cases, the affected boy appears to represent a new mutation for this gene and the mother is not the carrier.

Complementary deoxyribonucleic acid (cDNA) probes for the Duchenne muscular dystrophy gene and antibodies to dystrophin now allow identification of many affected male children or fetuses and obligatory female carriers of this gene. The allelically related Becker's muscular dystrophy is also characterized by calf pseudohypertrophy but is much more slowly progressive. The genetic abnormality resides in the same site of the X chromosome as with the Duchenne's type, but the muscle protein dystrophin is present in skeletal muscles, although in altered form.

Myotonic dystrophy. Myotonic dystrophy is very slowly progressive and affects the muscles of the face, neck, and hands. It usually begins in early adulthood. In addition to progressive weakness and wasting of the affected muscles, these individuals also

exhibit myotonia, that is, a delayed relaxation of a muscle after forceful contraction. Myotonia is most readily demonstrated in the hands, where there is a slow relaxation of the fingers after forceful grasping. However, the abnormality in myotonic dystrophy is not restricted to the skeletal muscles. Mental retardation, frontal balding, cataracts, and gonadal degeneration are common. Incidence of diabetes mellitus and cardiac disease is higher in patients with myotonic dystrophy than in the general population. The close linkage to genetic markers on chromosome 19 allows the identification of presymptomatic or asymptomatic members of some families.

Treatment. Treatment remains largely symptomatic. However, the identification of the Duchenne muscular dystrophy gene and the resultant absence of dystrophin raise the possibilities of treatment by genetic manipulation or replacement of dystrophin. In the meantime, these findings provide more accurate prenatal diagnosis and genetic counseling. See MUSCULAR SYSTEM DISORDERS.

S. M. Sumi; Thomas D. Bird

Bibliography. G. L. Barnett, *Muscular Dystrophy*, 1995; A. E. Emery, *Duchenne Muscular Dystrophy*, 2d ed., 1993; A. G. Engle and C. Franzini-Armstrong, *Myology*, 2 vols., 2d ed., 1994; R. N. Rosenberg and A. E. Harding (eds.), *Molecular Biology of Neurological Disease*, 1988; J. Walton, G. Karpati, and D. Hilton-Jones (eds.), *Disorders of Voluntary Muscle*, 1994.

Muscular system

The muscular system consists of muscle cells, the contractile elements with the specialized property of exerting tension during contraction, and associated connective tissues. The three morphologic types of muscles are voluntary muscle, involuntary muscle, and cardiac muscle. The voluntary, striated, or skeletal muscles are involved with general posture and movements of the head, body, and limbs. The involuntary, nonstriated, or smooth muscles are the muscles of the walls of hollow organs of the digestive,

circulatory, respiratory, and reproductive systems, and other visceral structures. Cardiac muscle is the intrinsic muscle tissue of the heart. *See* MUSCLE.

In this article the comparative embryology of the voluntary and involuntary muscles of the vertebrates will be outlined, followed by the comparative anatomy of the muscular system.

Comparative Anatomy

Phylogenetically speaking, muscles are very plastic organs. They have undergone considerable change during the evolution of vertebrates, correlated in large part with changes in the organisms' environments and in their methods of support and locomotion. Establishment of homologies among muscles is not easy. Adult relationships can be misleading because muscles have subdivided during their evolution, and parts have migrated far from their original positions. Nerve supply is a more reliable criterion, because nerves have tended to follow the muscles through their evolutionary gymnastics, but often homologies cannot be established without recourse to embryonic development. Determination of the development of the thousands of individual muscles among the vertebrate classes is a monumental task. Comparison of muscles among vertebrates is greatly facilitated if the muscular system is subdivided into groups whose homology can be more easily established in the various classes.

Muscle groups are particularly distinct in elasmobranchs and other primitive fishes, and they are generally defined on the basis of their embryonic origin in these animals. Two major groups of skeletal muscles are recognized, somatic (parietal) muscles, which develop from the myotomes, and branchiomic muscles, which develop in the pharyngeal wall from lateral plate mesoderm. The somatic musculature is subdivided into axial muscles,

which develop directly from the myotomes and lie along the longitudinal axis of the body, and appendicular muscles, which develop within the limb bud from mesoderm derived phylogenetically as buds from the myotomes.

The vertebrate muscular system is the largest of the organ systems, making up 35–40% of the body weight in humans. The movement of vertebrates is accomplished exclusively by muscular action, and muscles play the major role in transporting materials within the body. Muscles also help to tie the bones of the skeleton together and supplement the skeleton in supporting the body against gravity. *See* SKELETAL SYSTEM.

Axial musculature. Most of the axial musculature is located along the back and flanks of the body, and this part is referred to as trunk musculature. But anteriorly the axial musculature is modified and assigned to other subgroups. Certain of the occipital and neck myotomes form the hypobranchial muscles, and the most anterior myotomes form the extrinsic ocular muscles.

The trunk musculature of cyclostomes consists of a long series of segmental myomeres, each consisting of many longitudinal fibers attaching onto the myosepta (**Fig. 1**). Each is folded in such a way as to appear approximately zigzag-shaped on the surface. The arrangement in jawed fishes is essentially the same, but the folding of the myomeres is more complex, and each is divided by a horizontal connective-tissue septum into dorsal (epaxial) and ventral (hypaxial) portions. A spinal nerve passes to each myomere, the dorsal ramus going to the epaxial portion and the ventral ramus to the hypaxial portion. This pattern of innervation persists in all higher vertebrates.

Epaxial musculature. The epaxial musculature remains powerful in most cases. In amphibians it consists of a group of medial and deep fibers that interlace the vertebrae, and a larger group of superficial fibers (dorsalis trunci). Segmentation is retained and undulations of the trunk and tail still play a role in the locomotion of many amphibians (**Fig. 2**).

In many reptiles, the epaxial musculature is more complex. A medial and deep group of small, largely segmental muscles bind the vertebrae together and constitute the transversospinalis system; more laterally the musculature is arranged in two more extensive longitudinal groups, the longissimus dorsi, which lies dorsal to the transverse processes, and the iliocostalis, which is attached to the ribs.

These three main divisions persist in mammals, but posteriorly there is a union of the iliocostalis, longissimus, and sometimes the more superficial part of the transversospinalis system to form a powerful erector spinae (sacropinalis) complex that helps to support the vertebral column. In mammals the body is held off the ground by the limbs; thus the backbone is sometimes compared to a girder supported anteriorly and posteriorly by pillars. Much of the epaxial musculature functions as tie members resisting tension stresses along this girder. Anteriorly there is a cleavage of the epaxial divisions into a host

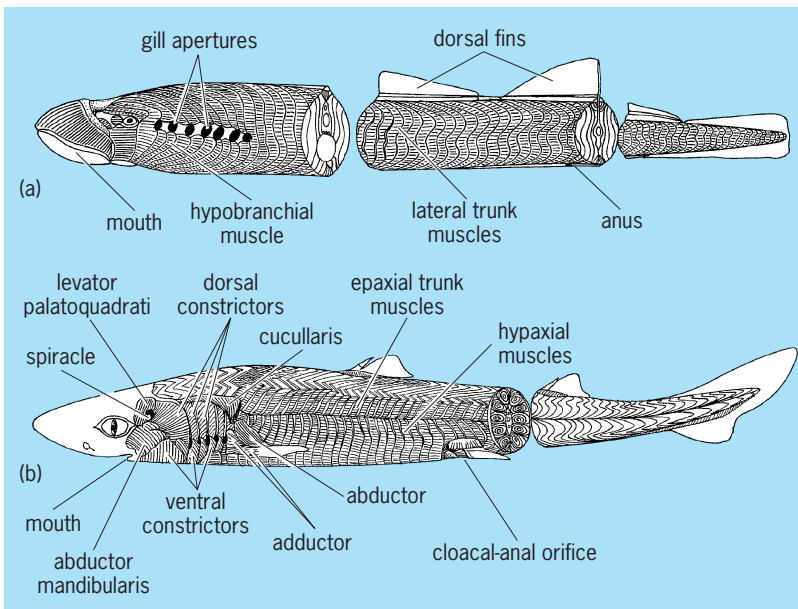


Fig. 1. Superficial muscles. (a) Cyclostome (*Petromyzon*). (b) Elasmobranch (*Squalus*). (After H. W. Rand, *The Chordates*, Blakiston, 1950)

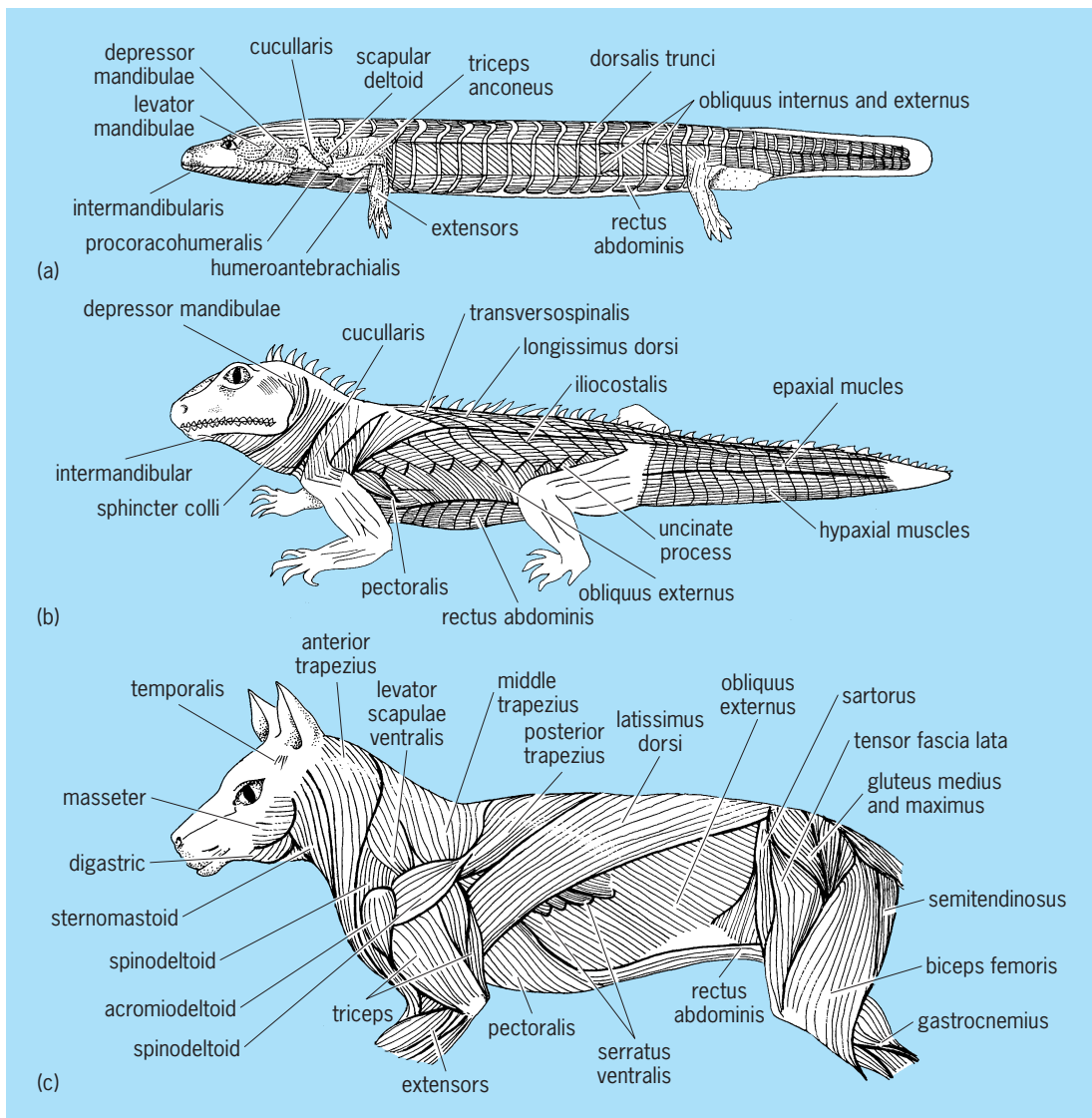


Fig. 2. Superficial muscles of three vertebrates, showing segmentation. (a) Amphibian (*Necturus*). (b) Reptile (*Sphenodon*). (c) Mammal (*Felis*). (After H. W. Rand, *The Chordates*, Blakiston, 1950)

of muscles associated with the complex head and neck movements.

In birds, the epaxial musculature in the trunk is greatly reduced, correlated with a fusion of many of the trunk vertebrae.

Hypaxial musculature. The hypaxial musculature of tetrapods can be subdivided into three groups: (1) a subvertebral (hyposkeletal) group located ventral to the transverse processes and lateral to the centra of the vertebrae, (2) the flank muscles forming the lateral part of the body wall, and (3) the ventral abdominal muscles located on each side of the midventral line.

The subvertebral musculature assists the epaxial muscles in the support and movement of the vertebral column. In mammals, it consists of longitudinal bundles—the longus colli in the neck and the anterior thorax, the quadratus lumborum, and psoas minor more posteriorly.

Most of the flank musculature takes the form of broad, thin sheets of muscle that form much of

the body wall and support the viscera. The ancestral, segmental nature of this musculature is retained throughout the trunk in salamanders, but is lost in higher tetrapods except in those parts of the trunk where ribs are well developed (Fig. 2). Three layers can be distinguished in the abdominal region of most tetrapods: a superficial external oblique, whose fibers extend caudally and ventrally; an internal oblique with fibers at right angles to the preceding; and a deep transversus abdominis. This pattern is much the same in the costal region, external intercostals, internal intercostals, and a reduced transversus thoracis being present in mammals. In reptiles the pattern is more complex; the external layer is represented by supracostals, external intercostals, and sometimes a subcutaneous muscle.

Respiratory movements of reptiles and birds are accomplished by the costal and abdominal muscles described above, but in mammals, which have a higher metabolic rate, additional respiratory muscles have evolved from the hypaxial muscles: the

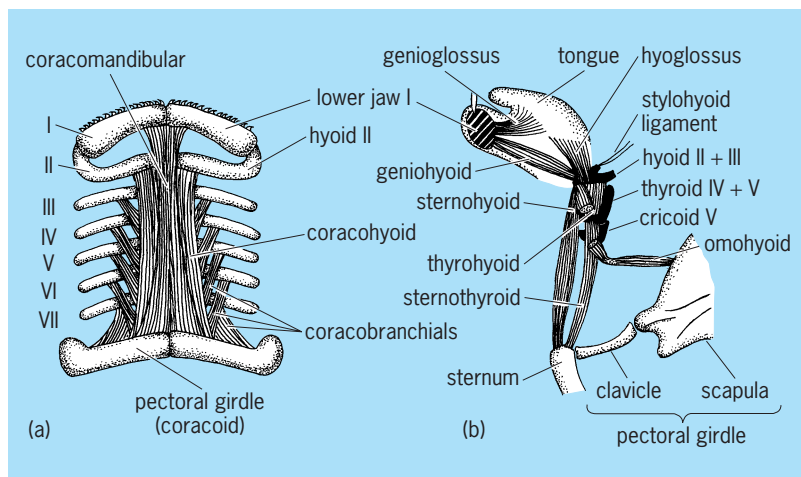


Fig. 3. Diagram of the major hypobranchial muscles. (a) Shark, in ventral view. (b) Mammal, in lateral view. (After H. W. Rand, *The Chordates*, Blakiston, 1950)

diaphragm (a derivative of cervical myotomes), serratus dorsalis, scalenes, and transversus costarum. See RESPIRATORY SYSTEM.

Other parts of the hypaxial flank musculature have gained an attachment to the pectoral girdle where they help to transfer body weight from the vertebral column to the girdle and appendage and help to regulate the movement of the girdle. Only a few muscles of this type, the thoracoscapularis and levator scapulae, for example, are present in primitive tetrapods such as salamanders, and the body is not held far off the ground. In mammals, however, this group includes such large and powerful muscles as the serratus ventralis, rhomboideus, and levator scapulae ventralis. In the pelvic region of tetrapods, weight is transferred to the appendage directly across the sacroiliac joint and not by muscles.

The midventral hypaxial musculature in all tetrapods consists of the rectus abdominis, a longitudinal muscle on each side of the midline that extends from the pelvic region to the anterior part of the trunk (Fig. 2). It has evolved from the oblique flank muscles and in some salamanders remains closely associated with them. Transverse tendinous inscriptions are often present and are believed to represent persistent myosepta.

Hypobranchial musculature. The hypobranchial musculature extends from the pectoral girdle forward along the ventral surface of the neck and pharynx to the hyoid arch, chin, and into the tongue. It is regarded as a continuation of part of the hypaxial trunk musculature, because it develops ontogenetically in most vertebrates from the ventral portion of several occipital and neck myotomes that grow around the back of the gill region and then into the neck and tongue. The innervation of the hypobranchial muscles in amniotes by the cervical nerves and the hypoglossal nerve, which itself has evolved from certain of the spinal and occipital nerves of fishes and amphibians, further indicates the myotomic origin of this group.

The hypobranchial musculature of cyclostomes (Fig. 1) retains its primitive myomeric character, but the myomeres fuse to form longitudinal muscles in

higher vertebrates. Traces of myosepta are evident in fishes and some amphibians, but these disappear in amniotes.

In all gnathostomes, the hypobranchial musculature can be divided at the level of the hyoid arch into prehyoid (infrahyoid) and posthyoid (suprahyoid) groups (Fig. 3). The prehyoid group of elasmobranchs consists of a single pair of muscles, the coracomandibulars, extending caudally from the jaw symphysis to attach to the posthyoid group slightly anterior to the pectoral girdle. The posthyoid group consists of a superficial mass, which can be subdivided into a coracoarcual and coracohyoid, extending between the pectoral girdle and hyoid arch, and a deeper mass, the coracobranchials, extending from the pectoral girdle to the ventral surface of the branchial arches. The coracobranchials act to expand the pharynx and gill pouches; the others help to support the floor of the pharynx and help to move the hyoid arch and open the mouth.

During the evolution of terrestrial vertebrates, loss of most of the coracobranchials occurs, along with the reduction and loss of many parts of the branchial arches, but the rest of the hypobranchial musculature remains and becomes modified in correlation with the evolution of a muscular tongue and the more complex problem of deglutition in a terrestrial environment. The prehyoid group of amphibians consists primarily of a geniogyoid extending from the chin to the hyoid, but a few muscle fibers have separated from it and enter the tongue. These represent the beginning of a complex group of muscles that manipulates the tongue of amniotes: genioglossus, hyoglossus, styloglossus, and probably the intrinsic lingualis. The posthyoid group of primitive terrestrial vertebrates consists primarily of a rectus cervicis extending between the ventral part of the pectoral girdle and hyoid arch, but several slips separate from it to go to other parts of the girdle or to the remnants of the branchial arches. In higher vertebrates, the rectus cervicis has split into several muscles acting upon the larynx and hyoid: sternohyoid, sternothyroid, thyrohyoid, and omohyoid.

Extrinsic ocular muscles. The extrinsic ocular muscles develop from the prootic somites (head cavities). All vertebrates have six in common (Fig. 4). The median (internal), superior, and inferior recti, together with the inferior oblique, develop from the first somite and are innervated by the oculomotor nerve. The superior oblique and lateral (external) rectus develop from the second and third somites, respectively, and are supplied, respectively, by the trochlear and abducens nerves. All of these muscles insert on the eyeball and move it. In addition, most terrestrial vertebrates, with the exception of birds, have a retractor oculi, a cone-shaped muscle lying deep to the recti, that pulls the eyeball deeper into its socket. The retractor oculi has evolved from the lateral rectus and continues to be innervated by the abducens. In many reptiles and in birds, one or more small muscles have also separated from the lateral rectus, or from the retractor oculi, and act upon the nictitating membrane. A levator palpebrae superioris, present in mammals,

completes the ocular group. This muscle elevates the upper eyelid in opposition to the action of certain facial muscles. It has evolved from the superior rectus and is innervated by the oculomotor nerve. *See EYE (VERTEBRATE).*

Appendicular musculature. Limb muscles are often classified as intrinsic if they lie entirely within the confines of the appendage and girdle, and extrinsic if they extend from the girdle or appendage to other parts of the body. This scheme has certain merits, but is misleading from the phylogenetic point of view, because the extrinsic muscles are not all appendicular in the sense in which this group has been defined.

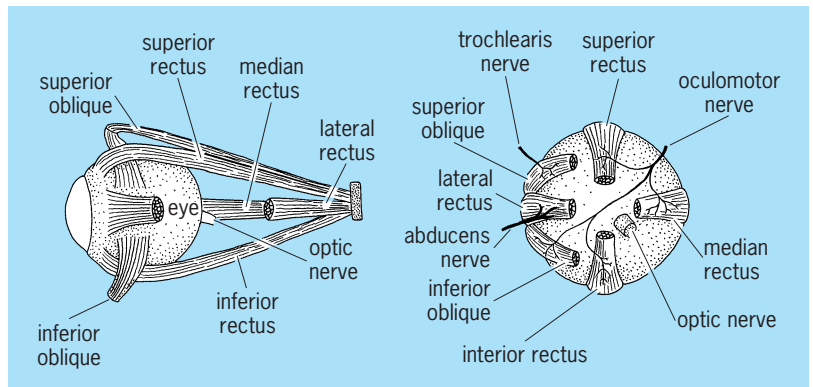


Fig. 4. Human extrinsic ocular muscles, which develop from prootic somites. (After H. W. Rand, *The Chordates*, Blakiston, 1950)

| Homologies of appendicular muscles* | | |
|-------------------------------------|--|--|
| | <i>Necturus</i> | Cat |
| Pectoral muscles | | |
| Dorsal | Latissimus dorsi | Citamepis trimci (part) |
| | | Latissimus dorsi |
| | Subcoracoscapularis | Teres major |
| | Scapular deltoid | Subscapularis |
| | Procoracohumeralis | Spinodeltoid |
| | | Acromiodeltoid |
| Triceps | Clavodeltoid | |
| | Teres minor | |
| Forearm extensors | Triceps | |
| | Epitrochlearis | |
| Ventral | Pectoralis | Forearm extensors |
| | | Cutaneous trunci (part) |
| | Supracoracoideus | Pectoralis complex |
| | Coracoradialis | Supraspinatus |
| | | Infraspinatus |
| | Humeroantibrachialis (or brachialis) | Biceps brachii |
| | Coracobrachialis | Brachialis inferior |
| Forearm flexors | Coracobrachialis | |
| | Forearm flexors | |
| Pelvic muscles | | |
| Dorsal | Iliotibialis | Sartorius |
| | | Iliacus |
| | Puboischiofemoralis internus | Psoas major |
| | | Pectineus |
| | Ilioextensorius | Vasti(?) |
| | | Iliofibularis |
| Iliofemoralis | Gluteus maximus | |
| | Shank extensors | Tensor fasciae latae(?) |
| | Gluteus medius | |
| | Gluteus minimus | |
| | Shank extensors | |
| Ventral | Puboischiofemoralis externus (adductor femoris, in <i>Necturus</i> not clearly separated from the preceding) | Obturator externus |
| | | Quadratus femoris |
| | Pubotibialis | Adductor brevis of longus |
| | | Adductor magnus |
| | Ischiofemoralis | Obturator internus |
| | | Gemelli |
| | Caudofemoralis | Crucrococcygeus (absent in some mammals) |
| Pyramiformis | | |
| Puboischiotibialis | Gracillius | |
| | Ischioflexorius | Semimembranosus |
| Shank flexors | Semitendinosus | |
| | | Biceps femoris(?) |
| | Shank flexors | |

*From W. F. Walker, Jr., *Vertebrate Dissection*, Saunders, 1970.

Some are appendicular muscles that have developed directly from mesoderm in the limb bud, but others are trunk, hypobranchial, and branchiomic muscles that have secondarily become associated with the girdle. These muscles are considered with the group to which they phylogenetically belong.

Fishes. The paired fins of fishes are primarily horizontal stabilizing keels, but they are also used in deceleration and steering, and to help control the depth at which the fish swims. Fin movements are not complex or powerful and the appendicular muscles in the strictest sense or morphologically simple. A single dorsal muscle (abductor) and a comparable ventral muscle (adductor) extend from the girdle into the fin (Fig. 1). Certain fibers of these muscles are so arranged that they can protract and retract the fin. The appendicular muscles are supplied by the ventral rami of spinal nerves.

Terrestrial vertebrates. In terrestrial vertebrates, the limbs become the main organs for support and locomotion, and the appendicular muscles become correspondingly powerful and complex. The muscles are too numerous to describe individually, but they can be sorted into dorsal and ventral groups, because tetrapod muscles originate embryonically in piscine fashion from a dorsal and a ventral premuscular mass within the limb bud (see table). In general, the ventral muscles, which also spread onto the anterior surface of the girdle and appendage, act to protract and adduct the limb and to flex its distal segments; the dorsal muscles, which also extend onto the posterior surface of the girdle and appendage, have the opposite effects (retraction, abduction, and extension). The limb muscles also serve as flexible ties or braces that can fix the bones at a joint and support the body.

Amphibians and reptiles. When at rest, the belly of most amphibians and reptiles is on the ground, and the proximal segment of each limb extends laterally and slightly dorsally from its articulation with the girdle. Locomotion involves the partial adduction of the humerus and femur to raise the body off the ground, as well as their protraction and retraction. Ventral adductor muscles, such as the pectoralis and supracoracoideus in the pectoral region and the puboischiofemoralis externus in the pelvic region, are

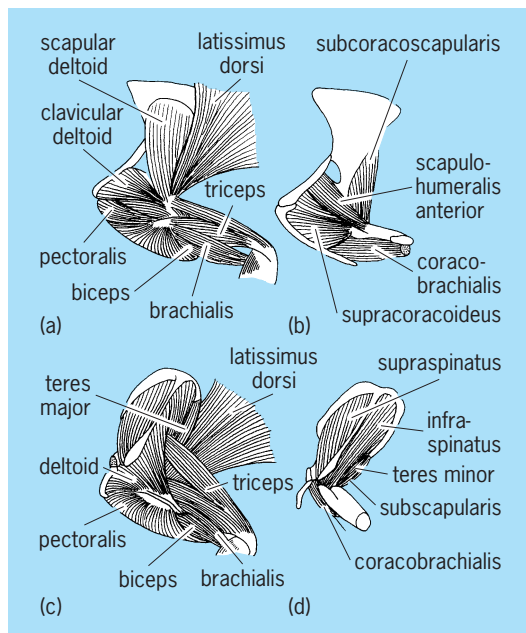


Fig. 5. Lateral views of the shoulder and upper arm muscle structure of two animals. (a, b) Lizard. (c, d) Opossum. Superficial muscles are shown in a and c; deep muscles are shown in b and d. (After A. S. Romer, *The Vertebrate Body*, 3d ed., Saunders, 1962)

relatively large and powerful. During the evolution of mammals, the limbs have rotated under the body and extend ventrally from the girdle. The body is held off the ground by bony columns braced by muscles. Adductor muscles are less powerful and certain ones of them have migrated to other positions and have assumed other functions. The supracoracoideus, for example, has extended dorsally onto the scapula to form the mammalian supraspinatus and infraspinatus (**Fig. 5**). These muscles now act as braces for the limb in its new position and play a role in its protraction and retraction.

Birds. Flight in birds has entailed a considerable modification of the musculature of the pectoral region. As one example, the ventral adductor muscles are exceedingly large and powerful, and the area from which they arise is increased by the enlargement of the sternum and the evolution of a large sternal keel. Not only does a ventral muscle, the pectoralis, play a major role in the downstroke of the humerus, but a ventral muscle, the supracoracoideus, is active in the upstroke as well. The tendon of insertion of the supracoracoideus has shifted so that it passes through a canal between the clavicle, coracoid, and scapula to attach to the upper surface of the humerus. Its action is analogous to pulling down on a rope that passes over a pulley and down onto a weight.

Branchiomic musculature. The branchiomic (branchial) muscles of fishes form a conspicuous part of the muscular system and are rather complex. In jawed fishes they can be subdivided according to the visceral arch with which they are associated. Mandibular muscles act upon the first, or mandibular arch, and are supplied by the trigeminal nerve. The

group includes such muscles as the levator palatoquadrati, which in the dogfish helps to support the palatoquadrata cartilage; the adductor mandibulae, the powerful muscles closing the jaws; and the intermandibularis, which together with certain hypobranchial and hyoid muscles, opens the jaws (**Fig. 1**).

Hyoid muscles act on the second or hyoid arch and are supplied by the facial nerve. The hyoid arch is modified in sharks and many other fishes to help support the palatoquadrata, and its musculature is correspondingly modified. The gill slits of bony fishes are covered by an operculum, which has developed from the gill septum of the hyoid arch, and part of the hyoid musculature controls its movement.

The remaining visceral arches (branchial arches) and their muscles are associated with the gills. The muscles of the third arch are supplied by the glossopharyngeal nerve; those of the fourth through seventh arches by the vagus. Muscles of a typical branchial arch include constrictors, interbranchials, adductors, and interarcuals, all of which act to compress the gill pouches and force water out through the gill slits. The gill pouches are opened primarily by the action of the coracobrachialis—a part of the hypobranchial musculature. In addition, each branchial arch has a levator, but the levators have united to form a single muscle, the cucullaris, and their insertion has shifted from all but the last branchial arch onto the pectoral girdle (**Fig. 1**).

During the evolution of terrestrial vertebrates, gills are lost and the visceral arches become reduced and greatly modified. Most of the mandibular muscles remain associated with the jaws and form the various muscles of mastication. In a mammal these are temporalis, masseter, pterygoids, anterior belly of the digastric, and the mylohyoid (**Fig. 2**). All but the last two close the jaws. The tensor palati, in the soft palate, and the tensor tympani, which attaches to the malleus (a derivative of the mandibular arch), also belong to this group. Only a few hyoid muscles remain associated with the hyoid arch or its derivatives: stylohyoid, posterior belly of the digastric, and stapedius. Most of the hyoid musculature has spread out beneath the skin of the face and neck to form the platysma and the numerous facial muscles (**Fig. 6**). Most of the musculature of the branchial arches is lost, but parts of it form the intrinsic muscles of the larynx and certain pharyngeal muscles. The cucullaris, in contrast, enlarges and subdivides to form the trapezius and sternocleidomastoid, muscles that act on the pectoral girdle and head (**Fig. 2**). The mammalian motor nerve to these muscles, the spinal accessory, is homologous to part of the vagus of fishes.

Integumentary musculature. In a number of terrestrial vertebrates, particularly amniotes, certain of the more superficial skeletal muscles of the body have spread out beneath the skin and inserted into it. These may be described as integumentary muscles, but it should be emphasized that they are not a natural phylogenetic group but are derived from several different groups.

Integumentary muscles are particularly well developed in mammals and include the facial muscles and

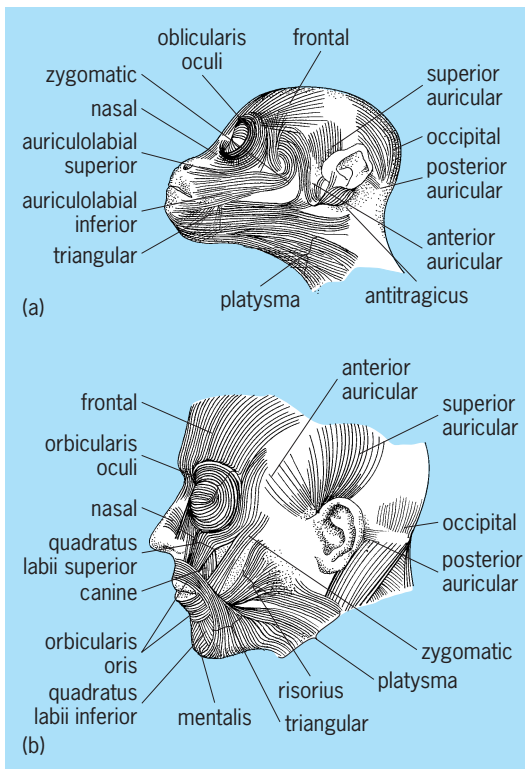


Fig. 6. Facial muscles. (a) Monkey. (b) Human. (After H. W. Rand, *The Chordates*, Blakiston, 1950)

platysma, derived from the hyoid musculature, and often a large cutaneous trunci. The last is derived from the pectoralis and latissimus dorsi and fans out beneath the skin of the trunk. The twitching of the skin of an ungulate is caused by this muscle.

Birds and reptiles have a sphincter colli (Fig. 2), a superficial neck muscle derived from the hyoid musculature and hence homologous to the platysma, but they lack facial muscles. Other integumentary muscles, derived from appendicular and trunk muscles, attach to the feathers, especially the large flight feathers on the wings and tails. In snakes, costocutaneous muscles extend from the ribs to the large ventral scales and play an important role in locomotion.

Histology

Three histological types of muscle are recognized: smooth, striated (skeletal), and cardiac. Smooth and cardiac muscle fibers generally occur in layers in the walls of organs, but striated muscle fibers are usually grouped into distinct entities, the skeletal muscles of gross anatomy. In some vertebrates, such as the rabbit, red and white skeletal muscles can be distinguished, but in most, an individual muscle contains a variable mixture of red and white fibers. Red fibers contain more sarcoplasm and myoglobin than white fibers. Myoglobin has a greater affinity for oxygen than hemoglobin, hence oxygen can be taken from the blood and stored by red muscle cells. The contraction of red fibers is more sustained, less subject to fatigue, and often slower than the contraction of white fibers. Red muscles, or muscles containing a preponderance of red fibers, tend to be found in situ-

ations where the muscles are particularly active in either moving the body or maintaining posture. Some examples are the diaphragm and other respiratory muscles, and the gluteus maximus. White muscles, or muscles with a preponderance of white fibers, are associated with a more intermittent and often a faster and more powerful movement. The biceps and digital muscles are examples.

Each skeletal muscle is individualized by a connective tissue sheath, the epimysium, which is a part of the deep fascia, and by its distinctive attachments onto skeletal elements. The epimysium is continuous with the connective tissue that invests the bundles of fibers within a muscle and with the connective tissue investing the individual fibers. Attachments to the skeleton are made by a continuation of this connective tissue into the periosteum surrounding the bone, and sometimes by connective tissue fibers that penetrate the bone. The connective tissue connection between muscle and bone may take the form of a cord-shaped tendon or a broad sheet known as an aponeurosis; or it may be relatively inconspicuous, in which case the muscle is said to have a fleshy attachment. The attachment of the muscle that tends to be stationary as the muscle contracts is its origin; the opposite attachment, which pulls on a structure that can be moved, is its insertion, but it must be recognized that the force exerted by a muscle during contraction is the same at each end.

Skeletal muscles vary greatly with respect to the number of fibers they contain and in the length and arrangement of their fibers. Important determinants of muscle architecture are the extent and nature of the movement that is to be brought about, the force that must be exerted, and the space available for the muscle in a particular area. The extent to which a muscle shortens is a function of the length of its fibers. Experimentally, muscle fibers can shorten to about one-half of their resting length, but they attach to bones so close to the joints that they seldom shorten this much. The strength of a muscle, on the other hand, is approximately proportional to the number of fibers it contains.

Many muscles, such as the sartorius on the thigh of mammals, have long fibers arranged parallel to each other (Fig. 7). An advantage of a strap muscle of this type is that it provides maximum excursion. Quite a different arrangement is seen in a muscle such as

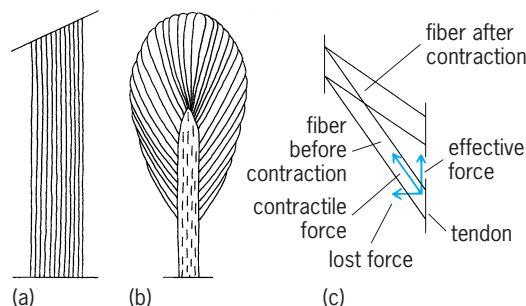


Fig. 7. Diagram showing two types of muscle architecture. (a) Parallel muscle. (b) Pinnate muscle. (c) Pinnate fiber before and after contraction.

the deltoid on the shoulder, which has a central tendon onto which many short, diagonal fibers attach at rather acute angles (Fig. 7). Pinnate muscles have many more fibers than a strap muscle of comparable mass, hence they can exert a greater force, but they shorten over a shorter distance. The full force of the contraction is not realized, however, because the muscle fibers pull on the tendon at an angle. The contractile force is resolved into an effective force along the axis of the tendon and a lost component at right angles to the tendon. This slight disadvantage is outweighed by major advantages. In this way much of the force of many muscle fibers can act through a common tendon upon a restricted area. This is possible because the strength of a tendon is 30–120 times that of a muscle of equal cross section. Another advantage is that the change in angle of the muscle fibers that occurs during contraction (Fig. 7) keeps bulging to a minimum. Some pinnate muscles, such as the tensor tympani in the middle ear of mammals, can act within a confined space.

The examples selected of muscles with fibers arranged completely parallel to one another or acutely pinnate are extremes of a continuum of muscle architecture. Some degree of pinnation is very common. The architecture of any particular skeletal muscle is dependent upon the requirements of packing within a given area, and the amount of force and extent of excursion needed.

Muscles are usually arranged so that one muscle or group of muscles will pull a structure in a certain direction, and an opposing muscle or group will pull the structure in the opposite direction. Several sets of terms describe these antagonistic actions. Flexion is the movement of a distal part of an appendage toward a more proximal part; this occurs at the elbow and knee. It also describes the bending of the head or trunk toward the ventral surface. Extension is the opposite movement. Flexion and extension are sometimes also applied to forward and backward movements of the appendage at the shoulder and hip, but because they have been used in conflicting senses by different authors, the terms protraction for a forward movement and retraction for a backward movement are more appropriate. Abduction is the movement of a part away from some point of reference, and adduction is movement toward it. For the appendages, the reference point is the midventral line of the body. Various types of rotary movement occur. For example, rotation of the bones of the forearm so that the palm of the hand faces up is supination; the opposite movement is pronation.

Muscle Mechanics

Many of the bones serve as lever arms, and the contractions of muscles are forces acting on these arms. The relationship between most muscles and bones is such that the lever systems are classified as third order (Fig. 8). The joint, of course, is the fulcrum and it is at one end of the lever. The length of the force arm is the perpendicular distance from the fulcrum to the line of action of the muscle; the length of the work arm is the perpendicular distance from

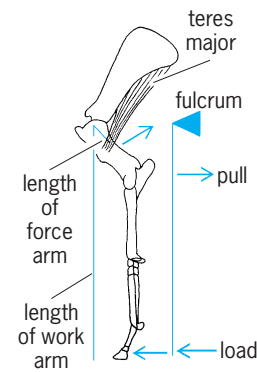


Fig. 8. Typical vertebrate lever system.

the fulcrum to the point of application of the power generated in the lever. Compactness of the body and physiological properties of the muscle necessitates that a muscle attach close to the fulcrum; therefore, the force arm is considerably shorter than the work arm. Most muscles are at a mechanical disadvantage, for they must generate forces greater than the work to be done, but an advantage of this is that a small muscular excursion can induce a much greater movement at the end of the lever.

Slight shifts in the attachments of a muscle that bring it toward or away from the fulcrum, and changes in the length of the work arm, can alter the relationship between force and amount of speed of movement.

In general, the force of a muscle is inversely related to the amount and speed of movement that it can cause. Certain patterns of the skeleton and muscles are adapted for extensive, fast movement at the expense of force, whereas others are adapted for force at the expense of speed. In the limb of a horse, which is adapted for long strides and speed, the muscles that move the limb insert close to the fulcrum and the appendage is long. This provides a short force arm but a very long work arm to the lever system (Fig. 8). In the front leg of a mole, which is adapted for powerful digging, the distance from the fulcrum to the insertion of the muscles is relatively greater and the length of the appendage is less, with the result that the length of the force arm is increased relative to the length of the work arm. See BIOMECHANICS.

Warren F. Walker

Locomotion. Different groups of muscles have different mechanical properties. Investigation of these properties in relation to their performance reveals the functional adaptation of the muscular system.

Running mammals. With increased running speed, the fraction of the stride for which the foot is on the ground decreases, so that the feet must exert larger forces while they are on the ground to make the average force over a complete stride match the body weight. Therefore, with increasing speed a greater muscle mass would need to be active to produce the necessary force.

The properties of extensor digitorum longus muscle and soleus muscles, which are components of the calf of the leg, are representative of fast and slow

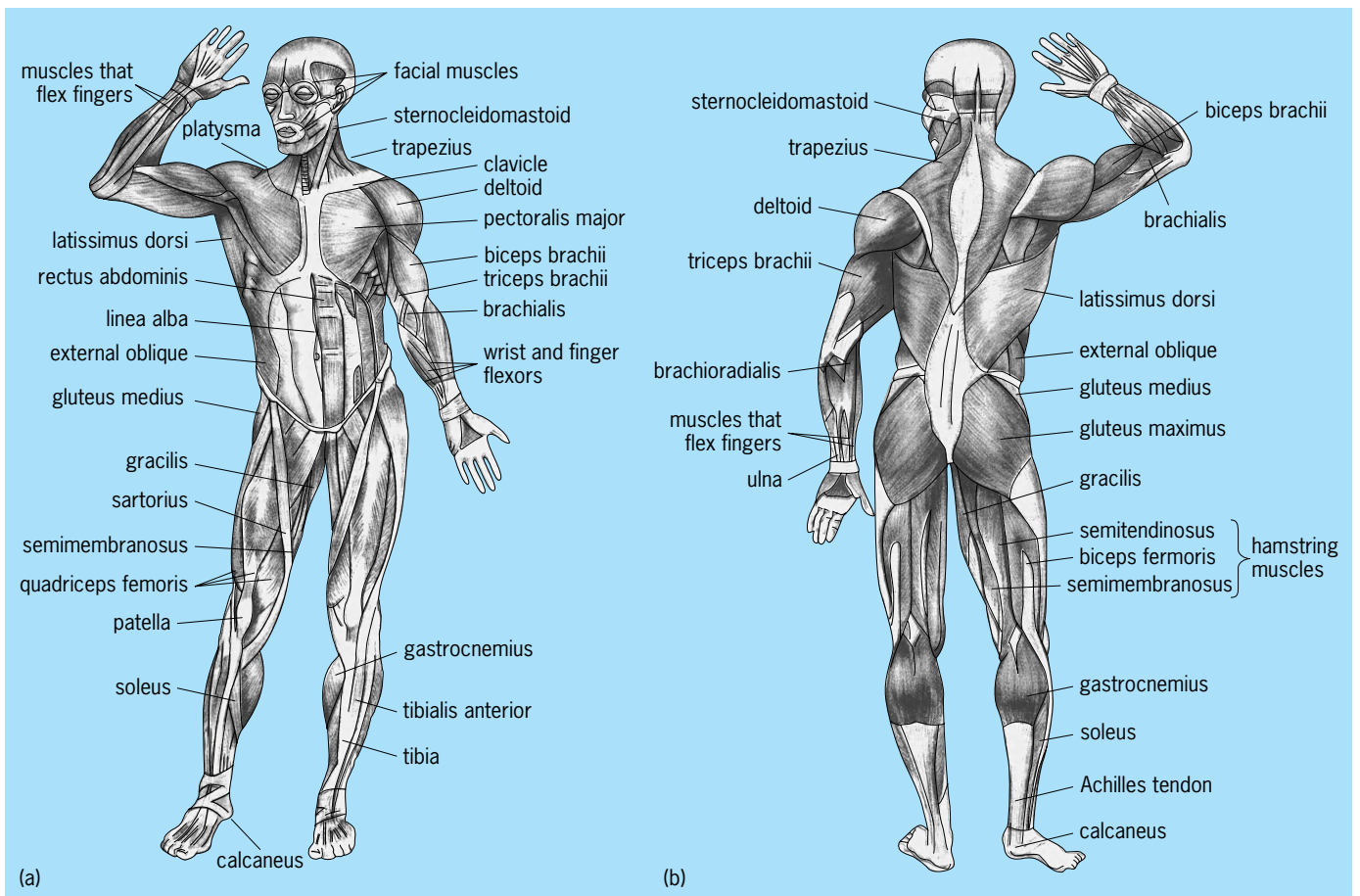


Fig. 9. Superficial muscles of the human body. (a) Anterior view. (b) Posterior view. (After C. A. Villee et al., *Biology*, 2d ed., Saunders, 1989)

muscle (**Fig. 9**). Extensor digitorum longus muscle consists wholly of fast glycolytic and fast-oxidative-glycolytic fibers in roughly equal proportions. In contrast, the soleus consists of only slow-oxidative and fast-oxidative-glycolytic fibers. These different fiber populations impart different properties to these two muscles. Extensor digitorum longus muscle has a faster contraction time, lower fatigue resistance, and higher power output than soleus. During normal function, these locomotory muscles undergo changes in length, doing work as they shorten and absorbing energy during reextension.

During walking and trotting, the soleus muscle produces 85–100% of its maximal power output. The frequency of maximum power output of extensor digitorum longus muscle coincides with the higher stride frequencies employed during fast galloping. This match between locomotion mechanics and muscle properties has been observed in other mammals and in reptiles—examples of the evolutionary optimization of design.

Fish swimming. As tailbeat frequency and swimming speed increase, there is a sequential recruitment of myotomal muscle (that is, skeletal muscle produced from a somatic cell) from superficial to deep muscle fibers. At slow, sustainable swimming speeds, only the slow muscle fibers close to the lateral line are active. The cost of locomotion is low at these speeds

and, consequently, this fiber type typically makes up just a few percent of the total myotomal muscle mass. slow muscle fibers generate maximum power at low tailbeat frequencies and can sustain activity for long periods of time. The power required for swimming increases rapidly with increasing speed and requires the recruitment of fast muscle fibers, which form the bulk of the myotomal muscle. Fast muscle fiber has a higher intrinsic power output that generates maximum power at high tailbeat frequencies, but it fatigues rapidly; therefore, fish can swim at these high speeds of only a brief period of time.

Most fish swim primarily by lateral oscillations of the body. The nature of these oscillations changes among species, and is related to body form (**Fig. 10**). For example, long, slender, round-bodied fish (such as the eel) have a swimming pattern described as anguilliform, that is waves of muscle contraction pass alternately down both sides of the body, causing a series of large-amplitude undulations. As the undulations pass down the body, they push against the water, generating the thrust which moves the fish forward. The power generated by the muscle is converted to hydrodynamic thrust all along the body. This thrust is generated on both sides of the body at any instant, because there is more than one undulatory wave passing down the body at a given time. Fish with a tailblade swim in a different way. In

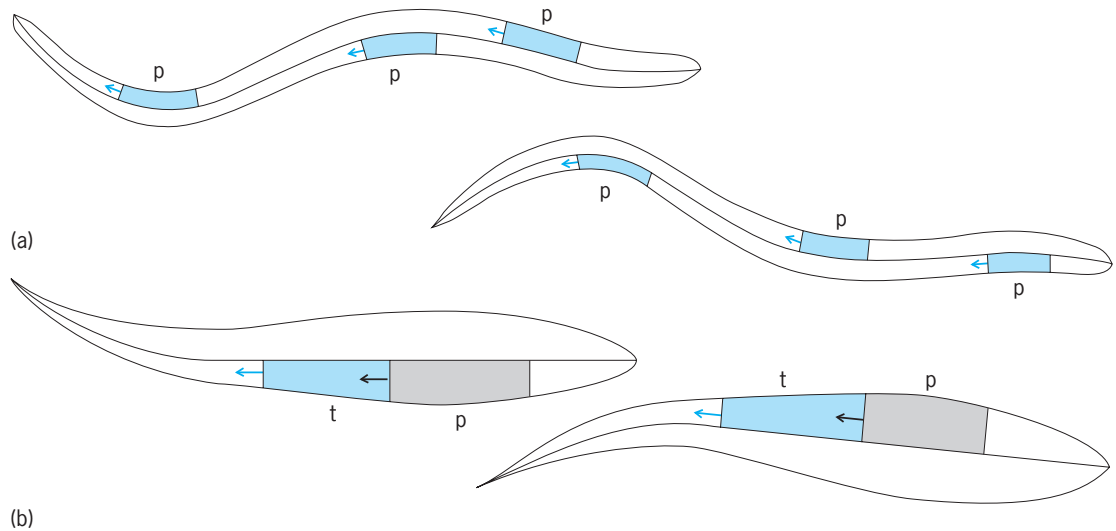


Fig. 10. Function of the myotomal muscle. (a) In anguilliform swimmers, waves of muscle contraction pass alternately down the sides of the body, throwing it into a series of large-amplitude undulations pushing against the water, generating power (p) thrust. **(b)** In carangiform swimming, one undulatory wave passes down the body at a given instant; thrust is generated alternately on the two sides and all the thrust is generated at the tailblade. The myotomal muscle generates power (p), which is transmitted down the fish toward the tail by muscle which acts temporarily as a tendon (t).

carangiform swimming, as seen in the mackerel and other fish with similar body forms, all of the thrust is generated at the tailblade.

To swim, the anguilliform and carangiform fish must use their myotomal muscle in different ways. In the eel, myotomal muscle functions in the same

way all along the body. As the wave of contraction passes down the body, the muscle generates power, which is passed directly to the water to generate hydrodynamic thrust. In contrast, the function of the active muscle in carangiform swimmers varies along the body. As the wave of contraction starts near the head, myotomal muscle generated power. However, this is not transmitted directly to the water, but down the fish toward the tail by muscle which acts temporarily as a tendon. The muscle transmitting this power is active, but is being stretched because of the complex interaction between the motions of the fish and the water. Under these conditions, muscle tissue is very stiff and is therefore a good power transmission element. As the active region passes down the body, muscle which previously transmitted power may generate power, which is transmitted to the tail by muscle located still further down the body. When wave of power is generated, it passes down the body, and it is transmitted by more posterior, stiffened muscle to the tail, which generates the thrust.

Given the enormous diversity of fish body forms and swimming modes, there are many variations on these basic patterns of muscle use. The scup, for example, is a fish with a short, deep body with high dorsal and ventral fins. When it swims, there is only 0.65 of an undulatory wave on the body at a given instant. Since it has a large tail blade, it might be expected to swim like a mackerel, generating all of its thrust at the tailblade. However, it appears to have some of the characteristics of the eel, and may pass power directly to the water all along its flattened body and through its high fins. A species of sculpin, a bottom-dwelling predator, shows that muscle recruitment patterns vary with the mode of swimming. Fast starts and slow turns, for example, require uses of the myotomal muscle.

Insect flight muscle. Insect flight muscle is divided on structural and physiological grounds into

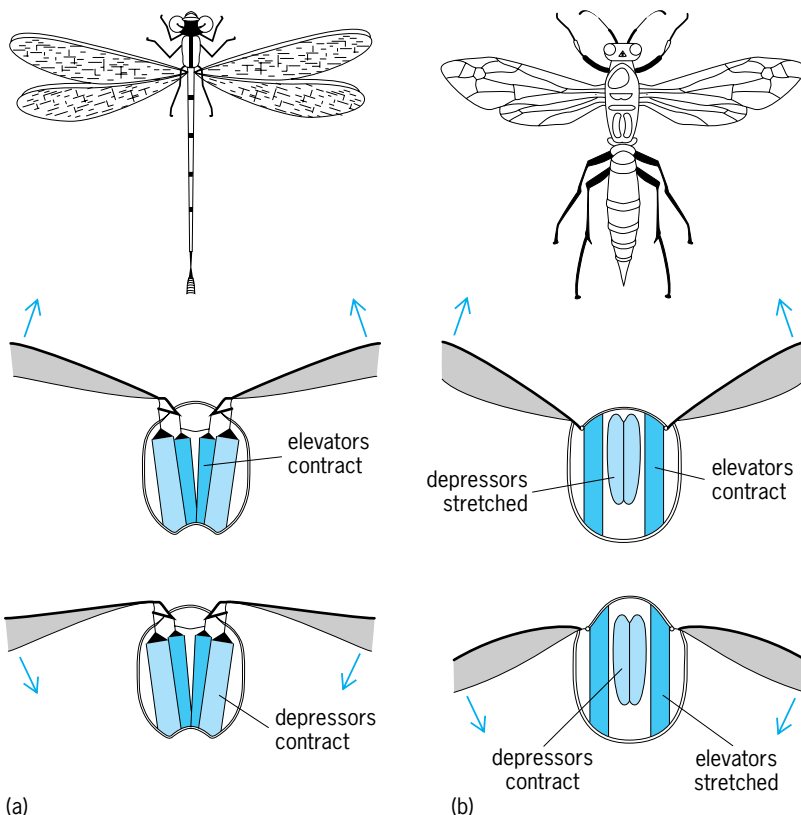


Fig. 11. Two types of insect musculature. (a) Direct flight muscle. **(b)** Indirect flight muscle. (After R. Eckert, D. Randall and G. Augustine, *Animal Physiology Mechanisms and Adaptations*, 3d ed., W. H. Freeman, 1988).

synchronous and asynchronous types, and on the basis of its mechanical operation within the insect into direct and indirect flight muscles. There is a large degree of structural diversity between types of synchronous muscle, but the functional distinction between it and asynchronous muscle is far more fundamental. The fibers of all insect flight muscles are rich in mitochondria and rely on aerobic metabolism despite the high energy demands of flight.

Direct flight muscle. This type of muscle is present in what are usually regarded as primitive insects, for example, the dragonflies (Fig. 11a). One end of each direct flight muscle is attached to the base of the wing, and the other end to the inside of the thorax. Contraction of the flight muscles, therefore, drives the wings directly. The wingbeat frequencies of insects with direct flight muscles are typically less than 100 Hz. All insects with direct flight muscles have synchronous muscle.

Indirect flight muscle. Indirect muscles attach to the inside of the thorax rather than the wing. Muscle contraction deforms the thorax and, through a complex hinge, moves the wings up and down. The indirect elevator muscles run between the roof and floor, and contraction pulls the roof downward and raises the wings. Contraction of the depressors, which run from front to back, buckles the thorax, raises the roof, and lowers the wings (Fig. 11b). The mass of wings, the aerodynamic forces acting upon them, and the elasticity of the thorax act as a resonant system that enables some insects to operate with wingbeat frequencies up to 1000 Hz. In these insects, both synchronous and asynchronous indirect flight muscles are found, but only those with asynchronous muscle can achieve the high wingbeat frequencies.

Synchronous muscle. Synchronous muscle is characterized by an equal and simultaneous neural input and muscular contraction, each mechanical contraction being caused by a burst of neural activity. The rhythm of the wingbeat is thus neurogenic in origin. Insects with this type of muscle rarely have wingbeat frequencies that exceed 100 Hz. This is a limitation imposed by the neurogenic nature, and the time required to activate and relax the muscle in each wingbeat. Some highly modified synchronous flight muscles are found, for example, in the singing muscles of cicadas.

Asynchronous muscle. Very high wingbeat frequencies are found in some species of insects (for example, the wingbeat frequency of some midges is of the order of 1000 Hz). These high frequencies are found in insects that possess asynchronous flight muscles, that is, muscles where the number of nerve impulses to the muscle is much lower than the high wingbeat frequency. The wingbeat frequency of blowfly is approximately 120 Hz, but the frequency of neural input is only 3 Hz. The nervous input to these muscles facilitates rather than controls the frequency of contraction.

The rhythmic contraction of asynchronous muscle that drives the wings results from a unique resonant coupling between the flight muscles and the elastic thorax. When the active muscle is stretched by the

recoil of the elastic thorax, it develops force after a short delay. However, this delay is long enough to allow a complete recoil of the thorax, before delayed force development deforms it. This property of a stretch-induced delayed force development is due to a unique myosin filament structure. The depressors and elevators work in opposition to oscillate the thorax and beat the wings. The time course of delayed force development enables the muscles and thorax to resonate at the same natural frequency. Because the system is operating at its resonant frequency, a minimum amount of energy is required to maintain the oscillations once the wingbeating has started. Because the muscle does not need to be repeatedly switched on and off, and because the thorax is relatively stiff, the system can operate at high frequencies.

Body size. If evolution has resulted in optimization of muscle function, size-dependent changes in these processes should be reflected in the properties of the muscles which drive them. The main function of muscle is to generate power, normally cyclically and repetitively, for example, contracting and relaxing of the diaphragm during breathing, or of the leg extensor muscles in walking. The body demands the greatest effort from the muscles at particular frequencies, for example, the stride frequency of a sprint. Therefore, under normal circumstances the muscles should perform best over those frequencies. See MUSCLE; RESPIRATORY SYSTEM.

Iain S. Young; John D. Altringham

Embryology

The muscles are derived from mesoderm, the middle germ layer. The exceptions are the sphincter and dilator muscles of the iris and the myoepithelial cells of the sweat and mammary glands, which are derived from ectoderm. The embryonic mesoderm that differentiates into muscle tissues includes the dorsal mesoderm, head mesenchyme, intermediate mesoderm, and lateral mesoderm (Fig. 12). The dorsal mesoderm that condenses into bilateral columns adjacent to the neural tube forms the segmentally arranged myotomes. Most intrinsic voluntary muscles of the neck and trunk are differentiated from these myotomes. Some voluntary head musculature (muscles of eye and tongue) and the limb musculature are derived from myotomes in the lower vertebrates; the limb muscles in higher vertebrates are mainly derived from lateral plate mesoderm. In tetrapods the myoblast component of all truncal and appendicular muscles originate from somites (somitic origin). The voluntary muscles of the branchial (visceral) arches of the head and neck are derived directly from head mesenchyme.

Cardiac muscle is derived from the splanchnic mesoderm. The involuntary musculature is differentiated from the intermediate mesoderm (mesomere, urogenital mesoderm, or nephrotomic mesoderm) and the lateral mesoderm (splanchnic mesoderm and hypomere). The intermediate mesoderm differentiates into much of the urogenital system, and the lateral mesoderm differentiates into the vascular,

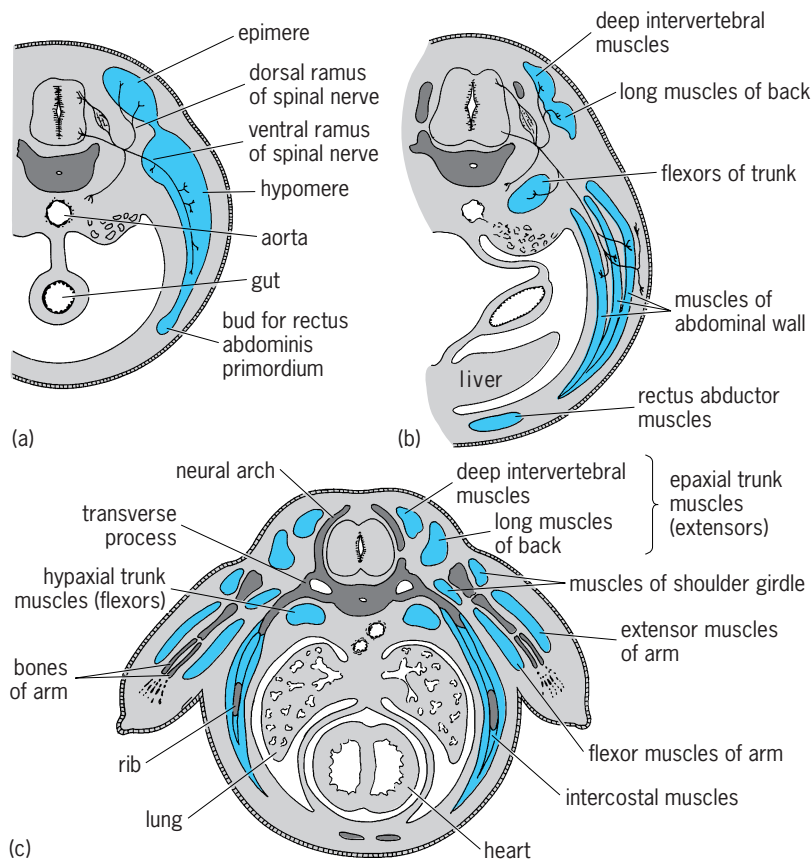


Fig. 12. Differentiation of myomeres and the establishment of the primordia of various muscle groups. (a) Early separation of the myomere into an epimere, supplied by the dorsal ramus of the spinal nerve of the level, and a hypomere, supplied by the ventral ramus of the nerve. (b) Later stage in the differentiation of muscle primordia at abdominal levels. (c) Primordial muscle masses at the level of the arm buds. (After C. E. Corliss, *Patten's Human Embryology*, McGraw-Hill, 1976)

digestive, and respiratory systems and related structures. See EMBRYONIC DIFFERENTIATION; EMBRYONIC INDUCTION.

Differentiation of striated muscles. Striated muscles differentiate from the myotomes of the somites and from mesenchyme of nonmyotomic origin. Muscle cells, fascial cells, tendon cells, and aponeurotic cells are derived from these structures. The premuscle cells (myoblasts) migrate as the organism develops. When the myotome is adjacent to the neural tube, it receives its initial innervation which is retained during subsequent migration of the developing muscles. The innervation of the primordial muscle masses and the retention of this innervation during development are significant as a means of determining the homology of muscles of different species. The nerves are probably not concerned with organizing these muscle masses, because muscles will develop without any innervation in certain monsters. See NERVOUS SYSTEM (VERTEBRATE).

The segmental pattern of the myotomes may be retained in the adult, as in the intercostal muscles of the thorax of mammals and the trunk muscles of fishes. The pattern may be modified as in the flat muscles of the mammalian abdominal wall. The segmental derivation of the muscle masses may be masked by migration as in the eye and tongue muscles, by

fusion of muscles from several segments as in the rectus abdominis muscle (the product of the fusion of muscle masses from successive myotomes), by the splitting of muscles into layers as in the flat abdominal muscles, and by fusion and realignment of muscles as in the back musculature of higher vertebrates where muscle fascicles extend through as many as six segments. During metamorphosis in frogs, the segmental patterns of many muscles in the tadpole are altered by the migration, fusion, and splitting of muscle masses to form many muscles in the adult frog.

Axial musculature. In all vertebrates, the muscles of the neck, trunk, and tail are derived from myotomes, although claims have been made that some myoblasts differentiate in dermatomes. The myotomes develop and migrate laterally and ventrally as the epimere to form the dorsal or epaxial muscle mass (dorsal and lateral back muscles), and as the hypomere to form the ventral or hypaxial muscle mass (lateral and ventral muscles) [Fig. 12]. The subsequent fate of this migration is said to differ in different animals. In the aquatic vertebrates, whose primary means of locomotion is swimming, the musculature is oriented to produce undulatory motion. Each myotome differentiates into a muscle mass which forms a band from the back to the belly. The muscle cells within each band are directed cephalocaudally. The epaxial musculature (muscle mass innervated by the dorsal ramus of a spinal nerve and located dorsal to the vertebral column) is approximately equal in size to the hypaxial musculature (muscle mass innervated by the ventral ramus of a spinal nerve and located ventral to the vertebral column). In most terrestrial animals the bilateral appendages assume primacy in locomotion. In these animals, the myotomes differentiate into the body-wall musculature pattern in which the epaxial musculature is restricted to the back and the hypaxial musculature is present on the lateral and ventral aspect of the body wall. In animals adapted for aerial locomotion, the primitive segmental pattern is modified. The epaxial musculature is greatly reduced in the trunk but is well developed in the nuchal region. The hypaxial musculature is mainly concentrated in the pectoral musculature.

Fate of myotomes. The fate of the myotomes during their differentiation into the axial muscles differs in the various vertebrates (Fig. 13). In the fishes the embryonic pattern is retained because the epaxial and hypaxial muscles are relatively equal in size. In the amphibians, two adult patterns are developed. In aquatic amphibians, such as *Necturus*, the embryonic segmental pattern is retained. In terrestrial amphibians, such as the frog, the myotomic derivatives are modified during metamorphosis by migration, fusion, and splitting. As a result, the embryonic pattern is modified as a functional adaptation to life on land. In the other terrestrial animals, reptiles and mammals, the embryonic myotomic segmentation is altered further than in the terrestrial amphibians. The epaxial muscles form the erector spinae back muscles. Some hypaxial muscles retain vestiges of the segmental patterns (intercostal muscles of

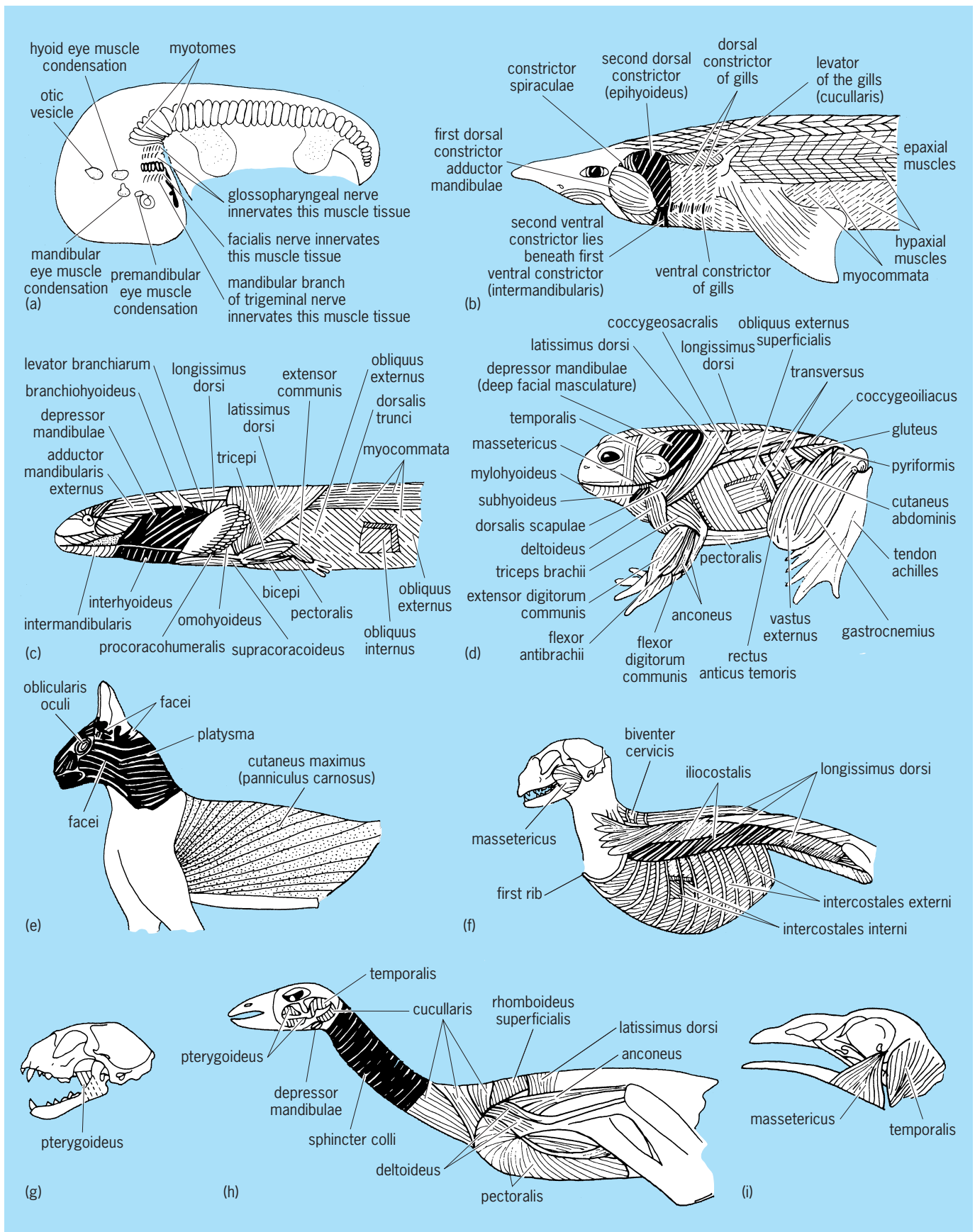


Fig. 13. Development of muscles from the visceral arches and from the myotomes in various vertebrates. (a) Basic areas of the embryo from which voluntary muscles develop. (b) Shark. (c) *Necturus*. (d) Frog. (e-g) Cat. (h, i) Goose. (After O. E. Nelsen, *Comparative Embryology of the Vertebrates*, Blakiston, 1953)

abdomen), whereas other muscles are the products of migration, fusion, and splitting during development (flat abdominal muscles). The muscle patterns are altered more drastically in birds.

The tail bud mesoderm differentiates into myotomes in tailed animals that exhibit lateral movements. In these forms—fish, tailed amphibians, crocodiles, and whales—the myotomes differentiate into well-developed epaxial muscles and hypaxial muscles that retain their embryonic metamerism. The musculature of animals with prehensile tails and tails adapted for grasping and wagging movements is partially the result of the migration of myotomes from the hindlimb area. Of significance in mammals is the derivation of the diaphragm from portions of the neck myotomes and their subsequent migration through the neck and thorax to the thoracoabdominal boundary.

Axial muscles. Each axial muscle in the vertebrates is derived from one or more specific myotomes which are numbered according to the vertebral level (cervical, thoracic, lumbar, sacral, caudal, or coccygeal regions) at which the myotome first differentiated. Because the innervation of each myotome occurs during early development, the number in a region of each spinal nerve is identical with that associated with each myotome. In the different classes of animals, the regional number of a myotome and the spinal nerve varies because the number of vertebrae in any region is not similar in the vertebrate classes. Because of this variability, only the segmental numbers of the myotomes in relation to the adult muscles in humans, as an example, will be presented below.

Each myotome expands by growing ventrally to form two myomeres: the epimere and hypomere. The epimere, innervated by the dorsal ramus of a spinal nerve, subdivides into a dorsal segment, which gives rise to the deep intervertebral muscles of the neck and back, and a more ventral segment, which gives rise to the long muscles of the neck and back (Fig. 12). The hypomere, innervated by the ventral ramus of a spinal nerve, differentiates into the flexor muscles of the trunk (hypaxial muscles); intercostal muscles the thorax; muscles of the abdominal wall; and rectus thoracic and abdominal muscles of the ventral body wall (Fig. 12).

Epaxial muscles. The epaxial muscles of humans, the intrinsic extensor muscles of the back, are derived from the dorsal portions of 29 myotomes—the first cervical segment through the fourth sacral segments inclusive. These include 8 cervical, 12 thoracic, 5 lumbar, and 4 sacral segments. The hypaxial cervical musculature is derived from the ventral portions of the 8 cervical myotomes. The myotomes differentiate into a prevertebral portion (immediately in front of the vertebral column), a lateral sheet, and a ventral or rectus column. As a result, the prevertebral portion forms the prevertebral muscles; the lateral sheet forms the scalene muscles; and the rectus column forms the geniohyoid and infrahyoid muscles. The diaphragm is derived from the hypaxial division of the third through fifth cervical myotomes from whence it migrates.

Hypaxial muscles. In humans, the muscles of the ventral and lateral thoracolumbar wall are derived from the hypaxial divisions of the first thoracic through the first lumbar myotomes inclusive. These myotomes differentiate into a main lateral sheet and a rectus column (ventral edge). The intercostal muscles, the oblique abdominal muscles, and the transversus abdominis muscle develop from the main lateral sheet, whereas the rectus abdominis muscle develops from the rectus column. The quadratus lumborum muscle of the posterior abdominal region is derived from the hypaxial divisions of the first through fifth lumbar myotomes. Although direct evidence is difficult to observe, it is probable that the muscles of the pelvic diaphragm (the muscular floor of pelvis, including the coccygeus and levator ani muscles) are derived from the hypaxial divisions of the last four sacral and all coccygeal myotomes.

In the chick, the skeletal muscles of the dorsal and dorsolateral trunk are derived from myotomic mesoderm, whereas the hypaxial muscles differentiate from the somatic mesoderm of the lateral plate.

Paired appendages. Theoretically the paired appendages were derived originally either from gill arches (gill-arch theory) or from fin folds (lateral-fold theory). On the basis of the embryonic development of the fin musculature, the lateral-fold theory is favored because, in fish, the fin musculature is differentiated from the myotomes of somites. Muscle buds of potential muscle cells differentiate at the lower edges of the myotomes and, as slips of tissue, invade the mesenchyme of each limb bud as the dorsal and the ventral pre-muscle masses of myoblasts. The dorsal pre-muscle mass differentiates into the dorsal, elevator, and extensor muscles, and the ventral pre-muscle mass into the ventral, depressor, and adductor muscles of the fins. These muscles are innervated by spinal nerves. Apparently the lateral plate mesoderm is not capable of forming the fin musculature.

In the tetrapods, including the amphibians, reptiles, birds, and mammals, the muscles of the limbs assume a relatively large bulk and a complexity of organization to cope with locomotion on land. The embryonic derivation of most of the myoblasts, which form this musculature in the tetrapods, differs from that in the fishes: much of the limb musculature of tetrapods is said to be derived in place from lateral plate mesoderm and only some from the myotomic mesoderm of somites and pharyngeal mesoderm. This mesenchyme is initially found surrounding the developing skeletal elements (future bones). The standard version is that this mesenchyme does not originate from the myotomic regions of the somites.

On the basis of their mesodermal embryonic precursors and their innervation, the appendicular musculature may be classified into three groups: (1) Mammalian muscles, derived from the mesoderm of the posterodorsal pharyngeal region, include the trapezius and the sternomastoid muscles, which differentiate from the mesoderm of the last branchial arch. These muscles have attachments which extend from the neurocranium and cervical vertebrae to the

proximal bones of the skeleton of the forelimb. They are innervated by the spinal accessory cranial nerve. (2) Muscles derived from myotomic mesoderm of somites include, among others, the rhomboids, pectorals, and serratus muscles of the forelimb and the quadratus lumborum and psoas muscles of the hindlimb. They are the true metameric derivatives of the embryonic myotomes. These muscles have attachments which extend from the vertebrae and ribs to the appendages; they are innervated by the anterior rami of the segmental spinal nerves proximal to where these rami form the brachial plexus of the forelimb and the lumbosacral plexus of the hindlimb. (3) Muscles derived from the nonsegmental core of the limb bud include all the intrinsic limb muscles, muscles with attachments located wholly within the limb (Fig. 14). They differentiate within each limb bud and are innervated by the nerves of the brachial plexus and of the lumbosacral plexus. In general, the muscles differentiating from the dorsal premuscle mass of the limb bud become the extensor muscles of the limb, those from the ventral premuscle mass become the flexor muscles, those from the dorsal premuscle mass near the trunk become the abductor muscles, and those from the ventral premuscle mass near the trunk become the adductor muscles.

Head and visceral arch musculature. The muscles of the head and visceral arches are derived from myotomic mesoderm and from nonmyotomic mesoderm (Fig. 15). The extraocular muscles of the eye and the tongue muscles are differentiated from myotomes or from a mesoderm that phylogenetically was originally derived from myotomes. The visceral arch musculature in all vertebrates is derived from mesenchyme of nonmyotomic origin. Some facial muscles are said to originate from neural crest cells.

The extraocular muscles of sharks are derived from the preotic somites of the premandibular myotomes (cranial nerve III), the mandibular myotome (cranial nerve IV), and the hyoid myotome (cranial nerve VI). The extrinsic muscles of the eye include six which are found in all vertebrates: the superior rectus, internal (anterior) rectus, inferior rectus, and inferior oblique muscles, which are innervated by cranial nerve III; the superior oblique muscle, which is innervated by cranial nerve IV, and the external (posterior or lateral) rectus muscle, which is innervated by cranial nerve VI. In addition, the retractor oculi of many mammals and the quadratus muscle and pyramidalis muscle of birds are in this category. The tongue musculature in the sharks develops from six postotic myotomes that migrate ventrally to the hypobranchial region. In higher vertebrates three postotic (occipital) myotomes appear to provide the mesodermal source of this musculature which is innervated in all vertebrates by cranial nerve XII (hypoglossal). Because direct myotomic origin of the extraocular muscles and the tongue muscles is difficult to demonstrate in the higher vertebrates, the literature disagrees.

The mesoderm of the branchial (gill) arches is

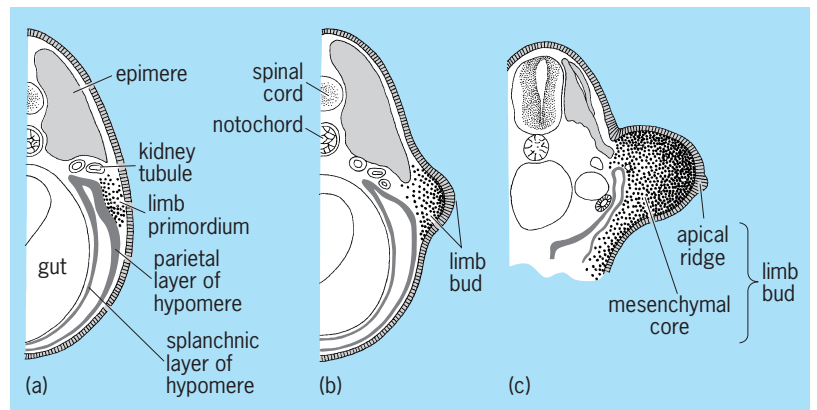


Fig. 14. Diagrams of the origin of a limb. (a, b) The development of a limb bud of an amphibian. (c) Limb bud of a chick embryo. (After T. Torrey, *Morphogenesis of the Vertebrates*, 2d ed., 1967)

derived from head mesoderm which develops in place and not from any myotome. The first branchial (mandibular) arch mesoderm differentiates into the muscles of mastication that are innervated by cranial nerve V (trigeminal). The muscles derived from this mesoderm in the fishes are the mandibular adductor muscle and the first ventral constrictor muscle; in the amphibians the temporal, masseter, pterygoid, and mylohyoid muscles; in birds the pterygotemporal, temporal, and digastric muscles; and in mammals the muscles of mastication (temporal, masseter, and pterygoid muscles), mylohyoid muscle, anterior belly of the digastric muscle, tensor palatini muscle, and tensor tympani muscle. The second branchial (hyoid) arch mesoderm differentiates into those muscles innervated by cranial nerve VII (facial). The muscles derived from this mesoderm in fishes are the hyoid gill arch muscles; in amphibians, the subhyoid and mandibular depressor muscles; in birds the sphincters of the neck and the mandibular depressor muscles; and in mammals the muscles of facial expression and other muscles such as the stylohyoid muscle, stapedius muscle, and the posterior belly of the digastric muscle. The mesodermal derivatives of this arch in mammals migrate to the scalp (occipitofrontalis muscle), ear region (auricular muscle), neck (platysma), and the face (orbicularis oculi, orbicularis oris, and others), collectively called the muscles of facial expression. The third visceral (first branchial arch) arch mesoderm differentiates into those muscles innervated by cranial nerve IX (glossopharyngeal). The muscles derived from this mesoderm in fishes are the gill constrictor muscles of this arch and in the higher vertebrates (mammals) the stylopharyngeus muscle and the upper constrictors of the pharynx. The mesoderm of the last three visceral arches (second, third, and fourth branchial arches) differentiates into the muscles innervated by cranial nerve X (vagus). These are the gill constrictor muscles in the fishes and the lower pharyngeal constrictor and laryngeal muscles in the higher vertebrates.

The sternocleidomastoid muscles and trapezius muscles of mammals are innervated by the spinal

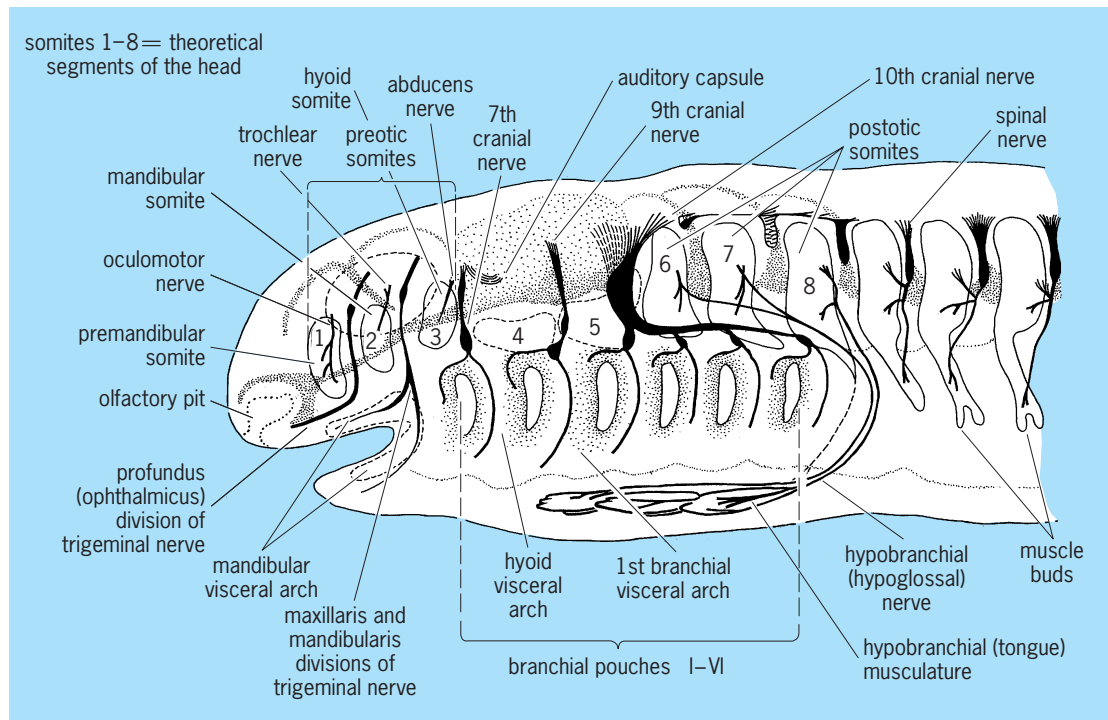


Fig. 15. Basic plan of vertebrate head. (After O. E. Nelsen, *Comparative Embryology of the Vertebrates*, Blakiston, 1953)

cord division of cranial nerve XI (spinal accessory). These muscles are derived either from the mesoderm of the last visceral arch or from postotic myotomes.

Voluntary skin muscles. The skin muscles, the voluntary muscles that move the skin, are divided into two groups, the muscles of facial expression innervated by cranial nerve VII, and the panniculus carnosus of the body-wall skin innervated by the anterior (ventral) thoracic nerves. The panniculus carnosus is derived from mesodermal cells that normally form the pectoral muscles. This muscle, found in such animals as the porcupine, dog, cat, horse, and guinea pig, may have both its origin and insertion in the skin of some species or its origin on the greater tuberosity of the humerus and its insertion in the fascia of the skin of the back and thigh of other species.

Involuntary muscles. Involuntary muscles arise independently of the segmental myotomes and visceral arches. The visceral mesoderm differentiates into the mesenchyme that forms the smooth muscles of the digestive system, respiratory system, and many blood vessels. The heart also differentiates from this mesoderm. The mesenchyme of the somatopleuric mesoderm of the body wall, head, and limb buds differentiates into the smooth muscles of the blood vessels of these regions. Mesenchyme, whatever its origin, is a potential source of smooth muscle. The smooth muscles of many organs develop in place from mesenchymal cells. The smooth muscles of the structures of the urinary system and the genital systems are derived from mesenchymal cells of the intermediate mesoderm (nephrotomic mesoderm). The smooth muscles of ectodermal origin are the dilator muscle

of the iris and the myoepithelial cells of the ducts of sweat glands.

Histogenesis of Muscle

Embryonic muscle cells (myoblasts) are derived from mesenchymal cells of mesodermal origin.

Voluntary muscle. The myoblasts of voluntary muscles of either myotomic or nonmyotomic origin are mononucleated spindle-shaped cells with clear cytoplasm. Embryonic stem cells from the mesoderm of the myotomes divide to form bipolar cells. These mononucleated myoblasts cease dividing and develop specialized cell surfaces that render them capable of fusing to form multinucleated elongated myotubes. Myosin and messenger RNA required for its synthesis are present in the myoblasts. After fusion the synthesis of myosin rapidly increases. This is followed by the appearance of contractile filaments, receptor sites for the neurotransmitter (acetylcholine), T-tubular system, and sarcoplasmic reticulum.

The first myofibrils are sparse and coarse. Cross striations are formed almost as soon as the myofibrils are visible. The elongated myotubes have centrally located nuclei and peripherally located myofibrils in the early stages of differentiation. Later, when the myofibrils fill the cells, the nuclei become peripherally located adjacent to the sarcolemma. In humans, at the time of birth, the muscle cells resemble those of the adult except that the nuclei are rounder, the myofibrils are more slender, and the cross striations are less prominent. The multinuclear striated muscle cells are the result of nuclear mitosis which is unaccompanied by cytoplasmic divisions. The postnatal growth of muscle cells is by hypertrophy.

Involuntary muscle. Smooth muscles arise from mesenchymal cells. During embryonic life the cells migrate and concentrate in the vicinity of the epithelial linings of the hollow organs. These myoblasts elongate and orient themselves as in the adult organ. Myofibrils (contractile elements) are visible in the early stages and become more numerous during later development. The iridic muscles (sphincter and dilator pupillae) and myoepithelial cells of sweat and mammary glands are derived from ectodermal cells.

Cardiac muscle. Cardiac muscle develops from the splanchnic mesoderm of the heart tube. The myoblasts adhere to each other, but unlike skeletal muscle, the plasma membranes do not disintegrate—rather the sites of adhesion give rise to the intercalated discs (gap junctions). As in skeletal muscles, myofibrils and the other structural muscle elements differentiate early.

Charles R. Noback

Bibliography. N. D. Agnish, Possible role of somites in developing mouse limbs, *in vitro*, *Anat. Rec.*, 184:340–341, 1976; R. McN. Alexander, *Exploring Biomechanics: Animals in Motion*, 1992; B. I. Balinsky, *An Introduction to Embryology*, 5th ed., 1981; A. Chevalier, Role of the somitic mesoderm in the development of the thorax in bird embryos, II: Origin of thoracic and appendicular musculature, *J. Embryol. Exp. Morphol.*, 49:73–88, 1979; K. L. Moore, *The Developing Human*, 6th ed., 1998; R. O’Rahilly and F. Muller, *Developing Stages in Human Embryos*, 1987; A. S. Romer, *The Vertebrate Body*, 6th ed., 1986; T. W. Torrey and A. Feduccia, *Morphogenesis of the Vertebrates*, 4th ed., 1979; W. F. Walker, Jr., *Vertebrate Dissection*, 9th ed., 2000.

Muscular system disorders

Disorders affecting skeletal (voluntary) muscle. Every skeletal muscle consists of a large number of muscle fibers, each of which is a multinucleated cell containing the usual types of cellular inclusions, but also containing the myofibrils, which are the actual contractile elements. The myofibrils have a banded structure which gives the muscle fiber its characteristic striated appearance on histologic examination. Muscle fibers have been separated into two main groups based on their histochemical reaction: type 1 fibers, which are rich in mitochondria and oxidative activity, and type 2 fibers. See MUSCLE.

The normal functioning of the skeletal muscle is dependent not only on the integrity of the muscle fibers themselves, but also on that of the motor cortex, the pyramidal tract, and the extrapyramidal system (including the cerebellum). It also depends on innervation by the motoneurons of the brainstem and the spinal cord. In addition, the proper functioning of the other organ systems, such as the endocrine system, and variations in the concentration of various electrolytes may also affect muscle function. See NERVOUS SYSTEM (VERTEBRATE).

Damage to the motor cortex or the pyramidal tract produces the type of weakness seen in humans after a stroke or spinal cord injury. Although the paralyzed

limb may initially be flaccid (hypotonic), spasticity (hypertonia) eventually develops. Despite the weakness, muscle atrophy is usually not striking. When the extrapyramidal system or the cerebellum is the site of damage, instead of weakness there are uncontrolled movements, difficulty with coordination, or both. In either of these situations there is no characteristic change in the muscle, either grossly or microscopically. At most, atrophy of type 2 fibers is seen.

Motoneuron damage. With damage to the spinal motoneuron or its axon, there is flaccid weakness of the muscle with proportionate wasting. Direct involvement of the spinal motoneuron was typically seen in poliomyelitis, but is now seen more commonly in the progressive spinomuscular atrophies of infancy and childhood, and in amyotrophic lateral sclerosis in adults. Spontaneous twitching of groups of muscle fibers (fasciculation) innervated by the same motoneuron (motor unit) is frequently seen in these disorders.

The axons of these motoneurons may also be damaged as they exit through the spinal nerve root foramina (nerve root compression, radiculopathy) or in the peripheral nerve (peripheral neuropathy). When muscle weakness is due to a radiculopathy, it is restricted to those muscles innervated by that nerve root. In peripheral neuropathy, muscle involvement is usually more diffuse and symmetrical or it occurs throughout the distribution of a particular nerve.

In all the following conditions, in which the motoneurons or their axons are damaged, the affected muscles undergo atrophy. Microscopically, marked variation in the size of individual muscle fibers is seen, in which the atrophic fibers of similar sizes occur in groups (neurogenic atrophy). Myosin adenosinetriphosphatase stain shows that both type 1 and type 2 fibers are involved in the atrophic process.

Poliomyelitis. This is an acute viral infection in which there is selective damage of the motoneurons, resulting in a flaccid paralysis of the muscles innervated by the affected nerve cells. Because the motoneuron damage is patchy and asymmetrical, the distribution of weakness is also quite patchy. Since poliomyelitis is an acute illness, the weakness and paralysis are nonprogressive. See POLIOMYELITIS.

Spinomuscular atrophies. These are progressive diseases of unknown cause and may occur in infancy (Werdnig-Hoffmann disease) or later in childhood (Kugelberg-Wielander disease). Both diseases result from degeneration of the motoneurons and appear to be inherited in an autosomal recessive pattern. In the infantile form, the baby is often floppy from birth with generalized weakness, a poor cry, and difficulty in sucking and breathing. Many children succumb in early childhood. In the later childhood forms, the rate of progression is slower and the outlook better. In these children weakness is more marked in the proximal muscles of the limbs.

Amyotrophic lateral sclerosis (ALS). This is a progressive disease of unknown cause, in which both the brainstem and spinal motoneurons, as well as the corticospinal tracts, undergo degeneration. This is a

relatively rapid progressive disease, with death usually occurring within 3 years of diagnosis due to the swallowing difficulty and respiratory failure. Although some cases of ALS appear to be inherited, most cases occur sporadically.

Neuromuscular junction. The most common example of disease at the neuromuscular junction is myasthenia gravis. Other, less common diseases are the myasthenic Eaton-Lambert syndrome and botulism. See MYASTHENIA GRAVIS.

In the Eaton-Lambert syndrome limb muscles are most frequently involved, and weakness is the main complaint. However, the muscle strength increases with repeated contractions, and, on repetitive nerve stimulation, there is a progressive increase in the amplitude of the evoked potential on electromyography. Thus, although this is called the myasthenic syndrome, the response to exercise and to repetitive nerve stimulation is opposite to that seen in myasthenia gravis. The neuromuscular defect appears to be due to a decrease in the number of acetylcholine quanta released by a nerve impulse. About 80% of patients with the Eaton-Lambert syndrome have been shown to have a small cell bronchogenic carcinoma. No consistent histologic abnormality has been described in the muscle, but electron microscopy has shown a marked increase in the number and complexity of the secondary synaptic clefts, with a resultant elongation of the postsynaptic membrane. This change also contrasts with the markedly simplified pattern seen in myasthenia gravis.

Botulism is another condition affecting the motor end plate and results from the ingestion of botulinum toxin, most commonly found in improperly canned food. The botulinum toxins block neuromuscular transmission in cholinergic nerve endings, probably by inhibiting the release of acetylcholine. The toxin produces rapidly spreading paralysis of voluntary muscles, including those of swallowing and respiration, leading to death due to respiratory failure. If the patient does not die, full recovery can be expected after a prolonged convalescence. No definite microscopic abnormality has been described in the muscle. See ACETYLCHOLINE; BOTULISM; SYNAPTIC TRANSMISSION.

Myopathy. Abnormalities of the muscle itself (myopathy) obviously result in muscle weakness, and muscle diseases fall into two large groups: those with a genetic basis and those which are nongenetic.

Genetic diseases. Congenital myopathies and the muscular dystrophies constitute the genetic muscle diseases. Congenital myopathies are characterized by a generalized weakness which is present at birth. The weakness is usually not progressive and often improves with time. In general, the different types of congenital myopathies cannot be differentiated on clinical grounds, but are distinguished on the basis of morphologic changes seen in muscle biopsy. Muscle fiber necrosis, a characteristic of the muscular dystrophies, is not seen in the congenital myopathies. Instead there are structural abnormalities, such as collections of rodlike particles (nemaline myopathy), central portions of muscle fibers devoid

of mitochondria (central core disease), or centrally placed nuclei in the majority of fibers (centronuclear myopathy). As more infants with muscle weakness are studied by muscle biopsy, other structural abnormalities will probably be identified. See MUSCULAR DYSTROPHY.

Nongenetic or acquired diseases. These are all characterized by rapidly progressive weakness of the proximal muscles of the limbs, and so resemble the limb-girdle form of muscular dystrophy in the distribution of weakness. They are often included as inflammatory myopathies. Individuals with these disorders have difficulty arising from a recumbent or sitting position, in climbing stairs, and in lifting heavy objects onto a shelf. They also often have tender and painful muscles, may be febrile, and may have other manifestations of a systemic illness.

In many individuals, the muscle disease appears to be primary (polymyositis), with or without associated skin involvement (dermatomyositis). The cause of this group of muscle diseases is unknown, but they are thought to be due to an altered immune state which may result in direct damage to the muscle, or secondary damage to the capillaries. Such a possibility is also supported by the frequent association of acquired muscle disease with collagen vascular diseases, particularly rheumatoid arthritis, scleroderma, and Sjogren's syndrome.

Histologically, there is considerable variation in the size of the individual muscle fibers. There are large numbers of necrotic or degenerating muscle fibers being invaded by phagocytic cells, as well as small basophilic fibers with large prominent nuclei. The latter are thought to be regenerating cells. These features are qualitatively similar to those seen in the muscular dystrophies, but these changes are quantitatively much more extensive in polymyositis, in which the disease process progresses much more rapidly. In approximately 75% of these patients, there is perivascular infiltration by mononuclear inflammatory cells, mostly lymphocytes and plasma cells.

Acquired muscle diseases may also be secondary to carcinoma, and the pathologic changes in the muscle may be quite similar to what has been described in polymyositis and dermatomyositis. Such an association with carcinoma has been noted in up to 20% of adults with polymyositis, and may be even higher in those with dermatomyositis.

Metabolic diseases. A number of metabolic diseases have been associated with muscle symptoms. These include thyroid diseases and certain endocrine diseases, particularly Cushing's syndrome, which are either spontaneous or secondary to therapeutically administered adrenocorticosteroid hormones. No specific histologic changes have been described in the muscle in these disorders.

Muscular symptoms are also seen with some of the glycogen storage diseases, in which specific enzymes necessary for the catabolism of carbohydrates are genetically absent. Muscle symptoms are prominent in Pompe's disease and McArdle's disease; both conditions are inherited in an autosomal recessive pattern. In Pompe's disease acid maltase is the deficient

enzyme. Glycogen accumulates in many organs of the body, including skeletal muscle, the heart, and the central nervous system. The excessive accumulation in the skeletal muscle is seen as a marked vacuolation of the muscle fibers. The child is often weak and floppy from birth and may die from cardiac failure, although milder forms of the disease may occasionally be seen. In McArdle's disease the enzyme myophosphorylase is missing, and the patient is unable to utilize glycogen as a source of energy. Glycogen storage is restricted to skeletal muscle and is seen as subsarcolemmal vacuoles in the muscle fiber. Symptoms are early muscular fatiguing and painful cramps after excessive activity. *See* METABOLIC DISORDERS.

Muscle weakness and paralysis may also be associated with alterations in the level of serum potassium. These are present with episodic or periodic muscle paralysis, and the individual is usually normal between these attacks. These diseases are all hereditary, and during the attack the serum potassium may be low (hypokalemic), elevated (hyperkalemic), or normal (normokalemic). Familial hypokalemic periodic paralysis is the most common form of potassium disorder. It is inherited as an autosomal dominant disease. The muscle fibers contain vacuoles which result from dilatations of the T-tubules and the sarcoplasmic reticulum. The mechanism for the depression of serum potassium during these attacks is unknown, but episodes of weakness and paralysis usually follow excessive exertion followed by rest, or after a heavy carbohydrate meal.

Myotonia. A delayed relaxation of the muscle after forceful contraction (myotonic) is another symptom of muscle disease. This phenomenon is a characteristic feature of myotonic dystrophy and of congenital myotonia. Congenital myotonia may occur as an autosomal dominant (Thomsen's disease) or as a recessive disorder. Myotonia is usually present from early life and is often associated with muscle hypertrophy, but muscle weakness is not a feature of this disease. The muscle shows no definite histologic changes.

Infections. Specific infections of the muscle are not very common and may be caused by bacterial, viral, or parasitic agents. The best known is trichinosis, an infestation by the pork tapeworm, *Trichinella spiralis*. The main muscle symptoms are pain and tenderness, but the systemic symptoms of fever, periorbital edema, and petechial hemorrhages may be more prominent. There is necrosis of muscle fibers and severe inflammation reaction with polymorphonuclear and eosinophilic leukocytes. The parasite can often be seen in the necrotic muscle fiber.

Tumors. Primary tumors of skeletal muscle are quite rare and are of two types: the benign rhabdomyoma and an extremely malignant rhabdomyosarcoma.

Injury. Traumatic injuries to the muscle may be due to laceration, blows, or prolonged pressure and result in focal areas of necrosis (rhabdomyolysis). If the injury is extensive, the muscle respiratory pigment,

myoglobin, may be released into the circulation and excreted in the urine (myoglobinuria), with subsequent renal failure. Rhabdomyolysis with myoglobinuria may also follow rigorous running, jumping, or long marches without previous training. Following the exercise, painful swelling of the anterior muscles of the lower legs occurs, accompanied by myoglobinuria. Histologically, there is necrosis of muscle and later healing with fibrous scar formation. It is thought that swelling of these muscles, which are contained in a rather firm fibrous and bony compartment, interferes with circulation and results in subsequent necrosis. *See* MUSCULAR SYSTEM. S. Mark Sumi

Bibliography. M. Adachi, *Neuromuscular Diseases*, 1989; J. R. Anderson, *Atlas of Skeletal Muscle Pathology*, 1985; R. Buschbacher, *Musculoskeletal Disorders: A Practical Guide for Diagnosis and Rehabilitation*, 1993; J. A. Walton, G. Karpati, and D. Hilton-Jones (eds.), *Disorders of Voluntary Muscle*, 1994.

Mushroom

A macroscopic fungus with a fruiting body (also known as a sporocarp). It has been estimated that over 1,500,000 species of fungi exist on the Earth; yet only about 69,000 species have been described to date. Approximately 14% (10,000) described species of fungi are considered mushrooms. Mushrooms grow aboveground or underground. They have a fleshy or nonfleshy texture. Many are edible, and only a small percentage are poisonous.

Reproduction. Mushrooms reproduce via microscopic spheres (spores) that are roughly comparable to the seeds of higher plants. Spores are produced in large numbers on specialized structures in or on the fruiting body. A 3-in.-diameter (7.5-cm) mushroom, for example, may produce as many as 40 million spores an hour. Large numbers of spores are necessary to ensure survival of the species because most spores do not land on a suitable medium that allows germination and growth. Spores that land on a suitable medium absorb moisture, germinate, and produce hyphae that grow and absorb nutrients from the substratum. If suitable mating types are present and the mycelium develops sufficiently to allow fruiting, the life cycle will continue. In nature, completion of the life cycle is dependent on many factors, including temperature, moisture and nutritional status of the substratum, and gas exchange capacity of the medium.

Commercial production of edible mushrooms. Fewer than 20 species of edible mushrooms are cultivated commercially. The most common cultivated mushroom is *Agaricus bisporus* [Fig. 1], followed by the oyster mushroom (*Pleurotus* spp.) [Fig. 2]. Other major species of commerce include *Lentinula edodes* (shiitake) [Fig. 3], *Auricularia* spp. (wood-ear), *Volvariella volvacea* (straw mushroom), *Flammulina velutipes* (winter mushroom), *Tremella fuciformis* (silver-ear), *Hericium erinaceus* (bear's-head), and *Pboliota nameko* (nameko). China is the leading mushroom-producing country; Japan leads



Fig. 1. Common cultivated mushroom (*Agaricus*). (Courtesy of D. Martinez-Carrera)



Fig. 2. Oyster mushroom (*Pleurotus*). (Courtesy of D. Martinez-Carrera)

the world in number of edible species cultivated commercially.

Cultivation technology. Mushrooms may be cultivated on a wide variety of substrates. The most commonly used materials include straw (wheat, oat, and rice), wood chips and bark, cotton waste (seed hulls and screenings), hay, banana leaves, corn stalks, coffee pulp, and other agricultural waste products. The biodegradation of lignocellulosic materials from agriculture or forestry confers ecological importance on mushroom cultivation, as millions of tons of these by-products are recycled every year as substrates for mushroom growing. Successful cultivation of some mushroom species, for example, *Agaricus bisporus*, requires a composting process to produce a selective medium. Other species usually are cultivated on a noncomposted substratum that has been treated to remove unwanted competitive microorganisms.

Mushrooms are grown from mycelium (the thread-like filaments or hyphae that become interwoven) propagated on a base of steam-sterilized cereal grain (usually rye or millet). This grain and mycelium mixture is called spawn, which is used to seed mushroom substrata. Most spawn is made with mycelium from a stored culture, rather than mycelium whose parent was a spore, because each spore is likely

to yield a new strain and its performance would be unpredictable. Spawn making is a rather complex task and is not practicable for the common mushroom grower. Thus, spawn usually is purchased from a number of commercial spawn makers. However, conventional breeding methods, now complemented by powerful molecular techniques, will produce new strains for spawn production capable of giving higher yields and better mushroom quality.

Agaricus bisporus is produced on compost made of straw-bedded horse manure or on synthetic composts made of straw, hay, corn cobs, or other fibrous materials, plus potash and gypsum. These raw materials are combined in piles that are periodically turned, watered, and formed. The mixing and watering serves to equilibrate the environment and promotes a uniform breakdown of the raw materials. The breakdown of materials is accomplished by the growth and reproduction of microbes on them. This breakdown is necessary to achieve a selective medium (one on which only the mushroom will grow) and to provide food for the mushroom in the most efficient form. Composting has developed from a relatively primitive process to a computer-controlled process. Bulk composting in tunnels with computer control of moisture, carbon dioxide, and temperature has allowed the development of a predictable and high-yielding compost. Research work is now focused on the production of environmentally acceptable compost, involving a considerable reduction of odor, ammonia, and water pollution generated by this process.

Other mushroom species may be grown on steam-sterilized or steam-pasteurized substrate. Several wood-inhabiting species of mushrooms that occur in the wild on dead trees or branches may be grown commercially on steam-treated wood chips. Shiitake, oyster mushroom, winter mushroom, wood-ear, silver-ear, bear's-head, and nameko are grown commercially on heat-treated wood chips. In general, sawdust supplemented with nutrient (rice and



Fig. 3. Shiitake mushroom (*Lentinula*). (Courtesy of D. Martinez-Carrera)

wheat bran, millet, sugar, and dried yeast) is mixed, moistened, and filled into steam-sterilizable containers such as polypropylene bottles or bags. The substratum is sterilized, cooled, and inoculated with spawn. After a period of growth, the colonized substratum is subjected to conditions conducive to the formation of mushroom fruiting bodies.

Nutritional and medicinal value. Mushrooms contain digestible crude protein, all essential amino acids, vitamins (especially provitamin D-2), and minerals; they are high in potassium and low in sodium, saturated fats, and calories. Although they cannot totally replace meat and other high-protein food in the diet, they can be considered an important dietary supplement and a health food.

Fungi have been used for their medicinal properties for over 2000 years. Li Shih-chen of the Ming Dynasty listed more than 20 species of medicinal fungi in his *Compendium of Materia Medica*, some of which are cultivated today. At present, there may be around 200 species of edible mushrooms having medicinal value; however, only a small proportion of these species is available to the consumer. Significant functional properties found in mushrooms are reduction of blood pressure and blood lipid concentrations, regulation of the immune system, and inhibition of inflammation, microbial action, and tumors.

Although there remains an element of folklore in the use of mushrooms in health and medicine, several important drugs have been isolated from mushroom fruiting bodies and mycelium. The best-known drugs obtained from mushrooms are lentinan from *L. edodes*, grifolin from *Grifola frondosa*, and krestin from *Coriolus versicolor*. These compounds are protein-bound polysaccharides or long chains of glucose, found in the cell walls, and function as antitumor immunomodulatory drugs. Administered orally or intravenously, they stimulate the body's immune system and enhance host-mediated resistance against some viral, bacterial, and fungal pathogens and various types of carcinomas. As immunostimulators, these drugs can enhance the activity of macrophages and T lymphocytes and raise interleukin-2 and antibody levels. Especially important is their ability to improve or stabilize the side effects of chemotherapy and radiation treatments, such as increasing appetite, restoring energy, and reducing nausea and pain. Krestin is commercially produced and sold in Japan for clinical use in the treatments of stomach, esophagus, colon, rectum, lung, and mammary gland cancers, usually in conjunction with chemotherapy or radiation. There are other species of mushrooms that have therapeutic effects similar to those elicited by lentinan, grifolin, and krestin, or unique effects. Another polysaccharide fraction, known as LEM, has been isolated from *L. edodes* mycelium. LEM inhibits human immunodeficiency virus (HIV) infection in vitro, and may be used against acquired immune deficiency syndrome (AIDS). The content and potency of bioactive ingredients are affected by the strain cultivated, substrates, growing conditions, and the way mush-

rooms are prepared and consumed. It is likely that additional chemically and biologically distinct compounds will be isolated which have medicinal value. See FUNGI; MEDICAL MYCOLOGY. D. Martínez-Carrera
Bibliography. C. J. Alexopoulos and C. W. Mims, *Introductory Mycology*, 1979; W. M. Breene, Nutritional and medicinal value of specialty mushrooms, *J. Food Protect.*, 53:883–894, 1990; L. A. Casselton and N. S. Olesnicky, Molecular genetics of mating recognition in basidiomycete fungi, *Microbiol. Mol. Biol. Rev.*, 62:55–70, 1998; S. T. Chang and P. G. Miles, *Edible Mushrooms and Their Cultivation*, 1989; P. Stamets, *Growing Gourmet and Medicinal Mushrooms*, 1993.

Musical acoustics

The branch of acoustics that deals with the generation of sound by musical instruments, the transmission of sound to the listener, and the perception of musical sound. It has fascinated scientists since the time of Pythagoras. It captured the interest of noted scientists such as Hermann von Helmholtz, Lord Rayleigh, Chandrasekhara Raman, John Tyndall, Félix Savart, and Frederick Saunders. It has become an interdisciplinary field joining physics with music. Indeed, important contributions have been made by psychologists, engineers, physiologists, architects, mathematicians, and materials scientists as well as by musicians and physicists.

Generation of musical sound. A main research activity in musical acoustics is the study of the way in which musical instruments vibrate and produce sound. The most common way of classifying musical instruments is according to the nature of the primary vibrator, into string instruments, wind instruments, and percussion instruments. The equivalent classification by musicologists Erich von Hornbostel and Curt Sachs is into chordophones (string instruments), aeorphones (wind instruments), idiophones (xylophones, marimbas, chimes, and so forth), and membranophones (drums; the last two categories are generally considered percussion instruments). To these may be added the electronic synthesizer, the digital computer, and the human voice.

Complex vibrations. The vibrations of a plucked string, a struck membrane, or a blown pipe can be described in terms of normal modes of vibration. Determining the normal modes of a complex vibrator is often termed modal analysis. Much of the progress in understanding how musical instruments generate sound is due to new methods of modal analysis, such as holographic interferometry and experimental modal testing. Holographic interferometry, for example, provides the investigator with a detailed contour map of the vibrational shapes at different frequencies from which the normal modes of vibration can be determined (**Fig. 1**). See CAVITY RESONATOR; INTERFEROMETRY; MODE OF VIBRATION; VIBRATION.

Nature of the feedback. In the case of most percussion, plucked string, and struck string instruments, the player delivers energy to the primary vibrator (string,

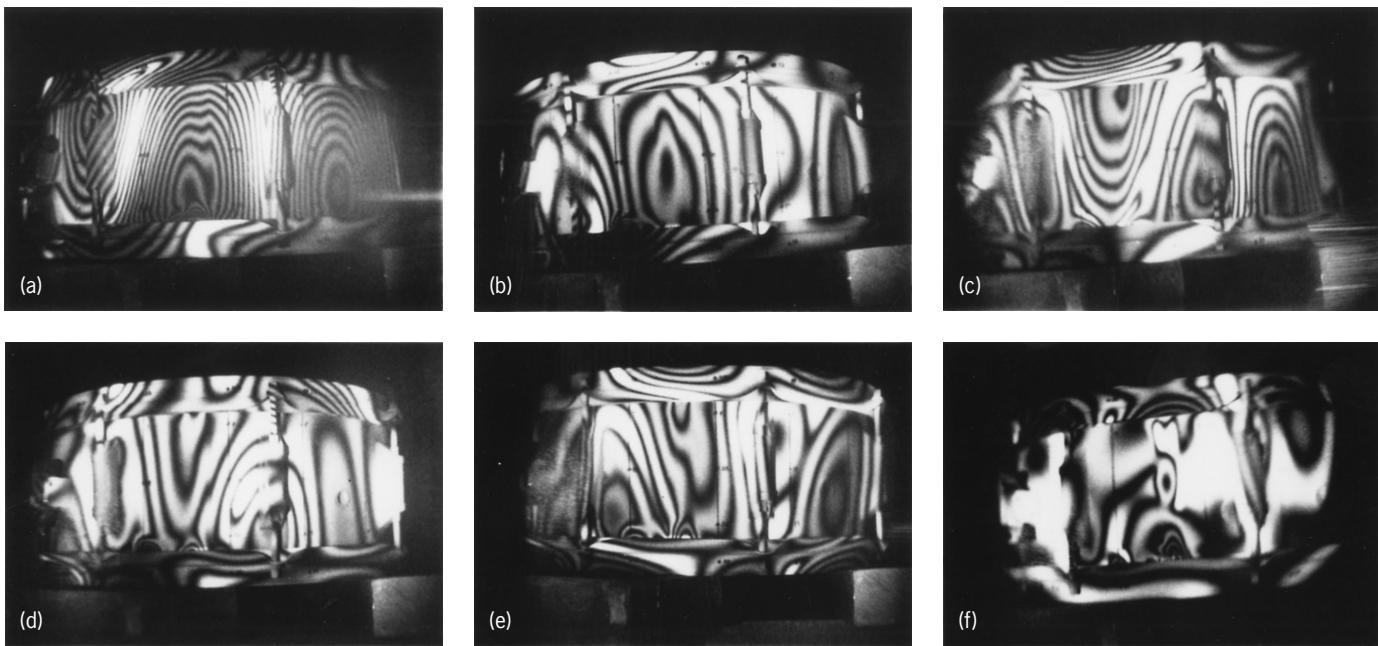


Fig. 1. Holographic interferograms of a snare drum. (a) 874 Hz. (b) 1064 Hz. (c) 1176 Hz. (d) 1271 Hz. (e) 1408 Hz. (f) 1670 Hz.

membrane, bar, or plate) and thereafter has little control over the way it vibrates. In the case of wind and bowed string instruments, however, the continuing flow of energy is controlled by feedback from the vibrating system. In brass and reed woodwinds, pressure feedback opens or closes the input valve. In flutes or flue organ pipes, however, the input valve is flow-controlled. In bowed string instruments, pulses on the string control the stick-slip action of the bow on the string.

Acoustics of singing. In singing, as in speaking, voice sounds are produced by vocal fold vibrations that interrupt the air stream from the lungs, producing a train of air pulses. This signal is modified when passing through the vocal tract, consist of the pharynx and the mouth. The vocal tract, acting as a tube resonator, filters the sound from the vocal folds, adding several strong resonance peaks or formants to the spectrum of the singer's sound. The formants of the singing voice are quite similar to those of the speaking voice, but there are some notable differences, such as the concentration of energy between 2000 and 3500 Hz, referred to as the singer's formant, which gives the trained singer's voice a "shine" and allows the opera singer to be heard in the presence of a loud orchestral accompaniment. *See* MUSICAL INSTRUMENTS; SPEECH.

Transmission of musical sound. Musical acousticians, like architectural acousticians, have always been very interested in the acoustics of concert halls. Since the average person now hears more recorded music than live performance, musical acousticians have paid increasing attention to the recording and reproduction of musical sound. It is impossible to divorce high-fidelity sound reproduction from room acoustics, since the acoustics of both the recording and the listening rooms have a strong influence on the musical sound that reaches the ear of the listener.

One of the problems in designing concert halls is the differences in the individual tastes of both performers and listeners. Some performers want the hall to be reverberant, others put a higher premium on clarity. Even getting the many parties concerned with planning a concert hall to speak the same language can be difficult. Leo Beranek (1996) selected and defined 18 terms that should be understood by both the musical and acoustical planners: intimacy (presence), reverberation (liveness), apparent source width (spaciousness), listener envelopment, clarity, warmth, loudness, acoustic glare, brilliance, balance, blend, ensemble, immediacy of response (attack), texture, freedom from echo, dynamic range, extraneous effects on tone quality, and uniformity of sound. Relating subjective qualities to physically measurable quantities, which the architect can attempt to control, is another important task. *See* ARCHITECTURAL ACOUSTICS.

Rapid advances in the technology of sound recording have had a profound effect on the distribution of music, both classical and popular. Although large libraries of music recorded using analog techniques still exist, nearly all recording now employs digital techniques. The most familiar format for digital recording of music has been the compact disc (CD) digital audio system, introduced in 1982, which now may be replaced by the DVD (digital versatile disk) with greater storage capacity. Digital audio tape recorders (DATs) appeared about 1987. *See* COMPACT DISK; OPTICAL RECORDING; SOUND RECORDING.

Perception of musical sound. Four attributes are frequently used to describe musical sound: loudness, pitch, timbre, and duration. Each of the subjective qualities depends on one or more physical parameters that can be measured. Loudness, for example, depends mainly on sound pressure but also on the

| Dependence of subjective qualities of sound on physical parameters | | | | |
|--|---------------------|-------|--------|----------|
| Physical parameter | Subjective quality* | | | |
| | Loudness | Pitch | Timbre | Duration |
| Pressure | +++ | + | + | + |
| Frequency | + | +++ | ++ | + |
| Spectrum | + | + | +++ | + |
| Duration | + | + | + | +++ |
| Envelope | + | + | ++ | + |

*+ = weakly dependent; ++ = moderately dependent; +++ = strongly dependent.

spectrum of the partials and the physical duration. Pitch depends mainly on frequency, but also shows lesser dependence on sound pressure and envelope. Timbre includes all the attributes by which sounds with the same pitch and loudness are distinguished. The **table** summarizes much of the knowledge about this dependence. Relating the subjective qualities of sound to the physical parameters is a central problem in psychoacoustics, and musical acousticians are concerned with this same problem as it applies to musical sound. See PSYCHOACOUSTICS.

Sound pressure level is measured with a sound level meter and is generally expressed on a logarithmic scale of decibels (dB) using an appropriate reference level and weighting network. From measurements of the sound pressure level at different frequencies, it is possible to calculate a subjective loudness, expressed in sones, which describes the sensation of loudness heard by an average listener. Musicians prefer to use dynamic markings ranging from *ppp* (very soft) to *fff* (very loud). See DECIBEL; LOUDNESS; SOUND; SOUND PRESSURE.

Pitch is defined as that attribute of auditory sensation in terms of which sounds may be ordered on a scale extending from low to high. Although various attempts have been made to establish a psychophysical pitch scale (using units called mels), the scale has never become very popular. Neither has the psychophysical scale based on critical bands of hearing, on which a critical bandwidth is designated one bark (which is nearly the same as 100 mels). Rather, pitch is generally related to a musical scale where the octave, rather than the critical bandwidth, is the “natural” pitch interval. See PITCH.

Timbre is defined as that attribute of auditory sensation in terms of which a listener can judge two sounds similarly presented and having the same loudness and pitch as dissimilar. Timbre depends primarily on the spectrum of the sound, but it also depends upon the waveform, the sound pressure, the frequency location of the spectrum, and the temporal characteristics of the sound. It has been found impossible to construct a single subjective scale of timbre (such as the sone scale of loudness or the mel scale of pitch); multidimensional scales have been constructed. The term “tone color” is often used to refer to that part of timbre that is attributable to the steady-state part of the tone, but the time envelope (and especially the attack) has been found to be very important in determining timbre as well.

Combination tones and harmony. Another subject relating to the perception of music is combination tones. When two tones that are close together in frequency are sounded at the same time, beats generally are heard, at a rate that is equal to their frequency difference. When the frequency difference Δf exceeds 15 Hz or so, the beat sensation disappears, and a roughness appears. As Δf increases still further, a point is reached at which the “fused” tone at the average frequency gives way to two tones, still with roughness. The respective resonance regions on the basilar membrane are now separated sufficiently to give two distinct pitches, but the excitations overlap to give a sense of roughness. When the separation Δf exceeds the width of the critical band, the roughness disappears, and the two tones begin to blend.

As the tone separation Δf continues to increase, a difference tone (sometimes called a Tartini tone after the Italian violinist who discovered it around 1714). This difference tone has a pitch corresponding to the frequency difference $f_2 - f_1$. Another combination tone that can usually be heard is the cubic difference tone, having a frequency $2f_1 - f_2$. Other combination tones may be audible, but generally the difference tone and the cubic difference tone are the most prominent. They are shown on a musical staff in **Fig. 2**.

Pythagoras of ancient Greece is considered to have discovered that the tones produced by a string vibrating in two parts with simple ratios such as 2:1, 3:2, or 4:3 sound harmonious. These ratios define the so-called perfect intervals of music, which are considered to have the greatest consonance. Other consonant intervals in music are the major sixth ($f_2/f_1 = 5/3$), the major third ($f_2/f_1 = 5/4$), the major sixth ($f_2/f_1 = 8/5$), and the minor third ($f_2/f_1 = 6/5$).

Why are some intervals more consonant than others? Helmholtz concluded that dissonance (the opposite of consonance) is greatest when partials of the two tones produce 30 to 40 beats per second (which are not heard as beats but produce roughness). The more the partials of one tone coincide in frequency with the partials of the other, the less chance of roughness. This explains why simple frequency ratios define the most consonant intervals.

More recent research has concluded that consonance is related to the critical band. If the frequency difference between two pure tones is greater than a critical band, they sound consonant; if it is less than a critical band, they sound dissonant. The

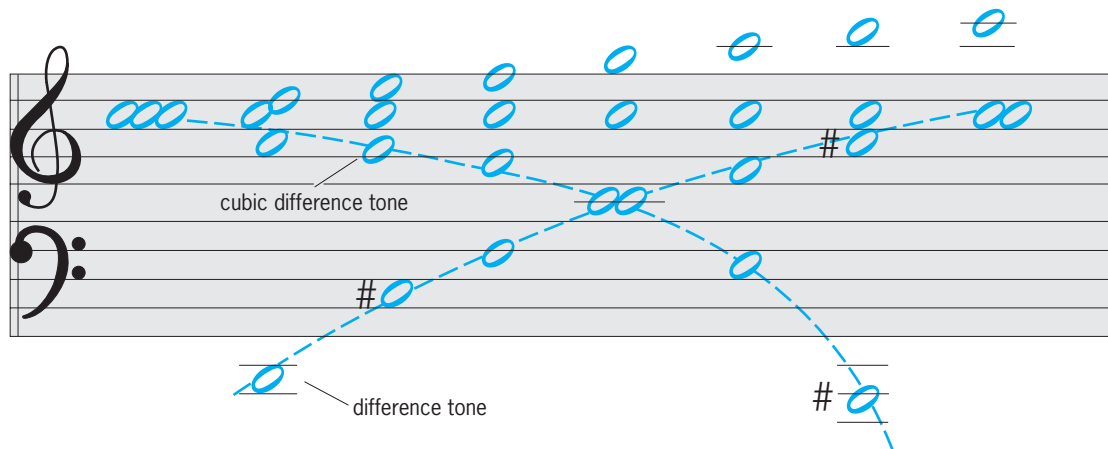


Fig. 2. Combination tones due to two tones of varying separation shown on a musical staff. As the difference tone rises, the cubic difference tone falls, so that they cross at $f_2 = 3/2 f_1$ (a perfect fifth).

maximum dissonance occurs when Δf is approximately $1/4$ of a critical band, which agrees reasonably well with Helmholtz's criterion for tones around 500 Hz.

Musical scales and temperament. Most musical composition is based on musical scales, the most common ones being those with five notes (pentatonic), 12 notes (chromatic), or seven notes (major and minor diatonic). Western music divides the octave into 12 steps called semitones. All the semitones in an octave constitute a chromatic scale or 12-tone scale. However, most music makes use of a scale of seven selected notes, designated as either a major scale or a minor scale.

There are many ways to construct musical scales. The three most important scales are the Pythagorean scale, the just scale, and the scale of equal temperament. The Pythagorean scale creates the greatest number of perfect fourths and fifths. An octave is a fourth plus a fifth ($3/2 \times 4/3 = 2$); therefore, going up a fourth leads to the same letter as going down a fifth, and vice versa. All notes on the scale (sharps and flats included) can be reached by going up or down 12 successive fifths or 12 successive fourths. If $3/2$ is multiplied by itself 12 times, the result is 129.7, which means that going up 12 perfect fifths takes one up seven octaves plus one-fourth of a semitone extra. The Pythagorean scale deals with this by using semitones of two different sizes. This leads to rather poor tuning of the thirds.

Because Pythagorean thirds sound out of tune, numerous alterations to the Pythagorean scale have been developed. Nearly all of them flat the third note so that the major third and minor third are close to the corresponding just intervals. Similar compromises form the bases of various meantone temperaments.

The scale of just intonation (or just scale) is based on the major triad, a group of three notes that sound particularly harmonious (for example, C : E : G). The notes of the major triad are spaced in two intervals: a major third (C : E) and a minor third (E : G). When these intervals are made as consonant as possible, the notes in the major triad are found to have fre-

quencies in the ratios 4:5:6. The scale of just intonation assigns this ratio to the three major triads (tonic, subdominant, and dominant).

The ultimate compromise is equal temperament, which makes all semitones the same. The scale of equal temperament consists of five equal whole tones and two semitones; the whole tones are twice the size of the semitones. Twelve equal semitones (each having a ratio $2^{1/12} = 1.05946$) make up an octave. Rather than deal with ratios, it is customary to compare tones by using cents. One cent is $1/100$ of a semitone in equal temperament. Thus an octave is 1200 cents, a tempered fifth is 700 cents, and so forth. See SCALE (MUSIC). Thomas D. Rossing

Bibliography. L. Beranek, *Concert and Opera Halls and How They Sound*, 1996; M. J. Crocker (ed.), *Encyclopedia of Acoustics*, vol. 4, 1997; F. A. Everest, *The Master Handbook of Acoustics*, 4th ed., 2000; N. H. Fletcher and T. D. Rossing, *The Physics of Musical Instruments*, 2d ed., 1998; W. H. Hartmann, *Signals, Sound, and Sensation*, 1997; K. C. Pohlmann, *Principles of Digital Audio*, 4th ed., 2000; T. D. Rossing, *The Science of Sound*, 2d ed., 1990.

Musical instruments

Musical instruments evolved from simple found objects: hollow logs, perhaps with animal skins stretched over the ends, bamboo tubes or large seashells, and stretched bowstrings. They are now classified as percussion instruments, wind instruments, and string instruments, but other divisions are possible.

Conventional Instruments

A useful distinction is between impulsively excited instruments (bells, drums, plucked strings, and so forth), in which the sound gradually decays after an initial input of energy, and sustained-tone instruments (flutes, trumpets, violins, and so forth), in which energy is supplied continuously. Because extended objects can vibrate in many characteristic ways, termed normal modes, the sounds they

produce are combinations of many different frequencies. The lowest of these frequencies is called the fundamental, and the others are called upper partials. In an impulsively excited instrument, the mode frequencies are determined by the shape of the vibrating object, and there is no necessary relation between them. In a sustained-tone instrument, the frequencies of the upper partials are all exact integer multiples of the fundamental frequency, and they are called harmonics. See HARMONIC (PERIODIC PHENOMENA); MODE OF VIBRATION; MUSICAL ACOUSTICS; VIBRATION.

Sustained-tone instruments generate sound by means of a mechanism, such as a moving bow or a pressure-controlled valve, which essentially inserts a negative acoustic resistance into the system. Because of feedback coupling and the fact that the generator mechanism is inherently nonlinear, an exactly repetitive waveform is produced in which all overtones have frequencies that are exact integer multiples of that of the fundamental. Such overtones are called harmonics, the fundamental being the first harmonic. This harmonic structure is, however, not exactly maintained during transients, particularly the attack transient, when the inharmonic normal modes of the resonator may be momentarily excited.

A system diagram for a sustained-tone musical instrument is shown in **Fig. 1**. For an instrument such as the violin all parts of the system are present, while for wind instruments there is no separate sound radiator. For impulsive instruments, such as the guitar or the drum, there is no nonlinear generator and the energy source operates only momentarily. In all cases the sound output power is only about 1% of the input power, which is dissipated in frictional, viscous, and thermal losses.

Percussion instruments. Instruments such as drums, bells, gongs, and xylophones are called percussion instruments because the sound is initiated by a blow from some sort of hammer.

Drums. A vibrating membrane has many closely spaced modes, and the sound of a drum does not give any strong impression of musical pitch. If the membrane is stretched over an air volume, however,

the resonances of this volume may emphasize certain frequencies, as in timpani or tom-toms, to produce a pitched sound. The player can vary the pitch by varying membrane tension, and can also modify the sound by striking the drumhead near the center or near the edge to preferentially excite different vibrational modes. See PITCH; RESONANCE (ACOUSTICS AND MECHANICS).

Tuned percussion instruments. If a heavy metal object is struck with a metal hammer, it stores a great deal of mechanical energy and can vibrate and radiate sound for many seconds. By careful shaping of the metal, usually bronze, by casting and subsequent machining, the vibrational frequencies can be brought into almost integer ratios, so that object becomes a bell with a pleasant sound and well-defined musical pitch. In a typical church bell, five or more modes are carefully tuned, one of which is in a minor-third frequency ratio 6:5 to the fundamental, to give a characteristic sound. Because cast bells are large and heavy, orchestras use sets of metal tubes to produce an approximation to a bell sound.

A rather flat metal shell struck with a soft hammer is generally called a gong. Only the first one or two modes are excited, and the sound has a soft and well-pitched note. If the shell is made from a rather thin sheet and is slightly conical in form, it is usually called a cymbal. Because the metal is thin, a blow with even a soft hammer can excite it into a quite large vibration, and this causes interactions between vibration modes, giving a characteristic shimmering sound.

Xylophones and marimbas consist of a graded set of slats of wood mounted on a frame and struck with a wooden mallet. The playing range is typically three to four octaves, though larger instruments have been built. Each slat is undercut with a circular arch to tune the first overtone, and in the marimba there is a tube resonator underneath each bar. Similar instruments are also made with metal bars.

Wind instruments. All wind instruments rely upon the vibration of a column of air enclosed by rigid walls. Sound waves propagate along this column and are reflected at its ends, the relative sign of the

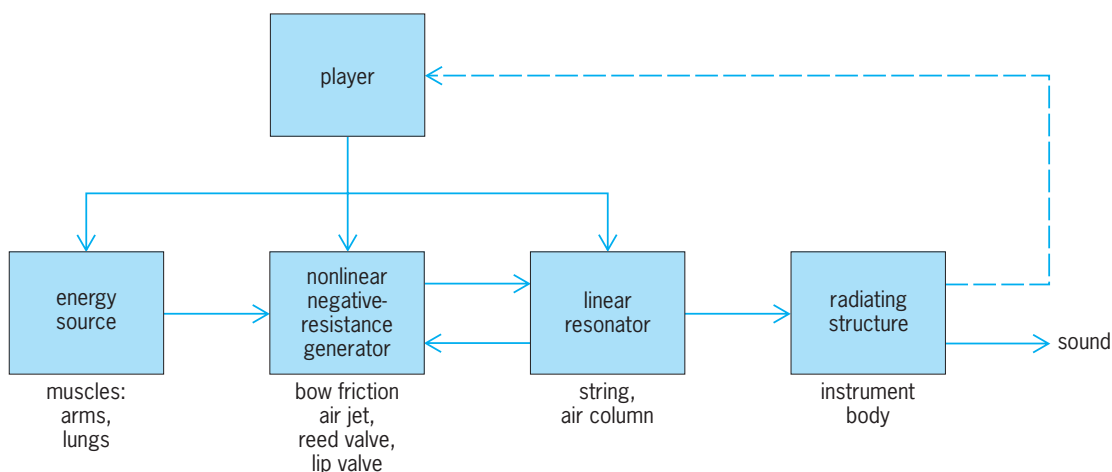


Fig. 1. System diagram for a typical musical instrument.

reflected pressure wave depending upon whether the end is open or closed. For a cylindrical air column open at both ends, the possible vibrations are those with an integral number of half wavelengths of sound along the tube, giving a frequency sequence 2, 4, 6, . . . times $c/4L$, where c is the velocity of sound in air (343 m/s or 1125 ft/s) and L is the length of the tube. The same result applies to an almost complete conical tube, irrespective of whether the narrow end is open or closed. For a cylindrical tube closed at one end, possible vibrations must have an odd number of quarter wavelengths along the tube, giving a sequence 1, 3, 5, . . . for a tube of the same length. These mode frequencies are, however, not quite exact, because of frequency-dependent end corrections. For brass instruments, the air column is nearly cylindrical for one-half to two-thirds of its length, beginning at the mouthpiece end, but flares to a bell at the open end. This flare raises the frequencies of all the modes, compared with a cylindrical tube of the same length, giving a sequence such as 1.4, 4, 6, 8, The first mode is not used in playing. See CAVITY RESONATOR.

In woodwind instruments the length of the air column, and thus the note to be played, can be changed by opening finger holes along the instrument body. Since there are 12 notes in an octave, and a human has only 10 fingers, this necessitates either compromise fingerings or else a complex system of keys as in modern instruments. In brass instruments all notes must emerge through the flared horn to maintain uniform sound quality, and pitch is changed as discussed below.

Air jet-driven instruments. In instruments of the flute family, a narrow jet of air is blown from the player's mouth, or from a narrow slit in the case of recorders and organ pipes, to cross an aperture near one end of a more or less cylindrical pipe and strike against a sharp edge. As the air jet blows into and out of the pipe mouth, it excites the pipe modes. The deflection of the jet is in turn governed by the airflow into and out of the pipe caused by the sound waves. For the system to work, there must be a match between the frequency of the mode that is excited and the travel time of waves on the jet from the player's lip to the sharp edge. The player can therefore select the mode to be played by varying lip position and blowing pressure. Once the oscillation is established, nonlinear effects generate all harmonics of the fundamental, and these are amplified because of their close match to the resonances of the tube. For a modern flute, the playing range is just over three octaves. Loudness is controlled primarily by varying the area of the lip aperture. See JET FLOW.

Reed-driven instruments. Many woodwind instruments use a reed valve to inject an oscillating flow of air into the instrument tube. In clarinets, which have a cylindrical tube, and in saxophones, which have a broad conical bore, the reed is a single piece of thinned cane clamped against an aperture in a carefully shaped mouthpiece. In the oboe and bassoon, which have a narrow conical bore, the reed consists of two similar pieces of thinned cane bound together.

The vibration of the reed is controlled by the sound pressure in the instrument, and it effectively closes one end of the tube. Its small vibration allows puffs of air to enter from the player's mouth to maintain the sound oscillation, which is locked to one of the modes of vibration of the air column. Nonlinear effects generate all harmonics of the frequency being played, and these are amplified by the well-aligned resonances of the air column—all harmonics for a conical bore and only odd harmonics for a cylindrical one. For all these instruments, the playing range is about three octaves. Loudness is increased by reducing lip pressure on the reed, allowing it to vibrate to larger amplitude, together with a modest increase in blowing pressure.

Lip-driven instruments. In brass instruments the player's lips provide a vibrating valve under the control of the resonances of the air column. Because of the geometry of the lip-valve motion, it is necessary for the player to adjust lip tension so that the natural frequency of vibration of the lips themselves matches that of the air column for the note to be played. Because of the necessity for this match, it is possible for the player to select any of the natural modes of the air column at will, thus allowing the playing of bugle calls without the use of any valves or slides. In "natural" horns and trumpets with no valves, a whole octave of notes can be played using the 8th to 16th modes of the horn. Again, nonlinear effects generate exact harmonics of the pitch being played, and the player controls loudness primarily by controlling blowing pressure, which can be very high. Notes between the natural modes of the horn are played either by using a slide to add extra cylindrical length, as in the trombone, or by using a set of finger-operated valves to insert extra tube lengths, as in the French horn or trumpet. A set of three valves that lower the pitch by one, two, or three semitones, respectively, is adequate.

Pipe organs. A pipe organ contains as many as 10,000 pipes. Most of these are flutelike flue pipes, but there are also pipes with single metal reeds. Since each pipe need produce only one note, a great range of shapes and tone colors is possible. In addition, depressing the key of one note can sound many pipes at suboctaves, octaves, and higher multiples of the note frequency, as controlled by drawstops. An organ has as many as five keyboards and one pedalboard, and the action either may be completely mechanical or may involve pneumatic, electric, or electro-pneumatic actuators. Loudness and tone quality are both varied by means of stop selection, but some pipe ranks are enclosed in chambers with louvred fronts controlled by a swell pedal, allowing additional control of loudness.

String instruments. A vibrating string clamped at its two ends has modes in very nearly exact harmonic relationship, the deviation being caused by string stiffness. In addition, the length of a vibrating string can easily be changed by pressing a finger against it. Strings are therefore a good basis for musical instruments. The problem is that, being very thin, a string radiates very little sound. Its vibrations must

therefore be coupled to a much larger structure, the soundboard or body of the instrument, to achieve adequate sound level. The musical quality of the instrument depends largely on the design and resonances of the radiating structure which, generally being of wood, varies considerably from one instrument to another that is outwardly similar. In impulsive instruments the string can be excited either by plucking, as in the guitar or harp, or by hitting with a soft hammer, as in the piano. In sustained tone instruments such as the violin, the string is excited by drawing a horsehair bow across it at a constant velocity.

Plucked string instruments. Since the modes of a thin string are very nearly harmonic, the overtones of plucked strings are also nearly harmonic. The relative amplitudes of different modes can, however, be varied greatly by changing the plucking point. A plucking position near the end of the string increases the relative amplitudes of high-frequency modes and thus brightens the tone. Loudness is increased by increasing the force of the plucking action. The relative amplitudes of the overtones, and thus the tone quality, are also affected by the resonance properties of the instrument body, and great attention is paid to this by the maker. In guitars the lowest body resonance involves motion of air through the guitar rose as well as motion of the upper plate and back, while for higher modes only the plate vibrations are important.

Bowed string instruments. Strings can be excited by the motion of a bow because the static frictional force between bow and string is greater than the sliding friction. The string successively sticks to the moving bow, or else slips back under the much smaller sliding friction. The vibration frequency of the string is determined only by its length and tension, and the sound quality varies only slightly with bowing position. Loudness is increased by increasing bow speed and contact force on the string. Violins, violas, and cellos have carved arched top and back plates, the shapes of which are carefully adjusted by the maker to achieve the desired body resonance frequencies which are important for tone quality. The shape of the traditional f-holes controls the lowest resonance, which also involves motion of the enclosed air.

Pianos and harpsichords. The notes on a piano have one, two, or three steel strings, the lower strings being overwrapped with brass or copper wire. Because of string stiffness, the octaves are stretched to a little greater than a 2:1 frequency ratio. String vibration is communicated through a bridge to a flat, ribbed wooden soundboard. Felted hammers are made to hit the strings through a complex mechanical action which also raises dampers from individual strings. Two pedals control damper motion, and a third allows only a single string of each note to be struck. Loudness is increased by increasing the force of the finger stroke on the keyboard.

The harpsichord is an earlier and mechanically simpler instrument in which single unwrapped metal strings are plucked by quill (now nylon) plectra. Most harpsichords have two sets of strings, usually with one set pitched an octave above the other. There is

also usually a lute stop which applies felt dampers to all the strings to give a soft effect. Large harpsichords have two keyboards and extra sets of strings, often including a set at suboctave pitch. Both tone quality and loudness are varied by varying the string sets used. No variation in plucking force is possible.

Human voice. While not really a musical instrument, the human singing voice is often involved in music. The larynx contains two vocal folds which can open and close like the human lips. When nearly closed, they can be made to vibrate by air pressure in the lungs—the frequency of vibration, and thus the musical pitch of the sound, being controlled by muscular tension. Vocal fold vibration releases successive pulses of air into the base of the vocal tract, giving a sound with high harmonic content. The spectral envelope of the sound is modified by resonances of the vocal tract, giving bands of emphasis called formants. The frequencies of the first three formants, typically around 500, 1500, and 2500 Hz, can be changed by movement of the tongue, jaw, and lips to change the resonance frequencies of the vocal tract, and thus produce characteristic vowel sounds. Consonants are produced as bursts of noise with shaped frequency envelopes. *See* SPEECH. Neville H. Fletcher

Amplification. The sound or vibration of any conventional musical instrument can be picked up, amplified, recorded, or reproduced. Many conventional musical instruments are amplified during live performance. Thus they can be considered electronic in the limited sense that all electronic musical instruments have amplifiers, and deliver their tones to listeners over loudspeakers or earphones. For instruments having mechanical vibrators such as tuned strings, reeds, or bars, vibration pickups can be placed in contact with (or in proximity to) the vibrators (**Fig. 2**). For resonant-air-column instruments such as brass, microphones can be placed in the sound field. The amplified sound can then reinforce or even dominate the direct sound from the instrument. *See* AMPLIFIER; AUDIO AMPLIFIER; MICROPHONE; SOUND RECORDING; SOUND-REPRODUCING SYSTEMS; TRANSDUCER.

Electronic amplification circuitry allows the performer to control the tone of the instrument in unconventional ways, such as changing the tone timbre or even the frequency. However, before such control can be effective, the direct tone from the instrument must be reduced to allow the modified tone to dominate. For this purpose, amplified instruments are usually redesigned to reduce direct sound

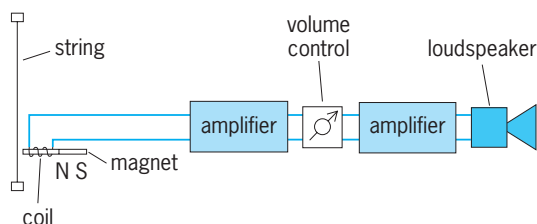


Fig. 2. Electrical scheme of an amplified electropiano. (After H. F. Olson, *Music, Physics and Engineering*, Dover, 1967)

output. Examples are solid-body electric guitars, electropianos with struck strings (or reeds) but without soundboards, amplified enclosed-reed organs, and brass instruments with mutes containing microphones. Such redesign also permits the instrument to be played more softly than normal, either for greater dynamic expression or for practice without disturbing other people, such as in a classroom where music students practice or receive instruction individually or in groups.

Changing the timbre of the conventional tone usually involves alteration of the relative amount of the harmonics in the tone spectrum. One method is to pick up the vibration or sound at different points along the string, for example, either by moving the pickup or by using several pickups mounted at different positions and selecting or mixing the outputs. Another method is to use a single pickup (or set) with tone controls or resonant filter circuits that can be varied or switched manually or by foot pedals during performance. Dynamic use of such controls can provide new timbre effects not possible in the original conventional instrument.

Another kind of dynamic timbre-changing effect is the Leslie effect, used often on the output of electronic keyboard instruments, such as Hammond organs. A Leslie effect is created by rotating one or more directional speakers inside a cabinet such that a mixture of Doppler-shifted reflections is generated. Many electronic “effects” can be considered imitations of the Leslie effect to varying degrees, such as a tremolo circuit, which imparts a simple periodic amplitude modulation on the sound, and the chorus effect, which is typically a sum of electronically delayed copies of the sound, slightly mistuned.

In addition to changing the harmonic mix, new harmonics are often created by using controlled amounts of distortion. Nonlinear distortion is often used by electric guitar players, and sometimes by keyboard players as well. When nonlinear distortion is applied to a harmonic, single-pitched sound, that is, a single note as opposed to a chord, the distortion is harmonic, and generally brightens the tone. If two or more notes are present, at different pitches, then intermodulation distortion occurs, giving rise to many new frequencies, and typically a more distorted sound.

Often in electronic amplification there is a feedback path from the loudspeaker to the vibrating source, such as acoustic coupling from an amplifier loudspeaker to the strings of an electric guitar. Such feedback can cause objectionable “squealing” noises, but it can also be used to provide indefinite sustain for certain harmonics. In effect, the amplifier becomes part of the instrument. *See* FEEDBACK CIRCUIT.

Electrical, Electronic, and Software Instruments

Electrical musical instruments produce electrical tone signals for amplification (and loudspeaker listening) without using tuned mechanical vibrators or air columns. Early electrical instruments predated electronics, generating their tone signals by electro-

magnetic or electrostatic rotating machinery. Later, most of the instruments became electronic, producing their tone waves from vacuum-tube circuits, then transistor circuits, and now integrated electronic circuits. In the 1980s, electronic instruments transitioned from analog to digital, offering much more precise tuning (and alternative tunings) and repeatability of configuration. The initial purpose of these instruments was to substitute for conventional musical instruments. However, over time, successful electronic instruments evolved their own repertoire, so that now electronic synthesizers typically emulate many synthesizers of the past as well as traditional musical instruments. *See* INTEGRATED CIRCUITS.

The purposes of these electronic instruments have been broadened from simple substitution for conventional instruments to (1) producing new musical sounds, (2) improving instrumentation for music education, (3) increasing the ease of music playing by amateurs, (4) extending the performance capability of trained musicians, and (5) facilitating music composition. Moreover, the means developed for each one of these purposes have also been adapted to other purposes. For example, inadequate electronic tonal substitutions have become new electronic sounds; amplified instruments have been found useful for group instruction; computer-generated tones and compositions have stimulated similar live, real-time tone generation and performances; and easy-play electronic aids for amateurs, such as automatic arpeggios, have extended professional performance beyond previous human manipulative capability.

Sound generation methods. Many different means of sound generation have been used during the evolution of electronic musical instruments, some now obsolete and others invented more recently.

In concept at least, the most simple electronic method would be to substitute directly for each conventional mechanical (string, reed) or acoustical (resonant air column) sound source an electronic circuit source or software algorithm that generates the same complex tone wave. However, in different parts of the musical scale the waveform typically differs for the same instrument. Moreover, keyboard instruments require that a number of tones be produced at the same time. Consequently, the substitution method is less feasible economically than other methods described below. *See* WAVEFORM.

Additive synthesis. Another method is to synthesize each desired complex tone by adding together an appropriate number of simple (sine) waves at harmonically related frequencies (for example, 100, 200, 300 Hz, and so forth), with the relative amplitudes of the harmonic waves adjustable to produce the desired resultant waveform (**Fig. 3**). Additive synthesis is the reverse of Fourier analysis, which enables any repetitive waveform, even a very complex one, to be analyzed into a series of simple sine-wave harmonics. *See* FOURIER SERIES AND TRANSFORMS; HARMONIC ANALYZER.

The additive synthesis method has the advantage that all of the tone-wave circuits generate the same mathematically simple waveform, and that the same

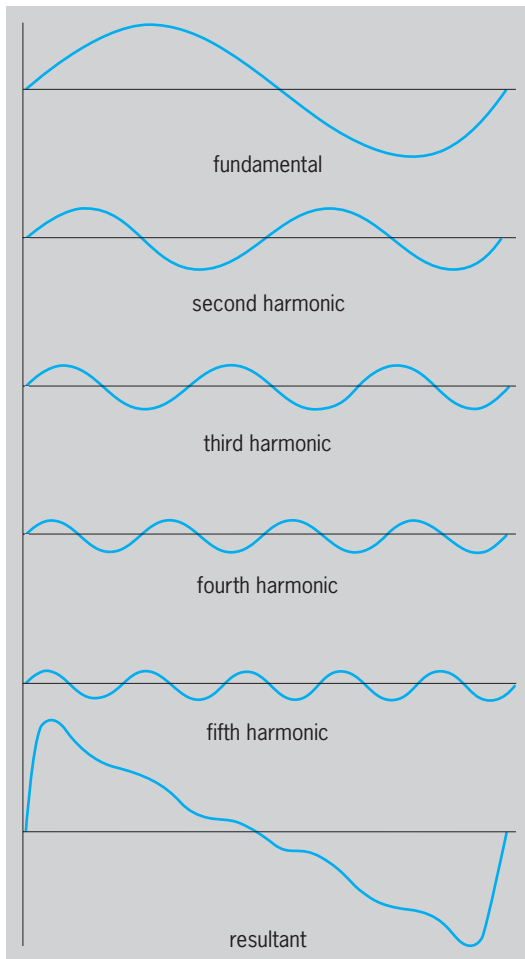


Fig. 3. Synthesis of resultant waveform from five harmonics. (After H. F. Olson, *Music, Physics and Engineering*, Dover, 1967)

harmonic generator can contribute to the synthesis of more than one tone in the musical scale. However, musical tone signals often require many harmonics (for example, string and reed tones), and thus the controls used to adjust the relative amplitudes of the harmonics to the desired values can be very complex and costly. This has typically led designers using this method to make compromises (for example, in the number of available harmonics per tone, and in borrowing inharmonic sine waves).

Additive synthesis has been used extensively in the field of computer music since at least the 1960s, and it has been the basis of a number of limited-production hardware devices. The Hammond organ can be viewed as an analog synthesizer based on additive synthesis, because its tone wheels produce quasi-sinusoidal tones that are summed with weightings provided by sliders. In current music synthesizers, however, there are typically more efficient ways than additive synthesis to achieve desired sounds, and mass-produced electronic musical instruments are generally not based on additive synthesis. However, the concept of additive synthesis remains important, and there are closely related descendants, such as sinusoidal modeling and group additive syn-

thesis, that are both efficient and general in their sound-generating capabilities.

Subtractive synthesis. Subtractive synthesis is similar in principle to that of speech (and singing) sound formation within the human voice system. At each desired musical frequency an electronic circuit produces a standard basic tone waveform (such as a sawtooth wave; Fig. 4), which already contains a very large number of harmonics at known relative amplitudes. Then a variety of electric or electronic filters is provided in the instrument wherein switch selection, individually or collectively, converts the basic tone signals into the desired musical tone waveforms. Subtractive synthesis is somewhat the opposite of additive synthesis, since it practically subtracts undesired harmonics instead of adding desired harmonics. See ELECTRIC FILTER; FUNCTION GENERATOR; WAVE-SHAPING CIRCUITS.

The subtractive synthesis method can produce a wide variety of tonal timbres using only a small number of signal generators, and it has less switching complexity than harmonic synthesis. However, it does require filters, and the filters are generally more effective on groups of harmonics than on individual harmonics. In order to improve further the variety of available timbres in formant-type instruments, the filter input signals may be derived from generators having more than one basic waveform (for example, square, sawtooth, or pulse waves) used either individually or in additive (or subtractive) combinations.

Subtractive synthesis was used extensively in the first analog synthesizers, and it has been the basis for most voice synthesizers for over half a century. It is also used in the CELP voice coding standard and in musical synthesizers. Several digital synthesizers are devoted to imitating the original analog synthesizers, using virtual analog technology.

FM synthesis. The first all-digital synthesizer was based on frequency-modulation (FM) synthesis.

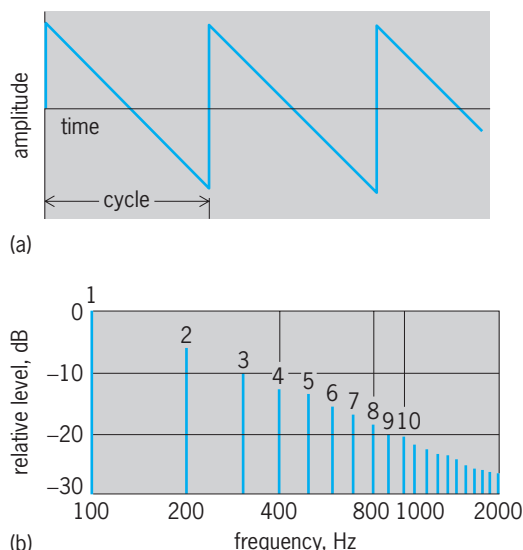


Fig. 4. Sawtooth wave. (a) Amplitude as function of time. (b) Harmonic spectrum analysis. Numbers above bars indicate harmonics. (After H. F. Olson, *Music, Physics and Engineering*, Dover, 1967)

Simple FM is carried out using two digital oscillators, with the output of one adding to the frequency (or phase) control of the other. In additive synthesis, two oscillators can provide only two harmonics. In FM synthesis, two oscillators can provide any number of harmonics. It is even possible for the output of a digital oscillator to modulate its own phase, giving so-called feedback FM. In the first all-digital synthesizer, several oscillators, called operators, could be combined in various ways, such as each oscillator modulating the next, or several FM pairs in parallel. A major advantage of FM is the synthesis of a wide variety of complex sounds from an extremely simple digital algorithm, thus making an all-digital synthesizer commercially feasible. Disadvantages of FM are its extreme sensitivity to its control parameters, and the fact that many parameter settings sound artificial. FM synthesis remains a valuable complexity-reduction technique in the context of additive synthesis. See FREQUENCY MODULATION.

Sampling synthesis. As memory costs came down in the late 1980s, sampling synthesis became more practical. Sampling synthesis can refer to any synthesis method based on playing back digitally recorded sounds. Usually, a sound is recorded at two dynamic levels and at selected pitches, such as every 3–6 semitones or even every octave. The velocity of the keystroke used to play a note is typically used to select a mixture of the two recorded dynamic levels, with a very slow stroke giving the softer recording and a very fast stroke giving the louder one. The duration is altered on playback, when appropriate, by looping playback somewhere in the middle of the recording. The nearest available pitch is altered to the desired pitch on playback either by adjusting the digital-to-analog conversion clock rate, or more commonly by resampling the stored waveform to change its pitch given a fixed sampling rate. In one type of sample-based synthesizer, a custom integrated circuit is used to provide many channels (such as 32 or 64) of simultaneous, high-quality, sampling-rate conversion.

Computer music composers manipulated digitally recorded sound for decades, but they used mainframe computers not available to most musicians. At first, sampling synthesis was available only in very expensive digital music systems, but later less expensive sampling synthesizers appeared, and finally highly affordable sampling synthesizers brought the technique into the mainstream. While all the early sampling synthesizers supported digital audio recording for creating sounds, most advanced synthesizers use samples stored in read-only memory (ROM), and there is no recording ability. Thus, while the synthesis technique is the same, it is better termed sample playback synthesis. It is also widely known as wavetable synthesis.

Today, ROM-based wavetable synthesizers dominate the field, from single-chip devices used in notebook computers to the most expensive synthesizers. With enough digital recordings of an instrument, it is possible to achieve any sound with sampling synthesis. However, approaching full expressive fidelity

for most instruments, particularly bowed strings and solo woodwinds, requires prohibitive amounts of memory, system development, and control complexity.

Physical modeling synthesis. A more recent approach to the synthesis of acoustic musical instruments, also made possible by digital technology, is known as physical modeling synthesis. In this approach, the algorithms are based directly on the mathematical physics of the instrument. One variant of physical modeling synthesis, which is especially effective for synthesizing string and wind instruments, is waveguide synthesis. Waveguide synthesis explicitly simulates traveling waves on a string or inside a bore or horn using digital delay lines. The theory is analogous to that of sampled (discrete-time) electric transmission lines. Filters are used in conjunction with digital waveguides to simulate losses due to bridge motion, horn radiation, air absorption, and the like.

The first commercial synthesizers based on waveguide synthesis appeared in 1994. Several companies now make products using this approach, including some implemented purely in personal-computer software. One model of electronic organ uses a combination of sampling synthesis, FM synthesis, and a variant of physical modeling called virtual lead (VL) technology, based on waveguide synthesis.

Simulation of inharmonic partials, noise, and modulation. Musical timbre involves more than the number and relative amplitude of harmonics. The parts (called partials) of a complex percussive tone may be slightly inharmonic (for example, with frequencies of 100, 200.3, 300.7 Hz, and so forth) as in a piano tone, or grossly inharmonic (for example, with frequencies of 100, 165, 210 Hz, and so forth) as in a bell tone. Successful electronic simulation of these tones may require additional frequency generators.

When the number of significant freely ringing partials is small, such as for the marimba, good results are obtained using a variant of subtractive synthesis called modal synthesis. A filter in subtractive synthesis normally provides a formant, that is, a gain boost or cut which affects several successive partials. In modal synthesis, filters are used to generate the partials themselves. The filters are typically organized into a parallel bank of resonators. Each resonator is tuned to an important, audible mode of the main vibrating element (such as the wooden bar in the case of the marimba), and it is excited by a short sound such as a noise burst (corresponding to the strike of the mallet). Modal synthesis is closely related to physical modeling synthesis since the filters are set according to physical parameters of vibration. However, it is usually better considered a parametric (filter-based) model of the time-varying spectrum. In some variants, the mode-simulating filters are excited according to an analysis of the physical excitation geometry; and in this case, modal synthesis behaves as an instance of physical modeling synthesis.

Both percussion tones and continuous tones typically contain small amounts of noise and tone-wave modulation resulting from the tone generation process (for example, blowing, bowing, or striking).

Fully authentic electronic tonal simulation requires that the electronic circuits (or algorithms) generate similar components (for example, broadly or sharply tuned noise) or produce similar modulations, either random or periodic, such as vibrato. Physical modeling synthesis typically obtains the best results along these lines, since the actual physical mechanisms causing noise and modulation are simulated explicitly. Noise due to air turbulence, however, is normally simulated using a digital pseudorandom number generator, filtered to provide the desired spectral content for the noise. See ACOUSTIC NOISE; MODULATION.

Envelope control. In conventional percussion tones the amplitude also changes during the course of the tone, rising suddenly at tonal onset and falling during tonal decay. Equivalent electronic tone waves thus require amplitude change, called envelope control. The different partials of a complex percussive tone would ideally have different envelope control. In modal and physical modeling synthesis, the proper amplitude envelopes are provided automatically by the filters used, since they resonate and decay in the same way that the physical instrument does.

The starting of a tone often includes a sound that lasts only for an instant. It may consist of one or more discrete frequencies, or it may be a puff of noise or a click. In pipe organs it is called chuff. In percussion instruments it is the strike. Authentic electronic simulation also requires these components to be under separate envelope control, or else a proper physical model is used as a basis.

Musical Instrument Digital Interface (MIDI). Since the 1980s, the Musical Instrument Digital Interface has emerged as the standard means for automatic control of electronic musical instruments. For example, the keyboard of one instrument can be used to control any number of additional synthesizers connected via MIDI cables. Electronic instruments typically have two or three MIDI connectors: one for MIDI input, one for MIDI output, and sometimes a MIDI "thru" connector which simply copies the MIDI input. A computer can also "play" one or more synthesizers via MIDI, and keyboard performances can be conveniently recorded to the computer. There are dozens of software products available for capturing, processing, and generating MIDI control information. The MIDI protocol has even become standard for interconnecting programs from different manufacturers running on the same computer. Many composers use a sequencer program for managing several MIDI tracks in a manner analogous to a multitrack recording device. Such programs support recording, playback, display, and editing of MIDI data; they provide multiple views such as common music notation (notes on a staff), a "piano roll" view, or simply lists of raw MIDI events, all of which can be conveniently edited.

Audio tracks and MIDI tracks are often manipulated together in a unified manner using a single sequencer program running on a personal computer. Digital audio data are far more demanding than MIDI data, since the former typically require 44,000 16-bit

samples per second, while three 8-bit bytes of MIDI data (a single MIDI message) can specify an entire note lasting several seconds on a synthesizer. However, standard personal computers have become powerful enough to handle several channels of digital audio data with ease. Moreover, personal computers are powerful enough to render MIDI data in real time as well, using software synthesis algorithms.

A digital recording from a sequencer program to a compact disk ideally involves mixing the digital audio tracks down to a stereo digital audio stream, and mixing in stereo digital audio streams from each MIDI synthesizer being controlled by the MIDI tracks. Creating a purely digital recording in this manner can be quite expensive if it requires a rack of MIDI synthesizers with digital outputs and a digital mixer. (One reason why digital mixers are expensive is that they need to be able to handle multiple sampling rates, and sampling rates from independent sources which are not synchronized.) If pure software synthesis is used for the MIDI tracks, a digital CD-ROM (compact-disk read-only memory) image file can be created on the computer's hard disk (not necessarily in real time) without any special hardware (aside from what is needed to record digital audio). Thus, a single personal computer can function as virtual all-digital recording studio environment, complete with virtual musical instruments and digital audio effects. See COMPACT DISK; COMPUTER STORAGE TECHNOLOGY; MICROCOMPUTER.

Daniel W. Martin; Julius O. Smith III

Bibliography. A. H. Benade, *Fundamentals of Musical Acoustics*, 2d ed., Oxford University Press, New York, 1976, reprint, Dover Publications, 1990; M. J. Crocker (ed.), *Encyclopedia of Acoustics*, vol. 4, sec. 130-139, pp. 1617-1695, John Wiley, New York, 1997; J. M. Eargle, *Music, Sound, and Technology*, 3d ed., Van Nostrand Reinhold, New York, 1996; N. H. Fletcher and T. D. Rossing, *The Physics of Musical Instruments*, 2d ed., Springer-Verlag, New York, 1998; P. Manning, *Electronic and Computer Music*, 2d ed., Oxford University Press, 1995; C. Roads, *Computer Music Tutorial*, MIT Press, 1996.

Musk-ox

An even-toed ungulate, *Ovibos moschatus*, which is a member of the family Bovidae in the mammalian order Artiodactyla. This single species is the northernmost representative of the family, ranging through the tundra areas and snowfields of Canada and Alaska, as well as Greenland.

The musk-ox derives its name from the musky odor it emits. It is a stoutly built animal, stands about 4 ft (1.2 m) high at the shoulder, and may weigh up to 700 lb (320 kg; see *illus.*). It has a coat of long dense hair that is resistant to the extreme cold of the windswept treeless tundra. The large, widely splayed feet are of great advantage to the animal when traveling over snow. These ruminants are herbivores; their food is principally willows during the short summer growing season, and in the winter, mosses and



Musk-ox (*Ovibos moschatus*). (Brent Huffman/Ultimate Ungulate Images)

lichens which they uncover from the snow by using their horns and hooves. They do not hibernate and are usually found in herds of 20–100 animals huddled together for warmth. They have 32 teeth and the dental formula is I 0/3 C 0/1 Pm 3/3 M 3/3. See DENTITION.

As protection against its natural enemy, the wolf, the musk-ox will form a circle with the young inside. It is hunted for its flesh, fur, and skins by both the Indians and Eskimos. Since a cow produces a single calf every 2 years, the numbers have been so reduced that they are now protected by the Canadian government. Projects to domesticate musk-oxen for wool production have been underway for a number of years at the University of Alaska and elsewhere. Wool produced from the soft undercoat is far superior to the best obtained from cashmere goats. See ARTIODACTYLA. Charles B. Curtin

Muskeg

A term derived from Chippewyan Indian for “grassy bog.” In North America ecologists apply diverse usage, but most include peat bogs or tussock meadows, with variable woody vegetation such as spruce (*Picea* mostly *mariana*) or tamarac (*Larix laricina*). The **table** covers a wide variety of organic terrain that is constituted of peat and covering vegetation, which contributes to the underlying peat. Plant remains accumulate when trapped in the water or in media such as sphagnum moss which inhibit decay. The peat might accrue indefinitely or might approach a steady state of raised bogs, blanket bogs, or forest if input became balanced by erosion or by loss as methane and carbon dioxide. See PEAT.

Classification. The communities of species making up the covering vegetation of the muskeg suggest kinds or orders of biological organization not treated here. The organization of the table is based on the physical structure observable by nonbotanists, de-

finer and symbolized by class letters for purposes of practical classification primarily for engineers. See TAIGA; TUNDRA.

Application of this system led to the identification of 17 major categories of peat structure, distinguished on the basis of relative amounts of granular and fibrous content. The texture of the peat is partly a function of its degree of woodiness, but it is also partly due to arrangement of the fibers. Texture, considered in relation to the entire fossilized mass, indicates the kind and order of characteristic homogeneity of the peat.

Vegetation cover formulas. The class symbols indicating structure in the cover may be combined into formulas. Thus cover formula AFI denotes association of trees more than 15 ft (4.6 m) high with a combination of grasslike plants and moss. The latter two elements of cover differ from the former and from each other as to stature and in being nonwoody. Also, the moss is of a velvety smooth texture in contrast, say, to a lichenaceous cover which is leathery in texture. The latter is symbolized in the reference system as class H and sometimes replaces I in the formula. Class H seldom associates with class F, and thus the formula AFH does not occur in nature. Class H is typically associated with E. Cover formulas AEH and AHE commonly characterize plant community organization in the Hudson Bay Lowlands. Class H usually signifies existence of ice in the underlying peat, indicating permafrost at these latitudes.

Qualitative expression. The position of a class symbol in a formula is important for a composite value of cover class and class predominance for a chosen area of muskeg. After the boundaries for the area are determined, the class offering most cover in the entire area is symbolized in the extreme left of the formula. Vegetation of short stature, if it covers more than that of higher, will be represented at the left in the formula only if there is no overlap of classes. Thus predominance of cover, but not necessarily stature, is progressively less from left to right in the formula.

| Properties designating nine pure coverage classes | | | | |
|---|-------------------------------|--------------------------|--------------------------|--|
| Coverage type (class) | Woodiness versus nonwoodiness | Stature (approx. height) | Texture (where required) | Growth habit |
| A | Woody | 15 ft (4.6 m) or over | — | Tree form |
| B | Woody | 5–15 ft (1.5–4.6 m) | — | Young or dwarfed tree or bush |
| C | Nonwoody | 2–5 ft (0.6–1.5 m) | — | Tall and grasslike |
| D | Woody | 2–5 ft (0.6–1.5 m) | — | Tall shrub or very dwarfed tree |
| E | Woody | Up to 2 ft (0.6 m) | — | Low shrub |
| F | Nonwoody | Up to 2 ft (0.6 m) | — | Mats and clumps or patches, sometimes touching |
| G | Nonwoody | Up to 2 ft (0.6 m) | — | Singly or loose association |
| H | Nonwoody | Up to 4 in. (10 cm) | Leathery to crisp | Mostly continuous mats |
| I | Nonwoody | Up to 4 in. (10 cm) | Soft or velvety | Often continuous mats, sometimes in hummocks |

Quantitative expression. Quantitative determination of class coverage is difficult to achieve accurately. Usually qualitative designation based on estimates made by inspection is sufficient for broad terrain analysis. In either case, if field representation of classes is less than 25%, the classes are not regarded as significant and are not symbolized in the formula. However, strips of small area (thin peat over sand or rocky ridges) could have importance for transport, or be impediments (for example, floating peat mats over open water).

Peat. Species of a given culminating community do not contribute equally to peat in the fossilizing process. Nevertheless, structural comparison between the peat and the culminating cover suggests formulas that indirectly symbolize peat structure and texture.

Aerial interpretation. It is possible to interpret subsurface structural constitution from aerial photographs through examination of patterns that the plant communities make in the living layer of vegetation over the peat. When trees occur in the pattern, they impart a stippled appearance and the pattern is called a Stipploid Airform Pattern. The peat beneath Stipploid is coarse and woody-fibrous and contains large chunks or logs (erratics) which on occasion may be as big as fence posts. See AERIAL PHOTOGRAPH.

From altitudes of 30,000 ft (9 km) it is possible to identify five different airform patterns which recur over wide expanses of muskeg. By the interpretive method, not only the peat categories but also the presence of microtopographic and water dispersement features can be determined.

Distribution. Organic terrain occurs on every continent. The largest expanses of it are in Russia (especially western Siberia) and Canada (500,000 mi² or 1,300,000 km²). Typical bog or spruce-larch muskeg develops in cool temperature and subarctic to arctic lands. In subtropics and the tropics there is considerable organic terrain, as in Paraguay, Uruguay, and Guyana in South America. Thus, despite differences in climate or floristics, the phenomenon of peat formation persists if local aeration or biochemical conditions hinder decay. Despite climatic and biotic differences, gross peat structure categories seem comparable the world over. Engineering prediction for operations on muskeg is thus possible, and design for operations can be standardized. See BOG.

Norman W. Radforth

Bibliography. H. Godwin, *Fenland: Its Ancient Past and Uncertain Future*, 1978; H. Godwin, *The History of the British Flora*, 2d ed., 1975; I. C. MacFarlane, (ed.), *The Muskeg Engineering Handbook*, 1969; P. D. Moore and D. J. Bellamy, *Peatlands*, 1973.

Muskmelon

The edible fruit of *Cucumis melo*, belonging to the gourd family, Cucurbitaceae, as do other vine crops such as cucumber, watermelons, pumpkin, and squash. The muskmelon appears to be indigenous to Africa. There are secondary centers of origin in India, Persia, southern Russia, and China. The muskmelon was a latecomer to the list of domesticated crops. It then exploded into numerous cultivars which were rapidly dispersed throughout Europe and, at an early date, into the Americas. American cultivars encompass the netted, salmon-fleshed cantaloupes; the smooth-skinned, green-fleshed Honey Dew; the green-skinned, bright-orange-fleshed Persian; the delicate-flavored, light-salmon-fleshed Crenshaw; and the wrinkle-skinned, white-fleshed Golden Beauty casaba. Other forms with very different plant and fruit characteristics are used in the Orient for pickling and in India for cooking. All these melons differ only in varietal characters and all intercross freely. See CANTALOUPE; HONEY DEW MELON; PERSIAN MELON; VIOLALES.

The plants are annual, trailing vines, with three to five runners that may attain a length of 10 to 12 ft (3 to 3.6 m). The branching vines have coarse, somewhat heartshaped leaves, with almost entire slight-angled, rounded, wavy margins. The runners produce short fruiting branches, which bear the perfect flowers and later the fruits. Most American varieties are andromonoecious, bearing male and perfect flowers (combined male and female). The pollen is heavy and slightly sticky, and therefore insects are required for fertilization. The domestic honeybee is the only known effective pollinator of muskmelon flowers.

Muskmelons maturing on the vine without becoming overripe are superior in quality to those harvested immature. The sugar content, flavor, and texture of the fresh fruit improves very rapidly as the fruit approaches maturity. When mature, the melon is sweet, averages 6 to 8% sugar, and has a slight to distinctly musky odor and flavor, depending upon

cultivar and environment. The flesh is rich in potassium, in vitamin C and, when deep orange, also in vitamin A.

Cultivation and culture. Muskmelon plants during all stages of development are easily killed by frost. They require fairly warm weather and are favored by bright sunshine, low humidity, and absence of rain, which tends to prevent certain diseases that often defoliate the plants in humid areas.

Muskmelons can be grown on several types of soil but not on muck. Peat, heavy clay, or adobe soil are not recommended. The soil should be fairly fertile and free from nematodes, wilt disease, and toxic amounts of alkali. Muskmelons are sensitive to acid soils. They thrive best on neutral or slightly alkaline soil.

Light-textured soils that warm up quickly in the spring favor early maturity. Maturity is also hastened by using glassine paper hot caps to increase soil and air temperature around the plants. In some locations the frost-free period may be too short to grow muskmelons from seed planted directly in the field. Muskmelons can be grown at such locations if the plants are started in greenhouses or hotbeds and transplanted to the field when the danger of frost has passed. Muskmelons are generally grown in rows 5–7 ft (1.5–2.1 m) apart, with a single plant 1 ft (0.3 m) apart in the row, or in hills of two or three plants 2 ft (0.6 m) apart. Under irrigation, the rows are on wide ridges with furrows for irrigation water between them. In the arid West, 2–3 acre-feet (2470–3700 m³) of water per acre is commonly required to grow a crop of muskmelons.

Harvesting. The time for planting to harvest is 85–125 days, depending upon variety and weather. Cantaloupes are harvested at the “full-slip” stage of maturity. At this stage a thin abscission crack encircles the stem where it is attached to the fruit, and the melon separates easily from the stem. Harvest maturity of Persian melons is determined primarily by skin color changes, because in this variety the abscission layer does not develop or is delayed until the fruit is over-ripe. The ground spot (area resting on soil) of Persian melons develops a pinkish color when the fruits are mature. Honey Dew melons have achieved harvest maturity when the skin color is white and no waxy skin coating is evident. The surface may feel prickly or hairy. With the casaba varieties, Golden Beauty, Crenshaw, and Santa Claus, maturity can be determined by applying firm pressure with the thumb to the blossom end. A slight yielding or softness indicates maturity.

California, Arizona, and Texas account for approximately 80% of the acreage and 85% of the production of muskmelons in the United States. Frank W. Zink

Muskrat

A large aquatic rodent, *Ondatra zibethicus*, in the family Cricetidae and in the vole subfamily (Arvicolinae). It is known as the common muskrat (see **illustration**). It is dark reddish-brown, with a long,

essentially naked, scaly, laterally compressed tail. The pelage is dense and shiny, and consists of a thick coat of underfur over which lies a covering of long, glossy guard hairs. The ears and eyes are relatively small for an animal of this size. The hindfeet bear webbed toes and are much larger than the front feet. The anterior faces of the upper incisors are yellowish-orange. The body length ranges about 14–16 in. (400–640 mm), with the tail about 7–12 in. (180–300 mm). The average weights of muskrats from Indiana are 1215 grams (2.67 lb) in adult males and 1247 grams (2.74 lb) in adult females, which is near average. The dental formula is I 1/1 C 0/0 Pm 0/0 M 3/3 = 16. The muskrat received its name because of its inguinal glands, which produce a musky odor. See DENTITION; MAMMALIA; RODENTIA; SCENT GLAND.

Habitat. The muskrat is found only in North America, but is common over much of the United States and Canada where suitable habitat is present. Marshes and other wet areas with an abundance of emergent vegetation, especially cattails, are the preferred habitat of muskrats, but the species also occurs along streams and ditches and about lakes and ponds. The largest populations are usually found in larger marshes. Draining of marshes and other wetlands has undoubtedly caused great decreases in muskrats. Muskrats will even take up residence in small ponds that may later dry up.

Houses. The muskrat is primarily nocturnal but is often active by day, especially in spring and fall. It spends most of the daytime in burrows or houses that it constructs in or near water. The houses are the most conspicuous indicators of the muskrat's presence on a marsh. Most muskrat houses are in water 1–2 ft (0.3–0.6 m) deep and are constructed of emergent and submerged vegetation, cattail being a favorite when it is available. The animals usually clear the emergent vegetation (including roots) from the immediate vicinity of the house, and much of this material goes into house construction. Cattails, burreed, bulrushes, rice cutgrass, watershield, wool grass, smartweed, hornwort, and other plants are used in building; considerable mud is sometimes included. Houses are roughly circular, average 4–5 ft (1.2–1.5 m) in diameter at the water level, and extend 2–3 ft (0.6–0.9 m) above the water surface. They have walls about 1–1.5 ft (0.3–0.5 m) thick. Inside



Common muskrat, *Ondatra zibethicus*. (Photo by Dr. Lloyd Glenn Ingles; © California Academy of Sciences)

the house is an irregularly shaped chamber above the water level. An underwater entrance leads to one side of this chamber; the entrance also serves as a plunge hole into which the muskrat dives when danger threatens. The chambers are about a foot in diameter, 6–8 in. (15–20 cm) high, and include nests of dry, shredded cattail leaves and other soft vegetation. Larger houses sometimes have two chambers. House sites are often on a stump, in or on brush, about a tree in the water, on or about some other object in the water, or along a fence. Where no such foundation is present, the muskrats cut vegetation and make a raft, which sinks to the bottom and serves as an anchor for the house. There are two major peaks of house construction—in spring and in late summer or early fall. Houses do not persist for more than a season unless they are well anchored to some durable object and constantly repaired. Rain, wind, and waves quickly destroy them, and the material from which they are made rots rapidly. In the spring, some new houses are built and others are repaired for rearing of the young. Depending on season, from one to ten animals may occupy a house, but there are usually four to five. This number can construct a house in one or two nights. In the spring, during breeding season, only one pair is found per house.

In deep water, bank burrows are used, and they are usually constructed in vertical banks. Burrows usually extend 15–20 ft (4.6–6 m).

Activity. Muskrats tend to enter and leave the water repeatedly at certain places, forming obvious “slides” in the soil along the water. In muddy areas, the footprints and the drag mark made by the tail between the foot tracks can be seen. Tagging studies have shown that many muskrats spend their entire lives within a radius of 200 yards (183 m) of their birthplace. In late summer and fall, however, individuals wander, and some have been observed more than a mile from water. During these overland travels into unfamiliar territory, muskrats are more vulnerable to predation and accidents, and many are killed by automobiles.

Muskrats remain active under the ice in winter, taking advantage of air trapped between the ice and the water. They enter and leave the water where the current prevents freezing, where the ice is broken, or where spaces occur in ice enclosing a tree or other object.

Food. Muskrats are primarily vegetarian but will eat animal matter, even carrion, under certain conditions. Probably, most emergent plants are eaten, and other common foods include cattails, burreed, bulrush, water lilies, pondweeds, smartweeds, duck potato, water plantain, rice cutgrass, hornwort, sedges, grasses, swamp loosestrife, buttonbush, and woolgrass.

Animal foods include winter-killed fishes, frogs, crayfishes, mussels, and other dead animals. Animal material is usually taken in greatest abundance in late fall, winter, and early spring, when succulent plants are not readily available.

Breeding and development. Pair formation is apparently initiated mostly by the female, which swims

back and forth before the male and utters squeaking notes. Courtship begins in February, and copulation has been observed as early as the second week of March. Gestation takes about 28–30 days. Most young are probably born in May and June, but there are records of gravid females in late fall and even in early winter. Litter size ranges from 3 to 10, averaging about 7.

Young muskrats at birth are blind, nearly naked, and helpless. They have round tails, which become laterally flattened at about the time of weaning and reach their full shapes at about 3 months of age. The young are reared in dry, warm nests of soft vegetation in the chambers of muskrat houses or burrows. The average weight at birth is about 22 grams (0.05 lb). The eyes open at about 2 weeks, when the young are covered with dark, lead-gray fur; this pelage is replaced by adult-colored hair at about 3 months of age. The young swim at about 3 weeks but have difficulty diving. At 4 weeks, when they average about 200 grams (0.45 lb), they are weaned and on their own. If the female produces a second litter, the young are driven from the house or burrow and must find new living quarters. Muskrats usually have two or three litters per season. The young from the last litter may stay with the parents until the next breeding season. The young grow rapidly and, in good habitat, equal the size of the adults when 4–5 months old. Weights of about 1.37 kg (3 lb) can be attained in 8 months. Reproductive success depends upon habitat conditions and population size. Overpopulation leads to smaller litters and fewer litters per female, because of stress and strife brought on by more frequent encounters between individuals. Some animals have been known to be alive in their fourth year, but only 10–15 percent reach an age of one year.

Threats. Minks can be important predators on muskrats. They sometimes enter muskrat houses by making a hole through the side into the central cavity. Raccoons, dogs, and red foxes prey on them, and dogs are especially destructive in late summer when water levels recede so that the houses are more accessible. Heavy rains, winds, and waves may destroy houses and drown the young. Highway kills may be numerous, especially where heavily traveled roads border good muskrat habitat. Raccoons obtain mostly young muskrats, which they take from their nests after tearing holes in the sides of muskrat houses. Drainage, however, remains the major threat to the species. Severe drought may dry up certain habitats, and the animals may be forced to abandon an area completely. Shallow-water areas that freeze to the bottom cause problems, since muskrats generally do not store food for winter. They may find it difficult or impossible to dig for roots and other foods when ice interferes, and may starve if ice persists for long periods.

Neofiber alleni. The common muskrat does not occur in Florida. However, in the grassy marshes or “prairies” (level, almost treeless bogs or marshes) of Florida and in the Okefenokee Swamp area of southeast Georgia, there is a much smaller

rodent, the round-tailed or Florida muskrat, *Neofiber alleni*. It is suggestive of a small muskrat in general appearance, and it has been characterized as a link between the meadow vole and the muskrat. The ectoparasites (parasites that live on the host's exterior) of these two animals are very different, suggesting that *Ondatra* and *Neofiber* have long been separated. A very young round-tailed muskrat in the grass looks much like a meadow vole. The round-tailed muskrat has a dense, soft, waterproof coat with very thick underfur and dark guard hairs. The small ears are nearly hidden in the fur. The scaly, round tail is almost devoid of hair, and the hindfeet are slightly webbed. The upper parts are generally a dark, rich, uniform brown. Juveniles are lead-gray. The hindfoot is much larger than the front foot. The skull is similar to that of *Ondatra* but smaller. The tooth formula is the same as that of the common muskrat.

Neofiber is much less aquatic than *Ondatra*, but it is an excellent swimmer and takes to the water readily, swimming and diving with ease. Sometimes, these animals live in burrows rather than houses. *Neofiber* is active primarily at night.

Neofiber breeds throughout the year. Gestation takes 26–29 days. Four to six litters of one to four are produced per year.

Dogs, cats, water moccasins, harriers (marsh hawks), and barn owls are important predators on this species.

Myocastor coypus. The nutria (family Myocastoridae) is a South American aquatic rodent but is included here because it is very similar and occupies a similar niche to the North American muskrat, *Ondatra zibethicus*. In addition, it has been introduced into the eastern United States where it competes with the muskrat for habitat. The tail is long, but is round rather than being laterally compressed as in the muskrat. As with the muskrat, eyes and ears are small, and the hindfeet are webbed and much larger than the forefeet. There are four mammae, which are far up on the sides of the chest. The nutria is strictly vegetarian.

John O. Whitaker

Bibliography. D. Birkenholz, A study of the life history and ecology of the round-tailed muskrat (*Neofiber alleni* True) in north central Florida, *Ecol. Monogr.*, 33:187–213, 1963; D. L. Bounds, M. H. Sherfy, and T. A. Mollett, Nutria *Myocastor coypus*, pp. 1119–1147 in *Wild Mammals of North America: Biology, Management, and Conservation*, ed. by G. A. Feldhamer, B. C. Thompson, and J. A. Chapman, Johns Hopkins University Press, Baltimore, 2003; J. Erb and H. R. Perry, Jr., Muskrats *Ondatra zibethicus* and *Neofiber alleni*, pp. 311–348 in *Wild Mammals of North America: Biology, Management, and Conservation*, ed. by G. A. Feldhamer, B. C. Thompson, and J. A. Chapman, Johns Hopkins University Press, Baltimore, 2003; R. M. Nowak, *Walker's Mammals of the World*, Johns Hopkins University Press, Baltimore, 1999; J. O. Whitaker, Jr. and W. J. Hamilton, Jr., *Mammals of the Eastern United States*, Cornell University Press, Ithaca, NY, 1998; D. E. Wilson and S. Ruff, *The Smithsonian Book of North American Mammals*, 1999.

Mustard

Any one of a number of annual crucifer species of Asiatic origin belonging to the plant order Capparales. Mustards eaten as greens are *Brassica juncea*, *B. juncea* var. *crispifolia*, and *B. birta*. Table mustard and oils are obtained from *B. nigra*. Cultural practices are similar to those used for spinach. Southern Giant Curled and Ostrich Plume are popular varieties for greens. Long days and high temperatures favor undesirable seed-stalk development. Harvesting is usually 1½–2 months after planting. Important production centers for mustard greens are in the South, where the crop is popular. Montana and the West Coast states are important sources of mustard seed. See CAPPARALES; SPINACH. H. John Carew

Mutagens and carcinogens

A mutagen is a substance or agent that induces heritable change in cells or organisms. A carcinogen is a substance that induces unregulated growth processes in cells or tissues of multicellular animals, leading to the disease called cancer. Although mutagen and carcinogen are not synonymous terms, the ability of a substance to induce mutations and its ability to induce cancer are strongly correlated. Mutagenesis refers to processes that result in genetic change, and carcinogenesis (the processes of tumor development) may result from mutagenic events.

Mutagenesis and carcinogenesis. Ionizing radiation was the first agent that was recognized as having the capacity to induce mutations and also to induce cancer. Induction of mutations by ionizing radiation has been recognized to be a product of ionization paths within cells that, when they occur in deoxyribonucleic acid (DNA), cause the breaking of chemical bonds and the alteration of the nucleotide bases. Alterations in properties of cells that occur as a result of exposure to radiation of chemicals are called phenotypic changes; if these changes are the result of DNA alterations and are transmitted to progeny cells, they are mutations. When such mutations affect genes that regulate cellular growth, they may result in cells called transformants, which can give rise to tumors.

Radiation can induce alterations in any segment of DNA, but some of the most important effects occur in genes that are required for survival of cells (injury to one of these genes may kill the cells) or for other important processes such as regulation of cell functions (injury to one of these genes may allow cells to grow uncontrollably). Cancers are disturbances in the growth of cells so that they are no longer under the control of the body. As the altered cells grow, they produce additional cells like themselves, leading to the development of a tumor.

Unlike radiation, many chemicals must undergo a process of enzymatic interaction (metabolism) in cells or tissues. As the cellular enzymes interact with chemicals, they often generate closely related chemical structures that possess mutagenic potential. The

chemical basis for induced mutations was found to be the presence of electrophilic groups in the mutagenic chemicals. See MUTATION; RADIATION BIOLOGY.

Mutagens. A mutation is any change in a cell or in an organism that is transmitted to subsequent generations. Mutations can occur spontaneously or be induced by chemical or physical agents. The cause of mutations is usually some form of damage to DNA or chromosomes that results in some change that can be seen or measured. However, damage can occur in a segment of DNA that is a noncoding region (that is, no gene product is produced), and thus will not result in a mutation. Mutations may or may not be harmful, depending upon which function is affected. They may occur in either somatic or germ cells. Mutations that occur in germ cells may be transmitted to subsequent generations, whereas mutations in somatic cells are generally of consequence only to the affected individual.

The impact of induced germ cell mutations on humans has been difficult to determine. Germ cell mutations are the basis of inherited human disorders, but there is uncertainty about what types or frequency of mutations are induced by exposure to mutagens. Most mutagens cause specific chemical changes in the informational content of DNA. Among the specific changes are alterations of purine or pyrimidine bases in DNA by rearrangement, deletion, or insertion events. The consequences of such changes can be the loss or altered activities of gene products.

Not all heritable changes result from damage to DNA. For example, in growth and differentiation of normal cells, major changes in gene expression occur and are transmitted to progeny cells through changes in the signals that control genes that are transcribed into ribonucleic acid (RNA). The mechanisms by which differentiation is controlled are not known. It is possible that chemicals and radiation alter these processes as well. When such an effect is seen in newborns, it is called teratogenic and results in birth defects that are not transmitted to the next generation. However, if the change is transmissible to progeny, it is a mutation, even though it might have arisen from an effect on the way in which the gene is expressed. Thus, chemicals can have somatic effects involving genes regulating cell growth that could lead to the development of cancer, without damaging DNA.

Carcinogens. Cancer arises because of the loss of growth control by dearrangement of regulatory signals. Included in the phenotypic consequences of mutations are alterations in gene regulation brought about by changes either in the regulatory region or in proteins involved with coordinated cellular functions. Altered proteins may exhibit novel interactions with target substrates and thereby lose the ability to provide a regulatory function for the cell or impose altered functions on associated molecules. Through such a complex series of molecular interactions, changes occur in the growth properties of normal cells leading to cancer cells that are not responsive

to normal regulatory controls and can eventually give rise to a visible neoplasm or tumor. While mutagens can give rise to neoplasms by a process similar to that described above, not all mutagens induce cancer and not all mutational events result in tumors. Many DNA changes may occur in genes that do not relate to cancer or do not affect gene function. Also, normal cells possess the ability to repair segments of DNA damaged by ionization, chemical adduction, or other events, thus erasing the damage from the DNA sequence. Alternatively, some chemicals that have the potential to form electrophiles may not be taken up by cells, may not be metabolized to an electrophilic intermediate, or may be rapidly detoxified and excreted from the organism.

However, there are carcinogens that do not give rise to electrophiles and are believed to induce tumors by mechanisms that do not involve direct DNA interactions and mutation. The mechanisms by which nonmutagenic carcinogens act are not well understood.

The identification of certain specific types of genes, termed oncogenes, that appear to be causally involved in the neoplastic process has helped to focus mechanistic studies on carcinogenesis. Oncogenes can be classified into a few functionally different groups, and specific mutations in some of the genes have been identified and are believed to be critical in tumorigenesis. Tumor suppressor genes or antioncogenes provide a normal regulatory function; by mutation or other events, the loss of the function of these genes may release cells from normal growth-control processes, allowing them to begin the neoplastic process. See ONCOGENES; ONCOLOGY.

Identifying mutagens and carcinogens. There are a number of methods and systems for identifying chemical mutagens. Because mutagenesis involves several steps and since there may be variation between organisms used for detecting mutagens, an agent may be mutagenic only in some systems. Mutations can be detected at a variety of genetic loci (genes) in very diverse organisms, including bacteria (*Salmonella typhimurium*, *Escherichia coli*), insects (*Drosophila melanogaster*), cultured mammalian cells, rodents, and humans. An important advantage in using cultured cells is that they can be used under conditions designed to specifically select for induced mutants. Spontaneous and induced mutations occur very infrequently, the estimated rate being less than 1 in 10,000 per gene per cell generation. This low mutation rate is probably the result of a combination of factors that include the relative inaccessibility of DNA to damaging agents and the ability of cellular processes to repair damage to DNA. By applying methods that can select for mutations occurring in specific genes, it is possible to use fewer cells to detect mutations, and to be able to generate enough mutants to study quantitative aspects of the mutation process.

Factors that contribute to the difficulty in recognizing substances that may be carcinogenic to humans include the prevalence of cancer, the diversity of types of cancer, the generally late-life onset

of most cancers, and the multifactorial nature of the disease process. These factors make the identification of causal factors in human cancer extraordinarily difficult and reduce the certainty of epidemiological methods that attempt to identify risk factors and causal associations. Adding to this uncertainty is the diversity of types and forms of cancers, which suggests different causes for what may be a group of diseases. Furthermore, like the problem of heritable mutations, it is difficult to discriminate between cancers that may be of spontaneous origin and those that may be induced by exogenous substances. With the exception of sunlight (UV), asbestos, and tobacco, the major causes of most human cancers are not known.

Approximately 50 substances have been identified as causes of cancer in humans, but they probably account for only a small portion of the disease incidence. See CANCER (MEDICINE); HUMAN GENETICS.

Raymond W. Tennant

Bibliography. J. Ashby and R. W. Tennant, Definitive relationship among chemical structure, carcinogenicity and mutagenicity for 301 chemicals tested by the U.S. NTP, *Mutat. Res.*, 257:229–306, 1991; J. M. Bishop, Molecular themes in oncogenesis, *Cell*, 64:235–248, 1991; L. M. Franks and N. M. Teich (eds.), *Introduction to the Cellular and Molecular Biology of Cancer*, 3d ed., 1997; R. W. Tennant and E. Zeiger, Genetic toxicology: Current status of methods of carcinogen identification, *Environ. Health Perspect.*, 100:307–315, 1993; H. Vainio et al., *Mechanisms of Carcinogenesis in Risk Identification*, 1992.

Mutation

An abrupt, heritable change in genes or chromosomes manifested by changes in the phenotype (the appearance) of an organism. It is theoretically preferable to define mutations as changes in deoxyribonucleic acid (DNA) sequences, but the classical definition remains the operational definition in most circumstances. Genetic segregation and recombination, however, are not mutational processes unless aberrant. See DEOXYRIBONUCLEIC ACID (DNA).

The word mutation has two common meanings, one being the process and the other the product (the altered gene or chromosome carries a mutation). The process is also called mutagenesis. An organism bearing a mutation is called a mutant. An agent that induces a mutation is called a mutagen.

The study of mutation has long occupied a central position in genetics. Mutations are the ultimate sources of variability upon which evolution acts, despite being random changes that are far more likely to harm than to improve a complicated and highly evolved organism. Laboratory reconstructions have shown that rapidly mutating microbial populations overtake slowly mutating populations when the two are mixed and placed in a new environment to which neither is fully adapted. Mutation has consistently been the most telling probe into the nature of the

gene, and understanding of most aspects of biology has benefited from studies of the properties of mutant organisms. Mutation is also an important component of disease, either causing it directly (for example, through birth defects) or predisposing humans to a vast array of disorders that together constitute a substantial fraction of illnesses. Finally, deliberate selection of mutant plants and animals for economic or esthetic purposes has long been practiced, and has grown into an important aspect of genetics. With advances in molecular genetics, it is now possible to construct specific mutations at will, rather than merely selecting among an array of random mutations for the infrequent useful ones. See BREEDING (ANIMAL); BREEDING (PLANT).

Anatomy. Although there are many sites where mutations can occur in even a single chromosome, and many mechanisms to generate them, the products of mutation can be simply cataloged in genetic terms.

Genome or ploidy mutations. The sum of a cell's genes is its genome, and ploidy refers in a general sense to the number of copies of each chromosome in a (nondividing) cell. A cell that has accidentally doubled its number of chromosomes from the normal diploid state (two copies of each chromosome) is called tetraploid; one that has lost a single member of a normal pair is called monosomic; and one that has gained a single member of a normal pair is called trisomic. In higher animals such changes are usually lethal or severely debilitating, the best-known human example being trisomy for chromosome 21 leading to Down syndrome (which occurs at a frequency of roughly 0.001 per birth). See CONGENITAL ANOMALIES; POLYPLOIDY.

Chromosome mutations. Chromosome mutations, which alter sections composed of many DNA base pairs, consist of partial losses (deletions or deficiencies), rearrangements, and additions. Like genome mutations, most deletions that remove many genes are highly deleterious. Rearrangements may be less deleterious if they shuffle genes about but do not interrupt them or relocate them to sites where they cannot function well. They involve either inversions (simple reversals of an internal segment of a chromosome) or translocations (transfer of a segment of a chromosome to a new location). Translocations can occur within or between chromosomes, the latter often reciprocally. Even when not directly deleterious, rearrangements lead to anomalies of genetic recombination; a common secondary consequence in humans is sterility. Addition mutations are of two types, duplications and insertions. Duplications usually consist of tandem repeats of a segment of a chromosome, and may range from innocuous to lethal depending upon their location and extent. Insertions occur through the movement of special DNA sequences (transposons) that range from hundreds to thousands of DNA base pairs in length. See CHROMOSOME; CHROMOSOME ABERRATION.

Gene mutations. Gene mutations affect only a single gene and consist of intragenic chromosomal mutations, additions or deletions of one or a few base

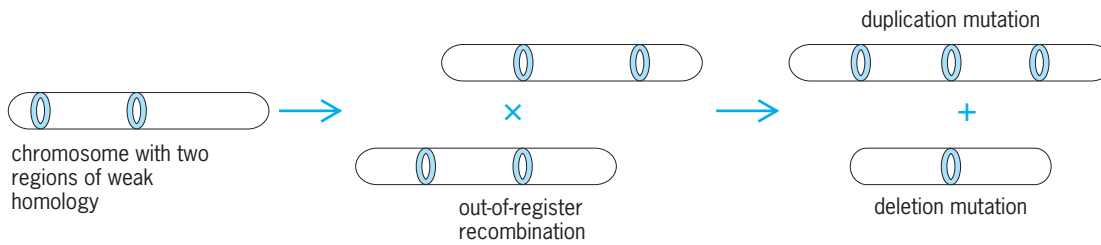


Fig. 1. Formation of duplication and deletion mutations by aberrant recombination.

pairs, base-pair substitutions (point mutations), and complex mutations comprising simultaneously arising clusters of any of the above. The severity of a gene mutation depends on its individual nature and on the importance of the affected gene, and can range from innocuous to lethal.

Base-pair substitutions are divided into two groups called transitions and transversions. In transitions, such as $A \cdot T \rightarrow G \cdot C$, the purine-pyrimidine orientation is maintained; in transversions, such as $A \cdot T \rightarrow C \cdot G$, this orientation is reversed. (Here the DNA bases are denoted by $A =$ adenine, $T =$ thymine, $G =$ guanine, and $C =$ cytosine, where A and G are purines and T and C are pyrimidines; the dots indicate hydrogen bonding between bases.)

Because the genetic code employs consecutive sets of three DNA base pairs to specify consecutive amino acids in proteins, the addition or deletion of multiples of three base pairs leads to the addition or deletion of one or more amino acids. However, the addition or deletion of one or two base pairs (or any nonmultiple of three) shifts the reading frame, so that everything from that point onward is read out of its normal frame, with drastic consequences for that particular gene. These occurrences are called frameshift mutations. See GENETIC CODE.

Forward and reverse mutation. Mutations from a normally functioning reference gene, chromosome, or organism to a mutant condition are called forward mutations. Their reversal by a new mutation that restores the original DNA sequence is called back mutation, or reversion. In addition, new mutations at a site distinct from a forward mutation can sometimes restore the nonmutant phenotype; these are called suppressor mutations. For instance, a base-pair addition in a protein-encoding sequence can sometimes be suppressed by the nearby deletion of a different base pair: the reading frame is restored, and the organism may no longer appear mutant if the associated amino acid changes are innocuous and the protein therefore functions normally.

Soma and germ line. In sexually reproducing multicellular animals, a mutation that arises in a somatic (body) cell cannot be passed to future generations, whereas a germ-line mutation can. Even though somatic mutations cannot harm future generations, they can be important to the individual that carries the mutant cell. They can be deadly (as by leading to cancer) or beneficial (as by generating new antibody molecules).

Scoring mutations. The rarity and sporadic nature of mutations render their study in wild organisms (that is, those in nature) very difficult. Instead, mutagenesis is studied in specialized laboratory organisms, among which microorganisms are favored because the ease of growing large populations (such as 10^9) overrides the low frequencies of mutants (such as 1 in 10^7 organisms). See BACTERIAL GENETICS; BACTERIAL GROWTH.

In the bacterial virus T4, the circular plaques formed by viral killing of a lawn of the bacterial host have a characteristic size and fuzzy edge. Visual screening readily detects *r* mutants that make larger, sharp-edged plaques. Among these, the *rII* mutants have been widely studied because their rare revertants (and also recombinants) can be selected by growing on special bacterial hosts resistant to the parental *rII* viruses. See BACTERIOPHAGE.

Among bacteria and yeasts, many biochemical traits have been used to score mutations. A frequent approach is first to obtain a mutant that requires for its growth a special nutrient, such as a vitamin or an amino acid. By growing the mutant population in a medium lacking the required growth factor, rare revertants to nutritional independence can be selected. For instance, the widely used Ames test for environmental mutagens uses mutants of the bacterium *Salmonella typhimurium* that require the amino acid histidine for growth, and the method tests for mutagens that can induce reversion.

In the fruit fly *Drosophila melanogaster*, special chromosomes are used to score induced lethal mutations on the X chromosome. Because these mutations are recessive (their effect being masked in the presence of a nonmutated chromosome) and because the X chromosome is a sex chromosome, this system screens for sex-linked recessive lethals. It is the most efficient mutation-scoring system in a higher eukaryote. See SEX-LINKED INHERITANCE.

Because of the difficulties of raising and examining mice by the tens of thousands under well-controlled conditions, mutation experiments are infrequently conducted with these or any other mammals. However, their chromosomal mutations can sometimes be scored by cytological (microscopical) analyses, and their gene mutations can be scored in a few systems, such as the specific-locus system in which mutations can be detected in any of seven specific genes determining, for the most part, coat-color traits. (A locus is the position of a gene on a chromosome.) The

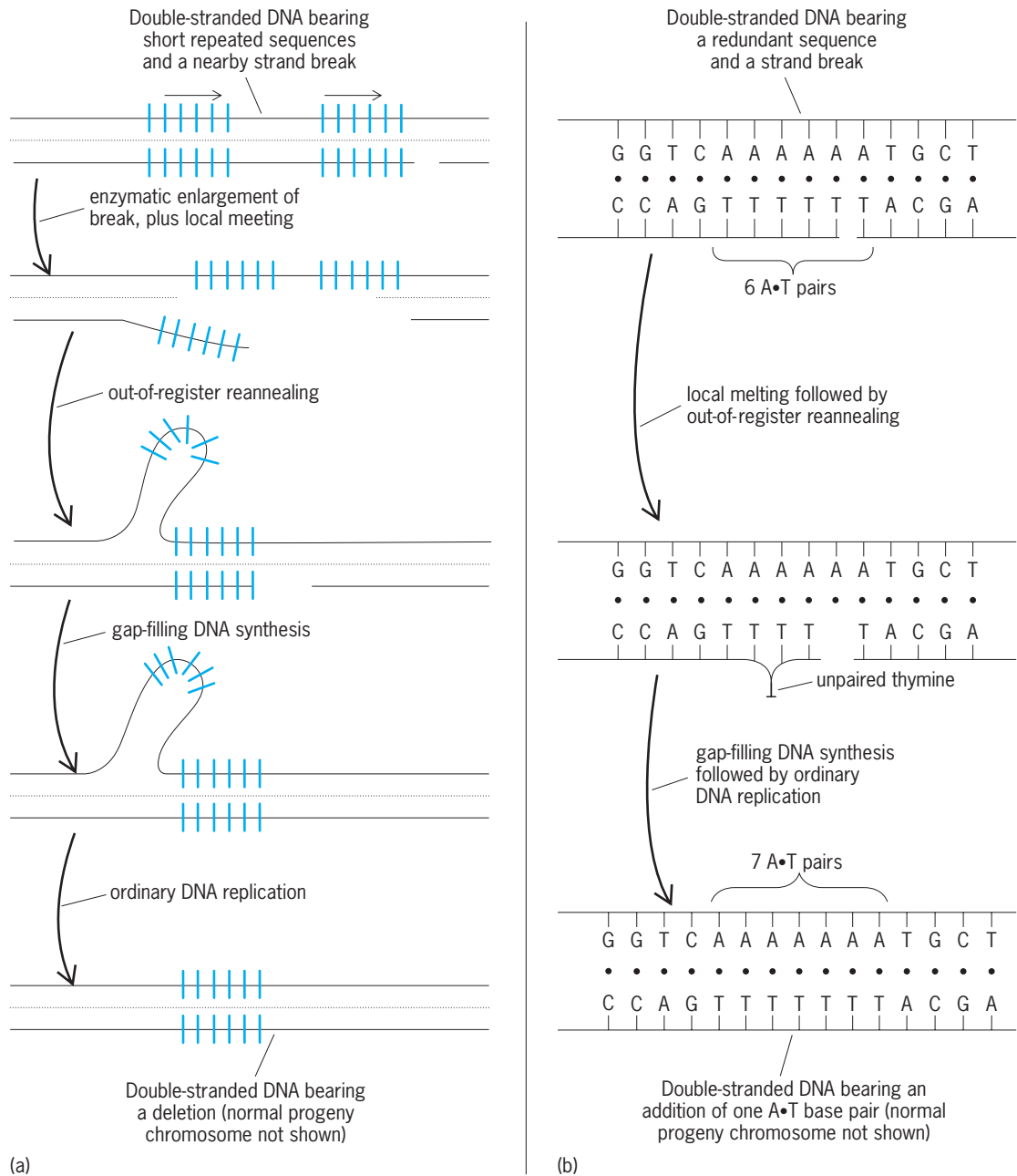


Fig. 2. Schemes for mutagenesis by misalignments within a single chromosome: **(a)** between distant repeated sequences; and **(b)** within a redundant sequence. Continuous lines represent DNA sugar-phosphate backbones, and dots indicate base pairing via hydrogen bonds.

systematic study of mutation in humans is difficult for obvious reasons, and has amounted to little more than recording incidences and examining patterns of inheritance.

Mechanisms. Although important mechanisms of mutagenesis undoubtedly remain to be discovered, many, and perhaps most, of the predominant mechanisms are now known, at least in outline.

Genome mutations. Mutations that alter the number of chromosomes in a cell usually result from the faulty distribution of chromosomes during mitotic or meiotic cell divisions. The fault probably often lies in the systems of spindle fibers that segregate daughter chromosomes into daughter cells; chemicals (such as colchicine) that interfere with such fibers

induce aneuploidy at high frequencies. See MEIOSIS; MITOSIS.

Chromosome mutations. The larger of these, and probably many of the smaller as well, can be formed by chromosome breakage followed by incorrect patterns of rejoining. Many agents, including ionizing radiations and numerous chemicals, can induce chromosome breaks. As might be expected from their topology, the frequency of chromosome mutations often corresponds to the square of the dose of mutagen, that is, as "two-hit" events (two breaks plus incorrect rejoining). The mechanism of efficient rejoining of broken chromosomes is not understood, but may involve base pairing between repeated DNA sequences in the chromosome. The frequency of

repeated sequences is high in the chromosomes of higher organisms, and chromosomal mutations, particularly deletions, are also frequent relative to point mutations in these organisms. In addition to events triggered by breaks, however, deletions and duplications are triggered by anomalies of genetic recombination between similar but nonhomologous DNA sequences. See RADIATION BIOLOGY; RECOMBINATION (GENETICS).

Insertion mutations. These usually arise, not following random chromosome breaks, but through the intrinsic mobility of highly specialized DNA sequences. Called transposons, they have been found to be a major factor in spontaneous mutagenesis, because their transposition into a gene is very likely to inactivate that gene. They seem to play at most a minor role in induced mutagenesis. Transposons come in several types and sizes. Many carry repeated DNA sequences at their ends. Their mobility is often engendered by one or more of their own genes. For example, a DNA copy may be produced and then inserted elsewhere by a specialized recombination mechanism. Alternatively, the transposon may be transcribed into ribonucleic acid (RNA), copied back into DNA by a reverse transcriptase, and then inserted. See TRANSPOSONS.

Misalignment mutagenesis. This is a set of mutagenic mechanisms that proceeds through correct DNA base pairing in an incorrect (misaligned) context, generating deletions, duplications, and point mutations.

Consider two DNA sequences, identical or nearly so but separated by several to many base pairs. If the repeated sequences are sufficiently long (perhaps dozens to hundreds of base pairs) to mediate genetic recombination, then “unequal” recombination can occur, generating a duplication or a deletion or both (Fig. 1). Even if the repeated sequences are too short for ordinary recombination, they may still mediate anomalies of DNA metabolism that lead to duplications and deletions (Fig. 2): a break occurs in one of the two DNA chains, the chains separate (melt) locally, and then they reform a double helix out of register (misanneal). Subsequent DNA synthesis then closes the gap, fixing the mutation in the chromosome. In the extreme example, the DNA sequence repeat lacks intervening bases, and the additions or duplications are of only one or a few bases (generating, if occurring within a protein-coding sequence, frameshift mutations).

If one of the components of a repeated DNA sequence is inverted with respect to the other (both end to end and top strand to bottom strand in order to preserve the chemical polarity of DNA), the result is a DNA palindrome (a sequence that is the same when read in either direction). Just as direct repeats can mediate misalignments, so can palindromes, but with sometimes quite different consequences. For instance, during the transitory stages when DNA becomes single-stranded, as during replication, excision repair, and recombination, palindromic DNA sequences can fold back upon each other to form “hairpin” structures (Fig. 3). Such an anomaly can

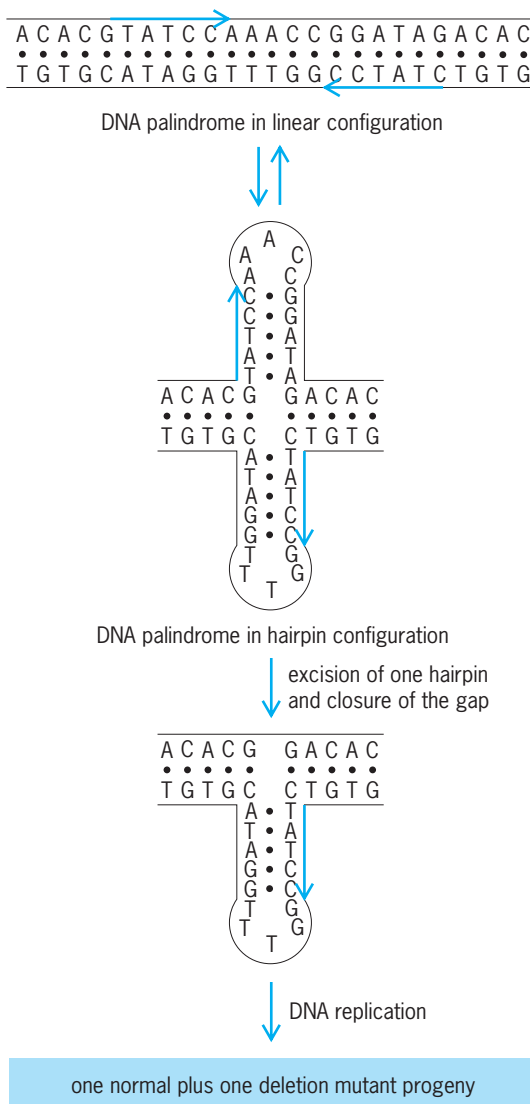


Fig. 3. Palindrome-mediated deletion mutagenesis. Color indicates palindromic sequences.

lead to a deletion, either because synthesis of the complementary strand passes by the hairpin or because the hairpin is recognized as an abnormal DNA structure and excised. In practice, while the ends of deletions often fall in either repeated or palindromic sequences, they also often fall in regions which contain both elements at once, thus providing enhanced misalignment stability.

Palindromes possess an additional property, not shared with direct repeats, that causes them to mediate the formation of point mutations. Consider a palindrome which is imperfect, the inverted repeats not being perfect complements. (Here the intervening bases between the palindromic elements are irrelevant.) If it assumes a hairpin structure, then its stem will encompass mispaired or nonpaired bases (Fig. 4). The repair systems that maintain the structural integrity of DNA by excising damaged, abnormal, or mispaired bases can then act upon this imperfect stem, “correcting” one of its strands. The result can be base-pair substitutions, base additions or deletions, and complex mutations. (It should be noted

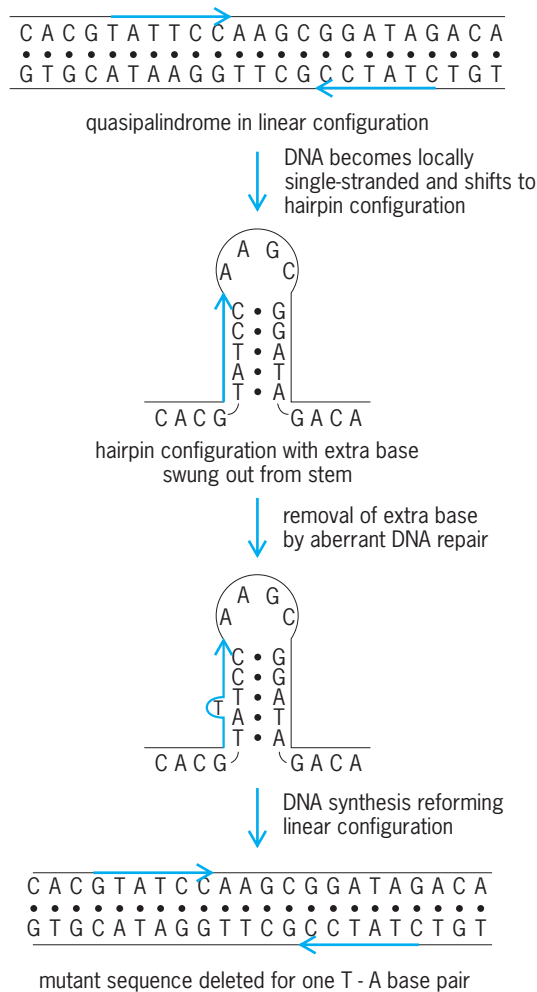


Fig. 4. Palindrome-mediated frameshift mutagenesis. Color indicates imperfectly palindromic sequences.

that the mechanisms discussed here have been set in the hairpin context, but may actually occur by topologically related misalignments at DNA replication forks.)

Mispairing mutagenesis. The normal pairing between the DNA bases (A · T and G · C) is shown in **Fig. 5a** and **b**. DNA bases can also pair incorrectly, usually because of transient changes in base structure. An early mispairing proposal invoking base tautomers created by proton migration is shown in **Fig. 5c**. Structural studies now suggest mispairing via “wobble” configurations (**Fig. 5d**) or ionized bases (**Fig. 5e**). Note, however, that the next round of pairing by these bases is likely to be normal, generating one mutant and one nonmutant progeny DNA double helix. See PURINE; PYRIMIDINE.

In order to generate a transversion, either two purines or two pyrimidines must mispair. In their usual configurations, however, neither such pair approximates the normal dimensions of a DNA base pair. In practice, most transversions arise via purine-purine mispairs; one example is shown in **Fig. 6**. In a normal base pair, the purines are in the *anti* configuration, which means that their hexagonal ring points toward the complementary pyrim-

idine. Occasionally, however, a purine (adenine in this example) rotates 180° around its glycosidic bond (the bond leading from the base to the sugar-phosphate backbone) into the *syn* configuration, thus presenting a different part of itself for potential pairing via hydrogen bonds.

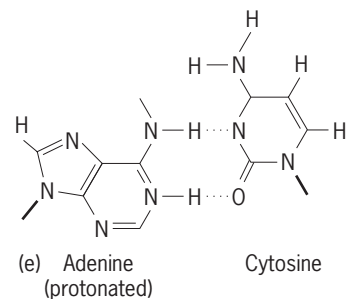
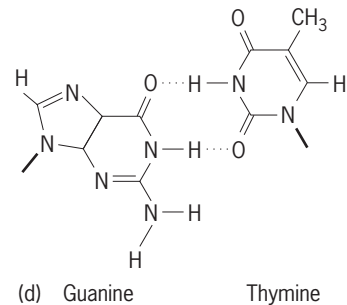
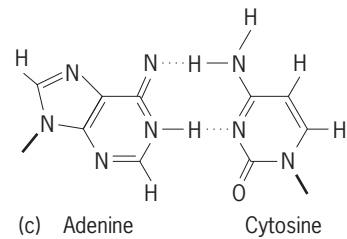
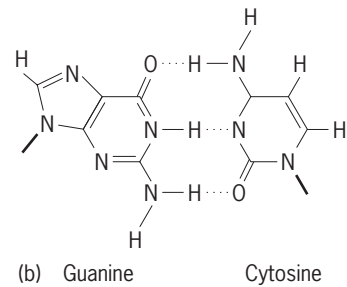
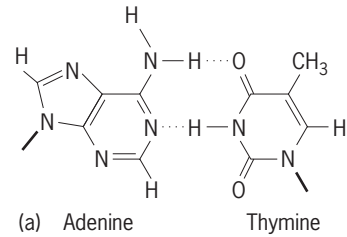


Fig. 5. Normal pairings and mispairings between the DNA bases. (a) The normal adenine-thymine base pair; hydrogen bonds are indicated by dotted lines and glycosidic bonds to the DNA backbone by heavy lines. (b) The normal guanine-cytosine base pair. (c) An adenine-cytosine mispair caused by proton migration in an adenine (arrow), leading to a transition mutation. (d) A guanine-thymine mispair involving wobble pairing. (e) An adenine-cytosine mispair involving both a protonated base and wobble pairing.

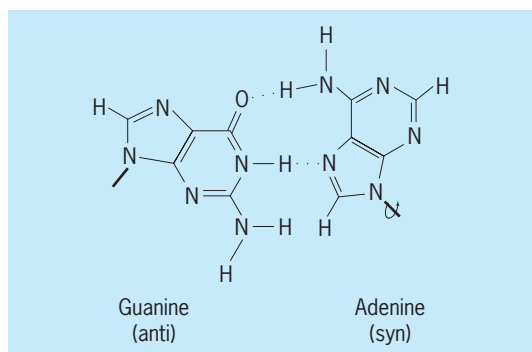


Fig. 6. A guanine-adenine mispair involving rotation of the adenine around its glycosidic bond. The mispair leads to a transversion mutation.

Chemical damage that alters the structure of a base can sometimes greatly enhance mispairing. Adding a methyl or ethyl group to the O⁶ position of guanine, for instance, renders the guanine able to pair incorrectly with thymine (Fig. 7a). The deamination of cytosine, which is promoted by heat and by low pH and which replaces the amino group with a keto group, generates uracil, which is identical to thymine in its base-pairing propensities (Fig. 7b). Thus, either of these treatments can induce G·C → A·T transitions, one via damaged guanine and the other via damaged cytosine.

Many kinds of base damage render bases unfit to pair at all. The dimerization of adjacent pyrimidines by ultraviolet irradiation, the addition of bulky groups to purines by mutagenic carcinogens such as the benzpyrines and aflatoxins, or the loss of a base (usually a purine) cause termination of DNA synthesis rather than mispairing. However, as noted below, mispairing can occur later at such a site.

Cause and prevention. Cells direct a number of repair operations against anomalies of DNA structure that would otherwise result in death or mutation. While some of these repair mechanisms circumvent death, others actually cause mutations.

Prevention. Most DNA damage is detected by excision repair systems before it can interfere with replication. These systems enzymatically remove the damaged bases and their associated sugar-phosphate backbones; the resulting gap is then filled in by DNA repair synthesis, using the complementary strand as a template. However, occasional damaged or transiently rearranged bases enter the DNA replication fork. There they encounter DNA polymerase, the enzyme responsible for DNA synthesis. Most DNA polymerases are able to discriminate against base mispairs, although by ways that are still poorly understood. Even when a mispaired base has been covalently incorporated, however, it is often removed. This kind of repair can occur at two times. First, a 3'-exonuclease may specifically remove just-incorporated but mispaired nucleotides (nucleotide = base + sugar-phosphate). This enzymatic activity is called proofreading. Second, the few errors that escape proofreading may still be caught by a subsequent mismatch correction system. This system is remarkable in distinguishing between parental

and progeny bases in newly synthesized DNA, so that the wrong rather than the right base can be excised.

Mutagenic translesion synthesis. When base damage renders pairing impossible and DNA synthesis terminates, it often reinitiates farther on, leaving behind a gap. Such a gap opposite a useless base is very apt to be lethal, certainly to the cell that inherits the damaged chromosome. Often, however, the cell is able to call up postreplication repair systems to save itself (SOS systems). Some such systems use recombination and are relatively error-free. Others, however, appear to use error-prone translesion synthesis, which is highly mutagenic because the requisite genetic information is unavailable at the damaged site. Such bypass synthesis seems to be responsible for most mutations induced by chemicals and radiations, although perhaps not contributing substantially to spontaneous mutations.

Past and future. Enzymatic systems for preventing (and sometimes causing) mutations have long evolutionary histories, presumably dating back to the genesis of cells. Therefore, mutation rates have evolved to levels that balance the few mutations that are useful, the much greater number that are deleterious, and the investment required to further reduce error rates. In DNA-based organisms ranging from small bacteriophages to the fungus *Neurospora crassa*, this balance produces about 0.3% mutations per genome per replication; organisms with larger genomes have smaller spontaneous mutation rates per base pair. In RNA viruses, the balance produces about one mutation per genome per replication. Spontaneous mutation rates are much less widely sampled in higher organisms, but in the fruitfly there

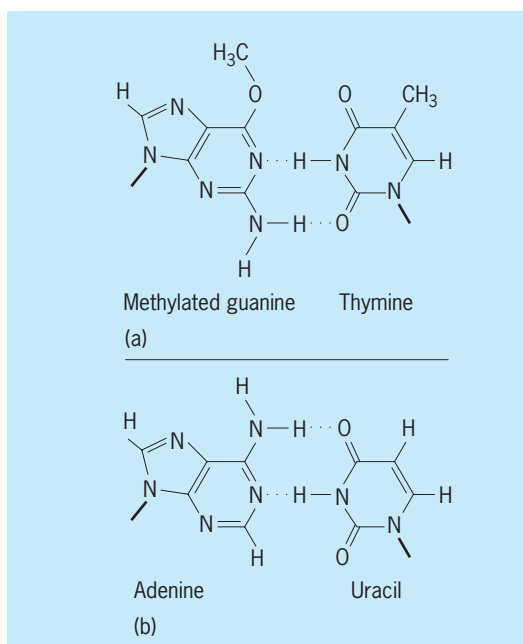


Fig. 7. Mispairing of DNA bases following chemical damage. (a) A guanine-thymine mispair promoted by methylation of a guanine (b) An adenine-uracil mispair in which the uracil was formed by the deamination of cytosine.

is roughly one new deleterious mutation per diploid genome per sexual generation. The large majority of these mutations, however, are only mildly deleterious, and would be unrecognizable in individual humans; they might, for instance, slightly shorten life-spans, slightly increase the frequency of infections by lowering body resistance, or affect health in other equally diffuse ways.

With the growing recognition (since about 1950) that many chemicals are mutagenic, and that many humans (between about 5 and perhaps 25%, depending on the criteria used) suffer from mutations of ancient or recent origin, much attention has been directed to discovering the mutagenic components of the environment. Many artificial chemicals are mutagenic, although these are still only a small percent of the total; and as a part of their ongoing war against predators, most plants elaborate chemicals that are mutagenic for at least some organisms. The impact of these exposures is largely unassessed, but intensive ongoing research is likely to provide some concrete answers. *See* MUTAGENS AND CARCINOGENS.

While the search for ways to prevent unwanted mutation continues vigorously, other searches seek methods to introduce specific, predetermined, desirable mutations into organisms of economic importance. By using the methods of molecular genetics, it is now usually possible to introduce specific mutations into specific genes in order, for instance, to increase the production of some marketable product. However, this step requires considerable prior knowledge about the organism's genetic and biochemical makeup. A more distant possibility is intervention in the human germ line in order, for instance, to cure specific genetic diseases. *See* GENETIC ENGINEERING; GENETICS; HUMAN GENETICS.

John W. Drake

Bibliography. J. W. Drake, B. W. Glickman, and L. S. Ripley, Updating the theory of mutation, *Amer. Sci.*, 71:621-630, 1983; M. E. Lambert et al. (eds.), *Eukaryotic Transposable Elements as Mutagenic Agents*, 1988.

Mutualism

An interaction between two species that benefits both. Individuals that interact with mutualists experience higher success than those that do not. Hence, behaving mutualistically is advantageous to the individual, and it does not require any concern for the well-being of the partner. At one time, mutualisms were thought to be rare curiosities primarily of interest to natural historians. However, it is now believed that every species is involved in one or more mutualisms. Mutualisms are thought to lie at the root of phenomena as diverse as the origin of the eukaryotic cell, the diversification of flowering plants, and the pattern of elevated species diversity in tropical forests.

Mutualisms generally involve an exchange of substances or services that organisms would find difficult or impossible to obtain for themselves. In many

cases, mutualists make limited but essential nutrients available to their partners. For instance, *Rhizobium* bacteria found in nodules on the roots of many legume (bean) species fix atmospheric nitrogen into a form (NH₃) that can be taken up by plants. The plant provides the bacteria with carbon in the form of dicarboxylic acids. The carbon is utilized by the bacteria as energy for nitrogen fixation. Consequently, leguminous plants often thrive in nitrogen-poor environments where other plants cannot persist. Another well-known example is lichens, in which fungi take up carbon fixed during photosynthesis of their algae associates. *See* NITROGEN FIXATION.

A second benefit offered within some mutualisms is transportation. Prominent among these mutualisms is biotic pollination, in which certain animals visit flowers to obtain resources (usually food, in the form of nectar) and return a benefit by transporting pollen between the flowers they visit. Biotic seed dispersal, in which animals disperse the seeds of the plants whose fruits they consume, is another example of transportation-based mutualism. A final benefit is protection from one's enemies. For example, ants attack the predators and parasites of certain aphids in exchange for access to the aphids' carbohydrate-rich excretions (honeydew).

Another consideration about mutualisms is whether they are symbiotic. Two species found in intimate physical association for most or all of their lifetimes are considered to be in symbiosis. Not all symbioses are mutualistic; symbioses may benefit both, one, or neither of the partners. Lichens and plant-*Rhizobium* associations are symbiotic mutualisms, whereas plant-pollinator, plant-seed disperser, and ant-aphid associations are nonsymbiotic mutualisms. In some cases, the symbionts become so closely integrated that their individuality becomes difficult and ultimately impossible to distinguish. The eukaryotic cell is now believed to have originated as a symbiosis between a primitive cell and a bacterium ancestral to modern-day mitochondria.

Mutualisms can also be characterized as obligate or facultative (depending on whether or not the partners can survive without each other), and as specialized or generalized (depending on how many species can confer the benefit in question). An obligate, highly specialized mutualism exists between fig trees and the fig wasp pollinators. There is a nearly one-to-one match between fig and fig wasp species, and neither can reproduce in the absence of its partner. In this case, there has been a long history of coevolution; that is, reciprocal evolutionary pressures have led to adaptations of both partners that facilitate the interaction. Most mutualisms are considerably looser than this, however, and the degree to which they are the result of coevolution is more questionable.

Clearly, mutualisms are extremely diverse interactions. However, there are two features common to most mutualisms. First, mutualisms are highly variable in time and space. For example, in the interaction in which ants protect aphids from their

enemies, the benefits to aphids are huge in times and places where their enemies are abundant, but essentially nonexistent where enemies are rare. Furthermore, under some conditions ants will consume aphids rather than protecting them, leading the costs of the interaction to exceed its benefits. Second, mutualisms are susceptible to cheating. Cheaters can be individuals of the mutualist species that profit from their partners' actions without offering anything in return, or else other species that invade the mutualism for their own gain. The best-studied cheaters are nectar-robbing bees, which obtain nectar quickly and efficiently by chewing holes in flowers and thus neither collect nor deposit pollen. Evolutionary biologists are increasingly interested in how mutualistic behaviors, and hence mutualisms, can persist, given the apparent benefits of being a cheater.

An understanding of mutualism has considerable practical significance. Certain mutualisms play central roles in humans' ability to feed the growing population. Cattle and sheep can consume grass because of their mutualisms with gut-inhabiting bacteria that degrade cellulose; soybeans, which make up half the world's grain legume production, are typically inoculated with *Rhizobium* to increase yields. It has been estimated that half the food consumed is the product of biotic pollination. Other mutualisms are critical in structuring natural communities and hence are of great interest from a conservation perspective. For example, fruits produced as a result of the obligate pollination mutualism between figs and fig wasps are relied upon by nearly every fruit-feeding bird and mammal in some tropical forests during times of food scarcity. These animals in turn disperse the seeds of other plants and thus help sustain populations of predators and parasites. Disruption of the pollination mutualism could therefore constitute a serious threat to local biodiversity. Unfortunately, mutualisms are apparently rather sensitive to human-induced environmental changes such as habitat fragmentation, atmospheric pollution, and the introduction of nonnative species. Rather than focusing solely on species in isolation, conservationists increasingly attempt to preserve the critical mutualisms in which endangered organisms are involved. *See* ECOLOGY; PLANT PATHOLOGY.

Judith L. Bronstein

Bibliography. D. H. Boucher (ed.), *The Biology of Mutualism*, Oxford University Press, 1985; J. L. Bronstein, Our current understanding of mutualism, *Quart. Rev. Biol.*, 69:31-51, 1994; A. E. Douglas, *Symbiotic Interactions*, Oxford University Press, 1994; J. N. Thompson, *The Coevolutionary Process*, University of Chicago Press, 1994.

Myasthenia gravis

A disease resulting from an abnormality in neuromuscular transmission, characterized by a fluctuating degree of muscle weakness. The weakness is usually aggravated by activity, and there is partial or complete restoration of strength after a period of rest

or the administration of anticholinesterase medications. *See* ACETYLCHOLINE.

Etiology. It has been shown that the basic defect in myasthenia gravis is a reduction in the number of acetylcholine receptor sites in the postsynaptic membrane of the neuromuscular junction. It has also been shown that in myasthenia gravis, the individuals often have immunoglobulins in the serum that partially block acetylcholine receptors, and that myasthenic changes can be produced in mice by the injection of serum immunoglobulin from myasthenic patients. The antiacetylcholine receptor antibody is thought to originate in the thymus gland, and the antigenic stimulus is believed to be the acetylcholine receptors in the myoid cells in the thymus gland. *See* AUTOIMMUNITY; IMMUNOGLOBULIN.

Pathology. Abnormalities have been demonstrated in the thymus gland and skeletal muscle in myasthenia gravis. There is an increased incidence of thymoma in myasthenia gravis, and in those without a thymoma, hyperplasia of the germinal centers is a common finding in the thymus gland.

Muscle from myasthenic persons shows variable histologic findings. In some it is completely normal. In others, collections of lymphocytelike cells are present in otherwise normal muscle. In still others patterns suggesting neurogenic atrophy or a myopathy are found. However, by electron microscopy, a simplification of the secondary synaptic clefts and a widening of the primary synaptic cleft of the motor end plate has been demonstrated.

Clinical features. Although the disease affects young women more commonly, usually in the third decade, it can occur in either sex at any age. In approximately 10% of those affected, myasthenia gravis is associated with a thymoma, and the prognosis in such individuals tends to be worse. In the majority of persons, weakness affects muscles of the head, neck, and limbs (generalized myasthenia), but in some the weakness is restricted to the muscles of the eyes (ocular myasthenia), in which case the disease is usually benign. Fifteen percent of babies born of myasthenic mothers have symptoms requiring treatment at birth, but these are transient and disappear in a few weeks (neonatal myasthenia).

The muscles most often affected in generalized myasthenia are those of the eyes, with ptosis and double vision (diplopia), of mastication, swallowing, and speech, and of respiration. Limb muscles are also frequently affected.

Initially, the affected muscles have normal or nearly normal strength. However, with repeated use, progressive weakness develops. Thus, ptosis and diplopia become worse as the day progresses, speech becomes slurred and nasal with continued conversation, or the individual may be unable to finish eating because of increasing difficulty with chewing and swallowing. These symptoms are often only partially reversed with rest. In some individuals myasthenia gravis is a life-threatening illness because of respiratory weakness.

Diagnosis. The diagnosis of myasthenia gravis is based on the demonstration of increasing muscle

weakness with exertion, and reversal of the weakness after the administration of anticholinesterase drugs.

Treatment. The standard treatment for myasthenia gravis has been the use of longer-acting anticholinesterase agents; thymectomy and immunosuppressive drugs are reserved for those individuals with generalized myasthenia that does not respond sufficiently to these agents. With the demonstration that myasthenia gravis is due to thymus-originated antireceptor antibody, thymectomy has been resorted to earlier in the course of the illness, often combined with corticosteroids or other immunosuppressive therapy. In those individuals with significant persistent symptoms, plasmapheresis has also become another mode of treatment. When a thymoma is detected, thymectomy is mandatory.

Persons being treated with anticholinesterase drugs may develop periods of severe muscle weakness with respiratory paralysis. These paralytic periods may be due to inadequate dosage (myasthenic crisis) of, or overmedication (cholinergic crisis) with, these agents, but most commonly they appear to result from development of resistance to the drugs. See SYNAPTIC TRANSMISSION.

S. Mark Sumi
Bibliography. M. H. De Baets and H. J. Oosterhuis (eds.), *Myasthenia Gravis*, 1993.

Mycobacterial diseases

Diseases caused by mycobacteria, a diffuse group of acid-fast, rod-shaped bacteria in the genus *Mycobacterium*. Some mycobacteria are saprophytes, while others can cause disease in humans. The two most important species are *M. tuberculosis* (the cause of tuberculosis) and *M. leprae* (the cause of leprosy); other species have been called by several names, particularly the atypical mycobacteria or the nontuberculous mycobacteria. This article deals mainly with nontuberculous mycobacteria. See LEPROSY; TUBERCULOSIS.

These bacteria are classified according to their pigment formation, rate of growth, and colony morphology. The organisms, and the diseases they cause, are summarized in the **table**. The most commonly involved disease site is the lungs. Nontuberculous mycobacteria are transmitted from natural sources in

the environment, rather than from person to person, and thus are not a public health hazard. Nontuberculous mycobacteria have been cultured from various environmental sources. For example, *M. avium intracellulare* has been found in soil and dust; *M. kansasii* has been isolated from water samples; *M. marinum* has been detected in fresh and salt water; and *M. chelonae* has been found in soil and dust. The environmental sources of many other species are unknown.

The diagnosis of disease caused by nontuberculous mycobacteria can be difficult, since colonization or contamination of specimens may be present rather than true infection.

Pulmonary disease. Pulmonary disease resembling tuberculosis is a most important manifestation of disease caused by nontuberculous mycobacteria. The symptoms and chest x-ray findings are similar to those seen in tuberculosis. *Mycobacterium kansasii* and *M. avium intracellulare* are the most common pathogens, followed by *M. fortuitum* and *M. chelonae*. The disease usually occurs in middle-aged men and women with some type of chronic coexisting lung disease such as chronic obstructive pulmonary disease (emphysema), chronic bronchitis, pneumoconiosis (for example, silicosis), or previous tuberculosis. Although the pathogenic mechanisms are obscure, it is presumed that there is an involvement with inhalation of droplet nuclei containing organisms, followed by mycobacterial proliferation in areas of destroyed lung.

Pulmonary infections due to *M. kansasii* can be treated successfully with chemotherapy. The treatment of pulmonary infections due to *M. avium intracellulare* complex is difficult. These organisms are highly resistant to chemotherapy even with multiple drug regimens. Surgical resection of localized disease is considered in selected individuals.

Bone and joint disease. Chronic infection involving joints and bones, bursae, synovia, and tendon sheaths can be caused by various species. Infection usually occurs by direct inoculation of the organism following trauma, surgery, or a joint injection. *Mycobacterium fortuitum*, *M. chelonae*, *M. kansasii*, and *M. avium intracellulare* are the usual pathogens. Treatment of bone and joint infection consists of surgical drainage and removal of tissue. Drug therapy may be useful for the relatively sensitive *M. kansasii*.

| Nontuberculous (atypical) mycobacteria | | | | |
|--|--|---|--|--|
| Group | Classification | Characteristics | Representative species | Type of disease |
| I | Photochromogens | Yellow pigment production when exposed to light | <i>Mycobacterium kansasii</i> <i>M. marinum</i> | Pulmonary or extrapulmonary Swimming pool granuloma |
| II | Scotochromogens | Pigment production in dark | <i>M. scrofulaceum</i> <i>M. goodii</i> | Cervical adenitis Saprophytic tap-water bacilli |
| III | Nonphotochromogens (<i>Battley bacilli</i>) | No pigment production | <i>M. avium intracellulare</i> complex | Pulmonary or extrapulmonary |
| IV | Rapid growers | Growth in 3–5 days | <i>M. fortuitum</i> <i>M. chelonae</i> | Saprophytic Wound abscesses |

Skin and soft tissue infection. Localized abscesses due to *M. fortuitum* or *M. chelonae* can occur after trauma, after surgical incision, or at injection sites. The usual treatment is surgical incision.

The most common soft tissue infection is caused by *M. marinum*, which may be introduced, following an abrasion or trauma, from handling fish or fish tanks, or around a swimming pool. Treatment is surgical.

Mycobacterium ulcerans causes a destructive skin infection in tropical areas of the world. It is treated by wide excision and skin grafting.

Lymphadenitis. Lymph-node enlargement and inflammation due to *M. avium intracellulare* or *M. scrofulaceum* may occur in children between the ages of 1 and 5 years. Usually, only the submandibular nodes in the neck are involved, but sometimes axillary or inguinal nodes are also affected. The child is usually asymptomatic, but node rupture with drainage and sinus formation can occur rapidly. The condition heals with fibrosis and calcification. Exact pathogenesis is unclear, but the mouth or throat might be the portal of entry. Therapy consists of excising the involved group of nodes.

Disseminated disease. Disseminated disease due to nontuberculous mycobacteria is usually associated with *M. avium intracellulare* and *M. kansasii*. Dissemination occurs most often in individuals with defects in immunity, such as those having leukemia, immunosuppressive drug therapy, chronic renal failure, or malignancy.

Disseminated *M. avium intracellulare* is one of the opportunistic infections seen in the acquired immune deficiency syndrome (AIDS). In individuals with AIDS, the organism has been cultured from lung, brain, cerebrospinal fluid, liver, spleen, intestinal mucosa, and bone marrow. No treatment has yet been effective in this setting. See ACQUIRED IMMUNE DEFICIENCY SYNDROME (AIDS).

George M. Lordi; Lee B. Reichman

Bibliography. J. McFadden (ed.), *Molecular Biology of the Mycobacteria*, 1990; J. L. Stanford, *The Biology of the Mycobacterium*, III. *Clinical Aspects of Mycobacterial Disease*, 1984.

Mycology

The study of organisms of the fungal lineage, including mushrooms, boletes, bracket or shelf fungi, powdery mildew, bread molds, yeasts, puffballs, morels, stinkhorns, truffles, smuts, and rusts. Fungi, in the traditional sense, are an ecological grouping of organisms found in every ecological niche. Mycologists estimate that there are 1.5 million species of fungi, with only 70,000 species now described. Fungi typically have a filamentous-branched somatic structure surrounded by thick cell walls; these structures are known as hyphae. See FUNGI; MUSHROOM.

Taxonomy. The names adopted for major taxa (groups) of fungi and fungi-like organisms are as follows:

Eumycota (true fungi)

Chytridiomycota (Chytridiomycetes)

Zygomycota (Zygomycetes, Trichomycetes)

Ascomycota (Archiascomycetes,

Hemiascomycetes, Euascomycetes)

Basidiomycota (Urediniomycetes,

Ustilaginomycetes, Hymenomycetes)

Deuteromycota (Agonomycetes, Balstomycetes,

Coelomycetes, Hyphomycetes)

Pseudomycota (fungi-like organisms)

Oomycota (Oomycetes)

Hyphochytriomycota (Hyphochytriomycetes)

Plasmodiophoromycota

(Plasmodiophoromycetes)

Myxomycota (slime molds)

Myxomycota (Myxomycetes)

See CLASSIFICATION, BIOLOGICAL; FUNGAL ECOLOGY.

Description. Fungi are heterotrophic, meaning they cannot make their own food as plants do. They are osmotrophic or nutrition-absorptive, obtaining their food by releasing enzymes into their environment to break down complex organic compounds. Fungi play a significant role in nature as decomposers of wood and wood products, and in forest ecosystems fungi release nutrients back to soil. Fungi also decompose fabrics, leather goods, petroleum products, and foodstuffs. They reproduce through spores produced sexually and asexually. See FOREST ECOSYSTEM; WOOD DEGRADATION.

Fungi are often viewed as detrimental to plants, animals, and humans. They cause nearly 80% of all plant disease and, despite the extensive use of fungicides and other control measures, the financial loss due to fungal activity remains enormous. In humans, fungal spores cause allergies in certain individuals; therefore, mold spore levels are carefully monitored and reported. With an increasing immunocompromised population due to human immunodeficiency virus (HIV), malignancies, chemotherapies, and organ transplants, the occurrence of new opportunistic fungal infections has risen. However, along with the increase in human mycoses comes realistic hope for improved diagnosis and therapy. See ERGOT AND ERGOTISM; FUNGISTAT AND FUNGICIDE; MEDICAL MYCOLOGY; PLANT PATHOLOGY.

Food. Because of their rapid growth rate and high protein content, fungi are an ideal source of protein for animals and humans. Many mushrooms are edible, and some are cultivated for the canned and fresh market. There are many foods, especially fermented products, that are produced by fungi. The food industry utilizes fungi for production of natural flavor and color compounds.

Uses. Secondary metabolites produced by fungi have been shown to have medicinal uses. Penicillin, an antibiotic extracted from fungi, revolutionized medicine, and many more antibiotics produced by fungi have been discovered. Many herbal medicines are derived from fungi. In addition to their antibiotic secondary metabolites, fungi have a broad potential for producing medically useful

compounds such as metabolic regulators, antitumor drugs, and immunomodulators that enhance host resistance against cancer, immunodeficiency disease, or generalized immunosuppression. See ANTI-BIOTIC.

Fungi do have some positive roles in agriculture. As an alternative to chemical pesticides, fungi are being tested as biocontrol agents to reduce the crop damage from insects, nematodes, weeds, and plant pathogenic microorganisms. Biofertilizers are specially formulated inoculants of mycorrhizal fungi used to improve crop growth in nutrient-deficient soils. See FERTILIZING.

There are many uses for fungi in industry. Biological delignification with fungi can remove the lignin component from wood without destroying the cellulose and hemicellulose important in paper production. Fungal polysaccharide-degrading enzymes can also be used instead of chemicals for removal of noncellulosic materials in the retting of flax fibers, producing high-quality yarns. A wide range of fungi are useful in converting coal to a liquid or to gaseous products that would normally require high temperatures and pressures which are expensive and difficult to achieve. Fungi are also a useful alternative to the conventional scavenging techniques for the removal of heavy metals from dilute solutions, such as electroplating and chrome tannery effluents.

The application of recombinant DNA techniques has greatly increased the potential use of fungi in biotechnology. New genetic material inserted into chromosomes or extrachromosomal elements directs the transformed fungus to metabolize or produce specific products. Because they are capable of secreting large quantities of certain proteins in liquid culture, fungi have proven to be useful cloning hosts for the production of recombinant proteins of both fungal and human origin. See DEOXYRIBONUCLEIC ACID (DNA); FUNGAL BIOTECHNOLOGY; FUNGAL GENETICS.

Shung-Chang Jong

Bibliography. C. J. Alexopoulos, C. W. Mims, and M. Blackwell, *Introductory Mycology*, 4th ed., Wiley, New York, 1996; Z. An (ed.), *Handbook of Industrial Mycology*, Marcel Dekker, New York, 2005; M. J. Carlile, S. C. Watkinson, and G. W. Gooday, *The Fungi*, 2d ed., Academic Press, New York, 2001; P. M. Kirk et al., *Dictionary of the Fungi*, 9th ed., CAB International, Wallingford, CT, 2001; G. M. Mueller, G. F. Bills, and M. S. Foster (eds.), *Biodiversity of Fungi: Inventory and Monitoring Methods*, Elsevier Academic Press, New York, 2004.

Mycoplasmas

The smallest prokaryotic microorganisms that are able to grow on cell-free artificial media. Their genome size is also among the smallest recorded in prokaryotes, about 5×10^8 to 10^9 daltons. The mycoplasmas differ from almost all other prokaryotes in lacking a rigid cell wall and in their incapability to

synthesize peptidoglycan, an essential component of the bacterial cell wall.

Taxonomically, the mycoplasmas are assigned to a distinct class, the Mollicutes, containing two orders, Mycoplasmatales and Acholeplasmatales. The distinction between the orders is based primarily on differences in nutritional criteria: members of the Mycoplasmatales require cholesterol or other sterols for growth whereas those of the second order do not. The main criteria used for the subdivision of the orders into families and genera are shown:

Class: Mollicutes

Order I: Mycoplasmatales (sterol required for growth; NADH₂ oxidase localized in cytoplasm)

Family I: Mycoplasmataceae (genome size approximately 5×10^8 daltons)

Genus I: *Mycoplasma* (70 species; do not hydrolyze urea)

Genus II: *Ureaplasma* (2 species; hydrolyze urea)

Family II: Spiroplasmataceae (helical organisms; genome size approximately 10^9 daltons)

Genus I: *Spiroplasma* (4 species)

Order II: Acholeplasmatales (sterol not required for growth; NADH₂ oxidase localized in membrane; genome size approximately 10^9 daltons)

Family I: Acholeplasmataceae

Genus I: *Acholeplasma* (9 species)

Genus of uncertain taxonomic position

Anaeroplasmata (2 species)

The term mycoplasmas is generally used as the vernacular or trivial name for all members of the class Mollicutes, irrespective of the classification in a particular genus. See PROKARYOTAE.

Morphology and reproduction. The individual cells are pleomorphic, varying in shape from spherical or pear-shaped structures (about 0.3–0.8 micrometer in diameter) to branched filaments varying in length from approximately 5 to 150 μm (illus. a); the smallest viable cells will pass a membrane filter of 450 nm pore diameter. The organisms are gram-negative, nonsporing, and usually nonmotile, although some species of genus *Mycoplasma* show gliding motility on liquid-covered surfaces. Members of the family Spiroplasmataceae are characterized under most conditions by the production of helical filaments (illus. b) exhibiting rotatory and undulating motility. Replication is basically by binary fission of spherical cells, but cell division is not necessarily synchronized with replication of the genome. Thus, the development of filaments and their transformation into chains of spherical cells may be ascribed to a lack of synchrony between division of the genome and the cytoplasm.

Growth in liquid medium produces a barely visible to rather heavy turbidity. Colonies on agar are minute [from less than 0.04 in. (1 mm) up to 0.15 in. (4 mm) in diameter]; under suitable growth conditions the colonies of many species have a characteristic "fried

egg" appearance, consisting of an opaque, more or less yellowish central area which grows down into the medium, and a translucent, flat peripheral zone (illus. c).

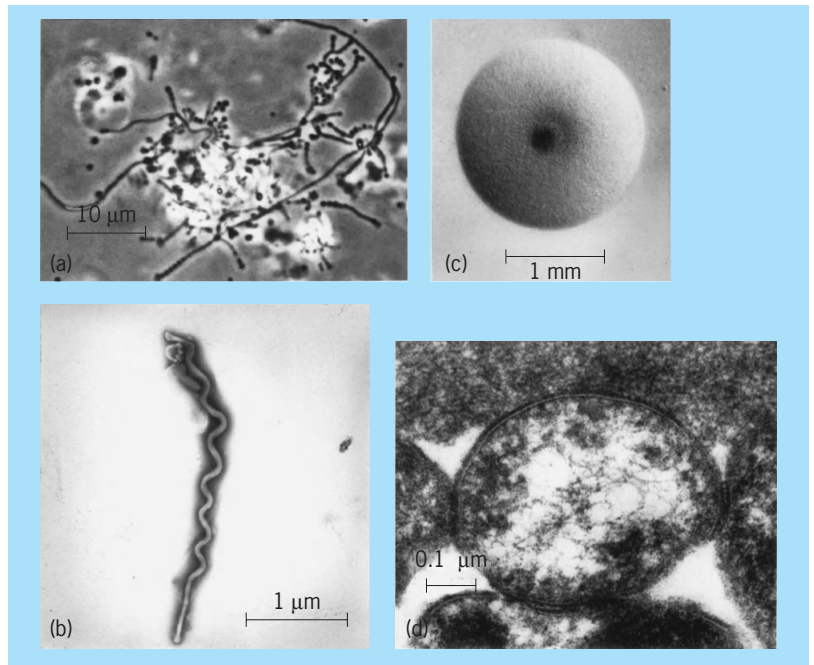
Ultrastructure. Electron microscopy of ultrathin sections of mycoplasma cells shows that the cytoplasm is bounded only by a plasma membrane, about 8 to 10 nm thick, while there is no cell wall as is found in other prokaryotes (illus. d). The cell membrane of some species is covered by a capsular substance. The cytoplasm consists of nuclear material (DNA strands), ribosomes, and intracellular granules. Specialized structures such as terminal organelles, which play a role in attachment of the organism to the surface of host cells, have been demonstrated in several species.

Though the mycoplasmas have much in common with the wall-defective or wall-less L-phase variants of bacteria, it is now generally accepted that there are several fundamental differences separating the two groups of organisms.

Metabolism and biochemistry. The mycoplasmas have limited biosynthetic abilities and therefore require complex growth media, consisting of beef heart infusion, peptone, yeast extract, and serum. Serum serves as a source of many essential compounds, especially fatty acids and cholesterol in an assimilable and nontoxic form as required for membrane synthesis in the vast majority of species. Almost all mycoplasma species catabolize glucose or arginine, or both. Glucose and certain other carbohydrates serve as an energy source for fermentative mycoplasmas possessing the Embden-Meyerhof glycolytic pathway. See FERMENTATION.

Arginine degradation occurs by the arginine dihydrolase pathway that may be the major source for adenosine triphosphate (ATP) in nonfermentative mycoplasmas. Urease activity is a unique property of genus *Ureaplasma*. Lipases are involved in the development of characteristic precipitates (known as "film and spots") formed by certain species on horse-serum agar or egg-yolk medium. Methylene blue, tellurite, and 2,3,5-triphenyltetrazolium chloride are frequently reduced. Proteolytic activity, as demonstrated by clearing and/or liquefaction of coagulated serum, gelatin and casein, is characteristic of some species. Several species show phosphatase activity. Pigmented carotenoids are produced by some *Acholeplasma* species. Weak to strong α and β hemolysis on blood agar containing erythrocytes from horse, sheep, guinea pig, or chicken, is a property shared by most mycoplasmas; the hemolysin has been identified as peroxide.

Mycoplasmas are generally highly resistant to benzyl penicillin and other antibiotics which act by interfering with the biosynthesis of peptidoglycan. They are usually susceptible to antibiotics that specifically inhibit protein synthesis in prokaryotes, as for example tetracyclines and chloramphenicol. Susceptibility to other antibiotics, such as erythromycin and other macrolides, is variable. Most mycoplasmas tolerate 1:2000 to 1:4000 thallium acetate, which is often incorporated, together with penicillin, as



Mycoplasmas. (a) *Mycoplasma hominis*, phase contrast micrograph (courtesy of W. Bredt). (b) *Spiroplasma citri*, electron micrograph of negatively stained specimen (courtesy of R. M. Cole). (c) Mycoplasma colony on solid nutrient medium. (d) *Ureaplasma urealyticum*, electron micrograph of thin-sectioned cell showing plasma membrane and network of DNA strands (courtesy of F. T. Black).

a selective agent in mycoplasma media to inhibit the growth of contaminating bacteria. See BACTERIAL PHYSIOLOGY AND METABOLISM.

Habitat and pathogenicity. The mycoplasmas are almost ubiquitous in nature. Several species are important pathogens of humans, animals and plants, while others constitute part of the normal microbial flora of, for example, the upper respiratory and lower urogenital tracts of humans.

Mycoplasma pneumoniae, previously regarded as a virus and known under the name of the Eaton agent, was found in 1962 to be the cause of cold agglutinin-associated primary atypical pneumonia. This disease is particularly frequent in the 5-15-year age group; it is probably endemic almost all over the world and often reaches epidemic proportions at intervals of 4 to 5 years. *Mycoplasma hominis*, which is a common inhabitant of the lower genital tract, is a potential pathogen and appears to be a relatively frequent cause of acute pelvic inflammatory disease (salpingitis and parametritis) and postpartum septicemia, and a less frequent cause of certain other infectious diseases, especially in an immunocompromised host. The human mycoplasma, *M. genitalium*, is another candidate as a possible cause of salpingitis and related disorders. A number of species, including *M. salivarium* and *M. orale*, are common members of the oral microflora. The role of *M. fermentans*, found rarely in the human genital tract, remains obscure. The same holds true of *Ureaplasma urealyticum* (previously known as T-mycoplasmas; T for tiny) that includes more than ten different serovars that are very common inhabitants of the genital tracts of males and females. Urethritis can be produced experimentally in

humans with *U. urea lyticum*, and the organism may be the etiological agent of some cases of nongonococcal urethritis.

A variety of the type species *M. mycoides* is the etiological agent of contagious pleuropneumonia of cattle and water buffaloes, a widespread infection of great socioeconomic importance, particularly in many countries in Africa, and in parts of India. Contagious pleuropneumonia and other infectious diseases of goats are caused by related organisms. *Mycoplasma agalactiae* causes contagious agalactia of goats and sheep, whereas *M. bovis* is an important cause of several diseases of cattle, including mastitis and arthritis. Lesions of the urogenital tract of cattle may be caused by this organism as well as by *M. bovigenitalium*. In pigs, *M. hyopneumoniae* is the most important pathogen, causing enzootic pneumonia, while *M. hyosynoviae* is a less frequent cause of a nonsuppurative, proliferative type of arthritis in piglets. *Mycoplasma gallisepticum* is responsible for chronic respiratory disease (tracheitis, air sacculitis, sinusitis) in chickens and sinusitis in turkeys and may cause severe economic loss; lesions in the brain associated with polyarteritis of the cerebral and meningeal arteries and arterioles have been reported in turkeys under natural and experimental conditions. Other pathogens for chickens and turkeys are *M. synoviae* and *M. meleagridis*. Though several species have been demonstrated in horses, dogs and cats, the possible role of these species in disease of their animal hosts is undetermined as yet. Among the mycoplasmas frequently occurring in small rodents, *M. arthritis* is a cause of purulent polyarthritis in rats, while *M. pulmonis* is highly pathogenic to both mice and rats, causing chronic respiratory disease, associated in up to 30% of female rats with oophoritis and salpingitis; acute purulent arthritis, that may lead to chronic lesions, can be produced in both rats and mice following intravenous inoculation. *Mycoplasma neurolyticum*, recovered from mice, is remarkable in that it produces a potent exotoxin having severe effects, under experimental conditions, on the central nervous system of mice and, to a lesser extent, of rats.

Members of the genus *Acholeplasma* were first isolated from sewage, compost and soil, but were later found to have a widespread distribution in a variety of tissues and secretions of almost every type of vertebrate examined. In addition, isolation of *Acholeplasmas* from plant material is being reported with increasing frequency. Yet there is no clear evidence at present that *Acholeplasma* species play a pathogenic role in either animals or plants.

The mycoplasmas classified in genus *Anaeroplasmata* are strict anaerobes, presumably non-pathogenic, found in the rumen of cattle and sheep.

The genus *Spiroplasma* includes a great variety of different species and serogroups found in vascular plant fluids (phloem), on the surface of leaves, flowers and fruits, and in the hemolymph and guts of insects that feed on plants. *Spiroplasma citri* is pathogenic for citrus plants (grapefruit and orange)

producing "stubborn" disease, a disease of considerable economic importance. Another spiroplasma is the cause of corn stunt disease. Although most spiroplasmas multiply in the tissues of insects without producing overt disease, one species, *S. apis*, is the cause of a devastating lethal infection of honeybees. *S. mirum* isolated from rabbit ticks produces ocular and nervous system disease on experimental infection of suckling small rodents and rabbits. See PLANT PATHOLOGY.

Mycoplasmalike organisms. Wall-less nonhelical mycoplasmalike organisms (MLO) can be demonstrated in great numbers by electron microscopy within sieve tube elements of plants affected by yellows diseases and in the salivary glands of insects serving as vectors of these diseases. Tetracycline antibiotics, to which mycoplasmas are known to be susceptible, have a marked suppressive effect on yellows plant diseases and on the number of mycoplasmalike organisms demonstrable in the phloem of plants and tissues of insect vectors. However, final classification of the mycoplasmalike organisms with the mycoplasmas (Mollicutes) must await cultivation of the organisms on artificial media. E. A. Freundt

Bibliography. M. F. Barile et al. (eds.), *The Mycoplasmas*, vols. 1-3, 1979; R. Dickesson, *Diagnostic Electron Microscopy: A Text-Atlas*, 1988; I. Kahane and A. Adoni (eds.), *Rapid Diagnosis of Mycoplasmas*, 1994; J. Maniloff et al., *Mycoplasmas: Molecular Biology and Pathogenesis*, 1992.

Mycorrhizae

Dual organs of absorption that are formed when symbiotic fungi inhabit healthy absorbing organs (roots, rhizomes, or thalli) of most terrestrial plants and many aquatics and epiphytes.

Mycorrhizae appear in the earliest fossil record of terrestrial plant roots. Roughly 80% of the nearly 10,000 plant species that have been examined are mycorrhizal. Present-day plants that normally lack mycorrhizae are generally evolutionarily advanced. It has been inferred that primitive plants evolved with a symbiosis between fungi and rhizoids or roots as a means to extract nutrients and water from soil.

Mycorrhizal dependency. The relative dependence of plants on mycorrhizal fungi has been viewed in different ways. Ecological dependence denotes that the plant must form mycorrhizae to survive, compete, and reproduce. Economic dependence means that the plant must form mycorrhizae to produce a commodity. The two types of dependence do not always coincide. A desert annual may need mycorrhizae to survive and produce seed, whereas some hay grasses may produce a large biomass without mycorrhiza formation.

The degree of dependence varies between species or groups of plants. In absolute dependence, characteristic of perennial, terrestrial plants, the host requires mycorrhizae to survive. Some plants are facultative; they may form mycorrhizae but do not

always require them. This group includes many of the world's more troublesome weeds. A minority of plant species characteristically lack mycorrhizae, so far as is known, including many aquatics, epiphytes, and annual weeds.

Functions. The three major types of mycorrhizae differ in structural details but have many functions in common. The fungus colonizes the cortex of the host root and grows its filaments (hyphae) into surrounding soil from a few centimeters to a meter or more. The hyphae absorb nutrients and water and transport them to host roots. The fungi thus tap far greater volumes of soil at a relatively lower energy cost than the roots could on their own. Moreover, many, if not all, mycorrhizal fungi produce extracellular enzymes and organic acids that release immobile elements such as phosphorus and zinc from clay particles, or phosphorus and nitrogen bound in organic matter. The fungi are far more physiologically capable in extracting or recycling nutrients in this way than the rootlets themselves.

Mycorrhizal fungi are relatively poorly competent in extracting carbon from organic matter. They derive energy from host-photosynthesized carbohydrates. Hosts also provide vitamins and other growth regulators that the fungi need.

Types. The major types are ectomycorrhizae, vesicular-arbuscular mycorrhizae, and ericoid mycorrhizae.

Ectomycorrhizae. Ectomycorrhizae, the most readily observed type, and some related variants form with members of the Pinaceae, Betulaceae, Fagaceae, Salicaceae, Myrtaceae, and several other families of woody perennials. Ectomycorrhizal hosts strongly depend on mycorrhizae to survive. Relatively few in number of species, they nonetheless dominate most forests outside the tropics.

The species of ectomycorrhizal fungi number in the thousands, being mostly mushroom or truffle formers in their reproductive stage (Ascomycotina, Basidiomycotina, and a few Zygomycotina). Many are host-specific, forming mycorrhizae only with certain trees such as alders, Douglas-fir, or eucalypts, respectively. Others colonize roots of a wide variety of hosts. Spores of ectomycorrhizal fungi are dispersed from mushrooms by air or, for the belowground fruiting truffles and trufflelike fungi, by being eaten by animals which later excrete them.

Ectomycorrhizae form on tiny host feeder rootlets. The fungus envelops the rootlet with a hyphal sheath, which lends its color to the rootlet surface. The color is often white but ranges to brown, yellow, pink, green, blue, or black, and is easily seen with a hand lens. Brightly colored hyphae can often be seen extending from the mycorrhiza into the surrounding substrate. They also grow from the sheath into the rootlet between the cells of the epidermis and cortex. In this region of intimate intergrowth, the fungi exchange nutrients and water for photosynthates.

Ectomycorrhizal fungi produce auxins and cytokinins that change morphology of host rootlets. A nonmycorrhizal rootlet of these hosts, uncommon

in nature, has a single tip and abundant root hairs and lives only a few months at best. When colonized by the ectomycorrhizal fungus, however, the rootlet usually branches and root-hair formation is suppressed. Ectomycorrhizae may be active for years if they do not shrivel from drought or freezing. The fungi also protect rootlets from pathogens by surrounding them with the hyphal mantle as a physical barrier and often by producing antibiotics as a chemical deterrent.

Vesicular-arbuscular. Vesicular-arbuscular mycorrhizae (sometimes simply termed arbuscular mycorrhizae) form with the great majority of terrestrial herbaceous plant species plus nearly all woody perennials that are not ectomycorrhizal (except for the Ericaceae, the heath family, as discussed below). Some ectomycorrhizal hosts, such as alders, eucalypts, and willows, can also form vesicular-arbuscular mycorrhizae. Vesicular-arbuscular mycorrhizal hosts range from strongly mycorrhiza-dependent, especially the woody perennials, to facultative, as are many grasses. Vesicular-arbuscular mycorrhizae are difficult to observe without special biological staining procedures, because the inconspicuous hyphae grow diffusely from the rootlets into the soil.

The fungal associates of vesicular-arbuscular mycorrhizae are members of the order Glomales of the Zygomycotina and, as presently known, number only about 200 species around the world. In general, vesicular-arbuscular mycorrhizal fungi show little or no host specificity, although different species may interact with hosts in different ways in different habitats. Such fungi sporulate as individual spores or clusters of spores produced among roots and are spread by movement of the soil in which they occur. The contrast between the low species numbers of vesicular-arbuscular-mycorrhizal fungi and the thousands of ectomycorrhizal fungi fuels speculation on evolution and ecology of the two kinds of associations. Judging from the fossil record, vesicular-arbuscular mycorrhizal fungi are more primitive than the ectomycorrhizal fungi and apparently have been relatively genetically stable for at least 400 million years. See ASCOMYCOTA; ZYCOMYCOTA.

Vesicular-arbuscular mycorrhizal fungi grow into the root cortex and penetrate root cells to form two kinds of specialized structures. All form arbuscules ("little trees") in some host cells, bushy structures that are organs for exchange of nutrients and carbon compounds between fungus and host. Arbuscules persist for several days, then are digested by the host cell. Many vesicular-arbuscular mycorrhizal fungi produce large, inflated bodies (vesicles) in other host cells. These fill with lipids, presumably as energy storage for the fungus. As the root senesces, vesicles may develop thickened walls to become asexual spores.

Vesicular-arbuscular mycorrhizal fungi seem neither to produce growth regulators nor to markedly change root configuration. Rather than suppress

root-hair formation, they often enter the rootlet through root hairs. They form no physical barrier to pathogens and show little or no antibiosis against pathogens.

Ericoid. Ericoid mycorrhizae are restricted to the Ericales, the heath order. The hosts are strongly mycorrhiza-dependent. Though relatively few in number, heath species dominate large areas around the world and are common understory plants in many forests. The fungi involved are poorly known but apparently are mostly Ascomycotina and probably some Basidiomycotina. They seem host-specific to the Ericales, and their spores are forcibly discharged to the air for dispersal. Ericoid fungi typically enter the outermost rootlet cells and fill them with masses of hyphae. These masses are sites of exchange of substances between host and fungus. The hyphae grow from these outer host cells into surrounding soil. See ERICALES.

Other types. Other mycorrhiza types include those special for the Orchidaceae (orchids) and Gentianaceae (gentians). Miscellaneous fungi commonly colonize roots to form various structures not mentioned above, often with nonmycorrhizal plants. The rootlets remain healthy, and such fungi may function with host roots as dual absorbing organs. Thus, they would fit the definition of mycorrhizae, but they are too poorly understood to be certain.

Ecology. The availability of propagules of mycorrhizal fungi dictates patterns of plant establishment and community development because of the differing degrees of dependency of different hosts. New soils, such as those in the front of retreating glaciers or emerging from the sea because of geologic forces, initially have few or no propagules of mycorrhizal fungi. The first plants to become established, therefore, are nonmycorrhizal or facultatively mycorrhizal species such as saxifrages, sedges, and rushes. Under these circumstances, no matter how abundant the seed source, mycorrhiza-dependent species cannot extract enough nutrients from the new substrate to survive. If abundant mushrooms of ectomycorrhizal or ericoid mycorrhizal fungi fruit in the vicinity, the new soil can soon become inoculated by aerial spore dispersal. Hence hosts such as the Ericaceae or ectomycorrhizal willows can establish along with or soon after the nonmycorrhizal species. Vesicular-arbuscular mycorrhiza-dependent species cannot gain a foothold until spore-bearing soil is introduced to the site by insects, animals, or erosion. After the eruption of Mount St. Helens in Washington covered a vast area with infertile, volcanic ash and pumice, patches of vesicular-arbuscular mycorrhiza-dependent plants soon appeared on gopher mounds. Protected underground when the mountain exploded, gophers later burrowed up through the ash layer to deposit the original, ash-buried soil from their tunneling on the surface. This soil contained spores of vesicular-arbuscular mycorrhizal fungi that enabled the vesicular-arbuscular mycorrhiza-dependent plants to survive.

The succession of plants from pioneering through seral to climax communities is governed by availabil-

ity of mycorrhizal propagules. When catastrophic fire, erosion, or clearcutting reduce the availability of mycorrhizal fungi in the soil, plants dependent on those fungi will have difficulty becoming established. In clearcutting of conifer forests, overstory ectomycorrhizal trees are removed, thereby removing the food base for the ectomycorrhizal fungi. In contrast, ericoid or vesicular-arbuscular mycorrhizal understory plants may be damaged but often not killed. These plants retain their mycorrhizal fungi and respond to release from the overstory trees. The ericoid or vesicular-arbuscular mycorrhizal fungi build up in the soil as their host plants flourish to strongly compete for soil nutrients and moisture with ectomycorrhizal fungi that must start anew on establishing conifer seedlings.

Fungal host specificity also can affect species establishment and community composition of ectomycorrhizal and ericoid hosts. Heathlands typically are slow to convert to conifer stands, in part because the ericoid fungi so dominate the soil that roots of ectomycorrhizal hosts may not contact ectomycorrhizal fungal propagules. Eucalypts associate with their own, specific array of ectomycorrhizal fungi. Their fungi will not form mycorrhizae with trees of the Northern Hemisphere, so where plantations of inoculated eucalypts are established outside their native range, their fungi soon fully occupy the soil. Native ectomycorrhizal hosts have great difficulty invading such plantations for lack of their own ectomycorrhizal fungi.

Biodiversity. Each mycorrhizal fungus has its own array of physiological characteristics. Some are especially proficient at releasing nutrients bound in organic matter, some produce more effective antibiotics or growth regulators than others, and some are more active in cool, hot, wet, or dry times of year than others. Healthy plant communities or crops typically harbor diverse populations of mycorrhizal fungal species. This diversity, evolved over a great expanse of time, is a hallmark of thriving ecosystems. Factors that reduce this diversity also reduce the resilience of ecosystems.

Applications. Mycorrhizal inoculation of plants in nurseries, orchards, and fields has succeeded in many circumstances, resulting in improved survival and productivity of the inoculated plants. Inoculation with selected fungi is especially important for restoring degraded sites or introducing exotics such as pines in the Southern Hemisphere. Because ectomycorrhizal fungi include many premier edibles such as truffles, seedlings can also be inoculated to establish orchards for production of edible fungi.

In general, however, management of nursery, field, or forest soils to preserve or enhance the vigor of desired mycorrhizal fungi is the sound approach. For example, fumigation of nursery soils to control pathogens may eradicate mycorrhizal fungi and other, beneficial soil organisms. Alternative pest-control measures that do not damage mycorrhizal fungi are preferred. If organic matter is maintained, erosion minimized, and fertilization rationally

applied in field and forest soils, healthy and diverse populations of mycorrhizal fungi can more likely be sustained. See FOREST SOIL; FUNGI. James M. Trappe

Bibliography. M. F. Allen, *Mycorrhizal Functioning: An Integrative Plant-Fungal Approach*, 1992; G. C. Carroll and D. T. Wicklow, *The Fungal Community, Its Organization and Role in the Ecosystem*, 1992; A. H. Fitter et al., *Ecological Interactions in Soil: Plants, Microbes and Animals*, 1985; V. Gianinazzi-Pearson and S. Gianinazzi, *Physiological and Genetical Aspects of Mycorrhizae*, 1986; D. J. Read et al., *Mycorrhizas in Ecosystems*, 1992; A. Varma and B. Hock, *Mycorrhiza: Structure Function, Molecular Biology, and Biotechnology*, 2d ed., 1998.

Mycotoxin

Any of the mold-produced substances that may be injurious to vertebrates upon ingestion, inhalation, or skin contact. The diseases they cause, known as mycotoxicoses, need not involve the toxin-producing fungus. Diagnostic features characterizing mycotoxicoses are the following: the disease is not transmissible; drug and antibiotic treatments have little or no effect; in field outbreaks the disease is often seasonal; the outbreak is usually associated with a specific foodstuff; and examination of the suspected food or foodstuff reveals signs of fungal activity.

Historically, the earliest recognized mycotoxicoses were human diseases. Ergotism, or St. Anthony's fire, which results from eating rye infected with *Claviceps purpurea*, has been known since the Middle Ages. Yellow rice disease, a complex of human toxicoses caused by several *Penicillium islandicum* mycotoxins, was reported in Japan during the late nineteenth century. In Russia, large-scale human mycotoxicoses involving massive hemorrhaging were reported in the early twentieth century. Large numbers of deaths occurred as a result of eating overwintered grain contaminated with *Fusarium poae*. See ERGOT AND ERGOTISM.

Aflatoxins. World attention was directed toward the mycotoxin problem with the discovery of the aflatoxins in England in 1961. The aflatoxins, a family of mycotoxins produced by *Aspergillus flavus* and *A. parasiticus*, can induce both acute and chronic toxicological effects in vertebrates. Aflatoxin B₁, the most potent of the group, is toxic, carcinogenic, mutagenic, and teratogenic. There appears to be a significant dose-response correlation between estimated ingestion of aflatoxin-contaminated foods and primary liver cancer, particularly in Thailand, Kenya, and Mozambique. Major agricultural commodities that are often contaminated by aflatoxins include corn, peanuts, rice, cottonseed, and various tree nuts. When cows consume aflatoxin-contaminated feeds, aflatoxin M₁, a derivative of the toxin, is excreted in the milk. This substance is both toxic and carcinogenic. The aflatoxin problem originates with fungal contamination of crops in the field, thus making control difficult. Moisture levels of 15% or more during commodity storage may exacerbate the contamination level. Regulations on allowable aflatoxin levels exist in most countries, ranging from 0 to 50 parts per billion. Detoxification using ammonia is effective, but has not been approved for commercial use by regulatory agencies. Promising studies on field control through use of biocompetitive agents and plant genetic engineering are under way. See AFLATOXIN.

Other mycotoxins. There are additional mycotoxins of varying medical and economic importance. Those of major concern are listed in the **table**. Additional mycotoxins occur, but detection is usually difficult because they are normally produced in very low concentrations in natural substrates. The problem is also compounded by synergistic toxicological activity when several toxins are present. See MEDICAL MYCOLOGY; TOXIN. Alex Ciegler; Maren Klich

Bibliography. B. Arora and L. Arora (eds.), *Mycotoxins in Ecological Systems*, 1991; V. Betina, *Mycotoxins: Chemical, Biological, and Environmental Aspects*, 1989; G. A. Bray, *Mycotoxins, Cancer, and Health*, 1991; K. Sinha and D. Bhatnagar (eds.), *Mycotoxins in Agriculture and Food Safety*, 1998.

Compounds involved in mycotoxicoses

| Toxin | Major producing fungi | Susceptible host | Biological effects |
|----------------------------|--|-----------------------|---|
| Aflatoxins | <i>Aspergillus flavus</i> , <i>A. parasiticus</i> | Mammals, fish | Hepatotoxic, cancer |
| Penitrem A | <i>Penicillium crustosum</i> , <i>P. clavigerum</i> | Cattle, horses, sheep | Tremorogenic, convulsant |
| T-2 | <i>Fusarium sporotrichioides</i> | Cattle, humans? | Dermal necrosis, hemorrhage |
| Zearalenone | <i>Fusarium graminearum</i> , <i>F. culmorum</i> | Swine | Vulvovaginitis, abortion |
| Sporodesmins | <i>Pithomyces chartarum</i> | Swine, sheep | Hepatotoxic, facial eczema |
| Ochratoxin A | <i>Penicillium verrucosum</i> , <i>Aspergillus ochraceus</i> | Swine, humans? | Nephrotoxic |
| Citrinin | <i>Penicillium citrinum</i> , <i>Aspergillus terreus</i> | Swine | Nephrotoxic |
| Deoxynivalenol | <i>Fusarium graminearum</i> , <i>F. culmorum</i> | Swine, humans? | Vomiting |
| Maltoryzine | <i>Aspergillus oryzae</i> | Cattle | Death |
| Secalonic acids D and F | <i>Penicillium oxalicum</i> , <i>Aspergillus aculeatus</i> | Farm animals | Death |
| Satratoxins | <i>Stachybotrys atra</i> | Horses, humans | Hemorrhage |
| Fumonisin | <i>Fusarium moniliforme</i> , <i>F. proliferatum</i> | Humans, horses, swine | Cancer, neurotoxic, edema |
| Patulin | <i>Penicillium expansum</i> , <i>Aspergillus clavatus</i> | Humans | Immunosuppressant, nephrotoxic, hemorrhage |
| Cyclopiazonic acid | <i>Aspergillus flavus</i> , <i>Penicillium commune</i> | Mammals | Neurotoxic, hemorrhage |

Myiasis

The infestation of vertebrates by the larvae, or maggots, of numerous species of flies. These larvae may invade different parts of the bodies of these animals or may appear externally (see **table**). Some invertebrates, such as spiders, may be attacked by species of Sarcophagidae, the flesh flies.

In cutaneous myiasis, the larvae are found in or under the skin. There may be a migration of some species of these larvae through host tissues, resulting in a swelling with intense itching. Such a condition is known as larva migrans, or creeping eruption, and may require surgical treatment.

Intestinal myiasis in humans is usually the result of accidentally swallowing the eggs or larvae of

Some species of the families of Diptera causing myiasis

| Common name | Scientific name | Host and site of infestation |
|--|--------------------------------------|--|
| Muscidae (houseflies) | | |
| Houseflies | <i>Musca domestica</i> | Humans; gastrointestinal tract |
| | <i>Muscina stabulans</i> | Numerous animals; intestinal tract |
| Lesser housefly | <i>Fannia canicularis</i> | Humans; intestinal tract |
| Latrine fly | <i>Fannia scalaris</i> | Humans; intestinal tract |
| Sarcophagidae (flesh flies, blowflies, scavenger flies) | | |
| Gray blowfly | <i>Sarcophaga haemorrhoidalis</i> | Humans and numerous animals; intestinal tract, cutaneous and genitourinary tract |
| Gray blowfly | <i>Sarcophaga carnaria</i> | Humans and numerous animals; intestinal tract, cutaneous and genitourinary tract |
| Flesh fly | <i>Wohlfahrtia vigil</i> | Skin of infants; cutaneous |
| Fox maggot | <i>Wohlfahrtia opaca</i> | Humans, fox, and mink; cutaneous |
| Old World screwworm | <i>Wohlfahrtia magnifica</i> | Eyes, ears, and nose of humans and animals; cutaneous |
| Tumbu fly, ver du cayor | <i>Cordylobia anthropophaga</i> | Humans, domestic and wild animals; cutaneous |
| Congo floor maggot | <i>Aucheromyia luteola</i> | Humans; suck blood |
| Calliphoridae (blowflies) | | |
| American screwworm | <i>Callitroga americana</i> | Humans; cutaneous |
| Secondary screwworm | <i>Callitroga macellaria</i> | Sheep and goat; cutaneous |
| Old World screwworm | <i>Chrysomya bezziana</i> | Humans and animals; cutaneous |
| | <i>Chrysomya megacephala</i> | Sheep; cutaneous |
| | <i>Phoenicia cuprina</i> | "Strike" of sheep; cutaneous |
| Surgical maggot | <i>Phoenicia sericata</i> | Humans, sheep, and goat; wounds |
| Bluebottle fly | <i>Calliphora vomitoria</i> | |
| Greenbottle fly, surgical maggot | <i>Lucilia caesar</i> | Humans; wounds |
| Blackbottle fly, fleece worm | <i>Phormia regina</i> | Sheep; cutaneous |
| None | <i>Apaulina</i> sp. | Birds, especially nestlings; suck blood |
| Gasterophilidae (horse bots) | | |
| Botfly | <i>Gasterophilus intestinalis</i> | Stomach of horse |
| Nose fly | <i>Gasterophilus haemorrhoidalis</i> | Stomach and duodenum of horse |
| Chin fly | <i>Gasterophilus nasalis</i> | Mouth, throat of horse |
| None | <i>Gasterophilus inermis</i> | Rectum of horse |
| None | <i>Gasterophilus pecorum</i> | Pharynx, stomach, and rectum of horse |
| Cuterebridae (robust botflies) | | |
| Gusano del monte | <i>Dermatobia hominis</i> | Cattle and humans; cutaneous |
| None | <i>Cuterebra</i> sp. | Rodents and small mammals |
| None | <i>Bogeria</i> sp. | Rodents and small mammals |
| Oestridae (botflies, gadflies, nose flies) | | |
| Sheep botfly | <i>Oestrus ovis</i> | Sometimes eyes of sheep and antelope; cutaneous |
| Nose fly | <i>Rhinoestrus purpureus</i> | Horse, hippopotamus, nose and eyes of humans |
| Warble fly | <i>Hypoderma lineata</i> | Warbles of cattle and reindeer |
| None | <i>Hypoderma bovis</i> | Warbles of cattle and reindeer |
| None | <i>Hypoderma crossi</i> | Goat, India; subcutaneous |
| None | <i>Cephalopina titillator</i> | Camel; nostrils and nasopharynx |
| None | <i>Kirkioestrus</i> sp. | African antelope; nasal passages |
| None | <i>Gadoelstia</i> sp. | African antelope; nasal passages |
| None | <i>Cephalopsis</i> sp. | Camel; nasal passages |
| Head botfly | <i>Cephenemyia</i> sp. | Deer; head bots |
| Stratiomyidae (soldier flies) | | |
| Soldier fly | <i>Hermetia illucens</i> | Recorded from intestine of humans |
| Syrphidae (flower flies) | | |
| Rat-tailed maggot | <i>Eristalis tenax</i> | Recorded from stomach of humans |

these flies. It occurs commonly in many herbivores who ingest the eggs when feeding on contaminated herbage. The larvae settle in the stomach or intestinal tract of the animal host.

Cavity, or wound, myiasis occurs when the larvae invade natural orifices, such as the nasopharynx, vulva, and sinuses, or artificial openings such as wounds. A study of wound myiasis led to the use of the larvae of *Lucilia sericata* and *Phormia regina* in treating osteomyelitis and suppurating wounds. These maggots commonly feed on necrotic tissue; there is a danger, however, that normal tissue may be attacked. Maggots used for this purpose are reared under sterile conditions. It was also discovered that allantoin, one of the compounds found in the urine of the maggots, has a healing effect on ulcers.

External myiasis includes infestation by those maggots which are blood feeders.

Of considerable economic importance is myiasis of such domestic animals as horses, sheep, or cattle. Cattle afflicted with cutaneous myiasis produce poor-grade hides full of small perforations. Heavy infestations of cattle by screwworms result in the death of the host if not treated. Horses afflicted with the stomach bot become emaciated and may succumb to the ravages of these larvae.

The Diptera of medical and veterinary importance are largely confined to the families Oestridae, Calliphoridae, and Sarcophagidae. An interesting ecological phenomenon has been observed in the life cycle of the human botfly. The female oviposits on the undersurface of the abdomen of mosquitoes as they emerge from their aquatic habitats. When the mosquito takes a blood meal from the vertebrate host, the eggs are in close contact with the skin of the host. Heat from the host's body causes the larvae to hatch from the eggs, and these tiny larvae then burrow into the skin and cause a creeping eruption. See DIPTERA; MEDICAL PARASITOLOGY. Charles B. Curtin

Bibliography. J. W. Beck and J. E. Davies, *Medical Parasitology*, 3d ed., 1981; M. W. Sloss and R. L. Kemp, *Veterinary Clinical Parasitology*, 6th ed., 1994; F. Zumpt, *The Stomoxysine Biting Flies of the World (Diptera, Muscidae): Taxonomy, Biology, Economic Importance and Control*, 1973.

Myliobatiformes

An order of batoid fishes (subclass Elasmobranchii), the typical members of which are described by the following characteristics: The disc is strongly depressed and varies from oval longitudinally to much broader than long; the tail is well marked off from the body sector, very short to long and whiplike, and equipped with a poisonous spine in some species; the pectoral rays are either continuous along the side of the head or separate from the head and modified to form rostral lobes or finlike rostral appendages (cephalic fins); the dorsal fin, if present, is near the base of the tail; and development is ovoviviparous. Sometimes an individual ray may have two or three, very rarely four, tail spines rather than the usual one.

Multiple spines result from the failure of older spines to shed before new ones develop. Most species are tropical or subtropical, although some species occur in warm temperate and cool temperate zones. The usual habitat is shallow shore waters and upper continental and insular slopes. Some, such as the eagle rays, may venture far out to sea, but none are considered truly oceanic rays. See BATOIDEA; ELASMOBRANCHII.

Phylogeny

In the past two decades, studies on the phylogeny of batoids have led to some major changes in their classification, one of which results in the inclusion of the platyrhinids and *Zanobatus* in Myliobatiformes. Although these fishes superficially resemble guitarfishes more than myliobatoids, they share certain esoteric skeletal features with myliobatoids. The following is the most recently proposed classification of the order.

Suborder Platyrhinoidei. The family Platyrhinidae (thornbacks) has only two extant genera (*Platyrhina*, two species; and *Platyrhinoidis*, one species). It is distinguished by a round disc, stout tail sector, two large dorsal fins, large caudal fin, and one or three rows of stout, knobby thorns on the disc and tail. It occurs in tropical to cool-temperate waters on the continental shelves of the North Pacific in Asia and North America, including California and Mexico. A Late Cretaceous fossil from Italy, *Tethybatis*, is assigned to this family. See CRETACEOUS.

Suborder Zanobatoidei. The family Zanobatidae (Panrays) consists of one species, *Zanobatus schoenleiniti*, which occurs in the eastern Atlantic from Morocco to the Gulf of Guinea. It is similar to Platyrhinidae but has a less robust tail sector.

Suborder Myliobatoidei. The family Hexatrygonidae (sixgill stingrays) consists of one species, *Hexatrygon bickelli*, and is very unusual in possessing the following features: The disc is longer than wide; there are six rather than five gill arches and gill slits as in other batoids; the snout is elongate, flattened, and translucent; large spiracles are placed well behind the eyes; the spiracle is closed by an external valve. In contrast, other rays have an internal valve; a barbed spine occurs on the tail; and maximum length of 100 cm (39 in.). This unusual ray occurs on the continental and insular slopes of the Indo-West Pacific.

The family Plesiobatidae (deepwater stingrays) has one species, *Plesiobatis daviesi*. Its round disc is bordered in black; the anterior profile is obtuse with a pointed snout; a barbed spine is present on the tail; the nasal curtain is incompletely united and does not reach the mouth, a character this ray shares with *Hexatrygon*; and maximum length is 2.7 m (8.8 ft). It occurs on continental and insular slopes of the Indo-West Pacific to a depth of 680 m (2230 ft) [Fig. 1].

In the family Urolophidae (round stingrays), the disc usually is only slightly broader than long; the tail sector is short to moderately long and bears a barbed spine; the caudal fin is well developed, with

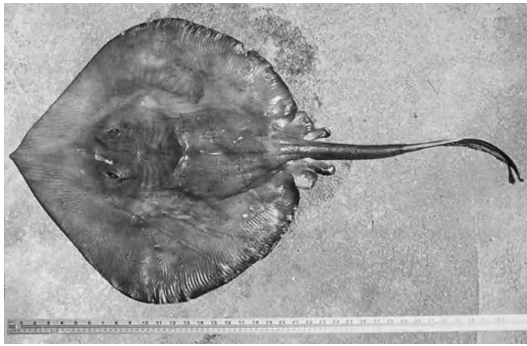


Fig. 1. *Plesiobatis daviesi*. (Photo courtesy of R. F. Myers)

cartilaginous radial supports; and a dorsal fin is present or absent. Round stingrays occur in the western Pacific on continental shelves and slopes, as well as inshore on reefs and grass beds. The family comprises two genera and 24 species (probably more), the smallest of which is 17 cm (6.7 in.) in maximum disc length, the largest 80 cm (31.5 in.).

Members of the family Urotrygonidae (American round stingrays) are similar to urolophids, but differ in having a slender tail, which is about equal to the length of the disc, a distinct caudal fin and poisonous spine, and no dorsal fin. They occur on continental shelves in tropical and warm temperate waters of the western Atlantic and eastern Pacific. The maximum disc length varies from 12 to 46 cm (4.7 to 18 in.), depending on the species.

Members of the family Dasyatidae (whiptail stingrays) have a long slender or whiplike tail armed with a serrated poisonous spine and lack a dorsal and caudal fin. As in round stingrays, the disc is not more than 1.3 times broad as long. Settled over the substrate, these fishes, as well as other bottom-dwelling rays, fan their "wings," sending up plumes of sand or mud that covers the body, leaving only the eyes, spiracles, and part of the tail exposed. Fanning the substrate exposes crustaceans, mollusks, and polychaete worms that provide the principal part of the rays' diet. Small fishes and squid are consumed in lesser quantity. The six genera and about 71 species of whiptail stingrays occur for the most part on continental and insular shelves of tropical and warm temperate zones of the Pacific, Indian, and Atlantic oceans, including the Mediterranean Sea. Some species are common in brackish water, and some make temporary excursions into freshwater. Species of the family vary in disc width from 11 to 220 cm (4.3 in. to 7 ft).

Urolophids, urotrygonids, and dasyatids, collectively known as stingarees, are potentially dangerous to humans. The piercing tail spine delivers a poison that causes immediate and excruciating pain, followed by swelling and most often a serious infection. When a stingaree is stepped on, the reflex action is to strike with the barb. Stings may be avoided by dragging the feet while wading. In this manner, the ray is nudged on its wing, causing it to move away. On clearwater beaches, search the bottom carefully before stepping from a boat.

Potamotrygonidae (river stingrays) is the only family of chondrichthyan fishes restricted to a freshwater environment. Known commonly as river stingrays, members have low concentration of urea in the body fluids and a rectal gland for salt excretion. The family is further characterized by the following: There is a round flat disc, the dorsal surface of which is usually marked with spots, ocelli, or reticulate and vermiform color patterns; the pre-pelvic process is greatly expanded; the disc and tail are usually armed with denticles, thorns, or tubercles; dorsal and caudal fins are absent; development is ovoviviparous, but essentially viviparous in that embryos are nourished by uterine "milk," a reproductive strategy characteristic of other myliobatoids. The three genera and 20 species of river stingrays occur in rivers of South America that drain into the Atlantic Ocean and Caribbean Sea. Maximum length of the disc is 25 to 100 cm (10 to 40 in.), depending on the species.

Members of the family Gymnuridae (butterfly rays) are very distinctive in having a disc more than 1.5 times broad as long and a short tail lacking a caudal fin, as well as lacking a dorsal fin and tail spine in some species (Fig. 2). They occur on tropical and temperate continental shelf waters of the Atlantic, Indian, and Pacific oceans. The largest species, *Gymnurus altavela*, of both sides of the Atlantic, including the Mediterranean Sea, reaches 400 cm (13.7 ft); disc width most of the 11 species (in two genera) are far smaller.

In the family Myliobatidae (eagle rays), the margins of the pectoral fins are deeply indented or entirely interrupted just posterior to the eyes, sharply distinguishing the anterior part of the head from the rest of the disc. Anterior subdivisions of the pectoral fins form a separate lobe (or lobes), which in the manta ray and the devil rays form cephalic fins. The eyes and spiracles are on the sides of the head. The family is represented in the tropical and warm temperate waters of the Atlantic, Indian, and Pacific oceans. It inhabits continental and insular shelves to offshore, but none of the species are considered oceanic inhabitants. There are three subfamilies.

In the subfamily Myliobatinae (eagle rays), the anterior subdivision of the pectoral fins forms a single subrostral lobe extending forward below the front of

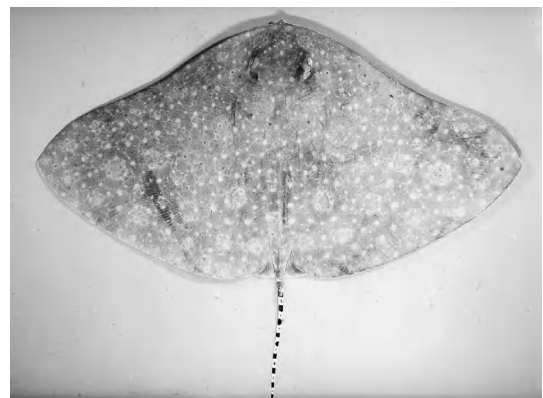


Fig. 2. *Gymnura poecilura*. (Photo courtesy of J. E. Randall).

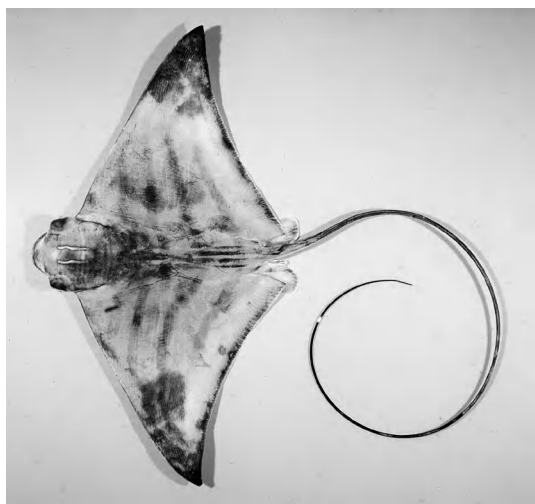


Fig. 3. *Aetomylaeus nichoffii*, a banded eagle ray. (Photo courtesy of J. E. Randall)

the head. The crown of the head is conspicuously elevated, and the eyes and spiracles are laterally placed. The tail is much longer than the disc and is armed with a long serrated spine in some species; a dorsal fin is present near the base of the tail, but a caudal fin is absent. Eagle rays have the ability to leap from the water and high into the air. The smallest and largest species are 65 cm (2 ft) and 300 cm (10 ft) in disk width (Fig. 3).

In the subfamily Rhinopterinae (cownose rays), the subrostral lobe extending forward from the lower surface of the head is deeply incised, leaving two distinct lobes, and the concave anterior contour of the cranium leaves the front of the head bilobed, resulting in a configuration reminiscent of a cow's nose. Cownose rays and eagle rays subsist chiefly on bivalve mollusks, especially oysters and clams, and to a lesser extent on crustaceans. Of the seven species in a single genus, the smallest attains a disc width of 86 cm (33 in.), the largest 213 cm (7 ft).

In the subfamily Mobulinae (devil rays), the anterior subdivisions of the pectoral fins are in the form of two tongue-like cephalic fins, which are widely separated from each other and curve forward from the front of the head. Unlike the eagle rays and cownose rays, which have several rows of large pavement-like teeth, the mobulids have minute teeth in many series. The disc is relatively thin; the head is broad and its dorsal surface is only slightly convex. Devil rays subsist on planktonic crustaceans and small fishes that are directed into the mouth by the cephalic fin and then strained by gill plates located on the inner surface of the gill slits, which are longer than those of eagle and cownose rays. These rays are famous for their large sizes and spectacular leaps out of the water. The family comprises two genera: *Manta*, with one species, *Manta birostris*, commonly known as the giant manta or giant devil ray; and *Mobula*, with nine species, collectively known as the devil rays.

Manta birostris has a very broad head (its interorbital distance is about one-quarter of the great-

est width of the disc), relatively large forward-projecting-cephalic fins, and a terminal mouth extending across the front of the head, with no teeth in the upper jaw. The dorsal surface of the disc is covered with denticles, and the tail is essentially without a spine. The ray ranges worldwide in tropical and warm temperate zones near coral and rocky reefs and sometimes over deep water, as well as in shallow bays and the mouths of large rivers. Up to 800-cm (26-ft) disc width has been reported.

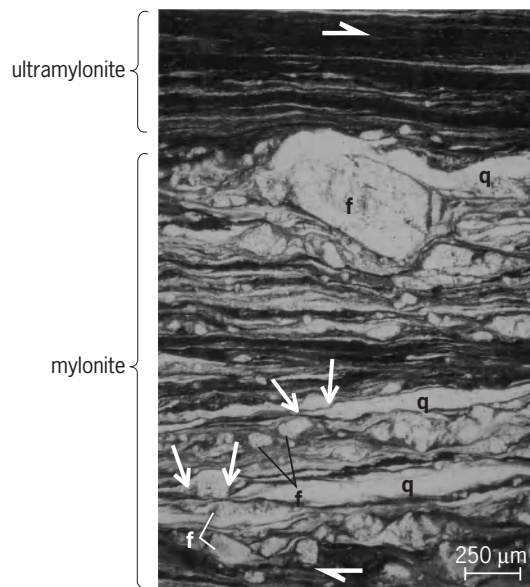
Mobula differ from the giant manta in that the head is not nearly as wide, the mouth is on the lower surface of the head, and teeth are in both jaws. Members of the genus occupy tropical and warm temperate belts of all oceans, including the Mediterranean Sea. The known maximum disc width varies from 100 to 520 cm (3 to 17 ft), depending on the species. Herbert Boschung

Bibliography. L. J. V. Compagno, Checklist of living elasmobranchs, pp. 471–498 in W. C. Hamlett (ed.), *Sharks, Skates, and Rays: The Biology of Elasmobranch Fishes*, Johns Hopkins University Press, Baltimore, 1999; L. J. V. Compagno, Checklist of Chondrichthyes, pp. 503–547 in W. C. Hamlett (ed.), *Reproductive Biology and Phylogeny of Chondrichthyes: Sharks, batoids and chimaeras*, Science Publishers, Enfield, NH, 2005; J. D. McEachran and N. Aschliman, Phylogeny of Batoidea, pp. 79–113 in J. C. Carrier, J. A. Musick, and M. R. Heithaus (eds.), *Biology of Sharks and Their Relatives*, CRC Press, Boca Raton, FL, 2004; J. D. McEachran and M. R. de Carvalho, Batoid Fishes, pp. 507–589 in L. K. E. Carpenter (ed.), *The Living Marine Resources of the Western Central Atlantic: FAO Species Identification Guide for Fishery Purposes*, vol. 1, FAO, Rome, 2002; J. D. McEachran, K. A. Dunn, and T. Miyake, Interrelationships of the batoid fishes (Chondrichthyes: Batoidea), pp. 63–84 in M. L. J. Stiassny, L. R. Parenti, and G. D. Johnson (eds.), *Interrelationships of Fishes*, Academic Press, San Diego, 1996; J. S. Nelson, *Fishes of the World*, 4th ed., Wiley, New York, 2004.

Mylonite

A rock that has undergone significant modification of original textures by predominantly plastic flow due to dynamic recrystallization. Mylonites form at depth beneath brittle faults in continental and oceanic crust, in rocks from quartzo-feldspathic to olivine-pyroxenite composition. Mylonites were once confused with cataclases, which form by brittle fracturing, crushing, and comminution. Microstructures that develop during mylonitization vary according to original mineralogy and modal compositions, temperature, confining pressure, strain, strain rate, applied stresses, and presence or absence of fluids.

At low to moderate metamorphic grades, mylonitization reduces the grain size of the protolith and commonly produces a very fine-grained, well-foliated rock with a pronounced linear fabric defined by



Photomicrograph of greenschist-grade mylonite and ultramylonite derived from granodiorite, Borrego Springs shear zone, southern California. Half-arrows indicate right-lateral shear sense. Feldspar porphyroclasts (f), some with recrystallized tails (indicated by arrowheads), appear in matrix of recrystallized quartz ribbons (q) and dark, fine-grained biotite. Characteristic darkness of ultramylonite zone results from extremely fine recrystallized grain size of component minerals.

elongate minerals. Lineations may be weak or absent, however, in high-strain zones that lack a significant rotational component. At high metamorphic grades, grain growth during mylonitization can produce a net increase in grain size, and the term mylonitic gneiss is used where there is a preserved or inferred undeformed protolith.

At low temperatures, high strain rates, and high applied stresses, most minerals are strong, and they deform by brittle failure. A mylonite forms at and above the conditions favorable for dynamic recrystallization of the weakest mineral phase present, provided there is sufficient weak phase to allow the bulk rock to flow. In quartzo-feldspathic rocks, mylonitization occurs above 480–570°F (250–300°C), whereupon quartz flows plastically at natural strain rates by the mechanisms of dislocation glide and dislocation creep. Under these conditions, feldspar remains brittle and is often preserved as broken megacrysts, or porphyroclasts, surrounded by fine-grained, recrystallized ribbons of quartz (see *illus.*).

At temperatures above about 840°F (450°C), feldspar also dynamically recrystallizes, to produce a narrow mantle of small recrystallized grains around a more rigid core. The tiny new grains are swept into tails around the porphyroclast; and, if asymmetrically developed, they record the sense of shear (see *illus.*). Other common asymmetrical microstructures in mylonites include shear bands, oriented crystallographic axes, recrystallized grain shapes, microfolds, and microfaults.

Protomylonites form at lowest strains and preserve some protolith textures. Blastomylonites con-

tain grains that have recrystallized and also grown in size significantly during or after mylonitization. Rocks of the type known as S-C mylonites contain two distinct but related foliations (S = schistosity; C = cisaillement, or shear bands) that give the overall shear sense for the rock. Ultramylonites form at highest strains and contain few relict porphyroclasts in an extremely fine-grained matrix (see *illus.*). See METAMORPHIC ROCKS. Carol Simpson

Bibliography. G. S. Lister and A. W. Snoke, S-C mylonites, *J. Struct. Geol.*, 6:617–638, 1984; C. W. Passchier and C. Simpson, Porphyroclast systems as kinematic indicators, *J. Struct. Geol.*, 8:831–843, 1986; A. Schedl and B. A. van der Pluijm, A review of deformation microstructures, *J. Geol. Educ.*, 36:111–121, 1988; C. Simpson, Determination of movement sense in mylonites, *J. Geol. Educ.*, 34:246–261, 1986.

Myodocopida

An order of marine organisms that forms an important part of the class Ostracoda (subphylum Crustacea) and comprises three suborders. It has a long geological history that extends back at least to the Early Silurian. However, the lightly calcified carapaces of the species in this order are a factor in their having a sparse fossil record. Indeed, of the sixteen families of myodocopids, three are known only from modern marine environments. Species of those families that have a fossil record are quite rare and discontinuous in their stratigraphical distribution. Myodocopids are only abundant as fossils from organically rich shale that was deposited in deep, anoxic environments, especially those of the Devonian seas. Such environments were devoid of benthic (bottom-dwelling) life, but the valves and carapaces of the nektonic (free-swimming) myodocopids, both immature forms (termed instars) and adults, sank into such environments and were preserved. The rapid evolution of myodocopids during the Devonian Period has led to their use as guide fossils for dating these otherwise poorly fossiliferous marine rocks. See CRUSTACEA; OSTRACODA.

The myodocopids are typically much larger than other ostracodes and may be more than a centimeter long, although adults of most species are unlikely to be longer than 3 mm. As is characteristic of ostracodes, the carapace of the myodocopids comprises two valves that are joined along the animal's dorsum by a hinge and ligament. In keeping with the generally light calcification of myodocopids, the hinge is simple and without the sort of morphological embellishments that make hinge morphology a useful feature in the systematics of many other groups of ostracodes.

Among the myodocopids, the carapace does not completely encase the soft anatomy of the animal when the valves are closed. Instead, the anterior portions of both valves are marked by a characteristic notch through which the setaceous antennae or antennules protrude.

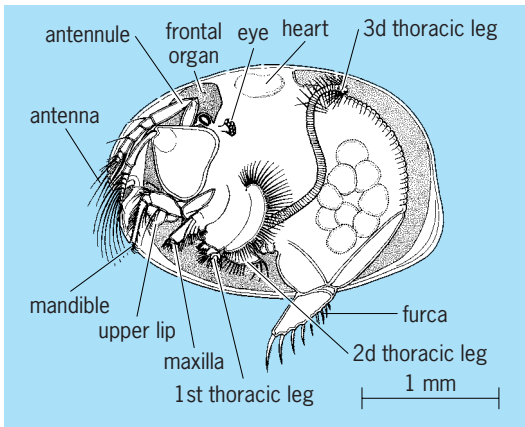


Fig. 1. Morphology of a typical myodocopid ostracode, an adult female of *Cypridina norvegica* Baird, viewed from the left side with the left valve removed. (Modified from R. C. Moore, ed., *Treatise on Invertebrate Paleontology*, pt. Q, *Arthropoda*, University of Kansas and Geological Society of America, Lawrence and Boulder, 1961)

The appendages of all ostracodes are homologous with those of other crustaceans and, as is characteristic of arthropods in general, have been highly modified during their evolution. The appendages of the

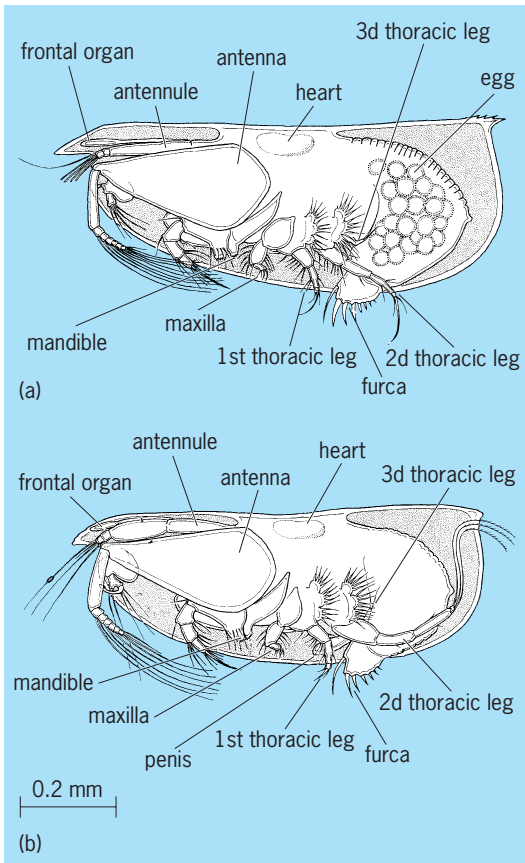


Fig. 2. Morphology of both sexes of *Conchoecia elegans* Sars, viewed from the left side with the left valve removed. (a) Adult female with eggs. (b) Adult male. (Modified from R. C. Moore, ed., *Treatise on Invertebrate Paleontology*, pt. Q, *Arthropoda*, University of Kansas and Geological Society of America, Lawrence and Boulder, 1961)

myodocopids are especially well adapted for swimming. The first two cephalic appendages, the antennule and antennae in particular, are characterized by long setae that facilitate swimming (Figs. 1 and 2). Sexual dimorphism is not as pronounced among the myodocopids as it is among most smaller, more heavily calcified ostracodes.

Ostracodes do not have a planktonic larval stage that is characteristic of so many kinds of marine organisms. Instead, the females carry a clutch of four to six fertilized eggs within their carapace until the tiny, first-instar ostracods hatch as typical, bivalved, nauplius larvae (Fig. 2). The larvae add appendages as they grow by molting. Most myodocopids have five or six instars followed by a terminal adult stage beyond which the ostracodes do not grow, molt, or otherwise change in morphology.

Myodocopids are not exclusively planktonic as is often assumed. They are typically strong swimmers, but many are nektobenthic, swimming in proximity to the substrate. Others, however, live near the surface of the open ocean and are not associated with the environments of deposition that lie far beneath them on the ocean floor, where their valves and carapaces will ultimately be deposited.

A fascinating characteristic of the myodocopids is the ability of many species to secrete bioluminescent material into the water. This bioluminescence functions as highly elaborate courtship displays that are species-specific and facilitate recognition of potential mates. See BIOLUMINESCENCE.

Roger L. Kaesler

Bibliography. A. C. Cohen and J. G. Morin, Morphological relationships of bioluminescent Caribbean species of *Vargula* (Myodocopa), in R. C. Whatley and C. Maybury (eds.), *Ostracoda and Global Events*, Chapman and Hall, New York, 1990; R. C. Moore (ed.), *Treatise on Invertebrate Paleontology*, pt. Q, *Arthropoda* 3, *Crustacea*, *Ostracoda*, University of Kansas and Geological Society of America, Lawrence and Boulder, 1961; R. C. Whatley, D. J. Siveter, and I. D. Boomer, *Arthropoda* (*Crustacea: Ostracoda*), in M. J. Benton (ed.), *The Fossil Record* 2, Chapman and Hall, New York, 1993.

Myricales

An order of flowering plants, division Magnoliophyta (Angiospermae), in the subclass Hamamelidae of the class Magnoliopsida (dicotyledons). The order consists of the single family Myricaceae, with about 50 species. Within its subclass the order is marked by its simple, resinous-dotted, aromatic leaves and unilocular ovary with two styles and a single ovule. The plants are trees or shrubs, and the flowers are much reduced and borne in catkins. The fruit is a small, waxy-coated drupe or nut. Several species of *Myrica* are occasionally cultivated as ornamentals. See HAMAMELIDAE; MAGNOLIOPHYTA; MAGNOLIOPSIDA; ORNAMENTAL PLANTS; PLANT KINGDOM.

Arthur Cronquist

Myrtales

An order of flowering plants in the core eudicots. The order consists of 10 families and approximately 9300 species. The circumscription of the order has been relatively stable, with only minor changes made on the basis of deoxyribonucleic acid (DNA) sequence data. The two largest families are Melastomataceae (approximately 4500 species) and Myrtaceae (approximately 3000 species). Thymelaeaceae are excluded in recent concepts of the order, being related instead to families of Malvales. Vochysiaceae, traditionally included in Polygalales, have been shown to be members of Myrtaceae with DNA sequence data.

Myrtales are chiefly tropical, but Onagraceae and Penaeaceae are predominantly temperate. Myrtales usually have opposite, simple, entire leaves and perigynous (situated around the pistil) to epigynous (having all floral parts conjoined near the top of the ovary) flowers with a compound pistil and most commonly axile placentation. The seeds have little or no endosperm. The stamens are normally numerous (Vochysiaceae being a marked exception), and many species have tetramerous flowers. Vascular bundles characteristically have internal phloem, which is otherwise rare in the rosoid dicots.

Economic crops in Myrtaceae include spice trees such as allspice (*Pimenta*) and cloves (*Syzygium*), and the timber trees *Eucalyptus*. Other important economic crops in the order include evening primrose (*Oenothera*, Onagraceae) and pomegranate (*Punica*, Lythraceae). Purple loosestrife (*Lythrum salicaria*) can be a noxious weed of North American waterways. See EUDICOTYLEDONS. Michael F. Fay

Mysida

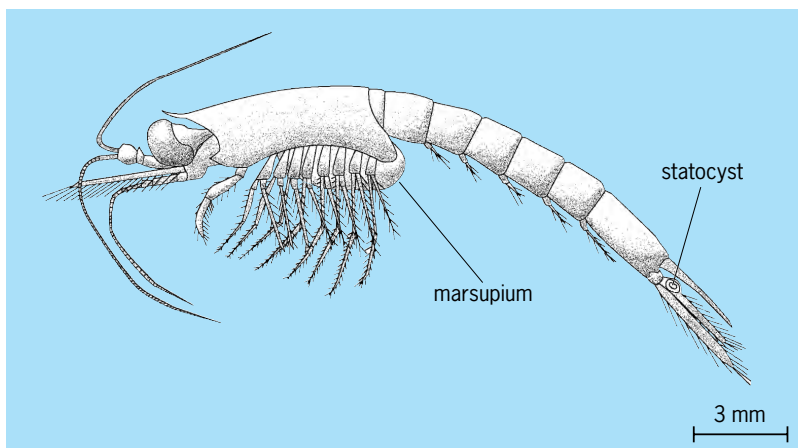
An order of free-swimming, shrimplike crustaceans, commonly known as opossum shrimps, belonging to the subclass Eumalacostraca. They occur in vast numbers in coastal and open oceanic regions of the

world. Mysida were once classified along with the Lophogastrida into an order Mysidacea, but the two, along with a closely allied extinct Paleozoic group, Pygocephalomorpha, are now generally considered to be separate orders. The mysidan adult body length generally averages about 0.6 in. (15 mm), and most species are distributed in shallow coastal and shelf waters of the oceans. A few mysidans live in the surface layers of the ocean bottoms. In addition, some species have invaded freshwaters, including specialized forms strictly confined to caves. See CRUSTACEA; EUCARIDA; EUMALACOSTRACA.

Biology. Mysida are among the most primitive of the peracarid eumalacostracans. The carapace generally covers the entire thorax, and it is not fused to the underlying segments. The thoracic limbs are biramous (having two branches), possessing both a well-developed medial endopod and a lateral natatory (adapted to swimming) exopod; the thoracopods generally bear well-developed gills on the base of the limbs. Although their first thoracic appendage is modified as a maxilliped to supplement the actions of the mouthparts, the first thoracic segment is not fused into the cephalon. Generally the pleopods are well developed, although in cave- and bottom-dwelling forms these abdominal limbs can be very reduced. The first pair of pleopods in males is often highly modified to facilitate sperm transfer in copulation. See PERACARIDA.

Mysidans feed on algae, detritus, and zooplankton. They pass through up to 12 molts before achieving sexual maturity. The time between successive molts depends upon environmental temperature and body size. In general, smaller species have shorter life cycles than larger species. Generation time in warm summer environments can be a few weeks; in colder seasons, the same species may have a generation time of several months. Most species have more than one generation per year, the maximum considered to be about five per year. Longer-lived species do exist. *Mysis relicta* is known to have a 2-year life cycle in high-latitude freshwater lakes.

Most species of mysids form aggregations. These are of different types for different purposes. Breeding aggregations occur in coastal and shelf taxa, and possibly some open oceanic species; these consist of loose groupings of sexually mature adults that come together seasonally for mating. Large aggregations called shoals can extend over large areas or volumes, but densities of animals within shoals vary. Swarms are often constituents of shoals and can be spherical or wafer-shaped; the greatest densities of mysids occur in swarming populations. Another type of aggregation is the school, in which swimming mysids are all orientated parallel to each other and are traveling in tandem through the environment. Schools, because they are moving, are usually elongated in shape. A single species will form a different kind of aggregation at different stages of its life history. The functions of these aggregations, except for breeding, are not always clear, although protection of the population from predators is important. The chance of an



Boreomysis arctica, a member of the Mysida showing the marsupium and the paired uropods with a statocyst in the inner branch or endopod. (From W. M. Tattersall and O. S. Tattersall, 1951)

individual surviving an attack when embedded in a swarm of thousands is significantly better than when on its own, although the open ocean school itself often attracts special predators such as whales. Aggregations occur in many coastal mysids, especially in estuarine or sandy beach habitats. The possession by mysidans of the unique balance organ, the statocyst, is probably linked with this swarming habit by facilitating the maintenance of proper orientation to the environment and with each individual's swarm mates.

As is the case with all peracarids, the young are carried within a marsupium, or brood pouch, formed by transparent concave plates attached to the insides of the thoracic legs. These plates, called oostegites, have interlocking setae that form the marsupium beneath the thorax. The eggs, which are fertilized while being laid, are deposited directly into the marsupium, develop to miniature adults, and emerge to swim freely in the water. The number of embryos depends on the species. Most species have brood sizes in the range 4–30, but broods of 50–100 are common. The largest brood recorded is 245 embryos in the marsupium of a *Mysis stenolepis*.

Distribution. Mysids are predominantly marine organisms, but many species occur in freshwater lakes and rivers, as well as freshwater or brackish-water caves and ground-water wells. One of the most widely distributed freshwater species is *M. relicta*. The geographical distributions of the 800 marine species are not well known: more than one-third live in the open ocean, and the remaining species are found in shallow coastal regions, including estuaries.

In contrast, many of the coastal species appear to have very limited distributions. Many species live actually on or immediately above the sediment or sand, not only in coastal regions. Some species are commensal, each with a different partner, ranging from sessile sponges and anemones to land and hermit crabs.

Classification. The order Mysida currently contains almost 1100 species in some 168 genera ascribed to four families: Petalophthalmidae, Mysidae (by far the largest family), Lepidomysidae, and Stygiomysidae. Except for the deep-sea petalophthalmids (six species) and the cave-dwelling lepidomysids (nine species), all mysidans have statocysts in the uropods of their tail fan (see **illustration**), a feature peculiar to mysidans.

Fossil record. Although most mysidans, based on living forms, are much too small and of so thin a cuticle that they would stand little chance of being preserved as fossils, there is a small record. These fossils mostly occur in the fine-grained lithographic limestones of Jurassic age in Germany and France, and consist of genera such as *Elder*, *Franco-caris*, and *Siriella*. Nevertheless, the larger, better-sclerotized lophogastridans and the Paleozoic pygocephalomorphs have more interesting and extensive fossil records.

Important species. Swarms of mysids in coastal waters are exploited commercially in tropical and

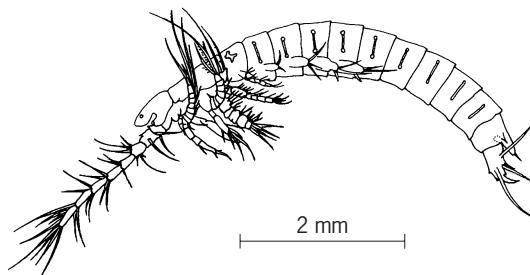
subtropical regions of the world, and are used to make shrimp paste and sauces. The brackish species, *Neomysis intermedia*, is marketed in Japan as a preserved, cooked food known as tsukudani. Other species, or several mixed species, are used in India and Southeast Asia as human food. Several species, notably the freshwater *M. relicta*, have been introduced into lakes in both North America and Europe to serve as food for resident fish such as salmon.

John Mauchline; Frederick R. Schram

Bibliography. J. Mauchline, The biology of mysids, *Adv. Mar. Biol.*, 18:1–369, 1980; T. Remerie et al., Phylogenetic relationships within the Mysidae (Crustacea, Peracarida, Mysida) based on nuclear 18S ribosomal RNA sequences, *Mol. Phylogenet. Evol.*, 32:770–777, 2004; F. R. Schram, Mysidaceans, *Crustacea*, pp. 107–127, Oxford University Press, New York, 1986; W. M. Tattersall and O. S. Tattersall, *The British Mysidacea*, Ray Soc. Monogr. 136, 1951.

Mystacocarida

A subclass of primitive Crustacea, discovered in the interstices of Massachusetts intertidal sandy beaches in 1939. Specimens of both sexes have since been found in similar habitats on shores of Europe, Africa, and Chile. The three species are in the genus *Derocheilocaris*. The body is wormlike and about 0.2 in. (5 mm) long (see **illus.**). The cephalothorax



Derocheilocaris typicus, lateral view.

bears first antennae, second antennae, mandibles, first maxillae, and second maxillae. Maxillipeds are on a separate segment, and four additional free thoracic segments bear platelike appendages. The six abdominal segments are without appendages, but large caudal rami are present. The labrum is enormous, mouthparts and nervous system are primitive, and the genital pore is on the thorax. See CRUSTACEA.

Robert W. Pennak

Myxiniformes

An order of Craniata in the class Myxini; commonly called hagfishes. There are old and new views regarding the classification of the Myxiniformes. The two classifications here comprise the extant myxiniform taxa.

New:

Phylum Chordata
 Subphylum Craniata
 Superclass Myxinomorphi
 Class Myxini
 Order Myxiniformes (Hyperotreti)—
 hagfishes
 Superclass Petromyzontomorphi
 Class Petromyzontida
 Order Petromyzontiformes (Hyperoartii)—
 lampreys
 Superclass Gnathostomata—jawed vertebrates

Old:

Phylum Chordata
 Subphylum Vertebrata (= Craniata)
 Superclass Agnatha
 Class Myxini
 Order Myxiniformes (Hyperotreti)—
 hagfishes
 Class Cephalaspidomorphi (Monorhina)
 Order Petromyzontiformes (Hyperoartii)—
 lampreys
 Superclass Gnathostomata—jawed vertebrates

In the old classification, the Vertebrata and Craniata are synonymous; extant hagfishes and lampreys are collectively called Cyclostomata (jawless fishes). This classification supports the monophyly of the cyclostomes (called the cyclostome hypothesis) and rejects the idea that lampreys are more closely related to gnathostomes than to hagfishes (called the vertebrate hypothesis). See CEPHALASPIDOMORPHA; CYCLOSTOMATA (CHORDATA); JAWLESS VERTEBRATES; PETROMYZONTIDA.

In the new classification, Vertebrata and Agnatha are not used as taxon names, and Craniata and Vertebrata are not synonymous. Myxinomorphi are the sister group of the vertebrates, and include the superclasses Petromyzontomorphi and Gnathostomata (vertebrate hypothesis). Hagfishes are excluded from the vertebrates primarily because they lack arcualia. Arcualia are embryonic cartilaginous elements that fuse to a notochordal calcification, the centrum, to form a vertebra. Proponents of the new classification assume that hagfishes are not degenerate forms of some vertebrate taxon.

Myxiniforms can be identified by the following set of characteristics: an eel-like scaleless body; 1 to 16 pairs of external gill openings; one semicircular canal; and no bones, no paired fins, no neuromasts, and no lens or extrinsic eye muscles. Based on a very scanty fossil record, it is probable that hagfishes have remained little changed since the Pennsylvanian age 300 million years ago. The order comprises one family, two subfamilies, seven genera, and about 70 species, all of which occur in the temperate seas of the world.

Myxinidae (hagfishes). Hagfishes (see **illustration**) have a low fin continuous from the middorsum, around the tail, and forward on the midven-



Myxine glutinosa. (Courtesy of Donald Flescher)

trum, but it is not differentiated as dorsal, caudal, and anal fins. The maximum total length is 18–116 cm, depending on the species. Copious amounts of slime are excreted through numerous pores on the ventrolateral sides of the body. The slime glands contain mucous cells and thread cells. Threads produced by the thread cells (unique to hagfishes) are thought to give tensile strength to the slime. The slime can aid in escaping the grip of a predator as well as clogging its tail and causing suffocation. The knotting maneuver in feeding depends on the slime. The eyes are degenerate and not visible externally. The nostril is on the tip of the snout, and the nasohypophyseal pouch opens into the pharynx. There are two barbels on each side of the nostril and one or two on each side of the mouth. The mouth is not funnellike or disklike and has a single tooth in its roof; the tongue is eversible and bears two rows of rasplike teeth. The cranium is an unroofed trough cradling the brain. Hagfishes have relatively large, capsule-shaped eggs, with hooks on each end; their ontogeny lacks metamorphosis; and they are the only known craniate whose body fluids are isosmotic with seawater.

Hagfishes are benthic, burrowing in mud that overlies substrates ranging in particle sizes from silt to rocks. Mud is an essential component of the habitat, as is full-strength seawater (35 parts per thousand). They feed on soft-bodied invertebrates, as well as the remains of crabs, shrimps, and fishes. Tough items such as the carcass of a large vertebrate require a technique called knotting. The hagfish grasps the flesh with its teeth and then throws its body into a simple overhand knot. Waves of muscular contractions move the knot toward the head, providing a mechanical means of tearing away the flesh. The same technique is used to clear the body of entrapment in its own slime, to avoid capture, and to escape from traps. Hagfishes support a rather large fishery worldwide, as thousands of tons are captured annually to supply manufacturers of leather goods with “eelskins.”

In the subfamily Myxininae, the efferent branchial ducts open by a pair of external gill openings, one on each side. The subfamily comprises four genera and about 25 species. In the Eptatretinae, the branchial ducts open separately to the exterior through 5 to 16 pairs of gill openings. This subfamily comprises three genera and about 45 species. Herbert Boschung

Bibliography. B. Fernholm, Hagfish systematics, pp. 33–44 in J. M. Jørgensen et al. (eds.), *The Biology of Hagfishes*, Chapman and Hall, London, 1998; P. Janvier, *Early Vertebrates*, Oxford Monog. Geol. Geophys. 33, Oxford University Press, 1996; J. M. Jørgensen et al. (eds.), *The Biology of Hagfishes*, Chapman and Hall, London, 1998; F. H. Martini and D. Flescher, Hagfishes: Order Myxiniiformes, pp. 9–16 in B. B. Collette and G. Klein-MacPhee (eds.), *Bigelow and Schroeder's Fishes of the Gulf of Maine*, 3 ed., Smithsonian Institution Press, Washington, DC, 2002; J. S. Nelson, *Fishes of the World*, 4th ed., Wiley, New York, 2006.

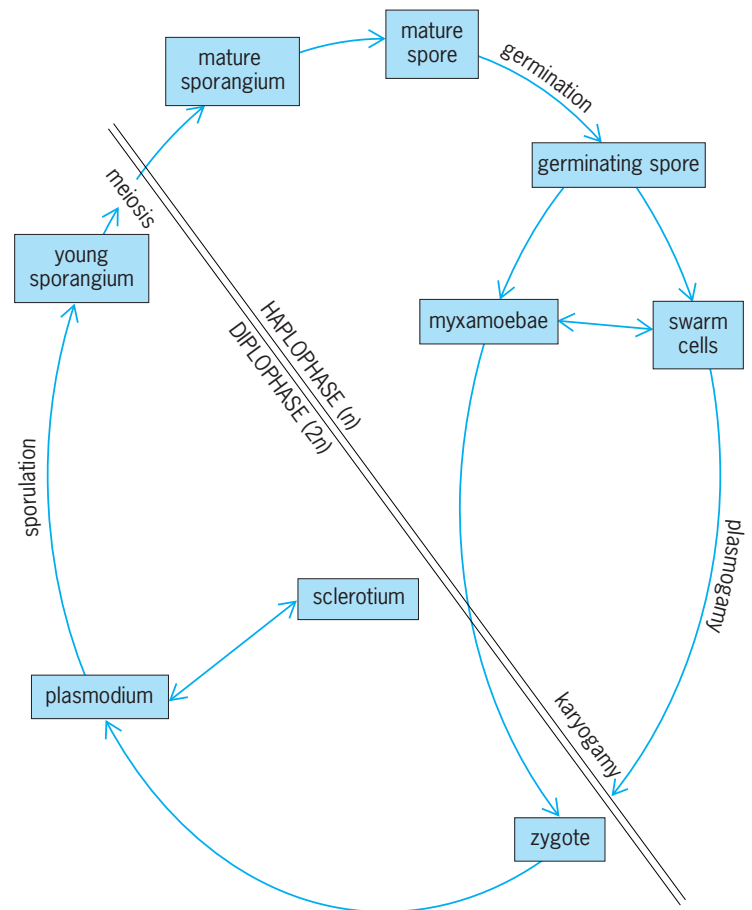
Myxomycota

Organisms that are classified in the kingdom Fungi and given the class name Myxomycetes, following the rules of botanical nomenclature; or classified in the kingdom Protista at various taxonomic ranks, as class Mycetozoa, following the rules of zoological nomenclature. Evolutionary origins are controversial, but many now believe, based on DNA sequencing techniques, that the Myxomycetes diverge early on the tree of life in the region where other protists are found.

The class consists of 3 subclasses, 6 orders, approximately 57 genera, and 600 species. Subclasses Ceratiomyxomycetidae, Myxogastromycetidae, and Stemonitomycetidae are distinguished by the type of sporophore development, type of plasmodium, and method of bearing spores (see **illus.**). The various orders, families, genera, and species are distinguished by characteristics of the fruiting bodies such as spore color, peridium, capillitium, calcium carbonate, or columella.

Distribution and habitats. Myxomycetes begin to appear in May and fruit throughout the summer until October in the north temperate regions. Many species are universally distributed and live in moist and dark places on decaying organic matter, such as decaying logs and dead twigs and leaves on the forest floor. Some species are restricted to more specialized habitats, which include on the bark surface of living trees and vines, under melting snowbanks in alpine regions, and on the dung of herbivorous animals.

Life cycle. Spores are released from the fruiting bodies when disturbed by animals, rain, or wind, and fall onto the substratum where, when water is present, they germinate and release protoplasts. The protoplasts may develop into either a myxamoeba or a flagellated swarm cell, both of which are haploid and behave like gametes (sex cells). The haploid (monoploid) gametes fuse in pairs forming diploid zygotes, which then divide mitotically without subsequent cell division, resulting in the formation of a multinucleated, free-living mass of unwalled protoplasm called the plasmodium. The diploid plasmodium is representative of the slime stage, and hence the common names sometimes used for this group of organisms include plasmodial, acellular, or true



Life cycle of a typical heterothallic myxomycete.

slime molds. The plasmodia ingest food as particulate matter (usually bacteria) by engulfment and are capable of growing to over 70 cm in diameter. The plasmodia may be colored red, orange, white, or almost black, but most are yellow, and are frequently seen in forested areas on moist, decaying logs, leaf litter, and fleshy mushrooms such as *Pleurotus*. The plasmodium undergoes a series of developmental stages, eventually becoming transformed into various types of usually colorful fruiting bodies, ranging in size from about 50 micrometers up to 70 cm.

Importance in research. The separate stages in their life cycle make myxomycetes ideal research laboratory organisms to study basic biological problems, ranging through protoplasmic streaming, the mitotic cycle, morphogenesis, aging, and cell division in cancerous cells. See EUMYCOTA. Harold W. Keller

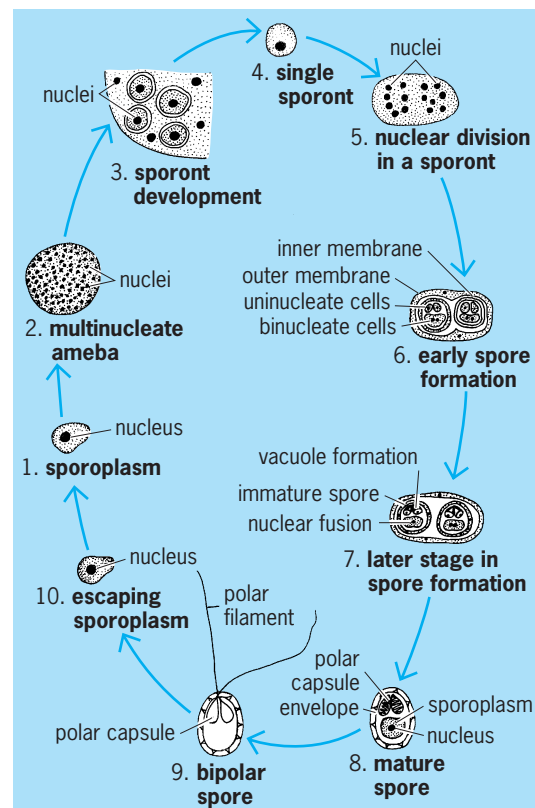
Bibliography. C. J. Alexopoulos, W. C. Mims, and M. Blackwell, Phylum Myxomycota: True slime molds, in *Introductory Mycology*, 4th ed., Wiley, New York, 1996; L. Frederick, Phylum plasmodial slime molds, Class Myxomycota, in L. Margulis et al. (eds.), *Handbook of Protozoa*, Jones and Bartlett, Boston, 1990; H. W. Keller, The Myxomycota, in S. P. Parker (ed.), *Synopsis and Classification of Living Organisms*, 1982; H. W. Keller, Biosystematics of Myxomycetes: A futuristic view, in C. Lado and J. C. Hernandez (eds.), *2d International Congress on the Systematics and Ecology of Myxomycetes: Abstract*

Volume, Real Jardin Botanico, Madrid, 1996; H. W. Keller and K. L. Braun, *Myxomycetes of Ohio: Their Systematics, Biology and Use in Teaching*, *Ohio: Biol. Surv. Bull., New Series*, vol. 13, no. 2, 1999; G. W. Martin et al., *The Myxomycetes*, 2d ed., 1983; S. L. Stephenson and H. Stempfen, *Myxomycetes: A Handbook of Slime Molds*, Timber Press, Portland, OR, 1994.

Myxosporida

An order of the protozoan class Myxosporidea (subphylum Cnidospora). It is characterized by the production of spores with one or more valves and polar capsules, and by possession of a single sporoplasm with or without an iodophilous vacuole. Myxosporidians are mainly parasites of fishes. They infect all parts of the body, including the heart and brain, and often induce considerable pathological changes in the host tissue.

Diagnostic features. Most species have two valves and two polar capsules, and are further identified by the size and shape of the spore. The size, shape, and position of the capsules are of diagnostic value, as are also various markings, prolongations, and appendages of the spore membrane. The exceptions to this general description are *Unicapsula*, with two valves and one polar capsule; *Chloromyxum*, with two valves and four polar capsules; *Kudoa*, with four valves and four polar capsules; and *Hexacapsula*, with six valves and six polar capsules.



Life cycle of *Myxobolus*, a fish parasite.

Parasitic species. These four genera contain species that cause extensive deterioration of the flesh in the Pacific halibut, the common swordfish, the North Atlantic herring, and the Pacific yellowfin tuna, often making such commercially important fishes unfit for human consumption. Members of the genus *Myxobolus* may induce massive tumor-like growths in the skin of fresh- and brackish-water fishes. None, however, is pathogenic to humans or other mammals.

Life cycle. Infection begins with the ingestion of the spore by a host fish. The digestive fluids cause the polar filaments to be extruded, and at the same time the sporoplasm is released from the spore (see *illus.*). The sporoplasm, or amebula, reaches the specific site of infection directly through the gut wall or by way of the bloodstream. The amebula becomes a trophozoite when it starts feeding on the host tissues. The trophozoite then goes through a series of nuclear divisions (nucleogony) and, by a process of budding, gives rise to a number of cells, each of which eventually develops into a sporont. A sporont is a monosporoblast if one spore is produced and a pansporoblast if two or more spores are formed. The sporont undergoes a series of nuclear divisions, in which the number of nuclei produced will determine the number of spores and polar capsules to be formed. That is, in every spore one nucleus is involved in the formation of each valve and each polar capsule. Two nuclei become the gametic nuclei, which then fuse to form the zygotic nucleus of the sporoplasm. See CNIDOSPORA; MYXOSPORIDEA; PROTOZOA. Ross F. Nigrelli

Myxosporidea

A class of the protozoan subphylum Cnidospora. Members of this class, which includes the orders Myxosporida, Actinomyxida, and Helicosporida, are parasites in fish, a few amphibians, and invertebrates. Actinomyxida with four families and Helicosporida with a single species (monotypic) do not contain subordinal categories. The Myxosporida, however, are divided into two suborders, the Unipolarina and Bipolarina.

Unipolarina is characterized by spores with one to six (never five) polar capsules located at the anterior end, except in some genera in which the capsules are widely separated or located in the central part of the spore but in which the polar filament is attached near the anterior end. The Unipolarina contains nine or more families that include important tissue (histozoic) destroying forms in fish such as *Unicapsula* (one capsule), *Myxobolus* (two capsules), *Chloromyxum* (four capsules), and *Hexacapsula* (six capsules). Most species have two anterior capsules and are included in such well-known and widely distributed genera as *Ceratomyxa*, *Myxosoma*, and *Henneguya*.

Bipolarina, containing a single family with three genera, is characterized by the presence of one capsule at or near each end of a fusiform or

ellipsoid spore. These parasites (coelozoic) are usually found in the gallbladder or urinary bladder of fish and a few amphibians. See ACTINOMYXIDA; CNIDOSPORA; HELICOSPORA; MYXOSPORA; PROTOZOA. Ross F. Nigrelli

Myzostomida

Small, soft-bodied marine worms associated with echinoderms, mainly crinoids. They are found in all oceans from subtidal to over 3000 m (9840 ft) depth. These worms are often considered to be an order of Polychaeta, although the first molecular analyses based on DNA comparison estimated that they would be a protostome group nested outside the Annelida. The Myzostomida include about 180 species within eight families (see **table**). Most of them are ectocommensals of crinoids (that is, they live on their outer body surface without affecting them), but some are parasites of crinoids, asteroids, or ophiuroids that infest the gonads, coelom, integument, or digestive system. The association between myzostomids and echinoderms is very old: signs of parasitic activities, similar to those induced by extant myzostomids, are found on fossilized crinoid skeletons dating back to the Carboniferous. The body plan of most myzostomids is singular and differs from the regular body plan of polychaetes as they are incompletely segmented, parenchymous, acoelomate organisms with chaetae. See ANNELIDA; POLYCHAETA.

For most myzostomids, the body consists of an anterior cylindrical introvert (also called proboscis) and a flat, oval or disklike trunk (**Fig. 1**). The introvert is extended when the individual feeds, but it is retracted into an anteroventral pouch of the trunk most of the time. The trunk ranges from a few millimeters to 3 cm (1.2 in.) long. Five pairs of parapodia are located lateroventrally in two rows, each parapodium containing a protrusible hook, some replacement hooks, and a support rod (or aciculum). Most species have four pairs of slit- or disklike lateroventral sense organs, commonly named lateral organs, and the trunk margin often bears flexible needlelike cirri (more than 100 in some species).

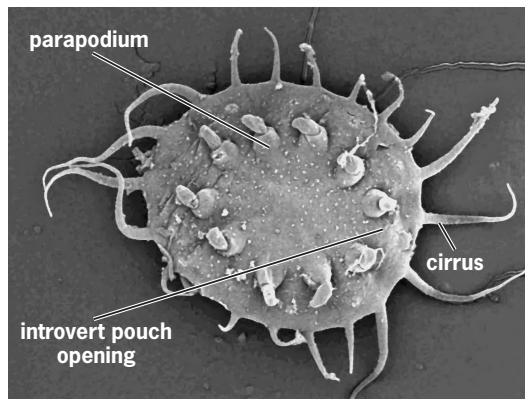


Fig. 1. Scanning electron microscopy view of the ventral side of the Indo-Pacific *Myzostoma ambiguum* showing the parapodia and the marginal cirri. The individual is 4 mm (0.16 in.) long.

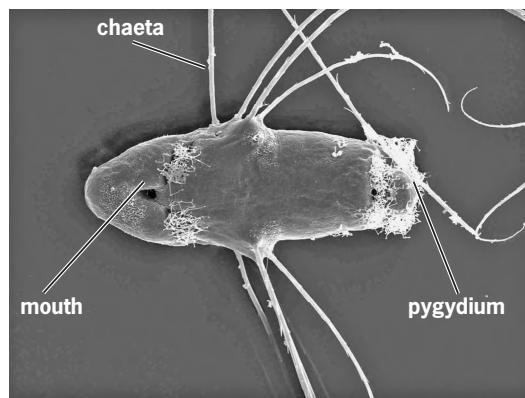


Fig. 2. Scanning electron microscopy view of the ventral side of the Atlantic *Myzostoma cirriferum* metatrochophore. The larva is 0.05 mm (0.002 in.) long.

Humplike or pointed cirri also occur at the base of each parapodium of about 20 species. Two male gonopores are located at the level of the third pair of parapodia, and the female gonopore opens close to the anus, posteroventrally. The body of parasites is often highly modified.

Myzostomids have a ventral nerve chain, circumpharyngeal connectives, and cerebral ganglia. The

Classification of the Myzostomida

| Family | Genera | Hosts | Distribution |
|-----------------------|--|------------------------|--|
| Myzostomatidae | <i>Myzostoma</i> , <i>Notopharyngoides</i> , <i>Hypomyzostoma</i> | Crinoidea, Ophiuroidea | Worldwide |
| Pulvinomyzostomatidae | <i>Pulvinomyzostomum</i> | Crinoidea | Mediterranean Sea |
| Endomyzostomatidae | <i>Endomyzostoma</i> , <i>Contramyzostoma</i> , <i>Mycomyzostoma</i> | Crinoidea | Worldwide |
| Mesomyzostomatidae | <i>Mesomyzostoma</i> | Crinoidea | Off Japan, Aru Island |
| Protomyzostomatidae | <i>Protomyzostomum</i> | Ophiuroidea | Russian Arctic, Bering Sea, Sakhalin, Japan, and Antarctic |
| Asteromyzostomatidae | <i>Asteromyzostoma</i> | Asteroidea | Russian Arctic, Antarctic, Atlantic |
| Asteriomyzostomatidae | <i>Asteriomyzostoma</i> | Asteroidea | Mediterranean Sea, off Southern California |
| Stelechopidae | <i>Stelechopus</i> | Crinoidea | Off Crozet Islands |

species that live on crinoids feed on particles carried by the ciliated host's ambulacral grooves. The digestive system is complete and made of a pharynx included in the introvert, a stomach, an intestine, and two to three digestive caeca. There are six pairs of protonephridia. Most species are simultaneous hermaphrodites and reproduce by transferring spermatophores, followed by hypodermic sperm penetration. Fertilization is internal, and eggs develop in the water column into trochophore larvae which later elongates into a metatrochophore (**Fig. 2**). There is no circulatory system. Igor Eeckhaut

Bibliography. T. Bartolomaeus and G. Pürschke, *Morphology, Molecules, Evolution and Phylogeny in Polychaeta and Related Taxa*, Springer, Dordrecht, 2005; P. L. Beesley, G. J. B. Ross, and C. J. Glasby, *Polychaetes and Allies: The Southern Synthesis*, in *Fauna of Australia: Polychaeta, Myzostomida, Pogonophora, Echiura, Sipuncula*, CSIRO Publishing, Melbourne, 2000; D. Lanterbecq et al., Molecular phylogenetic analyses indicate multiple independent emergences of parasitism in Myzostomida (Protostomia), *Systematic Biol.*, 55(2):208-227, 2006.

N

Nailing — Nitroparaffin

Nailing











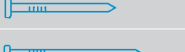







The driving of nails in a manner that will position and hold two or more members, usually of wood, in a desired relationship to each other. The contact pressures between the surfaces of the nails and the surrounding wood fibers hold the nails in position. Some types of nails are shown in the **illustration**.

Strength of a nailed joint. Factors that determine the strength and efficiency of a nailed joint are the type of wood, the nail used, the conditions under which the nailed joint is used, and the number of nails.

In general, hard, dense woods hold nails better than soft woods. The better the resistance of a nail to direct withdrawal from a piece of wood, the tighter the joint will remain. Nails driven into green wood tend to loosen slightly as the wood dries and shrinks. In seasoned material the resistance to withdrawal diminishes only slightly with time, unless moisture affects the wood. Withdrawal resistance is always higher when nails are driven into the side grain than when driven into the end grain. Because the lighter woods do not usually split as readily as the denser ones, more and larger nails can be used to offset the poorer nail-holding properties of the former. Hardwoods are more difficult to nail; they are sometimes used green or with holes drilled for nailing, to prevent splitting.

The surface condition of a nail affects its holding ability. The withdrawal resistance of a common nail increases directly with the distance it penetrates into the wood and increases almost directly with its surface area. A rusty nail may offer more resistance to withdrawal than a smooth one.

Means to increase withdrawal resistance. To increase resistance to withdrawal or loosening, nails may be coated, etched, spirally grooved, annularly grooved, or barbed. Grooved nails tend to hold well despite a change in moisture content. Coated

| | |
|--|---|
| spiral-threaded, insulated siding, face nail |  |
| annular-ring, gypsum board, dry-wall nail |  |
| asbestos shingle nails: annular-ring, spiral-threaded |  |
| annular-ring, plywood roofing nail for applying wood or asphalt shingles over plywood sheathing |  |
| annular-ring, plywood siding nail for applying asbestos shingles and shakes over plywood sheathing |  |
| spiral-threaded, casing head, wood siding nail |  |
| annular-ring roofing nail for asphalt shingles and shakes |  |
| spiral-threaded roofing nail for asphalt shingles and shakes |  |
| annular-ring roofing nail with neoprene washer |  |
| spiral-threaded roofing nail with neoprene washer |  |
| insulated siding nail |  |
| gypsum lath nail |  |
| wood shake nail |  |
| wood shingle nail |  |
| roofing nail |  |
| general-purpose finish nail |  |
| sinker head, wood siding nail |  |
| casing head, wood siding nail |  |

Special- and general-purpose nails.

nails usually provide a greater increase in withdrawal resistance when used in softer woods than in the denser woods. The increase in withdrawal resistance tends to decrease, however, with time.

In most cases, nails driven on a slant have more withdrawal resistance than nails driven straight into the wood. If a slant-driven nail is pulled in a direction which is at right angles to the surface, considerable resistance is encountered from the wood fibers on the pressure side. The nail may also progressively bend as it is pulled out. Both of these factors seem to offer continued holding power, even though the wood fibers are not gripping the entire surface of the nail. Nails slant-driven into the end grain of wood seem to gain proportionately more withdrawal resistance than those slant-driven into the side grain.

When members of a nailed joint tend to separate sideways, the nails are subjected to side loads. In this case doubling the diameter of the nail increases its lateral load capacity by nearly three times. This is true, however, only if the nail point has already been driven a suitable distance into the piece of material which is receiving it.

Blunt-pointed nails are often used to prevent the wood from splitting. Using nails of a smaller diameter also tends to prevent splitting but requires a greater number of nails per joint. Beeswax is sometimes applied to nail points to make them drive more easily, but it also reduces the holding power of the nail. *See* WOODWORKING. Alan H. Tuttle

Najadales

An order of aquatic and semiaquatic flowering plants, division Magnoliophyta (Angiospermae), in the subclass Alismatidae of the class Liliopsida (monocotyledons). The order consists of 10 families and a little more than 200 species. The Potamogetonaceae, with about 100 species, are the largest family of the order, and the name Potamogetonales is sometimes used instead of Najadales for the group. The Najadales are Alismatidae in which the perianth, when present, is not differentiated into evident sepals and petals. Usually the flowers are not individually subtended by bracts. The evolutionary history of the order is in large part a story of floral reduction, associated with progressive adaptation to aquatic and eventually marine habitats. The Zosteraceae of this order are unique among flowering plants in that they grow submersed in the ocean, albeit in shallow water near the shore. *Zostera marina*, or eelgrass, is a common member of the family. *See* ALISMATIDAE; FLOWER; LILIOPSIDA; MAGNOLIOPHYTA; PLANT KINGDOM. Arthur Cronquist; T. M. Barkley

Nanochemistry

The study of the synthesis and characterization of materials in the nanoscale size range (1 to 10 nanometers). These materials include large organic molecules, inorganic cluster compounds, and metal-

lic or semiconductor particles. The synthesis of inorganic materials of nanometer dimension is important because the small size of these particles endows them with unusual structural and optical properties that may find application in catalysis and electrooptical devices. Moreover, such materials may be valuable as precursor phases to strong ceramics. Approaches to the synthesis of these materials have focused on constraining the reaction environment through the use of surface-bound organic additives, porous glasses, zeolites, clays, or polymers. The use of synthetic approaches that are inspired by the biological processes result in the deposition of inorganic materials such as bones, shells, and teeth (biomineralization). This biomimetic approach involves the use of assemblies of biological molecules that provide nanoscale reaction environments in which inorganic materials can be prepared in an organized and controlled manner. Examples of biological assemblies include phospholipid vesicles and the polypeptide micelle of the iron storage protein, ferritin. *See* MICELLE.

Vesicles. Vesicles are bounded by an organic membrane that provides a spatial limit on the size of the reaction volume. If a chemical reaction is undertaken in this confined space that leads to the formation of an inorganic material, the size of the product will also be constrained to the dimensions of the organic host structure. This can be considered as analogous to producing inorganic materials in a soap bubble, except that the soap bubble is very, very small. If the chemical and physical conditions are not too severe to disrupt the organic membrane, these supramolecular assemblies may have advantages over inorganic hosts such as clays and zeolites because the chemical nature of the organic surface can be systematically modified so that controlled reactions can be accomplished. *See* SUPRAMOLECULAR CHEMISTRY.

Surfactant vesicles have been employed in a number of studies involving semiconductor, catalytic, and magnetic materials (see table). In the general method for preparing these materials, vesicles are formed spontaneously by sonicating aqueous solutions of phospholipids (**Fig. 1**). The presence of metal ions in the solution results in their encapsulation within the 25–50-nm internal volume of the enclosed aqueous space of the vesicle. The vesicle membrane is a bilayer of 4.5 nm thickness, and it prevents leakage of the metal ions back into the bulk solution. Thus, if metal ions are now exchanged from the bulk solution for inert cations such as the sodium ion (Na^+ ; by ion-exchange chromatography), the remaining entrapped ions can be subjected to chemical reactions solely within the intravesicular volume. The simplest procedure is to add a membrane-permeable coreactant such as gaseous hydrogen sulfide (H_2S). The H_2S rapidly diffuses through the phospholipid membrane and combines with the encapsulated metal ions to give an insoluble metal sulfide precipitate. As the number of entrapped metal ions is usually less than 10,000, semiconductor particles in the nanometer size range can be routinely produced. An extension of this method, in which an increase in the pH of

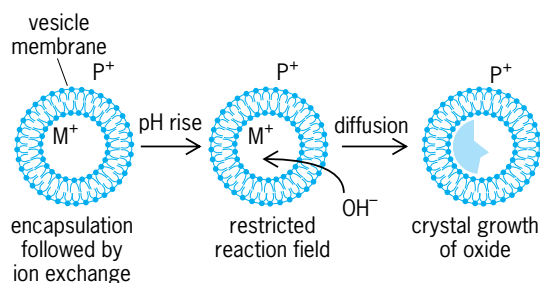


Fig. 1. Nanoscale synthesis of inorganic metal oxides by means of surfactant phospholipid vesicles. Cations (M^+) are encapsulated by sonication and replaced in the bulk solution by inert cations (P^+). Slow diffusion of hydroxide ions (OH^-) through the organic membrane results in intravesicular precipitation of metal oxides.

the bulk solution provides an excess of hydroxide ions (OH^-) that slowly diffuse through the surfactant membrane, results in the formation of metal oxide particles. In particular, nanophase iron oxides with catalytic and magnetic properties have been synthesized. See ION EXCHANGE; PHOSPHOLIPID; SONOCHEMISTRY.

The presence of the surfactant membrane in these nanoscale chemical reactions can have a profound effect on the structure and properties of the resulting inorganic materials. For example, the particles cannot come into direct contact, and the vesicles are usually charged; thus there is negligible aggregation or macroscopic precipitation. Under certain circumstances, the product is stable over many weeks as a monodisperse sol of finely divided particles.

Ferritin. One problem encountered with the use of phospholipid vesicles is their sensitivity to changes in temperature and ionic strength. Procedures have been developed in which the biomolecular cage of the iron storage protein, ferritin, has been used as a nanoscale reaction environment for the synthesis of inorganic materials (see **table**). Ferritin is a robust molecule constructed from 24 polypeptide subunits arranged into a hollow sphere of 8–9-nm internal diameter. The native protein contains a 5-nm-diameter core of a hydrated iron(III) oxide (ferrihydrite) within the internal cavity. Hydrophilic and hydrophobic channels penetrate the protein shell and provide the means by which iron atoms can

be accumulated within or removed from the protein cavity. In the laboratory vessel, the iron can be readily removed by reductive dissolution to give intact empty protein cages (apoferritin).

Several approaches utilize ferritin in the production of inorganic nanoscale particles (**Fig. 2**). In the simplest approach the native iron oxide core is transformed into another material by chemical exposure within the protein shell. For example, exposure of the red-brown protein solution to H_2S results in a green coloration due to the formation of amorphous iron(III) sulfide (FeS) cores, approximately 7.5 nm in diameter. No precipitation is observed because the FeS particles remain encapsulated within the protein shell.

Alternatively, the native iron oxide cores can be removed from their protein shells by the use of appropriate reducing and chelating agents. The resulting empty apoferritin molecules are structurally intact, and can be readily reconstituted at room temperature and pH 7 by incubation of the protein in aerated $Fe(II)$ -containing solutions. Moreover, other metal oxides can be formed within the protein cavity. For example, incubation of the empty protein cages with aqueous $Fe(II)$ but at a pH of 8.5 and temperature of $65^\circ C$ ($149^\circ F$) results in the synthesis of nanometer-size magnetic iron oxides such as magnetite (Fe_3O_4) and maghemite ($\gamma-Fe_2O_3$) [**Fig. 2**]. The resulting protein, termed magnetoferritin, is magnetic. Other metals oxides, such as manganese oxyhydroxide ($MnO \cdot OH$) and hydrated uranium trioxide (UO_3), can be synthesized within the protein cavity by incubation of apoferritin with the

| Biomimetic approaches in the nanoscale synthesis of inorganic materials | | |
|---|---|---|
| System | Materials | |
| Vesicles | Pt, Ag | |
| | Cadmium sulfide (CdS), zinc sulfide (ZnS), silver sulfide (Ag_2S), cobalt sulfide (CoS) | |
| | Silver oxide (Ag_2O) | |
| | Iron oxyhydroxide ($FeO \cdot OH$), Magnetite (Fe_3O_4) | |
| | Aluminum oxide (Al_2O_3) | |
| | Calcium phosphates | |
| | Ferritin | Iron(III) sulfide (FeS) |
| | | Manganese oxyhydroxide ($MnO \cdot OH$) |
| | | Uranium trioxide (UO_3) |
| | | Magnetite (Fe_3O_4) |

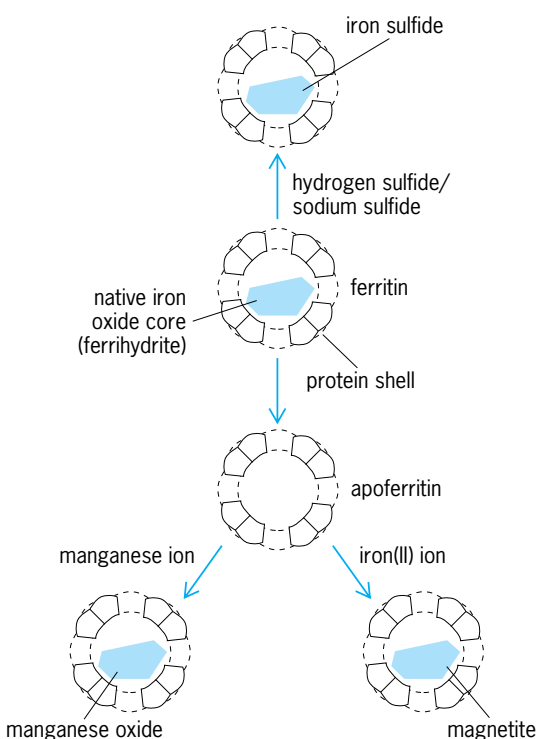


Fig. 2. Use of the supramolecular protein cage of ferritin in the synthesis of nanophase inorganic materials.

appropriate metal salt solutions in the presence of air.

Stephen Mann

Bibliography. G. Ozin and A. Arsenault, *Nanochemistry: A Chemical Approach to Nanomaterials*, 2005; G. B. Sergeev, *Nanochemistry*, 2006; H. Watarai et al. (eds.), *Interfacial Nanochemistry: Molecular Science and Engineering at Liquid-Liquid Interfaces*, 2005.

Nanoparticles

Synthetic particles that range from 1 to 100 nanometers in diameter. Semiconductor nanoparticles around 1–20 nm in diameter are often called quantum dots, nanocrystals, or Q-particles. These particles possess short-range structures that are essentially the same as the bulk semiconductors, yet have optical or electronic properties that are dramatically different from the bulk properties. The confinement of electrons within a semiconductor nanocrystal results in a shift of the band gap to higher energy with smaller crystalline size. This effect is known as the quantum size effect. In the strong confinement regime, the actual size of the semiconductor particle determines the allowed energy levels and thus the optical and electronic properties of the material.

Due to their finite, small size and the high surface-to-volume ratio, nanoparticles often exhibit novel properties. These properties can ultimately lead to new applications, ranging from catalysis, ceramics, microelectronics, sensors, pigments, and magnetic storage, to drug delivery and biomedical applications. Research in this area is motivated by the possibility of designing nanostructured materials that possess novel electronic, optical, magnetic, mechanical, photochemical, and catalytic properties. Such materials are essential for technological advances in photonics, quantum electronics, nonlinear optics, and information storage and processing.

Synthesis. A wide range of scientifically interesting and technologically important nanoparticles have been produced by both chemical and physical methods.

Colloidal. The synthesis of nanocrystals by colloidal methods involves nucleation (the initial formation of

the appropriate semiconductor bond), growth (the formation of a highly crystalline core), and passivation of the nanocrystal surface. The passivation step is important in stabilizing the colloid and controlling the growth of the nanoparticles, preventing the agglomeration and fusing of the particles, and allowing the solubility of the nanoparticles in common solvents. A common approach is to use polymeric surfactants (for example, sodium polyphosphate). The polymer attaches to the surface of the growing particles, usually electrostatically, and prevents their further growth. Another passivation approach is to use capping agents such as thiolates. *See* COLLOID.

In the liquid phase, particle size is controlled by confining the growing particles within micelles or reversed micelles, polymer films, glasses, or zeolites. In the case of micelles, the micellar reagent (for example, dihexadecyl phosphate or dioctadecyldimethylammonium) creates a physical boundary, that is, a well-defined region where the semiconductor is precipitated. In reversed micelles, a small amount of water is mixed with a large amount of an organic solvent and surfactant. The surfactant molecules tend to collect on the surface and stabilize the waterdrop. The size of the water droplets is directly related to the ratio of the water to the organic phase. Smaller droplets result in smaller nanoparticles. Nanoparticle formation occurs by the reaction of two reagents (one which is soluble only in water and another which is soluble only in the organic solvent). Many nanoparticles have been produced by reverse micelles such as cadmium sulfide, selenide, and telluride, lead sulfide and selenide, and zinc and titanium oxides. *See* MICELLE; SURFACTANT.

Vapor-phase. The common chemical synthesis of metallic and intermetallic nanoparticles includes the decomposition of organometallic precursors, such as metal carbonyls (by thermal, photochemical, laser pyrolysis, and ultrasonic methods), to yield the respective elements or alloys, and the reduction of inorganic or organometallic precursors by reducing agents. Different sources of energy can be used to decompose the precursor such as microwave plasma, laser pyrolysis, laser photolysis, and flame combustion. The size of the nanoparticle is determined by the particle residence time, temperature of the chamber, pressure, and precursor composition. Low-temperature flames can also be used to supply the energy to decompose the precursors. Flame synthesis is most common for the production of oxides. Pure metal particles are best produced by gas condensation.

The vapor-phase synthesis of metallic nanoparticles involves evaporation of the material of interest, followed by the condensation of clusters and nanoparticles from the vapor phase. The vapor may be generated with laser or electron-beam heating methods. Laser vaporization provides several advantages, including the production of a high-density vapor of any metal, the generation of a directional high-speed metal vapor from a solid target (which can be useful for directional deposition of the particles), control of the evaporation from specific spots

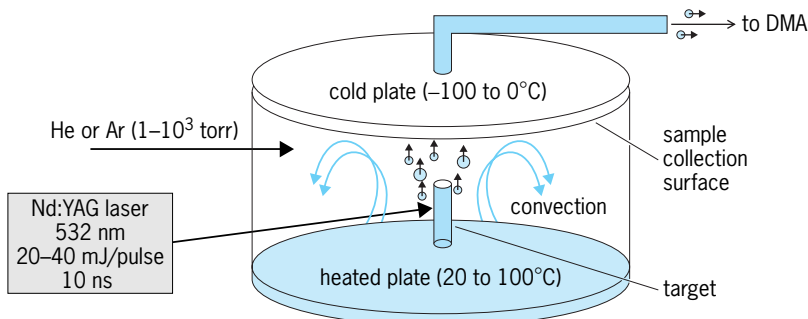


Fig. 1. Experimental setup for the synthesis of nanoparticles by the laser vaporization controlled-condensation method coupled with a differential mobility analyzer (DMA) for the size selection of the nanoparticles.

on the target, as well as the simultaneous or sequential evaporation of several different targets.

A technique that combines the advantages of pulsed laser vaporization with controlled condensation (LVCC) under well-defined conditions of temperature and pressure has been developed. It allows the synthesis of a wide variety of nanoparticles of metallic, intermetallic, oxides, and carbides of controlled size and composition.

The LVCC method consists of pulsed laser vaporization of a metal or an alloy target into a selected gas mixture in a convective atmosphere created by a temperature gradient between two plates separated by a glass ring (Fig. 1). A pure carrier gas such as helium or argon, or a mixture containing a known composition of a reactant gas (for example, oxygen for the synthesis of oxides, methane for carbides, and so on) can be used. A high-energy pulsed laser (532-nm Nd:YAG laser) with an intensity flux of about 10^6 – 10^7 W/cm² is focused on the target of interest. The resulting plasma causes highly efficient vaporization, and the temperature at the focusing spot can exceed 10,000 K (17,500°F). This high temperature can vaporize all known substances so quickly that the surrounding vapor stays at the ambient temperature. Typical yields are 10^{14} – 10^{15} atoms from a surface area of 0.01 cm² in a 10^{-8} s pulse. The large temperature gradient between the bottom and top plates results in a steady convection current which moves the nanoparticles away from the nucleation zone (once condensed out of the vapor phase) before they can grow into larger particles. The nanoparticles are deposited as weblike aggregates on the top cold plate (Fig. 2a), or transferred in a helium flow to a differential mobility analyzer (DMA) for size selections (Fig. 2b). Since the LVCC method produces a signif-

icant portion of charged nanoparticles (as ions or free electrons), the DMA classifies and separates the charged nanoparticles based on their electrical mobility in the presence of an electric field. See LASER.

Assembly of nanoparticles in electric fields. The assembly of nanoparticles into filaments and fibers, which retain the unique properties of the nanoparticles, holds promise for the development of novel functional materials and the engineering of a variety of nanodevices and sensors. The nanoparticles can be assembled into long-chain filaments (Fig. 3a) and fibers (Fig. 3b) by applying an electric field during their synthesis by the LVCC method. The filaments can be few centimeters long, and tangle together with neighboring wires to form bundles. Silicon nanoparticles have a tendency to form dendritic structures with unique fractal patterns (Fig. 3c, d).

The effect of the electric field on the formation of the chain aggregates acts through the polarization of the charges on the nanoparticles' surface. For larger particles, the effect of the electrostatic charge is overpowered by gravity, but for nanoparticles the electrostatic forces are dominant. The nanoparticles' surfaces have a mixture of positive and negative charges, and some of the nanoparticles may have net charges. The dipole force is very strong near the surface of the particle, while farther away from the surface the net charge or monopole force becomes important. These two effects, when combined, may lead to the sticking together of particles of the same net charge.

Properties and applications. Semiconductor nanoparticles have technological applications in many areas of optoelectronics such as light-emitting diodes, solid-state lasers, and optical switches. Silicon nanostructures have stimulated much interest

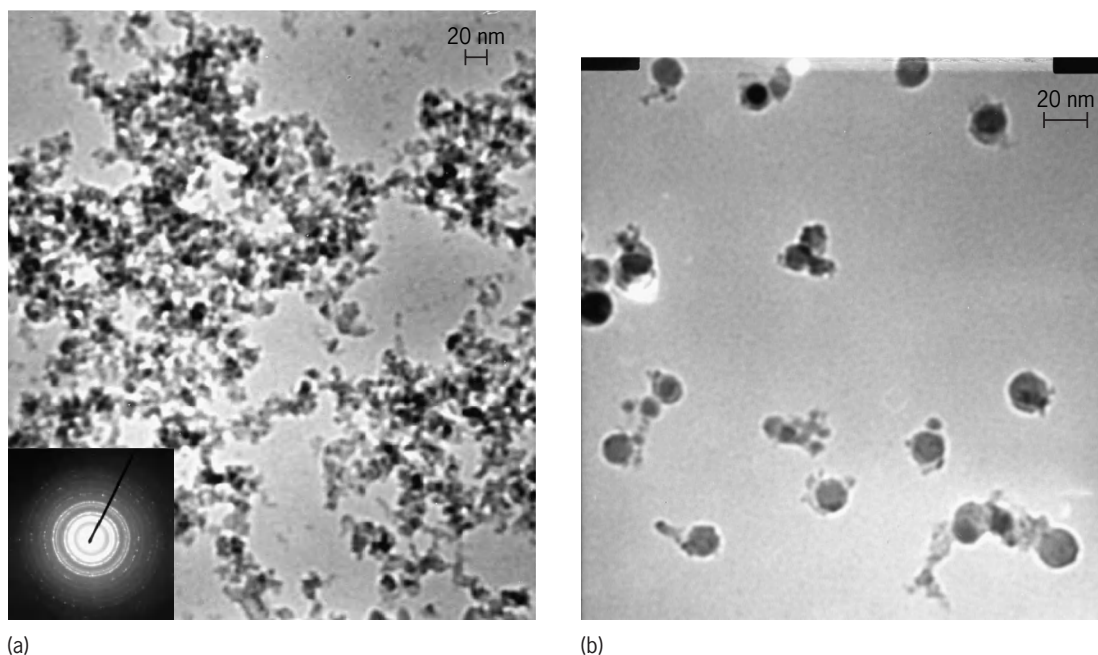


Fig. 2. Transmission electron micrographs. (a) Iron aluminide (FeAl) nanoparticles prepared with the laser vaporization controlled-condensation method with no size selection. The inset shows electron diffraction. (b) 20-nm-diameter FeAl nanoparticles selected by the differential mobility analyzer.

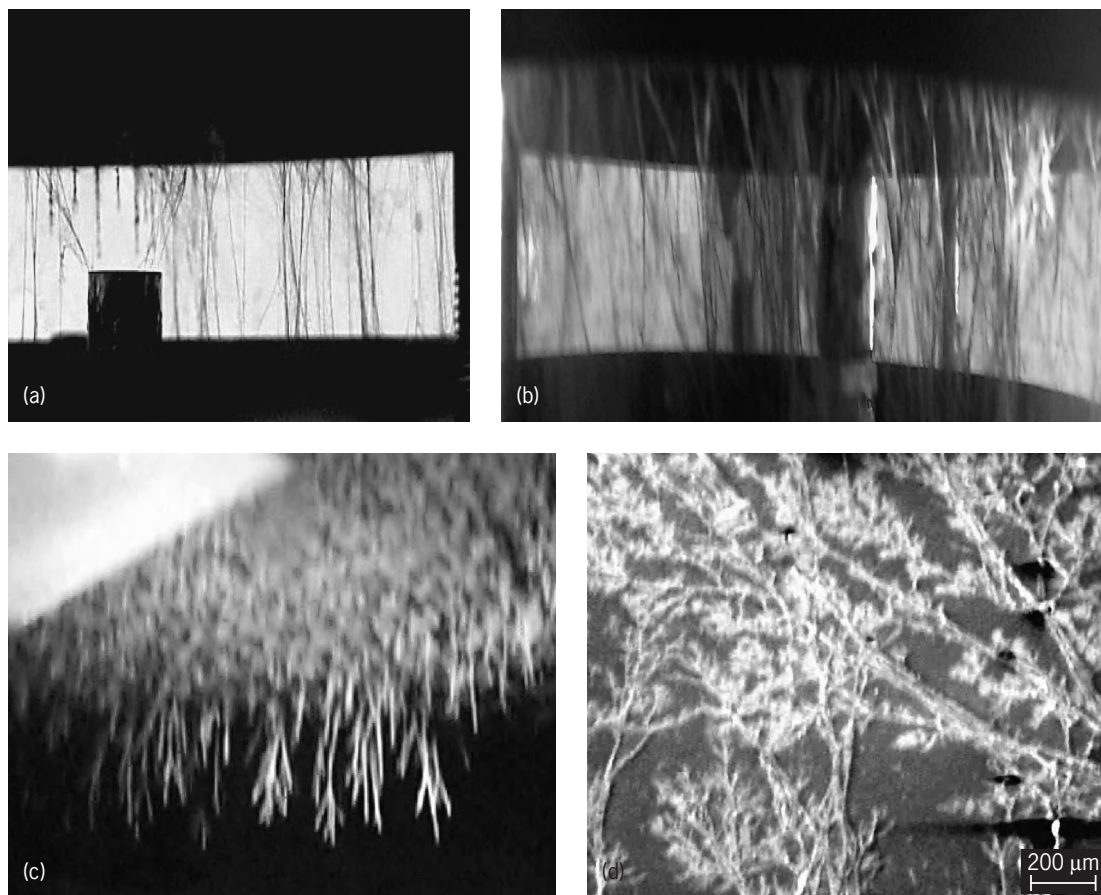


Fig. 3. Photographs of nanoparticle filaments of (a) iron aluminide (FeAl), (b) titanium aluminide (Ti₃Al), and (c) silicon, grown with the application of electric fields during the LVCC synthesis. (d) Scanning electron micrograph of a silicon nanoparticle dendritic assembly.

due to their unique properties, including single-electron tunneling, nonlinear optical properties, and visible photoluminescence. A transmission electron micrograph and electron diffraction pattern of silicon nanocrystals prepared by the laser vaporization with controlled condensation method are shown in **Fig. 4**. Because of its indirect band gap, bulk silicon does not exhibit visible photoluminescence.

The higher energy shift in the photoluminescence of silicon nanocrystals is attributed to the three-dimensional quantum size effect. *See* BAND THEORY OF SOLIDS; MESOSCOPIC PHYSICS; QUANTIZED ELECTRONIC STRUCTURE (QUEST); SEMICONDUCTOR.

The incorporation of silicon nanocrystals in polymer films may lead to the development of novel materials which combine several properties such as the visible photoluminescence and elasticity. Polymer films containing silicon nanocrystals exhibit the photoluminescence characteristic of the pure polymer and of the suspended silicon nanocrystals (**Fig. 5**). This may have interesting applications in the design of new materials for optical display and for silicon-based devices in optoelectronics.

Nanoparticles have important applications in catalysis and photocatalysis. The large number of surface and edge atoms provides active sites for catalyzing surface reactions. Highly nonstoichiometric oxide nanoparticles such as CeO_{2-x} provide a high oxygen-vacancy concentration, and active superoxide surface species. These nanoparticle oxides enable catalytic activation at significantly lower temperatures for the reduction of sulfur dioxide and oxidation of carbon monoxide. Other nanoparticles such as cadmium sulfide and zinc sulfide are efficient photocatalysts for generating hydrogen from water. *See* CATALYSIS.

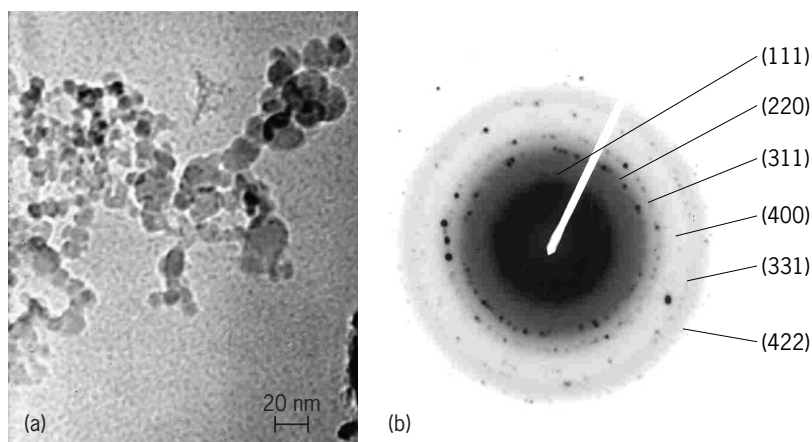


Fig. 4. Silicon nanoparticles. (a) Transmission electron micrograph; average particle diameter, 5–6 nm. (b) Electron diffraction pattern, showing diffraction rings corresponding to the planes of the randomly oriented nanocrystals.

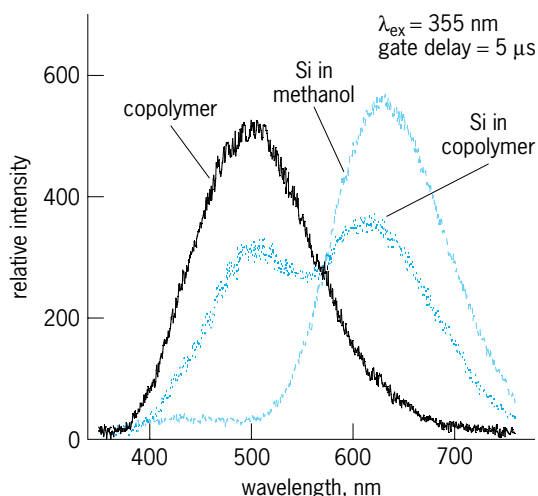


Fig. 5. Photoluminescence spectra of silicon nanocrystals suspended in methanol, sulfonated styrene-ethylene/butylene-styrene triblock copolymer film, and triblock copolymer film containing the silicon nanocrystals.

Molybdenum and tungsten trioxides are photochromic materials that change color upon oxidation from one state to another by the absorption of light. The photochromic effect is stronger in the nanoparticles than in the bulk material. The photochromic materials have potential practical applications in areas such as displays, imaging devices, “smart windows,” and solar energy conversion.

In general, nanostructured materials possess smoother surfaces and enhanced mechanical properties as compared to conventional-grain-size materials. Nanostructured ceramics possess increased ductility since they do not contain many dislocations, and nanoparticles can diffuse fast into the materials' cracks once they are formed. Consolidated intermetallic materials based on nanoparticles show enhanced plasticity; that is, they exhibit significantly better elongations as compared to cast and powder processed components. The filamentlike and tree-like assemblies of the nanoparticles may have some special applications as fillers (additives) to increase the elastic modulus (stiffness) and tensile strength (maximum stress before breaking) of low-strength oils and polymeric materials. See POLYMER; NANO-STRUCTURE.

Nanoparticles and functionalized nanostructured materials may have a significant impact in biological and biomedical applications. In drug delivery, nanoparticles could be used to deliver drugs where they are needed, avoiding the harmful side effects that often result from potent medicine. Prolonged residence times in blood have been achieved by nanoparticles, allowing the carriers to reach sites throughout the body without premature clearance. Because of the size of the nanoparticles, they can leave the vascular system, especially at sites of disease (such as tumors). Since nanoparticles are much smaller than cells, there is the possibility of subcellular targeting, that is, third-order drug targeting. This is key for gene delivery, where nuclear uptake is

a requirement. Since nanoparticles have similar dimensions to natural carriers (for example, lipoproteins and viruses), they may serve as structural models for natural carriers. See DRUG DELIVERY SYSTEMS.

Nanomedical technology is opening new avenues in medical diagnostics and disease therapy. Since the size of a quantum dot determines the color it emits after exposure to light, different sizes of dots attached to different biological molecules can be used to track the activities of many tagged molecules simultaneously. Magnetic nanoparticles have been investigated as a potential alternative treatment for cancer. The hyperthermic effect (heat produced by relaxation of the magnetic energy of the magnetic nanoparticles exposed to an alternating magnetic field) can be used to effectively destroy tumor tissue.

M. Samy El-Shall

Bibliography. M. S. El-Shall et al., Synthesis of nanoscale metal oxide particles using laser vaporization/condensation in a diffusion cloud chamber, *J. Phys. Chem.*, 98:3067, 1994; M. S. El-Shall et al., Synthesis and photoluminescence of web-like agglomeration of silica nanoparticles, *J. Phys. Chem.*, 99:17805, 1995; M. S. El-Shall and S. Li, Synthesis and characterization of metal and semiconductor nanoparticles, in M. A. Duncan (ed.), *Advances in Metal and Semiconductor Clusters*, vol. 4, chap. 3, pp. 115–177, JAI Press, London, 1998; I. N. Germanenko et al., Effect of atmospheric oxidation on the electronic and photoluminescence properties of silicon nanocrystals, *Int. J. Pure Appl. Chem.*, 72:245–255, 2000; I. N. Germanenko, S. Li, and M. S. El-Shall, Decay dynamics and quenching of photoluminescence from silicon nanocrystals by aromatic nitro compounds, *J. Phys. Chem. B.*, 105:59–66, 2001; S. Li and M. S. El-Shall, Synthesis and characterization of photochromic molybdenum and tungsten oxide nanoparticles, *Nanostruc. Mater.*, 12:215, 1999; S. Li, S. Silvers, and M. S. El-Shall, Surface oxidation and luminescence properties of weblike agglomeration of silicon nanocrystals produced by laser vaporization—Controlled condensation technique, *J. Phys. Chem. B.*, 101:1794, 1997; Y. B. Pithawalla et al., Synthesis of magnetic intermetallic FeAl nanoparticles from a non-magnetic bulk alloy, *J. Phys. Chem. B.*, 105:2085–2090, 2001; W. Wang, I. N. Germanenko, and M. S. El-Shall, Room temperature synthesis and characterization of nanocrystalline CdS, ZnS and $\text{Cd}_x\text{Zn}_{1-x}\text{S}$, *Chem. Mater.*, 14:3028–3033, 2002.

Nanostructure

A material structure assembled from a layer or cluster of atoms with size of the order of nanometers. Interest in the physics of condensed matter at size scales larger than that of atoms and smaller than that of bulk solids (mesoscopic physics) has grown rapidly since the 1970s, owing to the increasing realization that the properties of these mesoscopic atomic ensembles are different from

those of conventional solids. As a consequence, interest in artificially assembling materials from nanometer-sized building blocks, whether layers or clusters of atoms, arose from discoveries that by controlling the sizes in the range of 1–100 nm and the assembly of such constituents it was possible to begin to alter and prescribe the properties of the assembled nanostructures. (Many examples of naturally formed nanostructures can be found in biological systems, from sea shells to the human body.) See MESOSCOPIC PHYSICS.

Nanostructured materials are modulated over nanometer length scales in zero to three dimensions. They can be assembled with modulation dimensionalities of zero (atom clusters or filaments), one (multilayers), two (ultrafine-grained overlayers or coatings or buried layers), and three (nanophase materials) [Fig. 1], or with intermediate dimensionalities. Thus, nanocomposite materials containing multiple phases can range from the most conventional case in which a nanoscale phase is embedded in a phase of conventional sizes, to the case in which all the constituent phases are of nanoscale dimensions. All nanostructured materials share three features: atomic domains (grains, layers, or phases) spatially confined to less than 100 nm in at least one dimension, significant atom fractions associated with interfacial environments, and interaction between their constituent domains.

Multilayers and clusters. Multilayered materials have had the longest history among the various artificially synthesized nanostructures, with applications to semiconductor devices, strained-layer superlattices, and magnetic multilayers. Recognizing

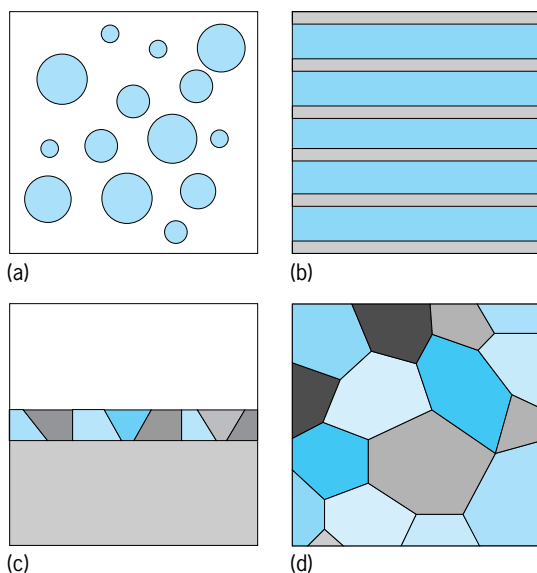


Fig. 1. Schematic of four basic types of nanostructured materials, classified according to integral modulation dimensionality. (a) Dimensionality 0: clusters of any aspect ratio from 1 to infinity. (b) Dimensionality 1: multilayers. (c) Dimensionality 2: ultrafine-grained overlayers (coatings) or buried layers. (d) Dimensionality 3: nanophase materials. (After R. W. Siegel, *Nanostructured materials: Mind over matter, Nanostructured Mat.*, 3:1–18, 1993)

the technological potential of multilayered quantum heterostructure semiconductor devices helped to drive the rapid advances in the electronics and computer industries. A variety of electronic and photonic devices could be engineered utilizing the low-dimensional quantum states in these multilayers for applications in high-speed field-effect transistors and high-efficiency lasers, for example. Subsequently, a variety of nonlinear optoelectronic devices, such as lasers and light-emitting diodes, have been created by nanostructuring multilayers. See ARTIFICIALLY LAYERED STRUCTURES; LASER; LIGHT-EMITTING DIODE; SEMICONDUCTOR HETEROSTRUCTURES; TRANSISTOR.

The advent of beams of atom clusters with selected sizes allowed the physics and chemistry of these confined ensembles to be critically explored, leading to increased understanding of their potential, particularly as the constituents of new materials, including metals, ceramics, and composites of these materials. A variety of carbon-based clusters (fullerenes) have also been assembled into materials of much interest. In addition to effects of confinement, interfaces play an important and sometimes dominant role in cluster-assembled nanophase materials, as well as in nanostructured multilayers. See FULLERENE.

A knowledge of the variation in cluster properties, both physical and chemical, with cluster size is important to both the fundamental understanding of condensed matter and the ability to use cluster-assembled materials in a variety of technological applications. The manner in which the structure and properties of collections of condensed atoms in a single cluster vary with cluster size, from atomic or molecular to bulk solid-state behavior, is fundamental in the development of realistic theoretical models for condensed matter. Theoretical areas that are impacted by this variation include understanding the forces acting among atoms, the structures of atom collections, electronic effects (quantum size effects) caused by spatial confinement of delocalized valence electrons, and cooperative (many-body) atom phenomena such as lattice vibrations or melting. See ATOM CLUSTER.

Synthesis and properties. A number of methods exist for the synthesis of nanostructured materials. They include synthesis from atomic or molecular precursors (chemical or physical vapor deposition, gas condensation, chemical precipitation, aerosol reactions, biological templating), from processing of bulk precursors (mechanical attrition, crystallization from the amorphous state, phase separation), and from nature (biological systems). Generally, it is preferable to synthesize nanostructured materials from atomic or molecular precursors, in order to gain the most control over a variety of microscopic aspects of the condensed ensemble; however, other methodologies can often yield very useful results. See CRYSTAL GROWTH.

Foremost in importance in nanostructuring is the ability to control the size and size distribution of the constituent phases or structures. The desirable sizes

are generally below 100 nm, since in this size range (and often below 10 nm) various properties begin to change significantly because of confinement effects. The chemical compositions of the constituent phases in a nanostructured material are also of crucial importance, as they invariably are to the performance of conventional materials. Finally, it is desirable to control the nature of the interfaces created between constituent phases and, hence, the nature of the interactions across the interfaces.

The properties of nanostructured materials are determined by the interplay among domain size, composition, and interfaces. In some cases, one or more of these features may dominate. Thus, it is desirable to synthesize nanostructured materials under controlled conditions, but with a focus on the particular property or properties of interest. The degree of control available, of course, depends upon the synthesis method used. The properties of these materials are significantly affected by their structural or compositional modulation, their spatial confinement, their interfaces, or a combination thereof. Spatial confinement can in general affect any property when the size of the atomic ensemble becomes comparable to or smaller than a critical length scale for the mechanism that is responsible for that property. Diverse examples include the blue (high-frequency, short-wavelength) shifts of the optical absorption in semiconducting clusters when their sizes fall below the Bohr radii (approximately 5–50 nm) of the excitonic (electron-hole pair) states responsible for absorption, and the increased strengthening of normally soft metals when their grain sizes fall below the critical length scales (less than about 50 nm) for the sources of dislocations (the defect responsible for easy deformation) to easily operate at conventional applied stresses. See CRYSTAL DEFECTS; EXCITON.

Examples. Examples of nanostructured materials that have been characterized include multilayers, individual and assembled atom clusters, and cluster-consolidated nanophase materials.

Magnetic multilayers. Magnetic multilayers, such as those formed by alternating layers of ferromagnetic iron and chromium, can be nanostructured so that the electrical resistance is significantly decreased (by up to a factor of 2 depending upon the chromium layer thickness) by the application of a magnetic field. Such an effect, called giant magnetoresistance, occurs when the magnetic moments of the neighboring alternating layers are arranged in an antiparallel fashion, so that application of the magnetic field overcomes the antiferromagnetic coupling and aligns the layers into a condition of parallel ferromagnetic ordering, strongly reducing the electron scattering in the system. Magnetoresistive materials have been introduced in the magnetic recording industry as read heads because of their lower noise and improved signal-handling capabilities. Other nanostructured materials besides multilayers also exhibit giant magnetoresistance, such as magnetic cobalt clusters embedded in a nonmagnetic matrix of copper or silver, or magnetic platelike nickel-iron deposits embedded in silver. See MAGNETORESISTANCE.

Optical properties of cluster assemblages. Noninteracting assemblages of small semiconductor clusters have optical properties of both scientific and technological importance. The optical absorption behavior of cadmium sulfide clusters with diameters in the nanometer size regime made by any of a variety of methods, including chemical precipitation in solutions or in zeolite supports, is different from that for bulk cadmium sulfide. The absorption edge is blue-shifted to appreciably shorter wavelengths because of the effects of quantum confinement in these nanoscale clusters. However, when these clusters are synthesized in zeolite supports with increasing loading such that they become close enough to begin to interact through quantum tunneling, the absorption edge begins to shift back toward bulk behavior (Fig. 2). A similar effect can be created by chang-

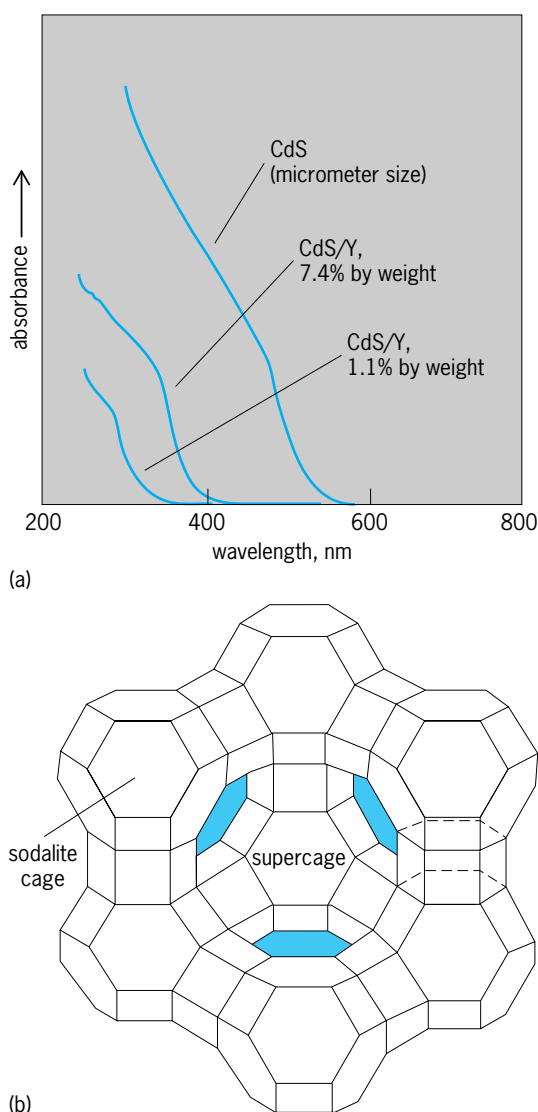


Fig. 2. Optical properties of cadmium sulfide (CdS) cluster assemblages. (a) Optical absorption spectra of CdS clusters in sodalite cages of zeolite Y for two different loadings compared to that for bulk CdS (micrometer size). (b) Cage structure of zeolite Y. (After G. D. Stucky and J. E. McDougall, *Quantum confinement and host/guest chemistry: Probing a new dimension*, *Science*, 247:669–678, 1990)

ing the sizes of the clusters in colloidal suspensions and thereby changing the degree of quantum confinement. Hence, control of the average distance between clusters in the zeolite cages, even though they are not actually in contact, or control of the cluster sizes in a suspension can enable control over a macroscopic property of the assembled cluster ensemble. Such quantum size effects provide a basis for verifying the understanding of the electronic structure of condensed matter, and may also provide for engineered optical properties. *See* ZEOLITE.

Chemical reactivity. Chemical reactivity of nanostructured materials, with their potentially high surface areas compared to conventional materials, can also be significantly altered and enhanced. Since clusters can be assembled by means of a variety of methods, there can be an excellent degree of control over the total available surface area in the resulting self-supported ensembles. Thus, it is possible to maximize porosity to obtain very high surface areas, remove most of it via consolidation but retain some to facilitate low-temperature doping or other processing, or fully densify the nanophase material. Also, control of chemical composition can be readily achieved, since rapid atomic diffusion paths are plentiful and diffusion distances are short in the clusters. Measurements of the decomposition of hydrogen sulfide over lightly consolidated, high-surface-area nanophase titania with a rutile crystal structure have clearly demonstrated the potential for enhanced chemical reactivity of nanophase materials. Nanophase rutile is far more reactive initially than any other available forms of titania and, more importantly, remains so even after extended exposure to the hydrogen sulfide at 500°C (932°F). This greatly enhanced activity results from a combination of unique and controllable features of the nanophase material, its high surface area combined with its rutile structure, and its oxygen-deficient composition, and makes this material suitable for a variety of catalytic and sensor applications. *See* NANO-CHEMISTRY.

Nanophase materials. The assembly of larger atom clusters into bulk nanophase materials can also have dramatic effects upon properties. In this case, the clusters interact fully with one another, yet the effects of cluster size are still very important. Clusters of metals or ceramics in the size range 5–25 nm have been consolidated to form ultrafine-grained polycrystals that have mechanical properties remarkably different and improved relative to their conventional coarse-grained counterparts. For example, nanophase copper and palladium assembled from clusters with diameters in the range 5–7 nm can have hardness and yield-strength values up to 500% greater than in the conventionally produced metal. This greatly increased strength arises from the increased difficulty in the spatially confined grains of nanophase metals in creating and moving dislocations, the defect normally responsible for the relatively easy deformation process in metals. In ceramics, however, which are normally difficult to deform and hence very brittle, cluster assembly yields a different benefit. Ceramics can be rendered ductile (ca-

pable of easier deformation) by being synthesized from clusters with sizes below about 15 nm. This ductility results from the increased ease with which the ultrafine grains created by the clusters can slide by one another in a process called grain-boundary sliding, owing to the short diffusion distances required for the necessary local healing of incipient cracks that could otherwise form during this grain-over-grain sliding process. Nanocomposites consisting of metallic phases, ceramic and metallic phases (cermets), and ceramic phases in a variety of modulation dimensionalities also have considerably enhanced mechanical properties, including increased strength and fracture toughness. *See* CERAMICS; COMPOSITE MATERIAL; GRAIN BOUNDARIES; METAL, MECHANICAL PROPERTIES OF. Richard W. Siegel

Bibliography. G. C. Hadjipanayis and R. W. Siegel (eds.), *Nanophase Materials: Synthesis, Properties, Applications*, 1994; P. Jena, S. N. Khanna, and B. K. Rao (eds.), *Physics and Chemistry of Finite Systems: From Clusters to Crystals*, 1992; B. H. Kear et al., *Research Opportunities for Materials with Ultrafine Microstructures*, National Materials Advisory Board, NMAB-454, 1989.

Nanotechnology

Techniques and products involving nanometer-scale structures, with dimensions ranging from 1 to 100 nanometers, especially those that transform matter, energy, and information using nanometer-scale components with precisely defined molecular features.

In the late 1980s, the term nanotechnology entered widespread use to describe anticipated technologies based on the use of molecule-based machine systems designed to build complex products with atomic precision. Since the mid-1990s, usage has broadened to embrace instruments, processes, and products in which key dimensions are in the 1–100-nm range. Technologies that fit this definition are extremely diverse, but many could potentially contribute to the development of new products and processes such as advanced molecular manufacturing.

Progress in nanotechnology can be judged by several metrics, including the increasing precision, complexity, cost-effectiveness, and scale of its products. The corresponding long-term objectives are atomic precision, arbitrary complexity, low-cost production, and large-scale products. This combination of objectives appears feasible, but only through a multistage process starting with the more limited capabilities of current nanoscale technologies.

Nanoscale technologies are extremely diverse, rapidly changing, and often only tenuously connected. Products include nanoscale particles, fibers, and films of diverse materials and structures; nanoscale lithographic structures for electronics (many integrated circuits now qualify); structures formed by spontaneous molecular aggregation (self-assembly); and solids containing nanoscale grains or pores. The means and materials used to produce

nanoscale and nanotextured structures often have little in common, and their applications range from stain-resistant clothing to state-of-the-art electronics. Many nanotechnologies are a continuation of pre-existing fields under a new label. What they share (particularly toward the lower end of the 1–100-nm range) is the emergence of novel properties, relative to the corresponding bulk materials, associated with surface and quantum effects, together with a distinctive set of instruments and computational modeling techniques. Grouping these diverse nanotechnologies together has fostered a vibrant cross-fertilization of disciplines. *See* MOLECULAR SIMULATION; MONOMOLECULAR FILM; NANOPARTICLES.

Looking forward, the metrics of complexity and scale define the chief frontiers. In small structures, precision has already reached the atomic limit. Examples include quantum dots, engineered biomolecular objects, self-assembled molecular structures, and sections of carbon nanotubes. For systems built with atomic precision, scale limits complexity. Great complexity is possible, even in nanosystems of microscopic scale. For example, a cubic micrometer of a typical material contains roughly 10^{11} atoms; with generalized atomic control on that scale, a cubic micrometer could contain roughly 10^9 distinct functional components. *See* CARBON NANOTUBES; NANOSTRUCTURE; SUPRAMOLECULAR CHEMISTRY.

Although complex systems with precise molecular features cannot be made with existing techniques, certain nanosystems can be designed and analyzed. Systems based on mechanical (rather than electronic) degrees of freedom are particularly tractable. These are of special interest, because programmable nanoscale mechanical systems could be used to produce atomically precise structures of arbitrary complexity. The development of productive nanosystems is a key strategic objective.

Productive nanosystems. Nature demonstrates a class of productive nanosystems based on polymeric components operating in a liquid medium. Ribosomes, for example, work as digitally controlled machine tools that read genetic information (six bits per codon) and use it to direct the assembly of sequences of amino acids. The resulting polymers (proteins) fold to make nanometer-scale objects with precise arrangements of atoms. Together with nucleic acids (made by other programmable machines), these molecular objects form the working parts of a wide range of biological molecular machines, including ribosomes themselves. These larger structures are made through a self-assembly process in which Brownian motion brings components together and complementary molecular surfaces cause selected components to bind in a precise manner. *See* MOLECULAR RECOGNITION; PROTEIN; RIBOSOMES.

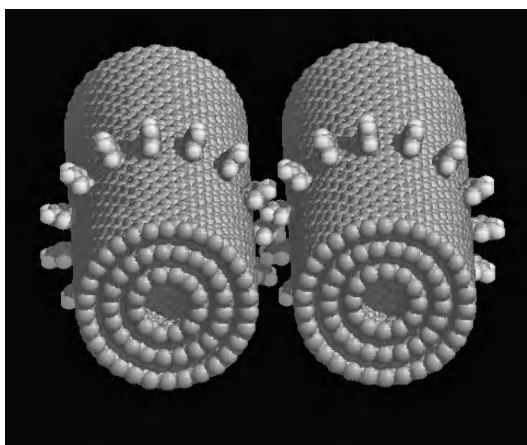
Biological systems provide not only examples of productive nanosystems but also tools to use and models for the development of next-generation systems. Developments in the design of novel protein and nucleic acid structures, including simple molecular machines, can be combined with a broad spectrum of current nanotechnologies in the construc-

tion of early-generation productive nanosystems. Self-assembly can join engineered biomolecular objects with atomically precise nanoparticles, fibers, and other products to make composite structures with properties beyond those found in biological systems.

As this path moves closer to fundamental physical limits, one direction of advance is toward better materials. Biomolecules have large monomers, low bond density, and low stiffness; advanced materials with higher bond densities can be stronger, stiffer, more fine-grained, and more regular. As with macroscopic machinery, the use of superior materials will enable better performance.

Although precisely structured devices with high bond densities cannot be made yet, they are amenable to computational design and modeling. Among the devices that have been analyzed are gears, shafts, bearings, belts, electric motors, computers, and programmable positioning mechanisms. These could be used to transform matter, energy, and information. Their components resemble those of conventional machine systems. For example, a design for a gearbox for transforming shaft power from higher to lower angular speeds would consist of a rigid framework supporting input and output shafts mounted on bearings. The gears would have teeth and obey the usual rules regarding gear ratios; however, in detail, the differences would be substantial. Each gear tooth typically would consist of a single row of atoms, with the smoothness of interatomic force fields enabling the gears and bearings to operate without an added lubricant (*see illustration*). Systems containing enough parts to serve the function of a programmable molecular assembly mechanism or a computer require several million to several billion atoms, and their volumes occupy a significant fraction of a cubic micrometer.

Basic principles. Nanotechnology based on productive nanosystems requires a combination of familiar molecular and mechanical principles in unfamiliar applications. In solution-phase reactions, molecules move by diffusion and encounter each other in all possible positions and orientations. The resulting



Simulation of a carbon nanotube-based gear. (Courtesy of NASA Ames Research Center)

molecular transformations are accordingly difficult to direct. Productive nanosystems, in contrast, could exploit mechanosynthesis, the use of mechanical devices to guide the motions of reactive molecules. By applying positional control to conventional molecular reactions, mechanosynthesis could cause structural changes to occur at precise locations in a precise sequence. Reliable positioning is required in order for mechanosynthetic processes to construct objects with millions to billions of precisely arranged atoms.

Mechanosynthetic systems are intended to perform several basic functions. Their first task is to bind raw materials from an externally provided source, typically a liquid solution containing a variety of useful molecular species. This process must separate molecules of different kinds and bind them reliably to specific sites. A second task (in advanced systems) is to transform bound molecules into highly active chemical species, such as radicals, carbenes, and strained alkenes and alkynes. Finally, mechanical devices can apply these bound, active species to a workpiece in controlled positions and orientations to deposit or remove a precise number of atoms of specific kinds at specific locations. *See* REACTIVE INTERMEDIATES.

To support these functions, it is necessary to provide appropriate mechanisms and conditions. Binding, transforming, and moving molecules can best be accomplished by nanoscale molecular machinery. To minimize friction, contamination, and side reactions, the ideal environment is free of fluids, which require a suitable enclosure and pumping mechanisms. If a mechanosynthetic system is to build complex products (rather than performing simple, repetitive operations), a programmable positioning mechanism could be used, requiring a source of instructions to guide its sequence of movements. Early-generation systems will sacrifice efficiency and range of products in exchange for simpler mechanisms, for example, by operating in an ambient fluid environment and using positioning mechanisms with fewer degrees of freedom.

Successful designs must overcome several challenges to reliable operation. Both quantum-mechanical uncertainty and thermal vibration cause random displacements in the positions of parts. For nanomechanical parts at room temperature, thermal vibration is overwhelmingly more important than quantum uncertainty. Because the mean-square displacement of a part is inversely proportional to the stiffness of the structure that holds it in place, the amplitude of thermal vibration could be limited in some circumstances by careful design. In mechanosynthetic systems, some vibrational amplitudes could be limited to less than one-tenth of an atomic radius. Because the probability of displacement has a Gaussian distribution, transient misalignments as large as an atomic diameter could be made extremely rare. *See* QUANTUM CHEMISTRY.

Infrequent, high-amplitude thermal vibrations can break even strong chemical bonds. Breaking a single bond in a molecular machine would typically cause

it to fail. In the terrestrial ambient background radiation environment, structures with reasonable thermal stability will experience bond breakage chiefly from ionizing radiation. The rate of device damage due to ionizing radiation is roughly proportional to device mass, and devices on a scale of hundreds of nanometers will last many decades in the terrestrial environment before encountering radiation damage. To be reliable, larger systems either must be made of parts large enough to tolerate some damage or must be organized redundantly so that the system itself can tolerate some failed parts. Photochemical damage can be prevented by enclosing systems in opaque shields; 0.25-micrometer-thick aluminum is ample for long-term protection in full sunlight.

Reliable molecular manufacturing systems have strong similarities to digital computers. In conventional materials processing, as in analog electronics, all operations are somewhat imprecise. Each object produced has a unique size, shape, composition, and microstructure, differing both from the ideal design and from all previous objects. In productive nanosystems, though, as in digital electronics, each operation is either entirely correct or clearly wrong. In digital logic, the result of an operation is a specific pattern of ones and zeros. No stable intermediate states are possible, and physical principles enable the design of circuits that produce the correct pattern with high reliability. In productive nanosystems, the result of an operation would be a specific pattern of bonded atoms. For suitable choices of product structure, no ambiguous states are possible, and physical principles would enable the design of processes that produce the correct pattern with high reliability.

Applications. Much as digital computers can produce an indefinitely large range of patterns of information, productive nanosystems could produce an indefinitely large range of patterns of matter. For both digital computers and productive nanosystems, unlike traditional devices such as an adding machine or a lathe, it is difficult to describe the range of applications.

Because productive nanosystem technology could make precise, nanoscale features, these methods could be used to fabricate improved circuitry for digital logic. It is widely recognized that diamond would be superior to silicon for this purpose, if it could be fabricated with comparable ease. Advanced productive nanosystems could make this practical. Design calculations based on simpler, mechanical nanocomputers place a lower bound on what could be achieved. It appears feasible to build a central processing unit for a computer that occupies a volume less than 1 cubic micrometer, consumes roughly 0.1 microwatt of power, and executes about 1 billion instructions per second.

Nanotechnologies based on productive nanosystems will enable superior molecular-scale sensing and manipulation and thus allow development of a broad range of novel scientific and medical instruments. A particularly important goal is to provide better instrumentation for probing the molecular

structure of cells and providing data regarding structure and function at the molecular level in a more direct manner than is presently possible. The application of this knowledge to medicine could enable the development of nanoscale medical devices of greater complexity and capability than modern pharmaceuticals, and of far greater precision than modern surgical instruments.

K. Eric Drexler

Bibliography. K. E. Drexler, *Nanosystems: Molecular Machinery, Manufacturing, and Computation*, Wiley/Interscience, 1991; D. S. Goodsell, *Bionanotechnology: Lessons from Nature*, Wiley-Liss, 2004; J. Storrs Hall, *Nanofuture: What's Next for Nanotechnology*, Prometheus Books, 2005; M. Wilson et al., *Nanotechnology: Basic Science and Emerging Technologies*, University of New South Wales Press, Sydney, 2002.

Naphtha

Any one of a wide variety of volatile hydrocarbon mixtures. They are sometimes obtained from coal tar, but more often they are derived from petroleum. Physical properties vary widely. The initial boiling point may be as low as 80°F (27°C), and end points may reach 500°F (260°C). Boiling ranges are sometimes as narrow as 20°F (11°C) or as wide as 200°F (110°C).

The main process for producing naphthas is fractional distillation. It may be of the extractive type when certain high-quality naphthas are desired. Acid treating, clay treating, and other techniques remove sulfur compounds and improve color, odor, and stability.

Strictly speaking, the refinery streams going into products like gasoline and kerosine are naphthas, and they are so designated within the petroleum industry. The final blended fuels, however, are sold under the more familiar names. The products sold as naphthas find their greatest use as solvents, thinners, or carriers.

Few naphthas are made up entirely of hydrocarbons belonging to one particular family. There is a fairly sharp differentiation, however, between aliphatic and aromatic types.

Aliphatic naphthas are relatively low in odor and toxicity and tend, also, to be low in solvent power, which in some cases is an advantage. In the processing of soybeans, for example, the aim is to extract the oil without extracting the less desirable materials. Naphthas used by dry cleaners likewise require only moderate solvent power. In printing ink, the naphtha is mainly a carrier of the carbon black or other pigments; resins requiring a solvent are present in only minor amounts.

The aromatic naphthas, often described as the high-solvency type, at one time came entirely from coal tar. The development of catalytic cracking and catalytic reforming made petroleum an alternative source. The main components are toluene and xylenes; benzene is less desirable because of the extreme toxicity of its vapors. A major use of

these naphthas is as thinners for paints and varnishes, to permit easy brushing. Both varnishes and enamels contain large amounts of gums and resins, and diluents with good solvent action are therefore needed.

The rubber industry also uses naphthas as solvents. The leather industry uses them to degrease skins, the metal industry to degrease metals. Naphthas in insecticides and weedkillers dissolve the toxic agents and often contribute toxic properties of their own. Floor waxes, furniture waxes, shoe polishes, metal polishes, and dry cleaners' soaps are among the many other products in which naphthas are used. See PETROLEUM PROCESSING AND REFINING; PETROLEUM PRODUCTS.

J. K. Roberts

Narcotic

A drug which diminishes the awareness of sensory impulses, especially pain, by the brain. This action makes narcotics useful therapeutically as analgesics. While they are the most powerful pain-relieving agents available, their use is complicated by a number of undesirable side actions. Indeed, much research in this area has been directed toward a search for an agent having the same degree of analgesic properties as morphine, the most widely used of the narcotics, but without its undesirable side actions. See ANALGESIC.

All of the generally used narcotics are in some way related to opium, and the term opiate is sometimes used interchangeably with the term narcotic. Opium is a gummy exudate obtained from the unripe seed capsules of the opium poppy. The narcotic effects of opium have been known from ancient times. Crude opium contains over a dozen alkaloids, all of which have been isolated and identified as to their structural chemistry. From this knowledge chemists have developed a number of synthetic chemical compounds, some of which have important advantages over the naturally occurring alkaloids. Therapeutically important natural alkaloids are morphine, codeine, and papaverine. Among the important synthetic narcotics are meperidine (Demerol), dihydromorphine (Dilaudid), oxymorphone (Numorphan), alphaprodine (Nisentil), anileridine (Leritine), piminodine (Alvodine), levorphanol (Levo-Dromoran), methadone (Dolophine), and phenazocine (Prinadol). See ALKALOID; OPIATES; POPPY.

Nalorphine (Nalline) is a narcotic antagonist and is used in the treatment of acute overdosage from narcotics; it is dangerous to drug addicts. Heroin is a highly addicting narcotic, and is so dangerous in this regard that the drug has been completely banned by both federal and state laws under all circumstances.

The pharmacology of narcotics is generally similar to that of morphine, the principal narcotic used for its analgesic effects. Differences among them lie in the potency of their action and in the degree and variety of the side actions which they produce. Effects are those of analgesia, accompanied

by a state of euphoria characterized by drowsiness and a change of mood from anxiety and tension to calmness and equanimity. There is a growing belief among practitioners that this state of euphoria is a valuable component of narcotic agents used for the relief of pain. For example, methadone which has high analgesic properties but little euphoria is not as widely used as morphine, which along with its relief of pain produces a relatively high degree of euphoria. It should be remembered that whatever narcotic is used, the effects are dose-related, and in higher doses all narcotics produce deep sleep and eventually general depression of all brain functions. Death from overdosage is due to depression of the respiratory centers with resultant failure of respiration.

Pharmacology

It is most convenient to discuss the pharmacological actions of the narcotics with specific reference to morphine, since it is the most widely used and typical narcotic analgesic.

Action. The predominant pharmacological effect of morphine (and the other narcotics) is on the central nervous system. From the standpoint of its medicinal use, its most important action is relief of pain. Researchers in the field of pain and analgesics have come to distinguish between pain produced by traumatic injury or physiological malfunction and the mental anguish which can be termed suffering. The former is a specific sensation resulting from sensory input, while the latter is the psychological reaction to this sensory pain. The narcotic analgesics relieve both pain and suffering, and indeed this double property is perhaps what has made them important for a long time in the practice of medicine. The action of morphine in the first regard is through depression of the pain reception centers of the primitive brain. Its action in the second regard is through a disruption of association pathways of the cortex, resulting in euphoria.

Side effects. Morphine classically illustrates the basic axiom in pharmacology that a drug always exerts multiple actions. Along with its valuable medicinal use morphine produces a great many undesirable side actions; the most frequent are worth mentioning, although they do not occur in every patient or at every dose level.

Depressed respiratory activity. Morphine is a depressant of the respiratory center of the brainstem. This depression is minimal in the usual doses used to produce analgesia, but it increases as the dose is increased. Death from ingestion of an overdose is from respiratory depression; treatment for overdose is directed toward maintaining respiratory function.

Effect on vomiting. Morphine and most of the other narcotic analgesics produce nausea and vomiting as the result of stimulation of the vomiting center of the brain. This illustrates the curious pharmacological effect of morphine on the brain, which is a mixture of depression of some brain centers and stimulation of others. One of the derivatives of morphine,

apomorphine, is predominantly a stimulant of the vomiting center with no significant analgesic action. It is used to induce vomiting, such as in the oral ingestion of an overdose of a poison. Curiously, as the dose of morphine is increased, the stimulation of the vomiting center is replaced by a profound depression of this center and obliteration of the vomiting reflex.

Inhibited defecation and urination. The general effect of morphine on the gastrointestinal tract is depressive, exerted directly on the smooth muscle and the nerve plexi within the intestinal wall. Peristaltic activity is diminished or abolished, producing constipation if morphine is used for any length of time. This effect is reflected in the use of opium from earliest times as a treatment for the relief of diarrhea and dysentery.

Morphine produces an increase in contractile tone of the detrusor muscle at the neck of the urinary bladder. This results in urinary retention, and may precipitate problems in prostatic hypertrophy.

Other effects. With morphine the cutaneous blood vessels dilate and the skin becomes flushed and warm with an increase in perspiration. The pupil is constricted, and the "pinpoint" pupil is often present in narcotic addicts.

Contraindications. Morphine, and indeed all of the narcotic analgesics, is contraindicated as premedication in childbirth. These small molecules pass freely across the placental barrier into the fetal circulation and produce respiratory depression at a time when the establishment of respiration is particularly important and the function especially vulnerable immediately following delivery.

Physicians use narcotics with great caution in the very young and the very old, since these groups are particularly sensitive to narcotics.

Medical Uses

Despite the undesirable and unwanted side actions and the possibility of addiction, narcotic analgesics remain among the most useful drugs in medical practice. These drugs can be used safely and effectively by a skillful analysis of the patient, and by selection of the proper narcotic analgesic beginning (as in the case of postoperative pain) with morphine, changing in a short time to Demerol or codeine, and then as the intense pain diminishes making use of one of the effective nonnarcotic analgesics such as Darvon or even acetylsalicylic acid (aspirin). The danger of producing addiction is minimal if the narcotic analgesics are used for relatively short periods for the control of intense pain. In instances such as terminal cancer, the pain is severe and of long duration, and the need for relief of pain and suffering prevails over the possibility of producing addiction.

Drug Dependence

All narcotics have the potential for producing dependence and addiction when used repeatedly over a period of time. Drug dependence results from compulsive, continued use of the drug, and is

characterized by one or more of the following conditions: habituation, tolerance, or addiction.

Habituation. Like any other habit pattern, habitual use of a drug can develop. Common examples are the use of nicotine in the form of cigarettes, or caffeine in the form of coffee or tea. Such habituation is generally regarded as innocuous. Stopping the use of a habituating drug may be attended with mental distress, but there is no major physical disruption.

Tolerance. Repeated ingestion of a drug in which the effect produced by the original dose no longer occurs results in tolerance. To produce the original effect, it is necessary to increase the dose. A common example is the rapid disappearance of the central stimulant effect of caffeine in the habitual coffee drinker.

Addiction. When the body develops a dependence for the drug, addiction occurs. If the drug is suddenly stopped after a period of frequent use, a withdrawal syndrome develops, which is characterized by physical pain and widespread body reactions. The addict comes to dread the development of such painful and distressing reactions, and is trapped into continuing the drug.

All narcotics can produce habituation, tolerance, and addiction to a greater or less degree. Addiction to codeine is relatively rare but possible. Addiction to heroin develops rapidly, and this narcotic is therefore exceedingly dangerous.

Unquestionably, the nonmedical use of narcotics, with the concomitant social problem of the drug addict, is the result of the euphoric property of narcotics. Some cases of addiction are the result of unwise use of powerful analgesic narcotics for the relief of pain in patients; but by far the greatest number of cases comes from the use of narcotics for "escape." Heroin is almost universally used by narcotic addicts; its potency is greater than that of morphine and its euphoric effect is high. It is most effective when injected intravenously. Morphine is easily converted to heroin by a simple process of acetylation.

The total treatment of narcotic addiction is far from satisfactory. It is fairly easy under proper hospital conditions to manage the withdrawal syndrome and to get the addict off the drug. The most simple and brutal way of doing this is the "cold-turkey" method, in which the addict is placed in a room and allowed to go through the 2 or 3 days of agonizing withdrawal. A longer procedure is that of gradually decreasing the amount of drug which the patient receives over a period of 2 or 3 weeks. At the end of that time the drug may be stopped altogether with a minimum of withdrawal symptoms. In the methadone treatment methadone is substituted for heroin, and after a period of time it may be stopped without development of the withdrawal syndrome. Regardless of the method of treatment, the addict must be rehabilitated after getting off the drug. This is the most difficult part of the treatment and involves psychotherapy and social readjustment. Too often after a successful management of the acute condition of addiction, the addict returns to his old environment

and habit patterns and once again becomes addicted. See ADDICTIVE DISORDERS.

Legal Regulations

Federal control of narcotics is embodied in the Harrison Narcotic Act, a basic law adopted by Congress in 1914. The purpose of the act was essentially twofold: to develop a source of revenue, and to bring the domestic control of narcotic agents into conformity with the nation's obligations to the Hague Convention of 1912, which was directed toward the control of international drug traffic. Because of the revenue aspects, the act was placed under the jurisdiction of the Treasury Department. It became the means for controlling domestic drug traffic and for preventing illicit sale and use of drugs by purveyors and addicts.

Drugs under control. Originally the federal legislation included opium and its derivatives, and coca leaves, from which cocaine was prepared. Later, additional drugs were included, some of which were newly discovered alkaloids obtained from opium; others were synthetic derivatives. Some of the drugs specified by the act are not narcotic; papaverine, for example, is a smooth-muscle relaxant; apomorphine is an emetic; and nalorphine is a narcotic antagonist. Of special interest was the inclusion in 1937 of marijuana and its derivatives. While marijuana is liable to abuse, it cannot be defined as a narcotic in a pharmacological sense. See COCA; MARIJUANA.

To add to the complexities of controlling the illegal use of dangerous drugs, the Food and Drug Administration was empowered under the Kefauver Law to control certain dangerous drugs. Among these are the barbiturates, tranquilizers, amphetamines, and lysergic acid diethylamide (LSD). To meet this responsibility, the Food and Drug Administration set up the Bureau of Drug Abuse Control. See BARBITURATES; PSYCHOTOMIMETIC DRUGS; TRANQUILIZER.

Jurisdiction. State laws concerning narcotics and other drugs liable to abuse vary extensively between the states and may be more stringent, but not less so, than the federal law. It can therefore be seen that the whole problem of drug abuse, from a legal standpoint, can become unwieldy and complex because of the division of responsibility between the Food and Drug Administration and the Treasury Department. In 1968 a reorganization was undertaken, and the whole problem of legal regulations and their enforcement was brought into the Justice Department.

Licensing provisions. The Federal Narcotic Act applies to different areas, such as import, export, manufacture, pharmacies, and hospitals, as well as to physicians, dentists, and veterinarians. The Bureau of Narcotics provides for four classes of licenses. Class I is generally used by physicians, and provides for the prescribing of the agents named in the Federal Narcotic Act on a properly written prescription to be dispensed by a licensed pharmacist. The class II license provides, in addition to the legal right to prescribe narcotics, the right to store and use these in the physician's office or clinic. Also, the physician

must have proper locked storage, keep an inventory of the stock on hand, and be prepared for an accounting to a federal agent at any time. Class III and class IV licenses apply to pharmacists, wholesalers, manufacturers, importers, and exporters. *See* MORPHINE ALKALOIDS; PAIN. James M. Dille

Native elements

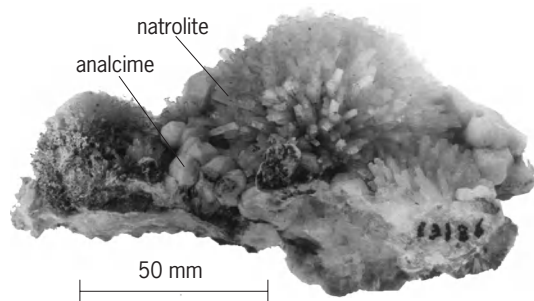
Those elements which occur in nature uncombined with other elements. Aside from the free gases of the atmosphere there are about 20 elements that are found as minerals in the native state. These are divided into metals, semimetals, and nonmetals. Gold, silver, copper, and platinum are the most important metals, and each of these has been found abundantly enough at certain localities to be mined as an ore. Native gold and platinum are the major ore minerals of these metals. Rarer native metals are others of the platinum group, lead, mercury, tantalum, tin, and zinc. Native iron is found sparingly both as terrestrial iron and meteoric iron. *See* ORE AND MINERAL DEPOSITS.

The native semimetals can be divided into the arsenic group, including arsenic, antimony, and bismuth; and the tellurium group, including tellurium and selenium. The members of the arsenic group crystallize in the hexagonal system, scalenohedral class; those of the tellurium group in the hexagonal system, trigonal trapezohedral class. Only rarely do the semimetals occur abundantly enough to be mined as ores of their respective elements.

The native nonmetals are sulfur, and carbon in the forms of graphite and diamond. Native sulfur is the chief industrial source of the element. *See* MINERAL; MINERALOGY. Cornelius S. Hurlbut, Jr.

Natrolite

A fibrous or needlelike mineral belonging to the zeolite family of silicates. It crystallizes in the monoclinic system in pseudo-orthorhombic prismatic crystals which are often acicular (see *illus.*). Most commonly it is found in radiating fibrous aggregates. There is perfect prismatic cleavage, the hardness is 5-5½



Natrolite crystals on analcime. (American Museum of Natural History specimen)

on Mohs scale, and the specific gravity is 2.25. The mineral is white or colorless with a vitreous luster that inclines to pearly in fibrous varieties. The chemical composition is $\text{Na}_2(\text{Al}_2\text{Si}_3\text{O}_{10}) \cdot 2\text{H}_2\text{O}$, but some potassium is usually present substituting for sodium.

Natrolite is a secondary mineral (low-temperature hydrothermal mineral) found lining cavities in basaltic rocks, where it is associated with other zeolites, calcite, apophyllite, and prehnite. Its outstanding locality in the United States is at Bergen Hill, New Jersey. *See* SILICATE MINERALS; ZEOLITE.

Clifford Frondel; Cornelius S. Hurlbut, Jr.

Natural fiber

A fiber obtained from a plant, animal, or mineral. The commercially important natural fibers are those cellulosic fibers obtained from the seed hairs, stems, and leaves of plants; protein fibers obtained from the hair, fur, or cocoons of animals; and the crystalline mineral asbestos. Until the advent of the manufactured fibers near the beginning of the twentieth century, the chief fibers for apparel and home furnishings were linen and wool in the temperate climates and cotton in the tropical climates. However, with the invention of the cotton gin in 1798, cheap cotton products began to replace the more expensive linen and wool until by 1950 cotton accounted for about 70% of the world's fiber production. Despite the development of new fibers based on fossil fuels, cotton has managed to maintain its position as the fiber with the largest production volume, although its use has fallen. *See* MANUFACTURED FIBER.

Cordage fiber consumption has also changed over the years. True hemp was largely replaced by abaca (Manila hemp) during the nineteenth century, and by 1940 essentially all heavy twine and rope in the United States was made of sisal, henequen, or abaca. Now, in their turn, synthetic fibers supply more than half of the United States cordage requirements.

The natural fibers may be classified by their origin as cellulosic (from plants), protein (from animals), and mineral. The plant fibers may be further ordered as seed hairs, such as cotton; bast (stem) fibers, such as linen from the flax plant; hard (leaf) fibers, such as sisal; and husk fibers, such as coconut. The animal fibers are grouped under the categories of hair, such as wool; fur, such as angora; or secretions, such as silk. The only important mineral fiber is asbestos, which because of its carcinogenic nature has been banned from consumer textiles. *See* TEXTILE.

Plant Fibers

The basic biological unit of plant fibers is the plant cell, with a length frequently more than a thousand times its diameter. The cotton fiber is a single cell. The bast-fiber and hard-fiber strands are composed of many overlapping, parallel fiber cells held together in a fiber bundle that may be as long as the leaf or stem from which it comes. *See* PLANT CELL.

The fiber cells are made up of bundles of microfibrils which in turn are made up of ordered bundles of cellulose chains which are highly crystalline in nature. The parallel cellulose chains are held together laterally by hydrogen bonds, which are also involved in holding together the thin microfibrils and the internal structure of the cell. *See* CELLULOSE.

The direction that the microfibrils take in the cell determines the elongation character of the fiber. The cotton cell structure is characterized by gross convolutions and a spiraling of the microfibrils which frequently reverse along the length of the cotton fiber. In the bast and hard fibers, the microfibrils are more nearly parallel to the fiber axis (Fig. 1).

Cotton. The cotton plant is a member of the mallow family, Malvaceae. Its genus, *Gossypium*, includes herbs and shrubs that attain heights of 10 in. to 30 ft (25 cm to 9 m).

After pollination, the cotton boll takes about 50 days to ripen. Within the boll the ovules mature simultaneously with the covering seed hairs. During the early stages of growth after pollination, the hair grows only in a longitudinal direction, becoming tubelike and elongated, and is surrounded by a wall of cellulose. The outside of the latter membrane consists of wax and fatty substances, which are impermeable to gas and water, and they give the fiber its silkiness and luster. After 24 days another layer of cellulose is deposited on the inside of the cell wall. Concentric rings are formed, corresponding to the days of growth. The lint hair lives until the time of the boll opening. As the lint hair dries, its tubelike form collapses and becomes very twisted in its length. The characteristic twists correlate with spiral fibril structures in the secondary wall (Fig. 2). There are reversals of this twist during the convolutory growing process.

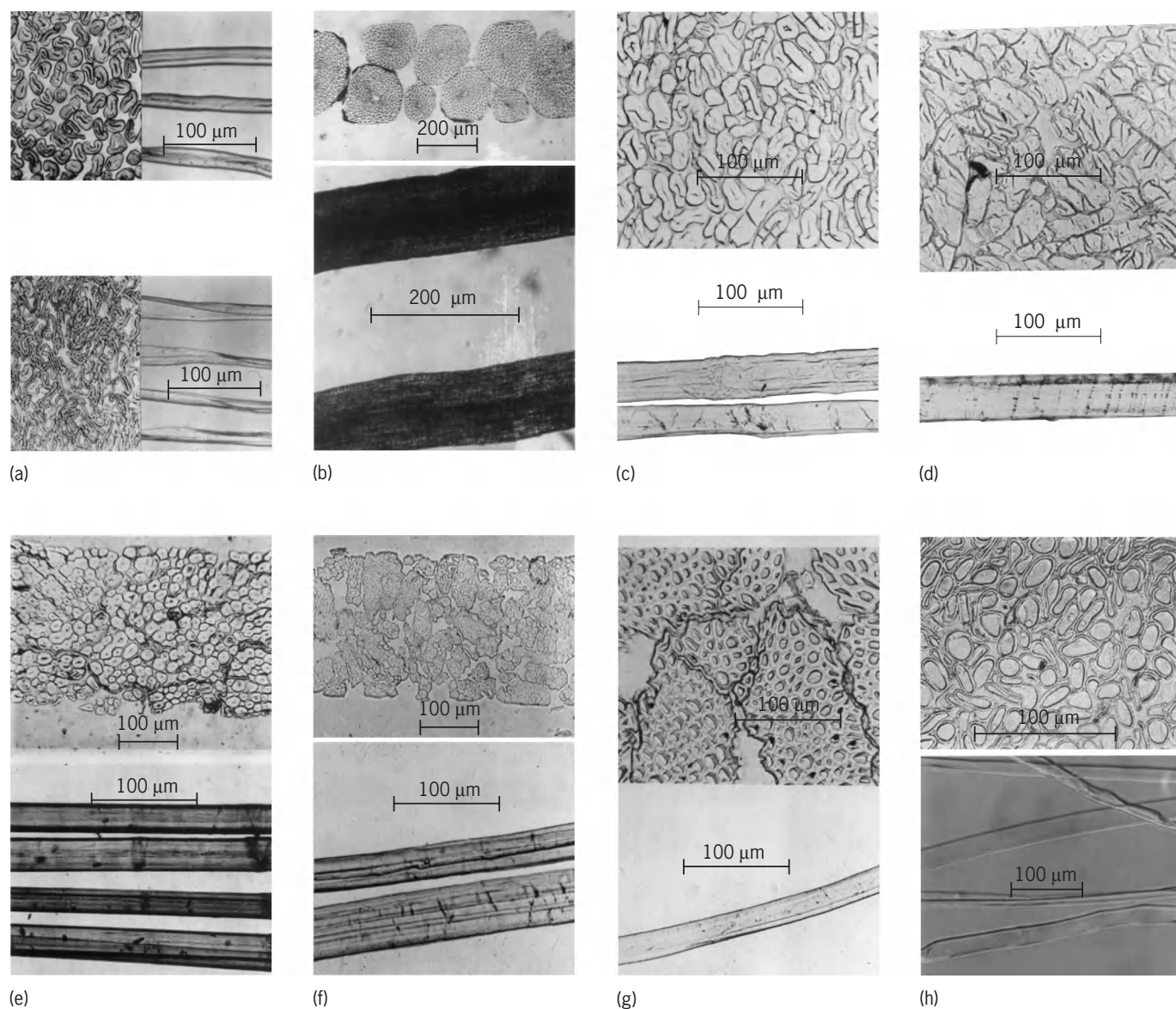


Fig. 1. Cross and longitudinal sections of important vegetable fibers. (a) Top: very immature cotton; bottom: average mature cotton. (b) Coir. (c) Ramie. (d) Hemp. (e) Flax. (f) Jute. (g) Sisal. (h) Kapok.

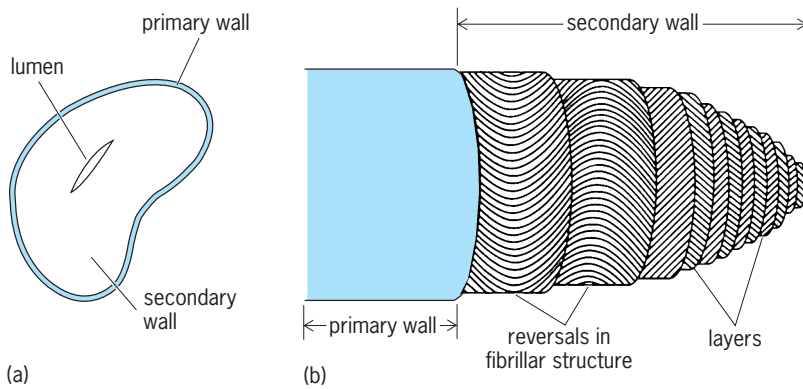


Fig. 2. Schematic of cotton: (a) cross section showing primary and secondary walls and lumen; (b) layered fibrillar structure of the walls. (After B. F. Smith and I. Block, *Textiles in Perspective*, Prentice-Hall, 1982)

After ripening, the cotton is harvested, ginned to remove the fiber from the seeds, and baled for transport to the spinsters. At the spinning plant, the graded and sorted fibers are further processed to remove trash and linters (very short fiber), and are twisted into yarn. Cotton yarn is either woven or knit into fabric which, depending upon its intended end use, may be subject to a variety of mechanical and chemical finishing treatments.

Fiber properties. The cotton fiber may be brown to creamy white; the whiter, more lustrous fibers are considered higher in quality. The fibers range in length from $\frac{1}{8}$ to $2\frac{1}{2}$ in. (0.3 to 6 cm), and may be from 10 to 20 micrometers in diameter. The longer, finer fibers are most prized. The fibers are moderately strong, with tenacities from 0.1 to 0.18 oz (3 to 5 g) per denier, dense (about 0.89 oz/in.³ or 1.54 g/cm³), stiff (elongation at break of 3–7%), and stronger wet than dry. Cotton is resistant to bleaches, most household chemicals, and sunlight, but is susceptible to attack by mildew and some insects. Cotton's major advantages are low cost, comfort (moisture regain is about 8.5%), good durability, and ease of dyeing and laundering. Its major disadvantages are poor wrinkle resistance and wrinkle recovery, with a concomitant requirement for ironing, and tendency to absorb waterborne stains, such as grass stains.

Uses. Cotton is the most widely used natural fiber. It is found in apparel, home furnishings, and industrial products. In particular, the development of durable-press treatments for cotton-polyester blends has made these cloths the most widely used of all the apparel fabrics. See COTTON.

Bast fibers. These fibers, which come from the phloem or bast of dicotyledonous plants, are also called soft fibers. The greatest quantity of these fibers goes into fabrics, but they have other uses such as in thread, twine, or paper. The more widely used bast fibers are hemp, jute, kenaf, linen, and ramie.

Bast fibers, as they go to the spinner, are in bundles of strands of varying lengths, with the longest ones approaching the length of the stem. Each strand of the commercial fiber is formed by long, overlapping ultimate fiber cells. The cells in the strand and the strands in the bundles are held together by non-

fibrous cementing substances composed of hemicelluloses, lignin, and pectinlike gummy substances which are alkaline-soluble.

The fibers are separated from the stem by a combination of mechanical and microbiological or chemical processes. The microbiological action, called retting, is of two types—water retting and dew retting. In water retting, the stems (or only the bark containing the fiber) are held beneath the surface of a tank, pool, or stream of water for several days, depending on the temperature, until bacteria break down the tissue surrounding the fiber and release it in the bark. In dew retting, the stems are spread on the ground after harvest and left for a few weeks while fungi invade the bark and release the fiber. Hemp and flax may be retted either way. Jute and kenaf are water-retted. Ramie is not retted, but the mechanically separated fiber is treated with chemicals.

Hemp. Hemp comes from the stems of the plant *Cannabis sativa*. The fiber is produced mainly in eastern Europe and central Asia. It is removed from the stem by retting and scutching. For centuries, hemp was the principal fiber used for marine cordage until replaced by abaca and sisal. It is still used extensively for twine and for many of the same products as linen.

The water-retted fiber is creamy white to brown; dew-retted fiber is gray to dark gray. The line fiber is usually about 4–7 ft (1–2 m) in length. The tow is also used for spinning, but goes into the coarser yarns. The quality of hemp fiber is based on factors similar to the ones for linen.

The fiber cells average about 22 micrometers in diameter and about 1 in. (25 mm) in length and are rather uneven in diameter. The lumen is larger than in linen and frequently, especially at the ends, is flattened so it appears as a line. The ends of the fibers are blunt and very thick-walled, and show some branching. This branching is a distinguishing characteristic of hemp fiber, which otherwise looks very much like linen under the microscope. Hemp nearly always exhibits some cells with forked or nodulated ends, whereas linen does not.

The manufacturing of hemp is similar to that of lower grades of linen, although the products are coarser. See HEMP.

Jute. Jute fiber is obtained from the stems of white jute (*Corchorus capsularis*) and from tossa and daisee (*C. olitorius*). The two species are similar, but *C. olitorius* fiber is slightly stronger and finer. The fiber is removed from the stem by retting and stripping the retted fiber from the stems by hand. Jute is grown mainly in Bangladesh and India, but Burma, Nepal, and Brazil produce sizable amounts.

Its suitability for uses requiring strong, bulky fabric or twine and its low cost compared to competitive fibers have made jute second only to cotton among the natural fibers in world consumption. The greatest quantity on a worldwide basis goes into fabric—hessian (known as burlap in the United States) and a heavier-weight fabric known as sacking. These are the standard wrapping and bag materials.

Jute fiber, usually 5–10 ft (1.5–3 m) or more in length, is creamy white to red-brown or gray, depending mainly on the species and the kind of water used for retting. The cells that make up the fiber strands average about 0.008 in. (2 mm) in length and about 20 μm in diameter, and are held together by natural plant gums. The longitudinal view shows the cells to be roughly cylindrical, with the ends varying from spearhead-shaped to tapering points. The cross-sectional view shows a polygon with pronounced oval lumen. Nodes or cross markings are usually absent, but are occasionally found.

While there is about as much cellulose in jute (about 75%) as in other retted bast fibers, it is a lignocellulose and the fiber is harsh.

Synthetic fibers, especially polypropylene, have made substantial inroads in the markets for jute. Polypropylene bags and prime back for tufted carpeting have displaced large quantities of jute. Also, bulk handling has eliminated much of the former market for grain bags, especially in the United States. See JUTE.

Kenaf. The name kenaf originally referred to the fiber *Hibiscus cannabinus*, but now also includes *H. sabdariffa*. Among some 125 common names for the fiber and plant, frequently used are Bimlipetam jute, mesta (or meshta), Thai (or Siamese) jute, Java jute, roselle, ambari, deccan hemp, stokroos, and dah.

The countries producing the most kenaf are India and Thailand. Other producing countries are Russia, Bangladesh, Indonesia, and Vietnam.

Kenaf is traditionally a substitute or alternate crop and fiber for jute. Its uses as a textile fiber are the same as for jute, and its products are usually sold as sacking, bagging, and so on, without reference to their containing kenaf. An important and increasing use is for paper.

Kenaf's appearance is similar to jute, although the fiber is somewhat lighter in color. Its strength is similar, but it is considered by some spinners to be less strong.

Linen. Flax is grown mostly in northern and eastern Europe. The fiber, linen, is separated from the stem of the plant by a system of retting and scutching. The line (long, parallel) fiber is used principally for fine damask, table linen, apparel fabrics, sheeting, lace, threads, and special twine. The tow (short, tangled fiber produced in the scutching and the hackling operations) and some of the coarser grades of line fiber are used for toweling, canvas, mail bags, meat wraps, twine, and high-grade papers such as cigarette, bond, and currency.

The factors that determine the quality of linen fiber are cleanliness, fineness, strength, density (as a hank of fiber held in the hand), color, uniformity, luster, length, and softness. Under the light microscope, the surface appears smooth to lightly marked longitudinally, with frequent transverse lines, often in the form of an X. In cross section, the cells appear somewhat polygonal to round, with cell walls that are fairly uniform in thickness (Fig. 3). The lumen is small and

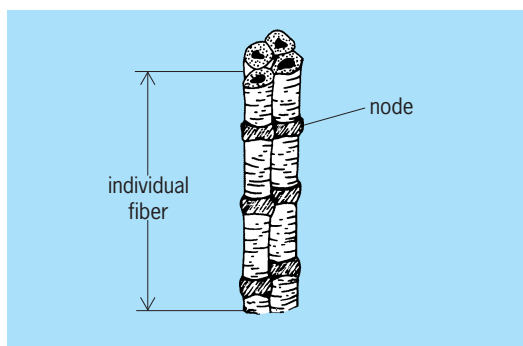


Fig. 3. Individual fibers in the flax stem. (After B. F. Smith and I. Block, *Textiles in Perspective*, Prentice-Hall, 1982)

distinct, and disappears near each end as the cell tapers to a point.

The fiber varies in color from light ivory to dark gray, in length from 1 to 3 ft (30 to 100 cm), and is about 20 μm in diameter. It is strong (tenacity about 0.21 oz or 6 g per denier), is very stiff (elongation at break from 2.7 to 3.3%), has a moisture regain of about 11%, and is very smooth. Its properties are very similar to those of the very best grades of cotton. See FLAX.

Ramie. Ramie fiber is removed from the stem of the plant (*Bohmeria nivea*) by a beating and scraping operation that leaves it in ribbons. These ribbons are dried and later chemically degummed, usually by the spinner. Thus, ramie is not retted as other bast fibers are.

The world's largest producer is believed to be the People's Republic of China. The only other countries that export the fiber are Brazil and the Philippine Republic.

The color of raw ramie (China grass) is creamy, green, or straw to light brown, and the length averages from 28 to 60 in. (70 to 150 cm). The degummed and bleached fiber and fabric are white.

Ramie fiber cells are longer than those of any other plant fiber, averaging more than 6 in. (165 mm) in length. The cell diameter averages 50 μm with considerable variation between and within varieties. Under the microscope, the diameter of the fiber appears uneven, sometimes with heavy cell walls and well-defined lumen, at other times broad and flat with an indistinct lumen but showing heavy striations along the fiber.

Ramie textiles have a natural luster, are very strong and absorbent, and tend to be stiff. They gain strength when wet and are highly resistant to mildew and rot. See RAMIE.

Hard fibers. Hard fibers, also known as leaf or cordage fibers, come from the leaves of monocotyledonous plants. The fiber of the leaves of many species is used, some in international trade, but the three consumed in greatest quantity are abaca, sisal, and henequen. On the basis of abaca strength as 100, sisal would be about 75 and henequen about 60.

Long, parallel, overlapping cells form the hard fibers as in bast fibers. However, the leaves are not

retted or treated chemically as are the stems of various bast fibers.

Abaca. Frequently known as Manila hemp, abaca comes from the pseudostem (modified leaf petioles that look like a tree trunk) of *Musa textilis*. Most abaca is grown in the Philippines.

The fiber is removed from the plant by methods that range from hand stripping to a factory-type of defibering (decortication).

Abaca fiber color ranges from almost white through cream, light brown, and very dark brown. The fiber is long, usually more than 9 ft (3 m), and ranges from about 3.3 to 16.5 ft (1 to 5 m) in length.

The cells of abaca average 0.2–0.24 in. (5–6 mm) in length and 25–30 μm in diameter. The fiber cells are uniform in diameter, lustrous, and rather thin-walled. The lumen is large and distinct. An unusual feature is the frequent appearance of silicified plates about 0.0012 in. (0.03 mm) in length and often in long chains. The fiber is a lignocellulose, about 75% cellulose and about 8% lignin. Abaca is traditionally used in marine cordage and in other applications that require a strong, water-resistant rope, although its use in cordage is being challenged by manufactured fibers such as nylon, polyester, and polypropylene. See ABACA.

Sisal and henequen. The sisal plant (*Agave sisalana*) is grown mainly in Africa and Brazil, while henequen (*A. fourcroydes*) is grown mainly in Mexico and Cuba. These are cordage fibers removed from the leaves by defibering. The greatest use for these fibers is for twine and rope.

Sisal fiber is stiff and harsh, usually 28–56 in. (70–140 cm) in length, and a creamy white color. Henequen tends to be lighter in color and slightly longer, but is not as strong as sisal.

The cells of sisal and henequen are about the same size and chemical makeup. They average about 0.1–0.12 in. (2.6–3.0 mm) in length and 20 μm in diameter. They show an unusual broadening toward the middle, and the ends are broad, blunt, thick, and at times forked. The width of the lumen is frequently greater than that of the cell wall. These lignified fibers have about 75% cellulose and 10% lignin. See HENEQUEN; SISAL.

Animal Fibers

Many animal species produce fibrous structures. These include the open webs of spiders, the more closed cocoons of moths, the fibrous part of bird feathers, and body hairs found on species from insects to mammals, including humans. The common characteristic of these fibrous structures is that they are all denatured proteins. Silk and mammalian hairs have achieved commercial and technical importance. See PROTEIN.

Mammalian fibers. Mammalian fibers have long been used by humans in the form of fur, all types of textiles, fiber fills, ropes, brushes, wigs, and so on. All mammalian fibers are so closely related that their chemical and physical fundamentals can be discussed together.

Fiber formation. The mammalian skin is densely populated with microscopic channels, the follicles, from which the fibers emerge. As the cell mass is pushed outward, its components are structurized and denatured in a process called keratinization. Hence, all mammalian hairs are called keratin fibers. By the time the fibers reach the surface of the skin, they are fully hardened and are no longer living tissue.

Many species produce a double coat, containing both long, coarse, and somewhat sparse guard hairs and a shorter, fine, and dense underfur. See HAIR.

Shapes and structure. The fibers are circular or oval in cross section, with slight tapering at the tip end. The coarser guard hairs can be circular, strongly flattened, or triangular in cross section, and vary widely in thickness along their length. Diameters can range from a few micrometers, as in vicuña, to millimeter size, as in elephant tail fibers.

Morphologically, the fibers are built from two important cell types: the central cortex and the outer-covering cuticular cells (Fig. 4). The thin, flat cuticular cells overlap each other in a fish-scale fashion, their outer ends pointing toward the tip of the fiber. The cuticular layer acts as a protective layer for the cortical portion of the fiber. The cortical cells are elongated spindles in shape.

In some wool fibers two different types, ortho and para, are differentiated by chemical means. In addition, the fibers may possess a fragmented or continuous medullary core, consisting of cells which can be partially filled with cellular debris. The medullary volume is high in some hairs, but is completely absent in fine wool.

Colored animal fibers contain a granular pigment, called melanin, which is distributed mostly within the cortex. The shade and strength of color of the otherwise colorless keratin fiber is determined by the amount of and particle size of this pigment.

Chemical composition and fine structure. Keratin is a polymer of 20 α -amino acids which differ from each other in the composition of their side groups (Fig. 5). The

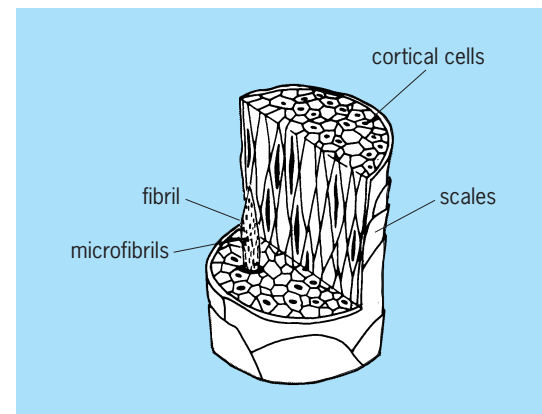


Fig. 4. Schematic of wool. The cuticle surrounds an inner structure, the cortex, made up of cortical cells which are composed of millions of small spindles (fibrils) which are composed of even smaller structures called microfibrils. (After B. F. Smith and I. Block, *Textiles in Perspective*, Prentice-Hall, 1982)

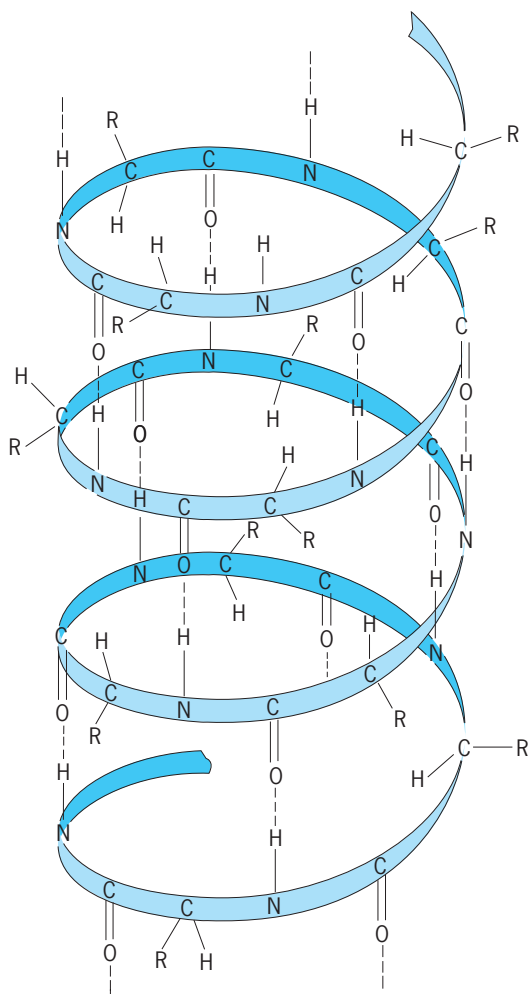


Fig. 5. Helix configuration of the α -keratin molecule. (After B. F. Smith and I. Block, *Textiles in Perspective*, Prentice Hall, 1982)

side groups can carry hydroxyl-, amino-, carboxylic-, and sulfur-containing groups which strongly influence the chemical and physical characteristics of the fiber. The sulfur content, mostly in the form of cystine, is between 3 and 5%. Neighboring polymer chains stabilize each other through hydrogen and salt bonds, and by disulfide cross-links of the cystine. Some of the polymer chains in the cortex fold into a regular helical pattern, and these group into fibrillar units with strong intermolecular bond formation. These fibrillar units, embedded in an amorphous random-coiled matrix, are well aligned with each other and contribute very significantly to the tensile behavior of the fibers. When the fiber is elongated, the α -helical structure reversibly unfolds into the so-called β configuration. The high extensibility and elasticity of the mammalian fiber is due largely to this structural deformability.

Physicomechanical characteristics. Mammalian hairs can reversibly sorb more water than any other commercially used fibers, with the exception of some regenerated cellulose. Part of the comfort factor of wool-based textiles is due to this fact. With the sorption of water the fibers swell and soften.

The dry- and wet-break extensions of mammalian fibers are about 30 and 60%, respectively. The fish-scale-type arrangement of the cuticular cells provides a unique differential frictional effect for fibers; the cut fibers move preferentially in their root direction under random forces, which is the basis of their felting property.

Wool. The most important commercial animal fiber is wool. The order of the largest producers is Australia, Russia, New Zealand, Argentina, South Africa, and the United States.

The finest and most valuable wool is produced by the Australian Merino sheep, developed from a Spanish breed, and is 15–30 μm in diameter and about 1.2–5 in. (3–13 cm) in length. It is used for the finest-quality wool fabrics. Medium- and coarse-variety fibers, 25–40 μm in diameter, are produced by other breeds. The coarser wools are used for heavier fabrics, blankets, upholstery, and carpets. See SHEEP.

Fiber properties. Wool fibers range in length from 1 to 20 in. (2 to 50 cm), and may be black, gray, brown, or white. The highest grades are fine (15–20 μm in diameter), long (about 12 in. or 30 cm), and creamy white in color. Wool is a weak fiber, with a tenacity of only 1–2 g per denier but, because of the long length and the high friction due to the scaly outer cuticle, makes strong yarn. The fibers exhibit very high moisture regain (13–16%), with exceptional extensibility (about 35% elongation at break). The cuticle also yields excellent abrasion resistance. Wool is quite resistant to household chemicals and sunlight, but will mildew if allowed to remain damp. It is also eaten by the larvae of various moths and beetles.

Wool's major advantages are high bulk, which permits the manufacture of warm soft fabrics; excellent moisture absorption for comfort; high resistance to wrinkling and good wrinkle recovery; strong yarns; and good abrasion resistance. It is subject to felting shrinkage upon laundering, so that hand washing or dry cleaning are recommended for wool products.

Uses. In apparel, wool is recommended for its warmth, moisture absorption, and durability. For suiting, it provides excellent wrinkle resistance for a neat appearance. In home furnishings, upholstery fabric and carpets of 100% wool are noted for color retention, durability, resilience, and comfort.

Goat hairs. Mohair is produced by the Angora goat mostly in Turkey, South Africa, and the United States. In size it ranges from 10 to 90 μm in diameter for hair and guard fibers, respectively, and is 4–10 in. (10–25 cm) long. Luster is its most important quality. The Cashmere goat in Asia grows fibers in the 15- μm range. Softness is the most important characteristic. As with most other textile fibers, the fine underfur is the valuable component of goat hairs. See CASHMERE; MOHAIR.

Camel hairs. The fiber known as camel originates from the Bactrian two-humped species from Asia. For textile purposes only the fine underfur of about 17 μm is collected by plucking or shearing. The guard hairs ranging over 100 μm have only specific uses. The alpaca and vicuña are South

American members of the camel family. The domesticated alpaca produces hair 8–20 in. (20–50 cm) long in 2 years. The diameter of the fur fibers is 25 μm . The wild vicuña grows a shorter fleece, 0.5–3 in. (1.3–7.5 cm), but the finest fibers, about 13 μm . The softest wool-type materials are made from it. *See* ALPACA; VICUNA.

Rabbit hairs. The Angora rabbit hair is very fine, 10–15 μm , and relatively long. It is used mostly for blending with wool for soft fabric effects. The fur hair of common rabbits is still very fine but shorter. The most important use is in felted hats.

Silk. Although it has lost ground to new synthetics, silk is still an important luxury fiber. It is the secretory product of the larval stage of many moth varieties. The cultivated breed is the *Bombyx mori*.

Silk has a few unique properties: it is the finest animal fiber, with a diameter of 10–12 μm ; it has no cellular structure since it was never part of body tissue; and it is a continuous filament. Chemically, the silk is a protein consisting of 17 different amino acids. Since these polymer chains are in an extended zigzag form, not in the helical folds characteristic for mammalian hairs, they form extensive crystalline regions embedded in an amorphous matrix.

The fibers are triangular in cross section, with significant unevenness along the length. They lack the fine morphological structurization of mammalian hairs, and resemble more closely an extruded manufactured fiber. Silk binds slightly less water than wool at all humidities, and the effect of water-swelling on its physical characteristics is less. Wetting reduces the fiber strength by only 15%. The fiber strength is high, comparable with that of nylon. While its break elongation is high, about 20%, it can be stretched elastically only to a few percent, unlike mammalian hairs. This is due to the lack of coiled helical units in the ordered regions of the fiber. Silk is more resistant to alkali treatments than animal hair, but less so to acids. Silk yellows when exposed to direct sunlight.

The most important qualities of silk fabrics are softness, smoothness, luster, and drape. *See* SILK; TEXTILE CHEMISTRY; TEXTILE PRINTING. Ira Block

Bibliography. R. S. Asquith (ed.), *The Chemistry of Natural Protein Fibers*, 1977; D. Eby, *Chemistry of Textiles*, 1989; P. W. Harrison (ed.), *Cotton in a Competitive World*, Textile Institute, Manchester, England, 1979; M. L. Joseph, *Introductory Textile Science*, 5th ed., 1993; H. L. Needles, *Textile Fibers, Dyes, Finishes, and Processes: A Concise Guide*, 1986; B. F. Smith and I. Block, *Textiles in Perspective*, 1982.

Natural gas

A combustible gas that occurs beneath the surface of the earth, often found in conjunction with petroleum deposits. Its main use is for fuel, but it is also used to make carbon black, certain chemicals, and liquefied petroleum gas. *See* LIQUEFIED PETROLEUM GAS (LPG).

Natural gas consists predominantly of methane,

but also contains a mixture of hydrocarbons such as ethane, propane, and pentanes. Carbon dioxide, nitrogen, helium, and hydrogen sulfide may also be present. The types of natural gas vary according to composition and can be dry or lean gas (mostly methane), wet gas (considerable amounts of other hydrocarbons), sour gas (much hydrogen sulfide), sweet gas (little hydrogen sulfide), residue gas (higher hydrocarbons having been extracted), and casinghead gas (derived from an oil well by extraction at the surface). *See* METHANE.

Distribution and reserves. Natural gas occurs on every continent. Wherever oil has been found, a certain amount of natural gas is also present. In addition, natural gas is associated with coal. Natural-gas hydrates are solids composed of water molecules forming a rigid lattice of cages (clathrates), with most of the cages containing a molecule of natural gas. Methane hydrate is the dominant natural-gas hydrate. Methane hydrates are found in polar and other continental shelves, and in deep-ocean outer continental margins in all oceans, including polar oceans. *See* CLATHRATE COMPOUNDS; COALBED METHANE; HYDRATE; PETROLEUM GEOLOGY.

Production and delivery. Natural gas comes from reservoirs, usually in sedimentary rock lying below the earth surface at depths varying from a few hundred feet to several miles. Being in gaseous form, it may occur alone in separate reservoirs. More commonly, it forms a gas cap, or mass of gas, entrapped between liquid petroleum and an impervious capping rock layer in a petroleum reservoir. Under conditions of greater pressure, it is mixed with or dissolved in crude oil. Successful exploitation involves drilling, producing, gathering, processing, transporting, and metering the use of the gas. Before it is commercially distributed, natural gas usually is processed to remove propane, butane, and nonhydrocarbon gases such as hydrogen sulfide. Processed natural gas has no distinct odor, so an odorant is added since an undetected leak could result in an explosion or asphyxiation. Transmission pipelines move the gas from processing centers to the market, where distribution pipelines carry it to consumers. Natural gas is also distributed worldwide in liquid form (liquefied natural gas, LNG), which is produced by chilling the gas to below its boiling point. *See* LIQUEFIED NATURAL GAS (LNG); OIL AND GAS FIELD EXPLOITATION; OIL AND GAS WELL DRILLING; PETROLEUM RESERVOIR ENGINEERING.

Storage. Because of the wide seasonal variations in temperature and the relative constancy of gas production and supply, excess gas must be stored in the summer to meet the increased demand in the winter. Without a reliable means of storing this excess gas, demand would be difficult to meet throughout the year. Depleted gas reservoirs are usually the first choice for use as storage reservoirs. Other types of reservoirs are used as well, including depleted oil reservoirs, aquifers, and salt caverns. *See* OIL AND GAS STORAGE. Michael A. Adewumi; Michel T. Halbouty

Bibliography. K. Arnold and M. Stewart, *Surface Production Operations*, vol. 2: *Design of*

Gas-Handling Systems and Facilities, 2d ed., 1999; B. Guo and A. Ghalambor, *Natural Gas Engineering Handbook*, 2005; N. J. Hyne, *Nontechnical Guide to Petroleum Geology, Exploration, Drilling and Production*, 2001; W. C. Lyons and G. J. Plisga, *Standard Handbook of Petroleum and Natural Gas Engineering*, 2d ed., 2004; M. D. Max (ed.), *Natural Gas Hydrate in Oceanic and Permafrost Environments*, 2003.

Natural language processing

Computer analysis and generation of natural language text. The goal is to enable natural languages, such as English, French, or Japanese, to serve either as the medium through which users interact with computer systems, such as database management systems and expert systems (natural language interaction), or as the object that a system processes into some more useful form, such as in automatic text translation or text summarization (natural language text processing). See DATABASE MANAGEMENT SYSTEM.

In the computer analysis of natural language, the initial task is to translate from a natural language utterance, usually in context, into a formal specification that the system can process further. Further processing depends on the particular application. In natural language interaction, it may involve reasoning, factual data retrieval, and generation of an appropriate tabular, graphic, or natural language response. In text processing, analysis may be followed by generation of an appropriate translation or a summary of the original text, or the formal specification may be stored as the basis for more accurate document retrieval later. Given its wide scope, natural language processing requires techniques for dealing with many aspects of language, in particular, syntax, semantics, discourse context, and pragmatics. (Analysis and generation of spoken natural language, which are not discussed in this article, also involve techniques for dealing with acoustic phonetics, phonology, stress, and intonation.) See VOICE RESPONSE.

Parsing. The aspect of natural language processing that has perhaps received the most attention is syntactic processing, or parsing. Most current techniques for parsing an input string of words involve (1) a description of the allowable sentences of the language (the grammar); (2) an inventory of the words of the language with their inflectional, syntactic, and possibly semantic properties (the lexicon); and (3) a processor which operates on the grammar, the lexicon, and the input string (the parser). This processor (1) simply accepts the input string, if grammatically well formed, or rejects it (a recognizer); (2) associates the string, if well formed, with its structure (or structures, if ambiguous) according to the grammar (an analyzer); or (3) associates the string with some other representation, for example, a semantic characterization (a transducer). Syntactic processing is important because certain as-

pects of meaning can be determined only from the underlying structure and not simply from the linear string of words.

One of the oldest parsing techniques, called augmented transition networks (ATNs), grew out of a system for parsing context-free (CF) languages, called recursive transition networks (RTNs). A recursive transition network parser consists of a set of named graphs or networks, each consisting of a set of nodes or states connected by a possibly ordered set of directed labeled arcs. The labels correspond to (1) words or classes of words that can be recognized or "consumed" on the arc; (2) the empty symbol, indicating an arc that can be taken without consuming any input; or (3) the name of a network, which indicates that the next segment of the input string must be recognizable by that network. Each network has a start state and one or more end states. If the parser can move through a network from start to end by consuming a segment of the input string, that segment is said to be recognizable by that network. See GRAPH THEORY.

An augmented transition network adds to the basic recursive transition network framework the ability to set and test variables or registers, thereby giving it the power to recognize a wider class of languages than a recursive transition network. An augmented transition network is appearing because its grammar is relatively easy to specify. Its weaknesses lie in the simple, uniform control structure provided by the basic augmented transition network (that is, unguided backtracking) and in its power, felt to be more than is needed for recognizing a natural language.

Current trends are to construct parsers and grammars which appear to follow more closely human parsing strategies and which have less power. In particular, researchers have begun to give almost context-free descriptions of natural languages, thereby allowing them to use slightly extended versions of efficient context-free parsing techniques. Such descriptions include generalized phrase structure grammar (GPSG), immediate dominance/linear precedence (ID/LP) grammar, and tree adjoining grammar (TAG).

Semantic analysis. A second phase of natural language processing, semantic analysis, involves extracting context-independent aspects of a sentence's meaning. These include the semantic roles played by the various entities mentioned in the sentence. For example, in the sentence "John unlocked the toolbox," "John" serves as the agent of the unlocking, and "the toolbox" serves as the object; in "This key will unlock the toolbox," "this key" serves as the instrument. Context-independent aspects of sentence meaning also include quantificational information such as cardinality, iteration, and dependency. For example, in the sentence "In every car, the mechanic checked to see that the engine was working," the checking is iterated over each car; the identity of the engine depends on the identity of the car, while that of the mechanic does not; and the cardinality of engines per car is one. Thus there are as many engines as cars, but possibly only one mechanic. The

representational formalism used by the system for semantic analysis (for example, first-order predicate calculus, case grammar, conceptual frames, procedures, and so forth) is usually chosen for its ability to convey those aspects of semantics that the system requires for later processing. For example, if temporal position (past/present/future) is not significant, it will not be captured in the formalism.

Most semantic analysis is done by applying pattern-action rules either during parsing or afterward. The pattern part of a rule consists of clauses, each of which specifies the presence of a particular lexical item, usually the head of some syntactic substructure (for example, the main verb of a clause or sentence or the head noun of a noun phrase); or a particular syntactic substructure (for example, a relative clause, to be interpreted as a restriction on the class described by the rest of the noun phrase). A pattern clause may also specify a test on another part of the current substructure. The action part of a rule usually calls for building a piece of semantic representation, often requiring the semantic analysis of some other part of the syntactic substructure. For example, there may be a pattern-action rule associated with “unlock” as the main verb of a clause. A test in the pattern may require that the subject of the clause be interpretable as an animate agent. The rule’s action may call for the inclusion of a conceptual frame for the concept “unlock” as part of the semantic representation of the sentence. The rule’s action may further specify that the agent role of the frame be filled by the semantic interpretation of the subject of the clause, and that the object role be filled by the semantic interpretation of the direct object. In some systems, a rule can have optional pattern clauses and actions: thus, the rule pattern for “unlock” might optionally specify a “with” prepositional phrase whose noun phrase object can be interpreted as a tool. If so, the rule action might additionally call for inserting the instrument role in the frame with the semantic interpretation of the prepositional phrase object.

Contextual analysis. Given that most natural languages allow people to take advantage of discourse context, their mutual beliefs about the world, and their shared spatio-temporal context to leave things unsaid or say them with minimal effort, the purpose of a third phase of natural language processing, contextual analysis, is to elaborate the semantic representation of what has been made explicit in the utterance with what is implicit from context. Two major linguistic devices that contextual analysis must deal with are ellipsis and anaphora.

Ellipsis. Ellipsis involves leaving something unsaid. To handle ellipses computationally, techniques are required for recognizing that something is indeed missing and for recovering the ellipsed material. When the utterance is a sentence fragment and not a complete sentence, it is fairly easy to recognize that something is missing. An example appears in the following sequence:

User: What is the length of the JFK?
System: ⟨some number of feet⟩.
User: The draft?

On the other hand, since parsers are usually designed for well-formed input, either the system’s grammar must be revised to accept sentence fragments or a special error-recovery routine must take over after the parser fails. When an utterance is syntactically well formed, it may still be elliptic in that some needed conceptual material is missing, as in the following example:

User: What maintenances were performed on plane 3 in May 1971?
System: ⟨list of maintenances⟩.
User: What maintenances were performed on plane 48?

In the user’s second question, the time period of interest is missing, and the question should not be answered until it is recovered. (It is clearly May 1971.)

The primary technique for recovering ellipsed material is a simple one, based on semantic features. For sentence fragment ellipses, the previous discourse is searched for the most recent utterance containing a constituent with the same features as the fragment. The utterance minus that constituent is taken as the ellipsed material. For conceptual ellipses in a syntactically well-formed sentence, the previous discourse is searched for the most recent utterance with a constituent having the required semantic features. That constituent is taken to be the ellipsed material. In each case, a new well-formed sentence is then constructed and processed as if the ellipsis had never occurred. For instance, in the first example, both “length” and “draft” are properties of ships. Thus, given the fragmentary utterance “The draft?,” What is ____ of the JFK?” is found as the ellipsed material. The question “What is the draft of the JFK?” is then interpreted and answered normally. This technique works often, but does not constitute a general solution. A more powerful solution has been developed based on recognizing a user’s goals in producing an utterance, but has been found computationally efficient only in very narrow task-oriented domains. This will be discussed below.

Anaphora. Anaphoric expressions are very simple words or phrases which cospecify something previously evoked by the discourse or are strongly associated with something so evoked. Instances of anaphora include definite pronouns such as “he,” “she,” “it,” “they,” and definite noun phrases such as “the mechanic” and “the cars.” The problem is that anaphora can be interpreted only in context, and the semantic interpretation of a sentence is not complete until all anaphoric expressions are resolved and the cospecified entities identified.

Early computational approaches to anaphora resembled those for dealing with ellipses: entities described in previous sentences were searched for the most recently mentioned one with appropriate semantic features. Now recency has been replaced by the notion of focus as a basis for anaphora resolution. Immediate focus reflects the particular thing the speaker is talking about; global focus involves things associated with it or in which it participates, and gives a sense of what may be talked about next.

Techniques have been developed for tracking immediate focus, projecting ahead from the current utterance what may be focused on in the next one. This is useful for resolving anaphora, in that it predicts what entities are likely to be respecified anaphorically.

Pragmatics. A fourth phase of natural language processing, pragmatics, takes into account the speaker's goal in uttering a particular thought in a particular way—what the utterance is being used to do. In an interaction, this will influence what constitutes an appropriate response. For example, an utterance which has the form of a yes/no question or an assertion may have the goal of eliciting information (for example: “Do you know how to delete a control-Z?” “I can't get the set file protection command to work.”). Because it is inappropriate (and possibly at times dangerous) to take a user's utterances literally or to assume that the user will take those of the system literally, computational techniques must be devised for relating the syntactic shape and semantic content of an utterance to its pragmatic function.

Plan recognition. One important approach to this problem has been to view language understanding as plan recognition. The actions (either communicative or physical) that constitute the plan may be motivated in one of two ways: goals in the world that the person wants to accomplish, for which he or she needs to elicit or offer aid or information; or aspects of an already ongoing interaction that need attention—for example, confusion over the speaker's foregoing utterance may lead the listener to seek clarification.

A user seeking particular information from the system illustrates clearly a plan-recognition approach to language understanding. The user's utterance—a well-formed sentence or an ellipsed fragment—is taken as a request for information that the user believes he or she needs in order to accomplish some goal, a goal which is not as yet presumed to be known to the system. Just as a medical diagnosis system uses rules which link findings back to those diseases which commonly manifest them, the plan-recognition system uses rules which link utterances back to those domain goals which need the intended information in order to be achieved. For example, consider the utterance “The train to Windsor?” made to a system serving as train information clerk. The system interprets this as a request: the user wants to know some property of that train in order to fulfill his or her goal. The system then tries to figure out what that goal is, in order to figure out what information the user might be requesting. There are only two possible goals considered: meeting a train and boarding one. The description “train to X” does not match that of incoming train, so the user's goal is taken to be boarding the Windsor train. To board a train, one needs to know its departure time and track. Since the system does not have evidence of the user's knowing either of these properties, it responds with both: “It leaves at 3:15, from track 7.” Currently, it is only by limiting the domain, and hence the range of possible goals the system needs to consider, that such a plan-based approach to pragmatics and natural language processing becomes feasible.

Cooperative principle. Because pragmatics acknowledges language use, it also acknowledges expectations that speakers and listeners have about the form and content of utterances, based on normal conventions of use. This has been well described by the philosopher Paul Grice, who noted that speakers acknowledge a “cooperative principle” of conversation (by either upholding it or purposely flouting it), which he further specified in terms of conversational maxims of quantity, quality, manner, and relation. For example, the quantity maxim states: “Make your contribution as informative as is required (for the current purposes of the exchange). Do not make it more informative than is required.”

The cooperative principle and its maxims are important to natural language processing because, if a system does not behave in accord with normal conventions of use, the user is likely to be confused or misled by the system's behavior. Conversely, if the system does not interpret the user's behavior in terms of normal conventions of use, the system is unlikely to understand the user correctly, if at all. In particular, the cooperative principle and its maxims reveal a method of implicit communication which Grice termed implicature. An implicature is basically an aspect of an utterance's interpretation which makes no contribution to its truth value (that is, semantics) but constrains its appropriateness in discourse. For example, consider the following discourse:

Q: Is there a gas station on the next block?
R: Yes.

The simple “yes” answer implicates to Q that, as far as R knows, the gas station is able to provide its normal services and hence fulfill Q's probable goal. Q reasons that if R knew that the gas station was closed and hence could not fill Q's needs, R would have said so: that is, R would have said “Yes, but it's closed.” Thus a system must be as aware of implicatures (both the user's and its own) as it is aware of what is communicated explicitly.

Overall organization. As for fitting the pieces together, there is no single way that natural language analysis is done. Some systems have a single processor for syntactic, semantic, contextual, and pragmatic analysis, with no distinction made as to the source of that knowledge. Some systems keep the knowledge sources separate but apply them simultaneously, extracting whatever can be derived at the moment and using whatever information is available. Other systems are very modular, separating the knowledge sources and specifying when they should be applied. Efficiency, extensibility, and transportability are some of the important issues to be considered when evaluating a system for natural language analysis.

Natural language generation. The bulk of early research in natural language processing was formerly directed at natural language analysis. Now researchers have taken up seriously the task of natural language generation. Generation is not just the reverse of analysis because the status of user

and system are fundamentally different. Systems can be developed which tolerate users' mechanical errors (for example, spelling, typing, and grammatical mistakes), treating them as insignificant variations. Users, on the other hand, may not be able to figure out which aspects of the system's natural language behavior reflect simple nonfluencies (for example, those due to limited lexical or grammatical options) and which embody significant aspects of communication. Moreover, the system's sense of language must be more highly developed for generation, lest it confuse or mislead the user by what it communicates or how. Work on explanation is also a significant aspect of natural language generation. See ARTIFICIAL INTELLIGENCE. Bonnie Webber

Bibliography. M. A. Covington, *Natural Language Processing for Prolog Programmers*, 1993; F. Pereira and B. J. Grosz (eds.), *Natural Language Processing*, 1994; *Proceedings of the Association for Computational Linguistics; Proceedings of the International Conference on Computational Linguistics*; H. Tennant, *Natural Language Processing*, 1981; T. Winograd, *Language as a Cognitive Process*, 1982.

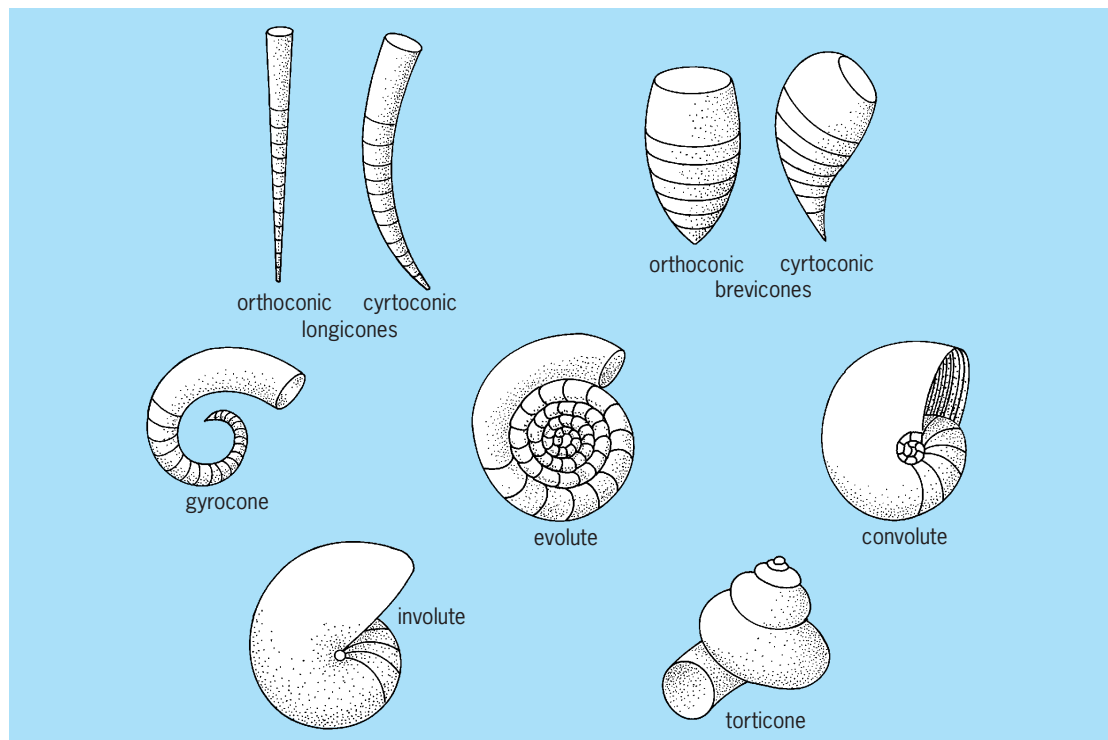
Nautiloidea

A group of externally shelled cephalopods, represented by the single extant genus *Nautilus*. The formal designation of this group as a subclass is now generally used only for those externally shelled cephalopods that resemble *Nautilus* in having com-

pletely coiled shells (and then the subclass includes *Nautilus* itself). In living forms, the basic structural plan includes a shell consisting of a septate phragmocone, a living chamber, and a siphuncle. In fossil nautiloids, this simple pattern is modified in great variety with respect to shell form and size, structure and size of the siphuncle, and the large number of devices to counteract the buoyancy of the phragmocone. The shape of fossil nautiloids may deviate in many ways from the simple *Nautilus* model (see **illus.**); the length of straight or slightly curved shells varies from less than 1 in. (2.5 cm) to more than 30 ft (10 m). Few coiled shells are larger than *Nautilus*. Also, the siphuncle may vary considerably in size and shape (see **illus.**). The aperture of the living chamber may be constricted or contracted into various shapes, and the interiors of siphuncle and camerae may be partially filled by layers of aragonite and conchiolin.

The nautiloids are represented by approximately 1000 fossil genera comprising an estimated 5000 species and grouped into more than 130 families, half of which lived in the Ordovician Period. The nautiloids are distributed among 12-14 orders, with the Oncocerida (183 genera), Nautilida (179), and Orthocerida (135) being the most numerous.

Nautiloids first appeared about 520 million years ago, very late in the Cambrian Period. Most early nautiloids had small shells, about 1 in. (2.5 cm) in length, although some reached 4 in. (10 cm) and were either straight or slightly curved. They reached the acme of their diversity in the Ordovician, with nearly 70 families, and then declined to 13 families in



Common types of nautiloid conchs. (After C. Teichert et al., *Cephalopoda: General features*, in R. C. Moore, ed., *Treatise on Invertebrate Paleontology*, pt. K, Mollusca 3, Geological Society of America and University of Kansas Press, 1964)

the Permian, 4 in the Tertiary, and 1 at present. The evolution of the nautiloids suffered severe setbacks at the end of the Cambrian and the end of the Triassic, when they came close to extinction.

Fossil nautiloids are found on all continents, including Antarctica. Especially noteworthy are the rich Ordovician and Silurian faunas of North America, northern Europe, Czechoslovakia, and central and southern China.

Coiled nautiloids almost certainly moved around by jet propulsion like *Nautilus* and lived close to the sea floor, at moderate to intermediate depths in many different environments. Other types may have included agile swimmers as well as slow-moving benthic adaptations. See CEPHALOPODA; NAUTILUS.

Curt Teichert

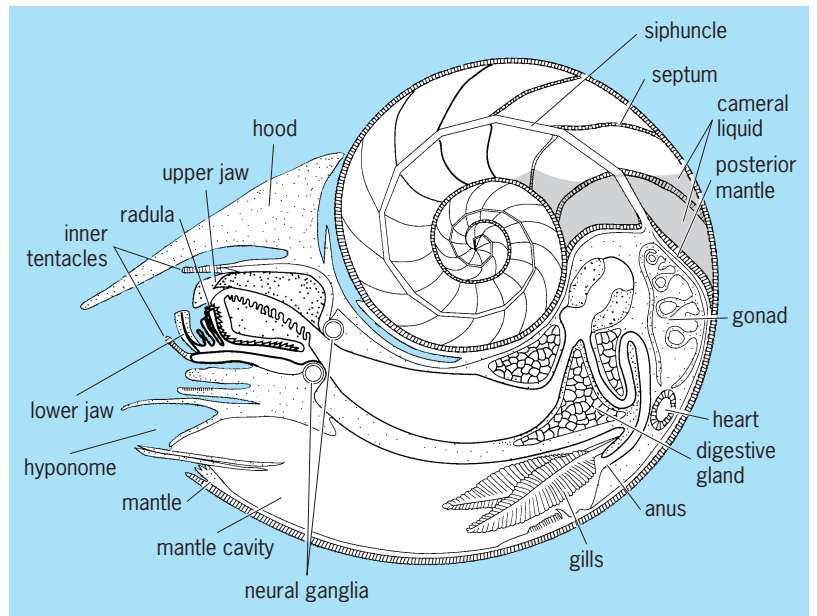
Bibliography. M. R. Clarke and E. R. Trueman (eds.), *The Mollusca*, vol. 12: *Paleontology and Neontology of Cephalopods*, 1988; R. E. Crick, Buoyancy regulation and macroevolution in nautiloid cephalopods, *Senckenbergiana Lethaea*, 69:13-42, 1988; C. H. Holland, The nautiloid cephalopods: A strange success, *J. Geol. Soc. (London)*, 144:1-15, 1987; R. C. Moore (ed.), *Treatise on Invertebrate Paleontology*, pt. K, *Mollusca* 3, 1964.

Nautilus

The only living genus of the molluscan subclass Nautiloidea, containing the only living cephalopods with an external chambered shell and numerous cephalic tentacles (see **illus.**). Six species of *Nautilus* live in moderately deep water in the western Pacific and around the East Indies.

The perfectly symmetrical planospiral shell is divided into about 25 chambers by transverse partitions. Only the outermost and largest chamber is occupied by the animal, the rest being gas-filled and serving as a buoyancy organ. Each partition, or septum, has a perforation reinforced with a shelly "neck" through which passes the siphuncle, a cord of living tissue containing blood vessels and surrounded by cornified layers, which can adjust the buoyancy by absorption of water from the penultimate chambers balanced by gas secretion. Healthy specimens of *Nautilus* can move to and from the surface water from depths of 1150 ft (350 m) by such adjustment alone. The mechanical strength of the chambered shell is necessarily high, and a gas-filled *Nautilus* shell of average size will successfully resist crushing implosion down to depths of about 1800 ft (550 m), that is, to external pressures of approximately 55 atm or 5.5 megapascals.

Apart from its buoyancy control, which allows it to live near the surface at night and on the bottom or at greater depths during the day, *Nautilus* is capable of horizontal movements by jet propulsion. As in squid and cuttlefish, this is achieved by water propelled by muscular contractions of the mantle cavity. However, *Nautilus* is a relatively slow-moving collector of carrion and moribund animals, whereas squid are among the fastest swimmers in the sea and are



Internal structure of *Nautilus macromphalus*. (After P. D. Ward, *The Natural History of Nautilus*, Allen and Unwin, 1987)

predacious carnivores capturing active prey.

The mantle cavity in *Nautilus* is unique among living mollusks in that the normally paired pallial and cardiac structures are doubled. There are four ctenidia (gills), four osphradia, and four excretory organs. The ventricle of the central heart receives blood from four auricles, and (unlike other cephalopods such as squid and octopus) there are no accessory pumps (or branchial hearts) in the gill circulation. The head of *Nautilus* bears rather primitive, lenseless, "pin-hole" eyes and 38 retractile, prehensile tentacles, which lack the characteristic suckers found on the 8 or 10 arms of other modern cephalopods. Compared to octopus, squid, and the rest, *Nautilus* obviously represents the persistent line of a more primitive cephalopod stock. In fact, fossil Nautiloidea date from the Upper Cambrian, while the stocks of modern cephalopods probably arose in the Cretaceous. See CEPHALOPODA; NAUTILOIDEA; OCTOPUS.

W. D. Russell-Hunter

Naval architecture

An engineering discipline concerned with the design of ships, boats, drill rigs, submarines, and other floating or submerged craft. The naval architect creates the initial overall concept for a new ship, integrates the work of other specialists as the ship design is developed, and is specifically responsible for the hull and superstructure shape, general arrangements, structural design, weights and centers calculations, stability analysis, hydrodynamic performance assessments, propeller and rudder design, and the arrangement and outfit of all living and working spaces, other than machinery spaces. The naval architect's ally, the marine engineer, is responsible for the design of the propulsion plant, the electric

plant, and other ship machinery and mechanical systems, including the so-called distributive systems: electric cabling, piping, and ventilation system ducting. The marine engineer also is responsible for ship control systems, including propulsion and electric plant controls and the steering system. *See* FERRY; HYDROFOIL CRAFT; ICEBREAKER; MERCHANT SHIP; NAVAL SURFACE SHIP; OIL AND GAS, OFFSHORE; SUBMARINE.

In the past, naval architecture was as much an art as a science, but research, coupled with advances in computer-aided design, has greatly enhanced the scientific basis of the profession. Naval architecture is a specialized form of mechanical engineering, as is marine engineering. Thus the education of naval architects is very similar in content to that of mechanical engineers, and the same types of degree programs are offered. Some colleges and universities combine naval architecture and marine engineering education and offer a combined degree. *See* MARINE ENGINEERING; MECHANICAL ENGINEERING.

Because naval architects have an overall ship design integration responsibility, they must be familiar with all the other engineering disciplines which contribute to ship design. Their outlook as ship design integrators must embody the principles of systems engineering: Overall ship cost and performance must be optimized, and this will generally require compromises in the designs of the individual subsystems or elements of the total ship. In addition, the naval architect must be knowledgeable of and infuse into the design the principles of human engineering and the "ilities": producibility, operability, supportability, reliability, and maintainability. Because of the importance of cost, the life-cycle costs (design, build, operation, and support costs) of the many alternatives considered during design must be carefully analyzed during the design process. This requires the naval architect to have a sound understanding of ship economics. National and international agencies establish regulations to which ship designs must adhere; the naval architect must know these requirements. *See* HUMAN-FACTORS ENGINEERING; RELIABILITY, AVAILABILITY, AND MAINTAINABILITY; SYSTEMS ENGINEERING.

The type of work done by naval architects varies widely. Some work alone or in a small team, developing the designs of yachts or other small craft in their entirety. At the other extreme, some work as members of very large teams designing complex warships. In this situation, the naval architect may focus on overall ship integration, coordinating the work of others, or may work in a specific functional area, such as structural design, stability, or general arrangements. The nature of the work done also varies with the design phase. In the very early stages of design, overall ship concepts are created and analyzed in their entirety using simplified sketches or computer models and engineering approximations. In the late stages of design, the focus is on the development of design details. Large amounts of data must be dealt with accurately, and detailed procurement and test specifications are prepared. Many naval ar-

chitects do not work as ship designers. Examples include researchers in a specialty area, teachers, the representatives of a ship owner who is having ships designed and built by others, ship production planners or inspectors, and employees of a government agency establishing safety regulations or managing the acquisition of ships for that agency, such as the U.S. Navy, Coast Guard, or National Oceanographic and Atmospheric Administration (NOAA). *See* SHIP DESIGN; SHIPBUILDING. Peter A. Gale

Bibliography. E. V. Lewis (ed.), *Principles of Naval Architecture*, 2d ed., 3 vols., Society of Naval Architects and Marine Engineers (SNAME), 1988; SNAME, *Careers in the Maritime Industry* (brochure), 1993; R. Taggart (ed.), *Ship Design and Construction*, 3d ed., SNAME, 1980.

Naval armament

A general term that covers the ordnance and control systems used by naval ships and aircraft. It includes a wide spectrum of weapons designed for use against targets in the air, on land or sea, or under the ocean surface.

Naval armament has always presented a formidable challenge to designers. Weapons must be launched from a rolling, pitching platform against targets that are frequently moving and often invisible. Accuracy and reliability are essential, because only a limited number of weapons can be carried on board ship. Naval weapon systems must be capable of a wide range of destructive force to meet the variable demands of "cold," limited, or all-out war. The spectrum of weaponry used by naval forces thus runs from small arms to nuclear warheads and includes weapons that are intended for use against a particular type of target as well as general-purpose weapons.

The combinations of weapons and control systems built into a warship are based on the requirements of the ship's intended missions and on the ship's ordnance-carrying capacity (**Fig. 1**). Any ship's weapon suite is a compromise between offensive and defensive armament requirements and such other characteristics as the vessel's endurance, speed, survivability, and habitability. *See* NAVAL SURFACE SHIP.

Naval weaponry is becoming more internationalized. Many developments abroad parallel those in the United States; others differ to suit the needs of the country involved. Many countries, including the United States, fill some of their needs by development in cooperation with other nations, or by foreign purchase. This article discusses United States naval armament.

Weapon systems. Modern naval weapons are not considered in isolation but in terms of weapon systems. A typical weapon system includes (1) devices capable of searching out, locating, and identifying the target such as radar and sonar; (2) launchers and weapon-handling equipment, such as gun mounts, missile launchers, and torpedo tubes; (3) control



Fig. 1. Modern naval ship employing a spectrum of weapons, sensors, and control systems. (U.S. Navy)

equipment, such as gun and missile director systems, to calculate the correct path for the weapon to intercept the target, and to transmit control information to the launcher or delivery vehicle; (4) a delivery vehicle, often self-propelled, such as a missile airframe or torpedo body, to carry the payload to the target; a gun's projectile does not require a vehicle, though a rocket-assisted projectile includes a rocket motor; (5) the weapon itself, containing the destructive force, such as a gun projectile or a missile or torpedo warhead. See RADAR; SONAR.

Nuclear weapons do not necessarily constitute separate systems. A nuclear capability can be designed into a weapon system as an alternative choice of payload, allowing either high-explosive or nuclear application of the same guidance and delivery systems. See ATOMIC BOMB; HYDROGEN BOMB.

Naval armament may be air-, surface-, or submarine-launched. It can be categorized as tactical or strategic, or by its intended primary target: surface attack, air defense, or antisubmarine. Many weapons can be used against different types of targets. Naval weaponry includes guns, guided missiles, rockets, bombs, depth charges, torpedoes, and mines.

Guided missiles. In the years since World War II, guided missiles have taken first place among families of naval weapons. Naval missiles may be adaptable to multiple launch modes: from ship, submarine, and aircraft. Modern missiles are more compact, saving critical space and weight, and their guidance systems have steadily become more sophisticated. Early shipboard launchers, mounted abovedecks and fed from magazines, could handle only one specific missile; newer ones can handle two or three different weapons, eliminating the need for sepa-

rate launchers. This, however, complicates ammunition stowage and loading, so that a particular launcher can engage only one type of threat at a time.

A form of vertical launch system (VLS) was first taken to sea in 1980 by the Soviet large missile cruiser *Kirov*. The American Mark 41 VLS entered the fleet in 1986 (Fig. 2). It is an advanced system in which missiles, in sealed shipping canisters, are stowed in launching tubes in the ship's hull. Weapons are protected from the weather and ready for immediate use, doing away with the need to load missiles in succession from magazines into an abovedecks launcher. Ships can be loaded with variable mixes of weapons to suit their mission. The vertical launch system arms newer *Ticonderoga*-class missile cruisers as well as *Arleigh Burke*-class destroyers, and has been backfitted into earlier *Spruance*-class destroyers. Some attack submarines have individual vertical launch tubes.

Advanced radars and information-handling technology have improved missile capabilities. Short-range, point-defense systems have been developed for ship defense against aircraft, missiles, and fast attack craft. Naval ordnance, because of the varied nature of the threats to be met, will continue to be a mix of missiles, torpedoes, guns, and antisubmarine weapons.

Guidance systems. Missiles are guided in a number of ways. In a preset system, the missile follows a flight plan programmed before launch; once the missile is launched, this cannot be changed. A command-guided missile follows course instructions transmitted by a control station. Homing systems can be active, emitting a radar or laser signal which is

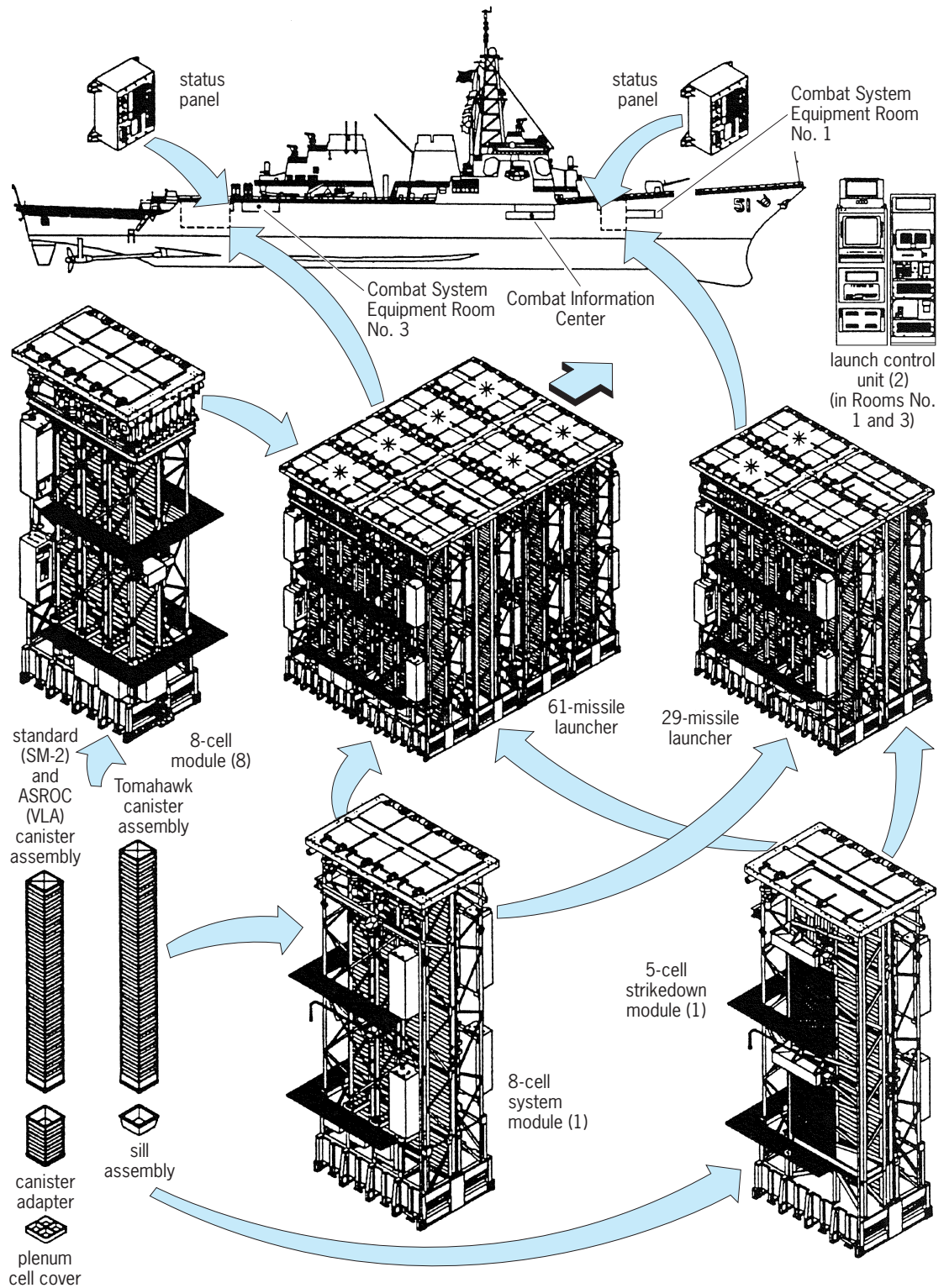


Fig. 2. Elements of the Mark 41 vertical launch system (VLS). This modular system is designed to be installed in ships of various sizes and configurations. (U.S. Navy)

reflected by the target and picked up by the missile's receiver. In a semiactive homing system, the missile receives target reflections of emissions from the launching ship or from another source. A passive homing system recognizes a physical characteristic of the target—heat or electronic emissions—and steers toward them. A composite guidance sys-

tem combines two or more of these types of guidance. See HOMING.

A missile's flight is divided into a boost phase, in which it is launched and attains flight altitude and direction; a midcourse phase, that is, the journey to the target area; and a terminal phase, the final run-in to the target. A missile may use, for example,

inertial guidance during its midcourse phase, shifting to homing guidance to attack its target. Short-range missiles as well as some torpedoes may be wire-guided, their command signals transmitted along a fine wire that unreels as the missile travels toward its object. See GUIDANCE SYSTEMS; INERTIAL GUIDANCE SYSTEM.

Warheads. Missile warheads may be designed for blast effect against underwater or ground targets, or for blast with fragmentation against surface or aerial targets; or they may contain nuclear explosives. Continuous-rod warheads, used in older missiles against aircraft, contain two sets of jointed rods which expand into semicircles, when the bursting charge is set off by a proximity fuze, and destroy or damage their target by cutting. Fragmentation warheads explode into an expanding sphere of casing fragments; controlled-fragmentation warheads are configured to spread their fragments in a pattern designed for maximum damage to the target.

Strategic missiles. An *Ohio*-class missile submarine carries 24 Trident fleet ballistic missiles (Fig. 3). Developed to replace the earlier Polaris and Poseidon, the Trident I C-4 missile arms the first eight *Ohio*'s, while the remaining ten ships use the more accurate and longer-range Trident II D-5. Four of the earlier submarines will later be backfitted to carry Trident IE. Four British submarines are also armed with it. See SUBMARINE.

Surface-to-air missiles. Standard, the Navy's principal air defense missile, replaced the first-generation Tartar, Terrier, and Talos. A supersonic solid-fuel weapon, it is produced in medium-range (MR) and extended-range (ER) versions. The current Standard



Fig. 3. Trident missile being fired from a submerged submarine. (U.S. Navy)

Missile 2 (SM-2) is designed for use with the Aegis system to counter large-scale missile attacks. It has been improved to enhance the missile capabilities of ships without Aegis. Newer versions of Standard, for use in VLS, are designed to cope with very fast targets at great altitudes; the most recent missile is intended for dual use against aircraft and ballistic missiles at lower-tier altitudes.

Sea Sparrow is an anti-aircraft adaptation of the airborne Sparrow III missile, developed as a relatively uncomplicated basic point-defense missile system (BPDMS) to protect ships without Standard missiles. It uses a radar director to acquire and follow its target. The missile uses semiactive homing, guiding itself on reflections from the fire control radar, and a proximity fuze. In its present version, Sea Sparrow is the principal North Atlantic Treaty Organization (NATO) point-defense missile, using a lightweight launcher with a digital fire control system. The internationally developed Evolved Sea Sparrow (ESSM), with greater speed, range, and maneuverability for use with VLS, is in production. It will be used in combination with the Rolling-Airframe Missile (RAM), defending against threats beyond RAM's range.

RAM, developed jointly by the United States, Denmark, and Germany, is stabilized by tail fins and helical ridges, or lands, around the body of the missile which cause it to rotate slowly in flight like a rifle bullet. A "fire-and-forget" counter to antiship missiles, RAM uses Sidewinder's rocket motor and warhead. It homes on the target's radiation until it picks up a heat signature with its infrared seeker. SEA RAM, under development, is a Phalanx radar-controlled anti-aircraft mount with a RAM launcher in place of a 20-mm automatic gun, designed for close-in defense against future high-performance antiship missiles.

Stinger, a portable Army-Marine Corps anti-aircraft missile, has been issued to ships when suicide air attack was considered possible. A shoulder-launched heat-seeking weapon, it has an explosive warhead with a proximity fuze.

Surface-to-surface missiles. Tomahawk, a long-ranged land attack cruise missile, was used in the Gulf War and in Kosovo. Capable of attacking targets at a range up to 1000 mi (1600 km), Tomahawk has greatly increased the striking power of the surface warship, which at one time was thought to have been relegated to a subsidiary role by the aircraft carrier (Fig. 4). It is also used by aircraft; submarines can carry them in torpedo tubes, and some submarines have been armed with vertical tube launchers. Newer Tomahawks have increased range and more powerful engines. They use Global Positioning System (GPS) satellite midcourse control and terrain-matching terminal guidance. The high-explosive warhead has an impact fuze; newer Tomahawks also have a variable-time fuze. A new control system will allow ships to program their own missiles, which until now had to be done ashore. Tactical Tomahawk, to enter service in 2003, will be simpler, more reliable, and less costly. See SATELLITE NAVIGATION SYSTEMS.

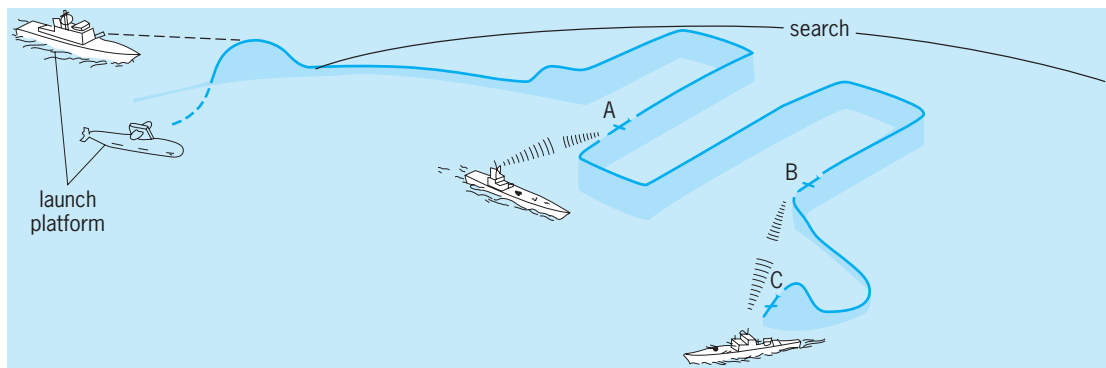


Fig. 4. Guidance of Tomahawk cruise missile. At A, Tomahawk detects ship, identifies it as nonhostile, and continues search. At B, it detects second ship, identifies it as hostile, and commences homing. At C, it acquires target and attacks.

Harpoon is a long-range antiship missile, originally designed as an air-to-surface weapon but now used in surface ships and submarines as well. Several other countries also use Harpoon for coastal defense. Introduced to the fleet in 1977, Harpoon is expected to serve well into the twenty-first century. Midcourse guidance is inertial, to be supplemented by GPS in future models; active radar terminal guidance can home on radar jammers. Later Harpoons have greatly increased range and can turn back to reattack a target missed on the first pass. Surface-ship and submarine Harpoons use a rocket booster; the air-launched version does not require this. Submarine Harpoons are launched in capsules from torpedo tubes; the capsule surfaces and releases the missile.

The Land-Attack Standard Missile (LASM) program will modify early models of the Standard missile to bomblet-dispensing surface fire support missiles for use against land targets. Launched from VLS in surface warships, it will use inertial guidance supplemented by GPS.

Air-to-air missiles. Phoenix was designed as a long-range task force air defense weapon to give the F-14 carrier fighter, with its digital missile control system, the ability to engage up to six targets simultaneously. It has inertial midcourse guidance with input from a semiactive homer; when the target is in range, an active seeker acquires the target and closes to a point at which a proximity fuze can detonate the fragmentation warhead. Newer Phoenix missiles have a speed over Mach 5.0.

Sidewinder is a heat-seeking weapon used in large numbers by the United States and other countries since its introduction in the mid-1950s, and has been continuously improved. Current models embody long combat experience, are faster and longer-ranged, and can better engage rapidly maneuvering targets. Earlier Sidewinders had impact fuzes for their fragmentation warheads; newer models are proximity-fuzed. AIM-9X, an improved close-combat "Evolved Sidewinder," is being developed, with an optical proximity fuze and its infrared seeker guided by the pilot's helmet sight. Early Sidewinders have been converted to the antiradar Sidarm for use by Marine helicopters and Harrier V/STOL ground-attack airplanes.

The original Sparrow I missile entered service in 1956 as a medium-range air-to-air weapon. The newest model of Sparrow III, from which Sea Sparrow was developed, is longer-ranged with much greater computer capability and a more powerful blast-fragmentation warhead with a proximity fuze. Semiactive midcourse guidance is assisted by a command data link. An infrared seeker searches for the target; if it cannot acquire the target, Sparrow continues in the semiactive mode.

Developed to replace Sparrow, though it will probably not completely do so for some time, AMRAAM (Advanced Medium-Range Air-to-Air Missile) has an autopilot which is programmed by the launching plane's computer to bring it within range of the target, when an active radar seeker turns on and engages. A longer-range version, being developed, is intended to replace Phoenix.

Air-to-surface missiles. These weapons can be used against ships; military installations and fortifications; and troops, vehicles, and artillery in the field. Some are designed for use by planes, others by helicopters; some can be used by both.

SLAM (Standoff Land Attack Missile) is designed for use by the F/A-18 Hornet carrier strike fighter. It uses the motor and warhead developed for Harpoon, the infrared seeker from models of Maverick, and Walleye's data link, and incorporates GPS. It was used successfully in the Gulf War. Its course can be corrected in flight, and its infrared seeker transmits a video image to the launching plane's pilot. The pilot visually identifies the target and locks SLAM's control system onto it. SLAM-ER (Expanded Response) has an improved warhead, GPS-assisted inertial midcourse guidance, and an infrared terminal seeker. Its wings, originally developed for Tomahawk, open in flight to improve its range and maneuverability. Its improvements are being backfitted into earlier SLAM missiles.

HARM (High-Speed Antiradiation Missile) is used by carrier planes against ground anti-aircraft defenses, and replaced the earlier Shrike missile. Its terminal seeker homes in on a target's missile control radar.

The air-launched version of Harpoon does not use a booster. It is carried by P-3C patrol planes and A-6E carrier attack planes.

The Joint Standoff Weapon (JSOW) was designed to replace laser-guided ground attack weapons, which place the airplane in danger since it must illuminate the target after launching its missile. JSOW is an unpowered inertially guided glide bomb, able to operate in foul weather and capable of carrying a single explosive charge or of attacking multiple targets with bomblets.

The Joint Air-to-Surface Standoff Missile (JASSM) is being developed for long-range use against defended targets. The version being studied will have inertial GPS-assisted midcourse guidance and an infrared terminal seeker, with either a blast-fragmentation warhead or a penetrating warhead for fortified targets.

The Norwegian-designed Penguin Mark 2 lightweight antiship missile, tested for shipboard use, was ordered for use by Marine helicopters. It has inertial midcourse guidance and an infrared homing seeker, with a range of over 21 nautical miles (39 km). See HELICOPTER.

Sidearm is an early Sidewinder converted to an antiradar homing missile to arm Marine helicopters.

A heat-seeking version of Maverick is used by Navy F/A-18 attack planes; the Marine Corps uses a laser-designated model. Both versions have a blast-fragmentation warhead.

The lightweight Hellfire air-launched antitank missile is a highly maneuverable supersonic weapon. Designed for laser guidance, Hellfire also takes infrared and radio-frequency/infrared seekers.

Skipper is a relatively inexpensive ground-attack weapon, a 1000-lb (454-kg) bomb fitted with an infrared seeker and guidance head and a rocket motor. See AIR ARMAMENT; GUIDED MISSILE.

Antisubmarine weapons. ASROC (antisubmarine rocket), launched by surface warships, was originally designed to carry either a nuclear depth charge or a homing torpedo. All nuclear ASROC warheads were taken out of service by 1989. ASROC is an unguided rocket carrying a Mark 46 homing torpedo. Aimed by shipboard computers using target information obtained by sonar, the rocket is fired from a launcher and follows a ballistic trajectory to the target's predicted position. Torpedo and rocket then separate; the torpedo, slowed by a drag parachute, lands in the water and seeks the target. Vertical-launch ASROC (VLA) is launched by *Ticonderoga*-class cruisers and *Arleigh Burke*-class destroyers from VLS. See ANTISUBMARINE WARFARE.

Rockets. Naval rockets, as distinguished from guided missiles, are unguided weapons carrying explosive warheads. Their light weight, in proportion to explosive payload, and lack of recoil let them be used by attack planes and helicopters. Rockeye II Mark 20 delivers a warhead containing 247 small shaped-charge bomblets; APAM (antipersonnel/antimaterial munitions) is an improved bomblet-dispensing rocket. See ROCKET PROPULSION.

Torpedoes. The torpedo could be called the first guided missile (Fig. 5). Developed and used by the world's navies for more than a century, torpedoes travel underwater on their own power to attack the



Fig. 5. Homing torpedo fired from a destroyer's launcher. (U.S. Navy)

vulnerable hulls of surface ships and submarines. While an above-water hit from another weapon can sometimes be temporarily disregarded when other matters are urgent, a torpedo hit requires immediate, and often drastic, attention. Modern naval torpedoes are ingenious and effective weapons: fast, far-ranging, and armed with a powerful explosive warhead. Torpedoes may be homing (guiding themselves acoustically to the target); nonhoming (following a preset course); or wire-guided (controlled by signals from the firing ship, transmitted through a trailing wire). They can be launched from surface ships, submarines, or aircraft. Homing torpedoes are used as payload by the ASROC system. Methods for countering the homing torpedo, like the weapons themselves, have been worked on since World War II.

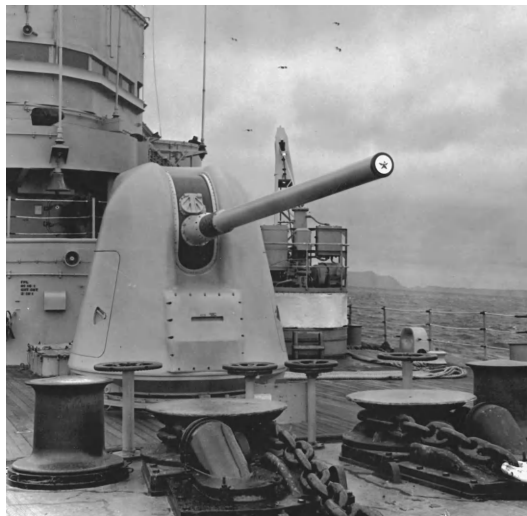


Fig. 6. Mark 45 lightweight 5-in. (127-mm) gun mount. (U.S. Navy)

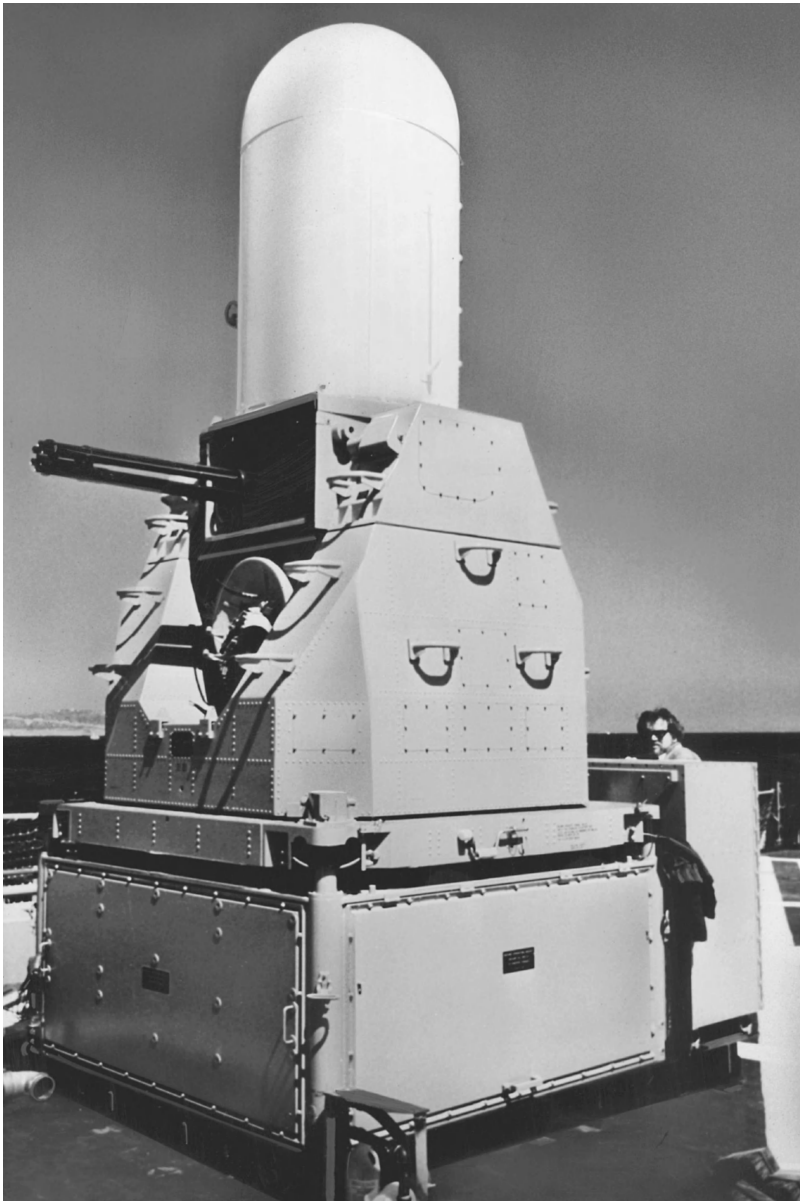


Fig. 7. Phalanx Close-in Weapon System (CIWS). The white dome contains antennas for the weapon's search and tracking radars. (U.S. Navy)

It remains a highly effective weapon, and will probably continue in service for a long time. See ACOUSTIC TORPEDO.

The submarine-launched 21-in. (533-mm) Mark 48 ADCAP (advanced capability) was designed to combat surface ships as well as fast, high-performance modern submarines. An improved version of the original wire-guided Mark 48 homing torpedo introduced in 1972, ADCAP has a higher-power active/passive homing sonar and increased fuel capacity. Many Mark 48's have been kit-upgraded, and an improved, quieter propulsion system is being developed. The lightweight Mark 46 and Mark 50 torpedoes are used by surface ships and aircraft. Introduced in 1963, the Mark 46 is used by many navies and, in its current improved form, is expected to serve for some years to come. The digitally controlled Mark 50 ALWT (advanced lightweight torpedo) can

dive deeper and move faster than the Mark 46; its homing seeker and ability to resist acoustic countermeasures are also improved.

Guns. Though missiles are widely used by ships and aircraft, guns remain significant naval weapons. Missiles are superior for most long-range attack missions and for defense against supersonic planes and missiles at high altitudes and long ranges; the opposite, however, is often true for such missions as shore bombardment, fire support of land forces, and defense against small attack craft. Renewed attention has been given, both in the United States and in other countries, to lighter guns, with high rates of fire, and to quick-reaction control systems for close-in defense against aircraft and missiles in combination with short-range anti-aircraft missiles (Fig. 6).

Guns in shipboard use range in size from 5-in. (127-mm) to 0.50-caliber (12.7-mm). The 5-in./54-caliber Mark 45 automatic gun can fire 16 to 20 70-lb (31.75-kg) shells up to 12.8 nautical miles (23.7 km) per minute. Used by cruisers and destroyers, it is loaded, controlled, and fired from belowdecks, normally with no one in the gun mount. An improvement program is under way to develop a new version of the Mark 45, with a 62-caliber barrel, to fire 40 99.2-lb (45-kg) or 70.5-lb (32-kg) shells per minute. An Extended Range Guided Munition (ERGM) round is under development for the new Mark 45 gun, designed to use inertial guidance and a miniature GPS receiver to carry a payload of bomblets up to 63 nautical miles (116.7 km) with a circular error of probability (CEP) of 11–22 yd (10–20 m).

The Italian-designed Mark 75 76-mm/62-caliber automatic gun can fire up to 85 14-lb (6.4-kg) shells per minute. This light, compact, remotely controlled gun fires explosive or high-performance fragmentation shells with impact or proximity fuzes. With a digital fire-control system, it arms missile frigates and Coast Guard cutters.

The Mark 15 Close-in Weapon System (CIWS), named Phalanx (Fig. 7), is used for rapid-reaction, short-range defense against missiles and planes. CIWS is a modular weapon that includes a search-and-track radar, a computer that identifies incoming high-speed targets and automatically engages them unless overridden by the operator, and an electrically powered six-barreled Gatling-type 20-mm gun, originally designed as the Vulcan for the Air Force. It fires inert, high-density bullets at a cyclic rate of 3000 per minute. The idea behind Phalanx is to produce enough of an umbrella of kinetic-effect projectiles to stand a good chance of hitting an aerial attacker during its final run against a ship.

A number of different types of light automatic guns, firing explosive shells, are used by ships and aircraft, or are being studied for such use. Developed by the Army for armored vehicles, the Mark 38 25-mm Bushmaster is used by combatant craft and is mounted in larger ships for quick-reaction defense against small attack craft and swimmers. A stabilized gun mount, used in *Cyclone*-class patrol craft, accommodates a 25-mm gun and a 40-mm Mark 19 rapid-fire launcher for small,

medium-trajectory, impact-fuzed explosive grenades. A similar stabilized mount is being developed to take a 25-mm gun and short-range air defense missiles. Improved versions of the Oerlikon-designed 20-mm automatic gun of World War II are also used for defense against small craft and swimmers. These fire 800 explosive or armor-piercing rounds per minute. The three-barreled 0.50-caliber (12.7-mm) GAU-19 automatic gun is used in small patrol and special-operations craft. The Browning 0.50-caliber machine gun, developed during World War I as an antitank weapon, was later used by the Navy as an antiaircraft gun. It has been reintroduced for short-range ship defense; the 0.30-caliber (7.62-mm) M60 light machine gun is also used in this role.

Two *Iowa*-class battleships, modernized during the 1980s, remain in the Navy's inactive fleet for potential use in gunfire support of troops ashore. Intensive work during that time showed that the 16-in. (406-mm)/50-caliber gun and its control system could deliver precise fire against fixed or moving targets at long ranges, and that the gun's potential was only then being realized. The 16-in. gun is a bag gun; that is, its powder charge is packed in heavy silk bags. It fires 1900-lb (862-kg) thin-walled explosive shells and 2700-lb (1225-kg) armor-piercing shells, as well as shells which dispense shaped-charge or fragmentation bomblets. Other types of shells have been studied. The *Iowa* has fired accurately at a range of 26 mi (42 km).

The 155-mm Vertical Gun for Advanced Ships (VGAS) is under development for naval fire support. This unusual weapon consists of two vertically mounted guns, automatically loaded to fire up to 10 rounds per gun per minute.

Aircraft guns include the 20-mm Vulcan Gatling-type gun, used by the F-14 Tomcat and F/A-18 Hornet, and the GAU-12 mutibarrel 25-mm gun used by the Marine Corps's AV-8B Harrier.

Bombs. These are free-falling weapons, unlike missiles, which are self-propelled. Bombs take many shapes and sizes, from small antitank and antipersonnel bomblets dispensed from a larger shell or bomb, to heavy weapons designed for blast effect. Navy planes can also deliver nuclear bombs, though these bombs are no longer in service. Most planes and helicopters carry arms externally to accommodate weapon-mix versatility and to keep aircraft size and weight down. High aircraft speeds led to development of streamlined, low-drag bombs. Their better aerodynamic shape is less degrading to aircraft performance, and tends to make them more accurate.

Bombs can be "dumb," that is, uncontrolled, or "smart." Smart bombs have guidance systems and movable control surfaces, and their trajectory can be adjusted to steer them toward a target. Smart bombs were used by Germany and the United States during World War II, but were not seen again until the growing cost of aircraft and sophistication of anti-air defenses demanded more accurate weapons that would reduce the number of sorties required to destroy difficult or heavily defended targets.

General-purpose bombs have a high-explosive filler for maximum blast effect. Fire bombs are loaded with fuel gel to ignite fires. Fuel-air explosive (FAE) bombs generate a cloud of fuel vapor in the air; this is detonated for a combination of blast and fire damage. Laser-guided bombs can be steered into specific targets. Cluster bombs carry a payload of small general-purpose, antitank, or FAE bomblets.

Bombs can be detonated by impact, time, or proximity fuzes. Fuzes may be mechanically or electrically operated, and set to explode instantaneously or to allow a delay between impact and explosion. Electric fuzes can be set in flight to suit the target to be attacked. Proximity fuzes automatically explode a certain distance above the target to produce air bursts. Some fuzes have an antidisturbance feature which arms on impact, to set off a bomb if any attempt is made to disturb it. Various safety features prevent fuzes from arming or detonating prematurely.

Navy aircraft use 500-, 1000-, and 2000-lb (227-, 454-, and 907-kg) unguided, aerodynamically shaped bombs. Conventional conical tail fins give them stability in flight. Snakeye fins open into a cross-shaped arrangement of drag plates, slowing a bomb's descent to permit dropping at low altitudes without endangering the delivering plane. Paveway modification kits convert bombs to laser-guided bombs (LGBs)—standoff weapons using semiactive laser guidance. LGBs steer themselves to their targets using illumination from the launching airplane or from another source. They were used to good effect against vulnerable parts of hardened targets during the 1991 Persian Gulf war.

Mines. A mine is a thin-cased, non-self-propelled weapon filled with high explosive and placed underwater, where it is designed to explode when struck, or closely approached, by a ship. Mines can be contact type (fired by actually striking the hull of a passing ship) or influence type (detonated by the close approach of a ship). An influence mine may be magnetic (actuated by a ship's magnetic field), acoustic (actuated by the underwater sound that a ship generates), or pressure (actuated by the change in water pressure caused by a ship's passage). It may also be fired by a combination of these influences. Influence mines are thus much harder to sweep than contact mines. Mines are planted by submarines or aircraft; some navies also use surface minelayers. *See* ACOUSTIC MINE.

Quickstrike mines are factory-modified 500- and 1000-lb (227- and 454-kg) aircraft bombs; the 2390-lb (1084-kg) Quickstrike Mark 65 was specifically designed as a mine. Destructor mines are also aircraft bombs, kit-converted to influence mines in the field. The Mark 56 antisubmarine influence mine floats at a preset depth, moored to an anchor, unlike most influence mines which are laid on the bottom. The deep-water Mark 60 CAPTOR (enCAPsulated TORpedo) releases a Mark 46 acoustic torpedo when triggered by the acoustic "signature" of a hostile submarine. The submarine-launched mobile mine (SLMM) Mark 67, based on the Mark 37 torpedo, is a shallow-water

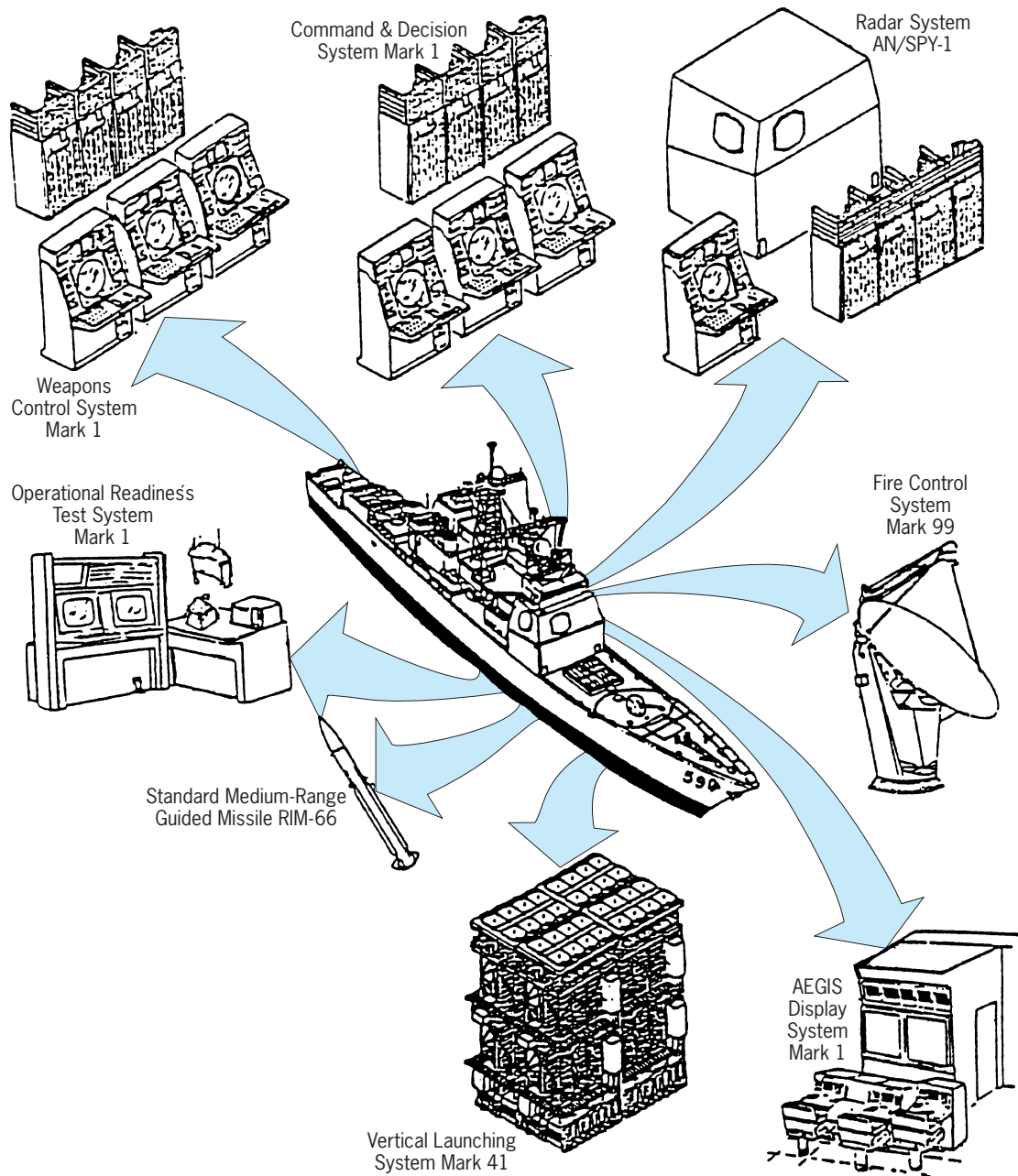


Fig. 8. Elements of the AEGIS Mark 7 weapon system. (U.S. Navy)

mine with its own propulsion system to take it into waters inaccessible to conventional minelaying.

Fire control. This is the method used to control guns and other weapons to ensure that shells, missiles, or other weapons hit the desired target. Major advances have been made in fire control since the beginning of World War II, as greatly improved equipment and sophisticated electronics have produced control systems capable of locking on to a fast, elusive target and hitting it. Different types of control systems are used for surface, air, and underwater targets; some of these are dual-purpose (surface and air). Older systems used electromechanical computers; modern systems are designed around digital computers. Emphasis has been placed on systems suitable for use with various weapons, and able to

engage multiple surface, air, underwater, or land targets in the rapidly changing environment of modern naval combat.

Several fire control systems are used in Navy ships. The most advanced of these is the AEGIS weapon system, an integrated, computer-controlled system using a multipurpose radar to detect contacts in any direction (Fig. 8). It can detect, track, and engage multiple threats of all kinds, while simultaneously maintaining constant all-round surveillance for new threats. It includes the three-dimensional SPY-1 radar, able to track hundreds of targets at once; a command and decision system complex that controls the AEGIS system; a weapon control system which controls the use of the ship's weapons; a fire control system to control the radars used with the Standard missile; command

displays; an automatic fault detection system; and an integrated combat training system. AEGIS also controls a ship's other weapons and electronic systems. The newest Mark 34 gun weapon system was designed around the 5-in./54-caliber Mark 45 gun to be fully integrated into the AEGIS system, and is used in *Arleigh Burke*-class destroyers.

The Mark 86 gun fire control system is used with the 5-in./54-caliber gun in earlier ships for surface and antiaircraft fire as well as shore bombardment. It is linked to such command systems as the Navy Tactical Data System (NTDS). Its digital computer, linked to a "track-while-scan" radar, makes it adaptable for use with the Standard missile. The lighter Dutch-designed Mark 92 fire-control system is used in *Oliver Hazard Perry*-class frigates. Like the Mark 86, it uses a track-while-scan radar and digital computer, and can control guns and missiles against multiple targets. See ELECTRONIC WARFARE. John C. Reilly

Bibliography. A. D. Baker III (ed.), *Naval Institute Guide to Combat Fleets of the World*, annually; D. R. Frieden (ed.), *Principles of Naval Weapons Systems*, 1985; N. Friedman, *Naval Institute Guide to World Naval Weapons Systems*, annually; *Jane's Naval Weapon Systems*, annually; D. G. Kiely, *Naval Surface Weapons*, 1988; T. M. Laur and S. L. Llanso, *Encyclopedia of Modern U.S. Military Weapons*, 1995; *Missile Systems of the World*, Raytheon, 1999; S. L. Morison, *Guide to Naval Mine Warfare*, 1995; N. Polmar, *Naval Institute Guide to the Ships and Aircraft of the U.S. Fleet*, annually; *Principles of Naval Ordnance and Gunnery*, NAVEDTRA 12970, 1992.

Navier-Stokes equation

A partial differential equation which describes the conservation of linear momentum for a linearly viscous (newtonian), incompressible fluid flow. In vector form, this relation is written as Eq. (1), where

$$\rho \left[\frac{\partial \mathbf{V}}{\partial t} + (\mathbf{V} \cdot \nabla) \mathbf{V} \right] = -\nabla p + \rho \mathbf{g} + \mu \nabla^2 \mathbf{V} \quad (1)$$

ρ is fluid density, \mathbf{V} is fluid velocity, p is fluid pressure, \mathbf{g} is the gravitational acceleration, μ is fluid viscosity, ∇ is the del or grad operator, and ∇^2 is the laplacian operator. The equation is named after its two principal developers, French engineer C. L. M. H. Navier (1823) and Irish scientist George G. Stokes (1845). When coupled with the conservation of mass relation, $\nabla \cdot \mathbf{V} = 0$, Eq. (1) can be solved for the space-time distribution of \mathbf{V} and p in a given region of viscous fluid flow. Typical boundary conditions are (1) the knowledge of the velocity and pressure in the far field, and (2) the no-slip condition at solid surfaces (fluid velocity equals solid velocity). See CALCULUS OF VECTORS; GRADIENT OF A SCALAR; LAPLACIAN; NEWTONIAN FLUID; VISCOSITY.

Equation (1) has been successfully applied to the prediction of both laminar flow and disorderly turbulent fluid flows. In the latter case, the evaluation

of the equation's adequacy has been limited to a comparison between numerical approximations to solutions of Eq. (1) and measured time-average values and higher statistics. A significant limitation of Navier-Stokes theory is the lack of any proof regarding uniqueness or existence of solutions of Eq. (1) for given boundary and initial conditions.

The primary dimensionless parameter which governs Eq. (1) is the Reynolds number, given by Eq. (2),

$$\text{Re} = \frac{\rho V L}{\mu} \quad (2)$$

where L is a characteristic body dimension. For small $\text{Re} \ll 1$, Eq. (1) can be simplified by neglecting the left-hand side, resulting in a linear approximation called Stokes flow, or creeping flow, for which many solutions are known. See CREEPING FLOW; LUBRICATION; REYNOLDS NUMBER.

For large $\text{Re} \gg 1$, viscous effects are often confined to a thin boundary layer near solid surfaces, with the remaining flow being nearly inviscid. Equation (1) reduces to a simpler boundary-layer theory, which can be solved in marching fashion along the surface, using the outer inviscid flow as a driving function for the viscous effects. See BOUNDARY-LAYER FLOW; D'ALEMBERT'S PARADOX.

The advent of large-scale, high-speed computers allows Eq. (1) to be modeled and solved by finite-difference techniques. Very fine grids are needed to resolve flow details, but parallel processors allow up to 10^7 mesh points. Laminar flows (moderate Re) are especially well modeled, while turbulent flows (high Re) may require further simplifications of the turbulence terms. See COMPUTATIONAL FLUID DYNAMICS; LAMINAR FLOW; TURBULENT FLOW.

For compressible viscous flow, where density ρ is variable, the term $\mu \nabla^2 \mathbf{V}$ in Eq. (1) is replaced by a more complicated set of stress gradients, including a second coefficient of viscosity, λ . Temperature is also variable, and Eq. (1) must be coupled to the partial differential equation of energy conservation. See COMPRESSIBLE FLOW; FLUID-FLOW PRINCIPLES; HEAT TRANSFER. Frank M. White; Arthur E. Bryson, Jr.; A. Gordon L. Holloway

Bibliography. J. A. Schetz, *Boundary Layer Analysis*, Prentice Hall, Upper Saddle River, NJ, 1992; H. Schlichting and K. Gersten, *Boundary-Layer Theory*, 8th ed., Springer, 2000; F. S. Sherman, *Viscous Flow*, McGraw-Hill, New York, 1990; F. M. White, *Viscous Fluid Flow*, 3d ed., McGraw-Hill, New York, 2005.

Navigation

The process of directing the movement of a craft from one place to another. Navigation involves position, direction, distance, time, and speed.

The process of keeping track of a craft's location by measuring and applying progress from a previous position is called dead reckoning. The location of a craft relative to external reference points such as landmarks or aids to navigation is called piloting.

Radio navigation involves determining distances or directions to radio transmitters. Celestial navigation involves the use of celestial bodies. *See* CELESTIAL NAVIGATION; DEAD RECKONING; PILOTING.

Types of Craft To Be Navigated

The craft to be navigated may be a ship, small marine craft, land vehicle, aircraft, missile, spacecraft, or any moving object requiring direction or capable of being directed, even an animal or bird. The characteristics of the craft have a significant influence upon the type of navigation and the equipment used. Size, mission, weight and space limitations, and economic factors are important considerations.

Ships. Ships of many types are navigated. Combatant military vessels generally require relatively high absolute and relative accuracy and great versatility to accommodate a variety of missions. The capability of coping with the effects of high maneuverability, short time lag between sensing and display of information, and extensive redundancy to maintain navigation capability in the event of damage in combat are important requirements. *See* NAVAL SURFACE SHIP.

Merchant vessels are characterized by voyages from one point to another involving essentially constant courses and constant speeds over relatively long periods of time, little reserve speed, maneuvers only when entering or leaving port or avoiding an obstacle such as another ship, high ratio of time at sea to time in port, and limited crew size. Simplicity of operation, minimum training requirements, good reliability, high maintainability, low obsolescence factor, and low cost of purchase and maintenance are important considerations. These characteristics are typical of freighters, less significant in the case of large passenger vessels, and acutely applicable in the case of fishing vessels. Supertankers with hundreds of thousands of tons of displacement have introduced requirements related primarily to lack of maneuverability because of their size and inertia and to the consequences of a disaster involving such a vessel. *See* MERCHANT SHIP.

Special-purpose vessels, such as hydrographic survey ships, missile tracking vessels, and submarines, have a wide diversity of requirements related to their size, speed, mission, and environment. *See* OCEANOGRAPHIC VESSELS; UNDERWATER VEHICLES; UNDERWATER NAVIGATION.

Small craft. Various types of small craft, particularly recreational craft, have varying requirements commensurate with their operating characteristics. Oceangoing yachts have requirements essentially the same as those of merchant vessels. However, except for the larger yachts, cost, size, weight, and power limitations are generally severe. Sailing vessels may also have other limitations, such as lack of a constant source of electric power. Small craft that do not venture out of sight of land generally have limited navigational capability to match their needs. *See* MARINE NAVIGATION.

Land vehicles. Like small marine craft, land vehicles vary widely in their navigational capability. The

equipment and techniques involved in navigation of vehicles constrained to using highways or rails, although of some diversity, are generally familiar. Vehicles used for transportation across deserts or polar ice fields, or for traversing a jungle or pathless forest, require more attention to navigation.

Inexpensive satellite navigation equipment is increasingly in use in automobiles and commercial vehicles. These applications support timely routing and monitoring of commercial and public safety traffic, automatic monitoring and collection of tolls, and the collection and dissemination of geographic information that allows a navigation-enabled user to find the preferred path from one location to another.

Aircraft. In contrast with surface ships and land vehicles, aircraft generally are characterized by higher speed, greater maneuverability, three-dimensional capability, greater limitations of weight and space, and severe limitations in the inability to hover or park in one spot while the navigation problem is being solved.

The effect of speed is to require less time lag between sensing and display of navigational information, more rapid interpretation of data, and the need for frequent position fixes. Speed is also an important factor in the need for some kind of air-traffic control to accommodate increased traffic density while maintaining safety of flight. Avoidance of other aircraft is an important function of air navigation.

Among different types of aircraft the diversity of characteristics is great. Military fighter aircraft involve high maneuverability, relatively small size, and limited crew. Bombers require high accuracy and, like fighters, have high speed. *See* MILITARY AIRCRAFT.

Intercontinental aircraft are generally large and of two types. The supersonic jet flies above most of the weather and favors a path that climbs as the aircraft becomes lighter through use of fuel. Because of the relatively small effect of wind on a fast-moving aircraft, the envelope of optimum paths at various times between any two points is of small lateral extent. Because of the very high speeds and the lesser effects of winds, a self-contained dead-reckoning type of navigation system providing continuous output of steering information is particularly favored.

The navigation requirements of subsonic turbojets are similar to those of supersonic jets but less critical. Favorable winds are a greater consideration because of the longer flight time.

The diversity and number of aircraft flying in domestic airspace of highly developed countries are considerably greater than is the case over the oceans. In addition to the types that fly over the ocean aircraft flying over land include smaller commercial aircraft, light pleasure and other general aviation craft, short takeoff and landing craft, helicopters, and airships. Because of the availability of numerous short-distance aids and the reduced separation standards, aircraft flying over these land areas make less use of long-distance aids, either of the station-reference piloting type or the self-contained dead-reckoning type. *See* AIR NAVIGATION; AIRSHIP; GENERAL

AVIATION; HELICOPTER; SHORT TAKEOFF AND LANDING (STOL).

Spacecraft. While involving the basic elements of navigation, spacecraft apply them in a somewhat different manner. Upon separation from its booster the spacecraft becomes a celestial body influenced in its motions by all other celestial bodies near enough and large enough to exert a significant gravitational pull. While terrestrial navigation involves motion from one fixed point on the Earth to another (except in the case of interception of a craft in motion), space navigation involves motion from one moving point to another. For successful completion of a flight, a spacecraft must arrive at a designated point in space at the same time as the celestial body to be intercepted. During flight a spacecraft might have long periods during which there is no application of propulsive power. It cannot be steered except by application of power; its attitude can be controlled only by reaction; and terrestrial forms of navigation are largely unavailable. *See* SPACE NAVIGATION AND GUIDANCE.

Elements of Navigation

As indicated above, navigation is concerned primarily with position, direction, distance, time, and speed.

Position. On the Earth, position is generally expressed in terms of geodetic latitude and geodetic longitude, although position of a craft may be stated relative to an aid to navigation, landmark, terminal, or a scheduled position. *See* LATITUDE AND LONGITUDE; TERRESTRIAL COORDINATE SYSTEM.

The position of primary interest to a navigator is that of his or her own craft. If this is determined by advancing a previous position for the direction and distance of travel, it is termed a dead-reckoning position. If it is determined by means of external references, it is called a fix. Traditionally, a fix is established at the intersection of lines of position, which might be arcs of small or great circles, hyperbolas, or other lines.

Direction. Horizontal direction is customarily expressed as angular distance from a reference direction. An adjective may be used to indicate the particular reference used. A "true" direction is based upon the direction of the north geographical pole from the observer. A "magnetic" direction uses the northward direction along the magnetic meridian which passes through the observer. The reference for a "compass" direction is north as indicated by a magnetic compass. This may differ from magnetic north because of magnetic influences within the craft itself. If a marine gyrocompass is used, a "gyro" direction results. Particularly in high latitudes, a somewhat arbitrary grid may be substituted for the graticule of meridians and parallels of latitude. The conventional "north" on this grid is the reference for "grid" direction. The forward direction along the longitudinal axis of the craft is the origin of "relative" directions. *See* GYROCOMPASS; MAGNETIC COMPASS; POLAR NAVIGATION.

The direction in which a craft is pointed is called its heading. The intended direction of travel is called the course, while the actual direction from a point

of departure to a subsequent point of arrival is called the course made good. The actual path followed, and sometimes also the direction of this path, is called the track. A bearing is the direction of one terrestrial point from another, generally the direction of an object as viewed from a craft. The horizontal direction to a celestial body is that body's azimuth, a term also used to indicate the horizontal pointing direction of an antenna.

Distance. Navigators usually express distance in nautical miles. One nautical mile is 1852 m or 6076.115486 ft. This is very nearly the length of 1 minute of arc of a great circle on the surface of the Earth. Yards, feet, and meters are often used for short distances. Depths of water are expressed in fathoms (6 ft or 1.8288 m), feet (1 ft = 0.3048 m), or meters, and heights are expressed in feet or meters.

Time. In navigation, time to the nearest integral minute is usually used for indication of the position of a craft. Celestial observations are usually timed to the nearest integral second. Navigators avoid the use of A.M. and P.M. designation, expressing time on a 24-h basis. Zone time (ZT) and Greenwich mean time (GMT) are used as applicable. Inexpensive quartz-crystal digital watches, exceeding the accuracy of marine chronometers, are now generally used for time both at sea and in the air. They are checked at intervals by radio time signals. *See* TIME.

Speed. The speed of a craft, current, or wind is generally expressed in knots. One knot represents 1 nautical mile per hour (0.5144 m/s). The speed of high-speed aircraft and missiles may be expressed in terms of the Mach number, which is the ratio of the speed of the craft to the speed of sound in the medium concerned. *See* MACH NUMBER.

Navigational Aids

Anything used in navigation, whether aboard the craft or external to it, is properly termed a navigational aid. Thus, in addition to onboard navigational equipment, the term includes such external aids as natural landmarks, prominent buildings, or other structures. Although sometimes used synonymously with "navigational aid," the expression "aid to navigation" is generally restricted to an object or device, external to the craft, established expressly to assist navigation.

Aids to navigation. In the restricted sense, aids to navigation for mariners consist of buoys, beacons, lighthouses, lightships, and navigation sound and electronic transmitters including Loran and navigation satellites. Aids to navigation for aviators consist primarily of radio ranges and beacons and radio position-fixing transmitters including navigation satellites. *See* BUOY; ELECTRONIC NAVIGATION SYSTEMS; HYPERBOLIC NAVIGATION SYSTEM; LIGHTHOUSE; TACAN; VOR (VHF OMNIDIRECTIONAL RANGE).

Charts and publications. A map intended primarily for navigation is called a chart, which displays such information as latitude and longitude scales, useful topographical features and aids to navigation, heights of land and depth of water, cultural features, obstructions, magnetic information, information on

electronic aids, and warning notes. Thus charts are useful for planning trips; for measuring courses, distances, and safe routes; and indicating the available aids to navigation. During the trip a chart is useful to indicate progress. Fixes can be determined by plotting lines of position directly on the chart and graphically moving those that need adjustment to a common time.

The selection of a suitable projection is an important consideration in the production of charts. An important property is conformality, permitting directions to be represented correctly. Nearly all nautical charts, except those for high latitudes, are on the Mercator projection. The Lambert conformal projection is widely used for aeronautical charts. The gnomonic projection, although nonconformal, is useful for representation of great circles, which appear on it as straight lines. The polar stereographic, transverse Mercator, modified Lambert conformal, and azimuthal equidistant projections are in limited use, particularly in polar regions. *See* MAP PROJECTIONS.

Much of the information shown on charts is in internationally standardized symbols and abbreviations. Nautical and aeronautical charts are similar in many respects, but they differ in the emphasis given certain information. Nautical charts are concerned primarily with water areas and information of interest to mariners, while aeronautical charts emphasize land areas, airports, heights of obstructions, and other information of interest to fliers.

Electronic charts. The electronic chart is an electronic facsimile of a nautical chart produced by a microprocessor. The chart becomes a basic part of an electronic chart display information system (ECDIS) when a suitable electronic positioning system such as the Global Positioning System (GPS) or the differential Global Positioning System (DGPS) is used to display a symbol of the ship at its position on the chart. Radar can be used to add useful information such as images of other vessels and buoys, and to confirm the accuracy of the positioning system. Color can be added to help in the identification of various types of information. The possibilities for adding or omitting data are almost limitless, making the combined display an effective presentation of the entire situation of the ship. The display is similar to that of a vessel traffic system (VTS), thus contributing to effective traffic coordination in harbors. Electronic charts can be updated by radio, so that vessels can enter harbors with up-to-date charts. *See* VESSEL TRAFFIC SERVICE.

Paper nautical charts are accepted as legal documents for litigation. In view of the numerous variables in the design and production of electronic charts, standards must be adopted if these charts are to have equal legal stature with paper charts. By the early 1990s, provisional standards were available.

A somewhat simplified electronic navigation system (ENS), using an electronic chart and differential Global Positioning System, has been suggested as more suitable than an electronic chart display information system for some applications, such as

traversing a dredged channel. Unlike the electronic chart system, the electronic navigation system would supplement, rather than replace, a paper chart. The position of the vessel, determined by the differential Global Positioning System where available, would be converted from geographical to channel or other pre-planned track coordinates.

Navigation Practice

Early navigation was essentially an art—the ability to interpret by human senses the elements of navigation by visual observation, sound, and smell.

As knowledge and understanding of the environment increased, the principles of science were utilized in order to develop instruments to improve the accuracy and extend the range of navigation. The application of science has made possible the development of celestial navigation and various new applications of the principles of piloting, dead reckoning, and radio navigation. Radio-navigation aids are now most frequently used.

Thus science has increasingly entered the realm of navigation, culminating in a guidance system that permits accomplishment of a mission without human intervention beyond the programming of the guidance equipment and the launching of the craft. Conventional navigation of craft bearing humans from one place to another generally consists of a blend of science and art. *See* GUIDANCE SYSTEMS.

Traditional methods. Throughout most of the nineteenth century, dead reckoning was generally performed manually by plotting on a chart or plotting sheet. This procedure provided a graphic indication of progress of the craft and a means of evaluating other navigational data received. When the craft was within sight of land or aids to navigation at sea, such as lightships, fixes were obtained from time to time by means of lines of position obtained generally by observing bearings of, or measuring the distance to, visible reference marks. Unless fixes were obtained at intervals of a very few minutes, as when traversing a channel, new dead-reckoning plots were started at each new, reliable fix.

At sea, navigation became more leisurely, consisting primarily of timed celestial observations by sextant during morning and evening twilight, observations of the Sun during the morning, at noon, and during the afternoon, and occasional observations of the Moon, and perhaps Venus. *See* SEXTANT.

Electronics. Traditional piloting and celestial navigation have severe weather limitations. Visual bearings are impossible during periods of heavy fog, and celestial observations with marine sextants can be made only when both the celestial bodies and the horizon are clearly visible. Electronics has increasingly been used to free the navigator from these limitations. The widespread use of computers has further revolutionized the navigation process.

Electronics was introduced into navigation early in the twentieth century when time signals were first broadcast by radio, thus increasing the accuracy and reliability of time indication at sea, an essential

element of longitude determination. The next application was the broadcast of marine weather forecasts, navigational warnings, and other useful information.

The earliest use of electronics for position fixing was in the early 1920s when a medium-frequency radio direction finder was developed, permitting measurement of direction of a known point out of sight, and during periods of restricted visibility. Other methods of determining direction followed. Directional antenna arrays were used extensively in the low-frequency four-course airway ranges established throughout the United States and elsewhere. Rotating patterns of electronic signals provided an additional method. The introduction of radar shortly before the start of World War II provided a means for determining both direction and distance simultaneously. *See* AIRBORNE RADAR; DIRECTION-FINDING EQUIPMENT; RADAR.

The next significant step in the application of electronics to navigation was the introduction of hyperbolic position fixing systems broadcasting synchronized radio signals from sets of two transmitters at known points. Measurement was made of the difference in arrival time of synchronized pulsed signals, or the phase difference of synchronized continuous-wave signals from the two stations, thus permitting the navigator to determine the difference in distances from the stations.

As the uses of electronics in navigation have proliferated, a wide variety of applications have appeared, such as underwater logs, echo sounders, sonar, compass repeaters, Doppler navigators, and electronic charts. Because of limitations on celestial observations, electronic methods of position determination have been particularly attractive. *See* DOPPLER RADAR; ECHO SOUNDER; SONAR.

Inertial navigation. The development of inertial navigators provided a self-contained, passive system continuously indicating the dead-reckoning position of a craft, with worldwide application in any kind of weather. It is widely used in long-range aircraft, submarines, and some special-purpose ships such as those engaged in geophysical exploration. When properly aligned, an inertial navigator uses inertial sensors to indicate speed, heading, and position. Gyroscopes are used to sense angular motions of the craft and to maintain accelerometers in the correct orientation to sense linear accelerations, or changes of speed. Single integration of acceleration provides a measure of speed, and double integrations a measure of distance. The term inertial indicates that measurements are made relative to inertial space; since output relative to the surface of the Earth is desired, provision must be made to eliminate effects related to rotation of the Earth. *See* INERTIAL GUIDANCE SYSTEM.

Satellite navigation. The NAVSTAR Global Positioning System provides very high-accuracy, worldwide, continuous, all-weather, three-dimensional fixes; time; and craft velocity indication. This system transmits signals used to measure pseudorange and pseudorange rate, and simultaneous data from four or more

satellites are converted by the user equipment to the desired information. Position accuracy is measured in meters in the best user equipment. Full operation capability was achieved in 1995.

Because of the desirable characteristics of the Global Positioning System and the development of versatile small user equipment units, the system has wide application in the air, at sea, and on land. As use of the Global Positioning System increases, other positioning systems can be expected to become less popular, and some of them eventually to be phased out. However, because malfunction or other sources of error of any electronic system is a possibility, some form of backup is needed. The integrity, that is, the ability of a system to indicate nonnormal operation, is an important consideration.

Because of the effect of the Earth's atmosphere on the propagation of electronic signals, and present intentional degradation of accuracy for military purposes, accuracy of the Global Positioning System can be improved by use of the differential Global Positioning System. Equipment at an accurately known position, such as in the vicinity of an airport, compares position determined by the Global Positioning System with the known position and transmits the difference as corrections to pseudoranges for use of other receivers in the area.

The Global Orbiting Navigation Satellite System (GLONASS) has been developed in Russia concurrently with the development of the Global Positioning System. It is similar to the Global Positioning System in concept but uses different frequencies and signal formats. *See* SATELLITE NAVIGATION SYSTEMS.

Computers. Development of the electronic computer eliminated much of the drudgery of navigation and enhanced safety by reducing the interval between procurement of data and its application. Further, the computations involved in the use of advanced systems such as inertial navigators and satellite positioning systems would render these systems totally impractical without the availability of fast, accurate electronic computers. Finally, computers make possible the development of integrated systems that provide greater accuracy and reliability by synthesizing the outputs of several independent positioning systems.

The development of fast, accurate computers, particularly the maturity of microcomputers, combined with advanced uses of electronics, has revolutionized navigation. Celestial navigation has essentially been relegated to the realm of recreational boating. The training needed to produce competent navigation has been simplified, and as a result the need for professional navigators has virtually been eliminated. The availability of continuous, accurate real-time position has contributed significantly to safety at sea and in the air. *See* DIGITAL COMPUTER; MICROCOMPUTER.

Traffic control. Traditionally, mariners have resisted attempts by shore-based personnel to restrict their movements at sea. Aviators, on the other hand, recognized early in the history of air travel that some form of traffic control, similar to that in operation on land,

was essential to their safety, and readily accepted the establishment of rules and a system of ground-based direction to help them avoid collisions with obstructions such as other aircraft. As ship speeds have increased, the number of vessels plying the oceans has multiplied, and the ability to communicate and locate the positions of one's own and other vessels has improved, modern ship masters have recognized the advantages of an orderly, prescribed flow of traffic in congested areas, resulting in a proliferation of ship routing systems. Additionally, the improvement in dissemination of reliable weather information has increased the reliance on weather routing of ships on the high seas to permit the selection of routes that are optimum in times of transit and ensure freedom from disabling and discomforting passages.

As the number of aircraft in flight has increased, the available airspace has become increasingly crowded, resulting in delays. The increase in accuracy of position determination of aircraft on a real-time basis makes possible reduction of separation standards, thus increasing airspace capacity. *See* AIR-TRAFFIC CONTROL.

Approach and landing. The most critical part of a flight is the approach and landing. High accuracy and avoidances of other aircraft are critical. Without some form of assistance, landings would be virtually impossible during periods of low visibility. To solve this problem, an electronic instrument landing system (ILS) was developed. Additional transmitters located alongside the runway provide guidance along a runway centerline approach path. In 1949 the International Civil Aviation Organization (ICAO) adopted the system as a worldwide approach and landing aid, but as the demands for precision landing in any condition of visibility increased, a higher-accuracy, increased-capability microwave landing system (MLS) using higher frequencies was developed and may replace the instrument landing system as the international standard. The differential Global Positioning System is also suitable for use as a high-accuracy, reliable landing aid. *See* INSTRUMENT LANDING SYSTEM (ILS); MICROWAVE LANDING SYSTEM (MLS).

Alton B. Moody

Bibliography. N. Bowditch, *American Practical Navigator, 2002 Bicentennial Edition*, National Imagery and Mapping Agency, 2002; T. Cunliffe, *Coastal and Offshore Navigation*, 2d ed., 1999; T. J. Cutler, *Dutton's Nautical Navigation*, 15th ed., 2004; R. R. Hobbs, *Marine Navigation*, 2 vols., 4th ed., 1998; M. Kayton and W. R. Fried, *Avionics Navigation Systems*, 2d ed., 1997.

Neandertals

A group of late archaic humans from Europe, the Near East, and central Asia that immediately preceded the first modern humans in those regions. The Neandertals (also spelled Neanderthals) are included by some within the species *Homo sapiens*, recognizing their close affinities to modern humans; others place them in their own species, *Homo nean-*

derthalensis, emphasizing the differences between them and modern humans.

The first recognized Neandertal remains were found in the Neander Valley near Dusseldorf, Germany, in 1856. Since then the remains of several hundred Neandertals have been discovered. Since the Neandertals were the first humans to bury their dead, a number of largely complete skeletons are preserved, providing detailed knowledge of their biology. Fossils that represent human populations of a similar evolutionary stage have been found in Africa and eastern Asia; these are sometimes referred to as Neandertal-like. Collectively, all of these predecessors of early modern humans can be called late archaic humans. *See* EARLY MODERN HUMANS.

In the early twentieth century, when Neandertals were the only archaic humans known, they were reconstructed as semihuman, dull-witted, and brutish. Hence their popular image was that of the archetypical cavemen. They are now recognized as relatively recent members of the human lineage; they lived between about 125,000 and 36,000 years ago (and as late as 30,000 years ago in certain isolated regions), as compared with earlier members of the genus *Homo* who extend back more than 2 million years. The Neandertals share many features with modern humans both anatomically and behaviorally. Yet, a number of important contrasts between them and more recent humans are recognized.

The Neandertals evolved gradually from more archaic humans during the last interglacial, 100,000–125,000 years ago, across Europe and western Asia. Fossils predating the Neandertals (between 500,000 and 200,000 years old) are becoming well known in these regions. These earlier archaic humans exhibit some but not all of the cranial, facial, and postcranial skeletal features that characterize the Neandertals. Except for the high mountains and the arctic areas, they inhabited most of that region during both the warm last interglacial and the cold phases of the first half of the last glacial period of the Pleistocene Epoch. The Neandertals disappeared 30,000–45,000 years ago across that region, after which time only the fossil remains of early modern humans (called Cro-Magnons in Europe) have been found. Other late archaic humans evolved into or were absorbed by early modern humans probably prior to 50,000 years ago in Africa and by 30,000 years ago in eastern Asia. In the Near East, the fossils of both Neandertals and early modern humans are found from the period between 100,000 and 40,000 years ago. The two groups alternated periods of occupation as the climate and environment of the area fluctuated during the later Pleistocene. Other late archaic humans evolved into and were absorbed by early modern humans, prior to 50,000 years ago in Africa, and by 30,000 years ago in eastern Asia.

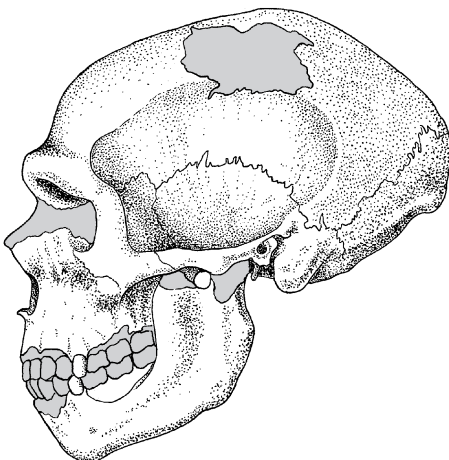
Physically, the Neandertals were about the same height as most modern humans, on the average 5 ft 5 in. (166 cm), but they were much more heavily built. They had heavy necks, broad and muscular shoulders, and extremely muscular arms, hands, and legs. Estimates of their strength show them to

have been about as strong as very athletic modern humans. Their leg bones show a marked thickening of their shafts, which is indicative of both marked strength and endurance—a necessary part of their survival. Even though this physique was energetically costly, it enabled them to exert great strength whenever needed and, more importantly, endure habitually high levels of physical activity.

This strength and endurance reflects their generally less efficient hunting and gathering way of life, as compared with early modern humans. Throughout their range, Neandertal fossils are generally associated with a Middle Paleolithic stone tool tradition called the Mousterian. Neandertal technology consisted of hand-held or wooden-handled stone tools and simple wooden or stone-tipped spears. They appear to have been semi-opportunistic hunters and gatherers, expending considerable time and energy searching for game and plant food, as well as scavenging from the kills of large carnivores. Symbolic systems for keeping track of information about the environment were apparently unknown among them.

The Neandertals are known for their long, low braincases and their projecting faces with large brows and prominent noses (see *illus.*). Their brains were larger than those of modern humans, about 90 in.³ or 1500 cm³ on the average (compared to 78–84 in.³ or 1300–1400 cm³ averages for modern humans). The large brain size was due in part, as with early modern humans, to their large body masses. The length and lowness of their braincases was due to relatively slow brain growth during infancy. There is no evidence that they were less intelligent than modern humans, only that their behavioral system was less elaborate.

The faces of Neandertals had large projecting browridges that led onto wide projecting noses (see *illus.*). Below their noses were dental arcades that held cheek teeth about the size of large modern human ones, and front teeth much larger than those of modern humans. From the rapid rates of wear on their front teeth, it is apparent that they used them for many purposes. As with their muscular hands



Reconstructed skull of a Neandertal from the cave at La Chapelle-aux-Saints, France. (After M. F. Ashley Montagu, *Introduction to Physical Anthropology*, 2d ed., 1951)

and arms, they used the front teeth to compensate for deficiencies in their technology. Although large and projecting, their faces were not especially massive, suggesting that they chewed with little more force than modern humans.

The Neandertals had relatively short forearms and lower legs, and wide trunks, as do modern arctic peoples, indicating an adaptation to minimize heat loss in a glacial climate. Their early modern successors had considerably longer forearms and lower legs, a condition that indicates more protection from the cold. In fact, shelters and hearths among the Neandertals were rudimentary at best, providing little warmth and protection.

There are other indications of the harshness of life among the Neandertals. The oldest lived only into their late 40s, and very few lived beyond their mid-30s. All who managed to survive to age 40 suffered at least one broken bone, and many had debilitating arthritis. About 75% of the Neandertals exhibit malformed tooth enamel, indicating periods of starvation during childhood. Despite these problems, they managed to keep several severely injured individuals alive for decades; to be the first humans to survive glacial climates in Europe and Asia; and to bury their dead.

The position of the Neandertals in modern human ancestry remains controversial. In some regions, the transition from Neandertals to early modern humans was rapid, with more changes in human anatomy than has occurred at any other similarly short time period in human evolution. Many scholars have argued that the time available to transform a Neandertal into a modern human was too short, and they conclude that the Neandertals contributed little, if at all, to populations of modern humans. Others see a natural progression from Neandertals to modern humans, with only an acceleration in the rate of change at the time of the transition. The truth may lie between these extremes, with the Neandertals being absorbed by early modern humans moving slowly into the Near East, central Asia, and Europe between about 100,000 and 30,000 years ago.

Whatever the extent to which Neandertals can be claimed to be ancestors of modern humans, they represent the most recent phase of premodern humans, one in which people were less efficient than modern humans at hunting and gathering, and compensated for their cultural limitations with biological attributes such as tremendous strength, large front teeth, and thermal adaptations. Yet they exhibited the beginnings of many of the attributes of modern humans. They were very successful for about 100,000 years, but they were eventually replaced by humans who were better able to exploit their environments. See FOSSIL HUMANS.

Erik Trinkaus; Steven Churchill

Bibliography. M. H. Nitecki and D. V. Nitecki (eds.), *Origins of Anatomically Modern Humans*, 1994; F. H. Smith, *The Origins of Modern Humans*, 1984; E. Trinkaus (ed.), *The Biocultural Emergence of Modern Humans in the Late Pleistocene*, 1989; E. Trinkaus and P. Shipman, *The Neandertals: Changing the Image of Mankind*, 1994; M. H. Wolpoff,

Human Evolution, McGraw-Hill, 1996; C. Stringer and C. Gamble, *In Search of the Neanderthals*, Thames & Hudson, 1993.

Nearshore processes

Processes that shape the shore features of coastlines and begin the mixing, sorting, and transportation of sediments and runoff from land. In particular, the processes include those interactions among waves, winds, tides, currents, and land that relate to the waters, sediments, and organisms of the nearshore portions of the continental shelf. The nearshore extends from the landward limit of storm-wave influence, seaward to depths where wave shoaling begins. *See* COASTAL LANDFORMS.

The energy for nearshore processes comes from the sea and is produced by the force of winds blowing over the ocean, by the gravitational attraction of Moon and Sun acting on the mass of the ocean, and by various impulsive disturbances at the atmospheric and terrestrial boundaries of the ocean. These forces produce waves and currents that transport energy toward the coast. The configuration of the landmass and adjacent shelves modifies and focuses the flow of energy and determines the intensity of wave and current action in coastal waters. Rivers and winds transport erosion products from the land to the coast, where they are sorted and dispersed by waves and currents.

In temperate latitudes, the dispersive mechanisms operative in the nearshore waters of oceans, bays, and lakes are all quite similar, differing only in intensity and scale, and are determined primarily by the nature of the wave action and the dimensions of the surfzone. The most important mechanisms are the orbital motion of the waves, the basic mechanism by which wave energy is expended on the shallow sea bottom, and the currents of the nearshore circulation system that produce a continuous interchange of water between the surfzone and offshore areas. The dispersion of water and sediments near the coast and the formation and erosion of sandy beaches are some of the common manifestations of nearshore processes.

Erosional and depositional nearshore processes play an important role in determining the configuration of coastlines. Whether deposition or erosion will be predominant in any particular place depends upon a number of interrelated factors: the amount of available beach sand and the location of its source; the configuration of the coastline and of the adjoining ocean floor; and the effects of wave, current, wind, and tidal action. The establishment and persistence of natural sand beaches are often the result of a delicate balance among a number of these factors, and any changes, natural or anthropogenic, tend to upset this equilibrium. *See* DEPOSITIONAL SYSTEMS AND ENVIRONMENTS; EROSION.

Waves. Waves and the currents that they generate are the most important factors in the transportation and deposition of nearshore sediments. Waves are

effective in moving material along the bottom and in placing it in suspension for weaker currents to transport. In the absence of beaches, the direct force of the breaking waves erodes cliffs and sea walls.

Wave action along most coasts is seasonal, responding to changing wind systems over the waters where the waves are generated. The height and period of the waves depend on the speed and duration of the winds generating them, and the fetch, or distance, over which the wind blows. Consequently, the nature and intensity of wave attack against coastlines vary with the size of the water body, as well as with latitude and exposure. Waves generated by winter storms in the Southern Hemisphere of the Pacific Ocean may travel 10,000 km (6000 mi) before breaking on the shores of California, where they are common summer waves for the Northern Hemisphere.

The profiles of ocean waves in deep water are long and low, approaching a sinusoidal form. As the waves enter shallow water, the propagation speed and wavelength decrease, the wave steepens, and the wave height increases until the wave train consists of peaked crests separated by flat troughs. Near the breaker zone, the process of steepening is accelerated so that the breaking wave usually attains a height greater than the deep-water wave. This transformation is particularly pronounced for long-period waves from a distant storm. However, the profiles of local storm waves and the waves generated over small water bodies, such as lakes, show considerable steepness even in deep water.

Wave shoaling, that is, the shallow-water transformation of waves, commences at the depth where the waves "feel bottom." This depth is about one-half the deep-water wavelength, where the wavelength is the horizontal distance from wave crest to crest. Upon entering shallow water, waves are also subjected to refraction, a process in which the wave crests tend to parallel the depth contours, and to wave diffraction, which causes a flow of energy along the wave crest. For straight coasts with parallel contours, refraction decreases the angle between the approaching wave and the coast, and causes a spreading of the energy along the crests. The wave height is decreased by this process, but the effect is uniform along the coast (**Fig. 1**). The amount of wave refraction and diffraction and the consequent change in wave height and direction at any point along the coast is a function of wave period, direction of approach, and the configuration of the bottom topography. *See* OCEAN WAVES; WAVE MOTION IN LIQUIDS.

Nearshore circulation. When waves break so that there is an angle between the crest of the breaking wave and the beach, the momentum of the breaking wave has a component along the beach in the direction of wave propagation. This results in the generation of longshore currents that flow parallel to the beach inside the breaker zone (**Fig. 2a**). After flowing parallel to the beach as longshore currents, the water returns seaward along relatively narrow zones as rip currents. The net onshore transport of water by wave action in the breaker zone, the lateral transport inside the breaker zone by longshore



Fig. 1. Longshore currents, generated when waves approach the beach at an angle. At Oceanside, California, the longshore current is flowing toward the observer. (Department of Engineering, University of California, Berkeley)

currents, the seaward return of the flow through the surf zone by rip currents, and the longshore movement in the expanding head of the rip current together constitute the nearshore circulation system. The pattern that results from this circulation commonly takes the form of an eddy or cell with a vertical axis. The dimensions of the cell are related to the width of the surfzone and the spacing between rip currents. The spacing between rip currents is usually two to eight times the width of the surfzone.

When waves break with their crests parallel to a straight beach, the flow pattern of the nearshore cir-

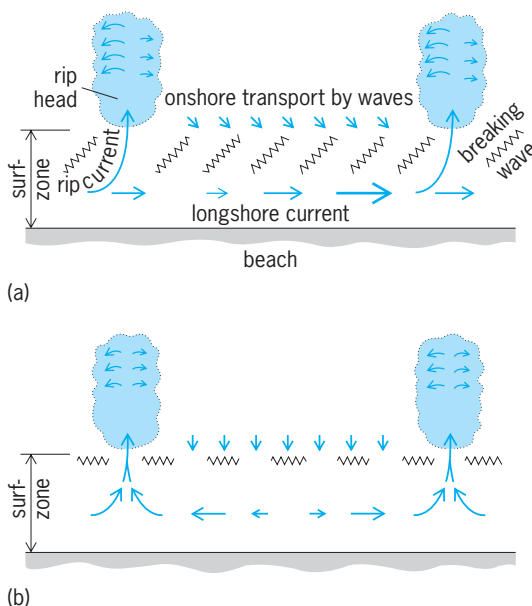


Fig. 2. Definition sketches for nearshore circulation cells. (a) Asymmetrical cell for breakers oblique to the shore. (b) Symmetrical cell for breakers parallel to the shore.

ulation cell becomes symmetrical (Fig. 2b). Longshore currents occur within each cell, but there is little longshore exchange of water or sediment from cell to cell.

The nearshore circulation system produces a continuous interchange between the waters of the surf and offshore zones, acting as a distributing mechanism for nutrients and as a dispersing mechanism for land runoff. Offshore water is transported into the surfzone by breaking waves, and particulate matter is filtered out on the sands of the beach face. Runoff from land and pollutants introduced into the surfzone are carried along the shore and mixed with the offshore waters by the seaward-flowing rip currents. These currents are a danger to swimmers who may be unexpectedly carried seaward. Longshore currents may attain velocities in excess of 2.5 m/s (8 ft/s), while rip current velocities in excess of 1.5 m/s (5ft/s) have been measured. Periodicity or fluctuation of current velocity and direction is a characteristic of flow in the nearshore system. This variability is primarily due to the grouping of high waves followed by low waves, a phenomenon called surf beat that gives rise to a pulsation of water level in the surfzone.

Formation of circulation cells. Nearshore circulation cells result from differences in mean water level in the surfzone associated with changes in breaker height along the beach. Waves transmit momentum in the direction of their travel, and their passage through water produces second-order pressure fields that change the mean water level near the shore. Near the breakpoint, the presence of the pressure field produces a decrease in mean water level (wave setdown) that is proportional to the square of the wave height and, for waves near the surfzone, has a maximum value that is about one-sixteenth that of the breaker height. Shoreward of the breakpoint, the onshore flow of momentum against the beach produces a rise in mean water level over the beach face, called setup.

If the wave height varies along a beach, the setup will also vary, causing a longshore gradient in mean water level within the surfzone. Longshore currents flow from regions of high water to regions of low water, and thus flow away from zones of high waves. The longshore currents flow seaward as rip currents where the breakers are lower. Pronounced changes in breaker height along beaches usually result from wave refraction over irregular offshore topography. However, on a smaller scale, uniformly spaced zones of high and low breakers occur along straight beaches with parallel offshore contours. It has been shown that these alternate zones of high and low waves are due to the interaction of the incident waves traveling toward the beach from deep water with one of the many possible modes of oscillation of the nearshore zone known as edge waves. Edge waves are trapped modes of oscillation that travel along the shore. Circulation in the nearshore cell is enhanced by edge waves having the period of the incident waves, or that of their surf beat, because these interactions produce alternate zones of

high and low breakers whose positions are stationary along the beach. It appears that the edge waves can be either standing or progressive. In either case, the spacing between zones of high waves (and hence between rip currents) is related to the wavelength of the edge wave.

Headlands, breakwaters, and piers influence the circulation pattern and alter the direction of the currents flowing along the shore. In general, these obstructions determine the position of one side of the circulation cell. In places where a relatively straight beach is terminated on the down-current side by points or other obstructions, a pronounced rip current extends seaward. During periods of large waves having a diagonal approach, these rip currents can be traced seaward for distances of 1.6 km (1 mi) or more.

Types of coasts. The geologic setting and the exposure to waves are the two most significant factors in determining nearshore processes. The large-scale features of a coast are associated with its position relative to the edges of the Earth's moving tectonic plates. Accordingly, plate tectonics provides a convenient basis for the first-order classification of coasts, with longshore dimensions of about 1000 km (600 mi). Tectonic classification leads to the definition of three general types: collision coasts, trailing-edge coasts, and marginal seacoasts. See PLATE TECTONICS.

Collision coasts. These coasts occur along active plate margins, where the two plates are in collision or impinging upon each other (Fig. 3a). This area is one of crustal compression and consumption. These coasts are characterized by narrow continental shelves bordered by deep basins and ocean trenches. Submarine canyons cut across the narrow shelves and enter deep water. The shore is often rugged and backed by seacliffs and coastal mountain ranges; earthquakes and volcanism are common. The coastal mountains often contain elevated wave-cut platforms and sea terraces representing former relations between the level of the sea and the land. The west coasts of the Americas are typical examples of collision coasts. See SUBMARINE CANYON.

Trailing-edge coasts. These occur on the trailing edge of a landmass that moves with the plate (Fig. 3b).

They are thus situated upon passive continental margins that form the stable portion of the plate, well away from the plate boundaries. The east coasts of North and South America are examples of mature, trailing-edge coasts. These coasts typically have broad continental shelves that slope into deeper water without a bordering trench. The coastal plain is also typically wide and low-lying and usually contains lagoons and barrier islands, as on the eastern coasts of the Americas. See BARRIER ISLANDS; COASTAL PLAIN; CONTINENTAL MARGIN.

Marginal seacoasts. These coasts develop along the shores of seas enclosed by continents and island arcs. These coasts are typically bordered by wide shelves and shallow seas. The coastal plains of marginal sea coasts vary in width and may be bordered by hills and low mountains. Rivers entering the sea along marginal seacoasts often develop extensive deltas because of the reduced intensity of wave action associated with small bodies of water. Typical marginal seacoasts border the South China and East China seas, the Sea of Okhotsk, the Mediterranean Sea, and the Gulf of Mexico. See BASIN; DELTA.

Other types. A complete classification would also include coasts formed by other agents, such as glacial scour, ice-push, and reef-building organisms, adding two other types of coast: cryogenic and biogenic. Common examples of these two coastal types are arctic coasts and coral reef coasts.

Beaches. Beaches consist of transient clastic material (unconsolidated fragments) that reposes near the interface between the land and the sea and is subject to wave action. The material is in dynamic repose rather than in a stable deposit, and thus the width and thickness of beaches is subject to rapid fluctuations, depending upon the amount and rigor of erosion and transportation of beach material. Beaches along collision coasts are essentially long rivers of sand that are moved by waves and currents and are derived from the material eroded from the coast and brought to the sea by streams. The coast may be cliffed (Fig. 3a), or it may contain a ridge of wind-blown sand dunes and be backed by marshes and water (Fig. 3b). Along many low sandy coasts, such as the east and Gulf coasts of the United States, the beach is separated from the mainland by water or by

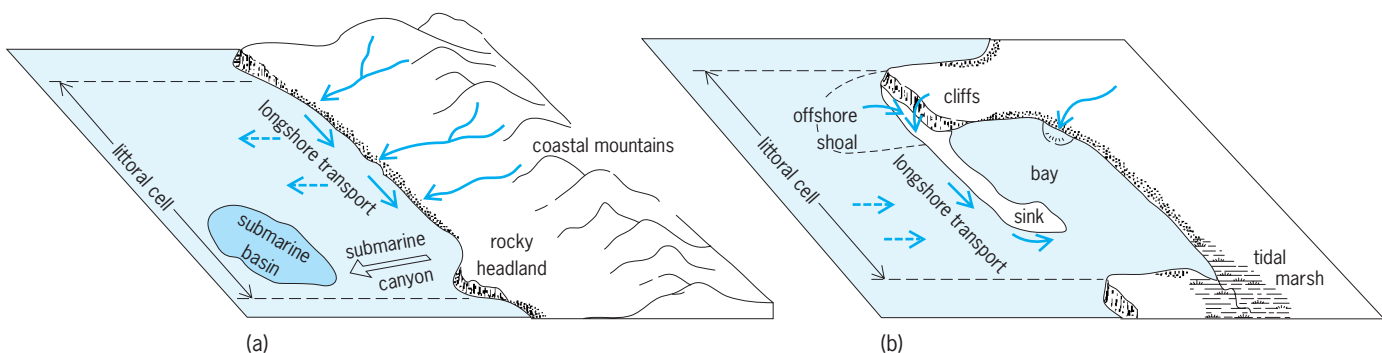


Fig. 3. Typical (a) collision and (b) trailing-edge coasts and their littoral cells. Solid arrows show sediment transport paths; dotted arrows indicate occasional onshore and offshore transport modes. (After D. L. Inman, *Types of coastal zones: Similarities and differences*, in National Research Council, *Environmental Science in the Coastal Zone*, National Academy Press, 1994)

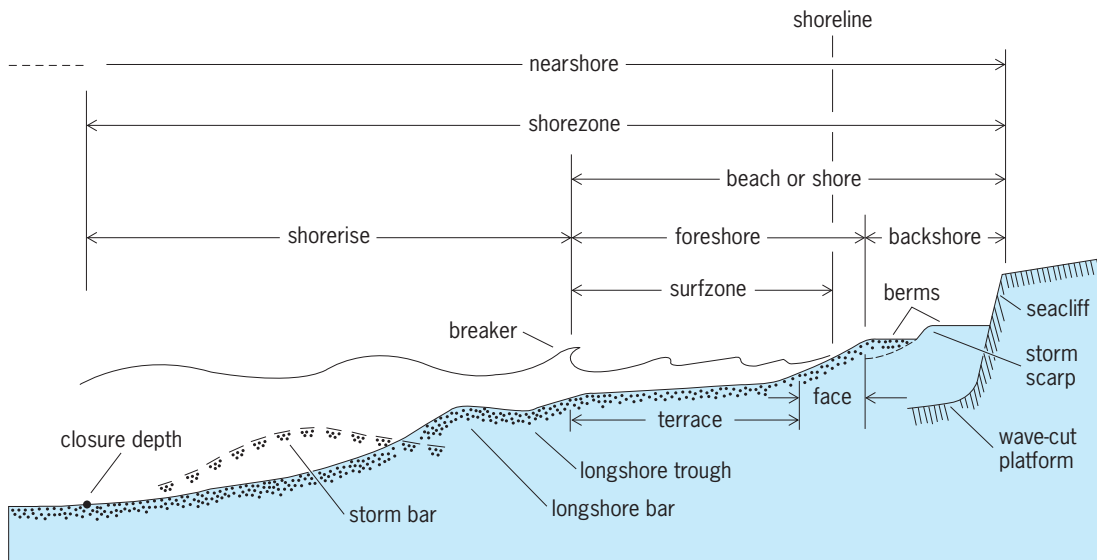


Fig. 4. Beach profile, showing characteristic features.

a natural coastal canal. Such beaches are called barrier islands. Barrier beaches are “braided” forms of the “river of sand” with transgressive rollover caused by sea-level rise and washover processes. A beach that extends from land and terminates in open water is referred to as a spit, while a beach that connects an island or rock to the mainland or another island is a tombolo.

Beach nomenclature. While differing in detail, beaches worldwide have certain characteristic features which allow application of a general terminology to their profile (Fig. 4). The beach or shore extends landward from the breakpoint bar to the effective limit of attack by storm waves. The region seaward is termed the shorerise; that landward is the coast. The coast is part of the coastal zone, which includes the continental shelf and the coast. The beach includes a backshore and foreshore. The backshore is the highest portion and is acted upon by waves only during storms. The foreshore extends from the crest of the berm to the breakpoint bar and is the active portion of the beach traversed by the broken waves and the uprush and backwash of the swash on the beach face. The foreshore consists of a steep seaward-dipping face, related to the size of the beach material and the rigor of the uprush, and of a more gentle seaward terrace, sometimes referred to as the low-tide terrace, over which the waves break and surge. In most localities the foreshore face and terrace merge into one continuous curve; in others there is a discontinuity at the toe of the beach face. The former condition is characteristic of fine sand beaches and of coasts where the wave height is equal to or greater than the tidal range. The latter is typical of coastlines where the tidal range is large compared with the wave height, as along the Patagonian coast of South America and portions of the Gulf of California. The foreshore frequently contains one or more bars and troughs that parallel the beach; these are referred to as longshore bars and longshore troughs. Longshore bars commonly form

at the plunge point of the wave, and their position is thus influenced by the breaker height and the nature of the tidal fluctuation. See TIDE.

The shorerise is the transition between the continental shelf and the beach, and is marked by the increase in slope leading from the gently sloping shelf up to the beach proper. The shorerise extends seaward from the breakpoint bar to the closure depth that marks the seaward extent of depth changes between winter (storm) and summer beach profiles (Fig. 4). Together, the shorerise, foreshore, and backshore comprise the shorezone, which is the zone of active transport of beach material and the resulting areas of beach accretion and erosion.

Beach cycles. Waves are effective in causing sand to be transported laterally along the beach by longshore currents and in causing movements of sand from the beach to the shorerise and back again to the beach. Although these two types of transport are interrelated, for convenience longshore movement of sand is discussed separately.

Along most coasts, there is a seasonal migration of sand between the beaches and the shorerise in response to the changes in the character and direction of approach of the waves. In general, the beach face builds seaward during the small waves of summer and is cut back by high storm waves in winter. There are also shorter cycles of cut-and-fill associated with spring and neap tides and with nonseasonal waves and storms. Bottom surveys indicate that most offshore-onshore interchange of sand occurs in depths of 10–15 m (33–45 ft) but that some effects may extend to depths of 30 m (100 ft) or more.

A typical summer beach is built seaward by low waves (Fig. 4). During stormy seasons, the beach face is eroded, sometimes forming a beach scarp. Subsequent low waves build the beach face seaward again. The beach face is a depositional feature, and its highest point, the berm crest, represents the maximum height of the runup of water on the beach. The height of wave runup above still-water level is

about equal to the height of the breaking wave. Since the height of the berm depends on wave height, the higher berm, if it is present, is sometimes referred to as the winter or storm berm, and the lower berm as the summer berm. The entire beach may be cut back to the underlying country rock during severe storms. Under such conditions, the waves erode the coast and form sea cliffs and wave-cut platforms. These features and their terrace deposits are frequently preserved in the geologic record and serve as markers for the past relations between the levels of the sea and the land.

Mechanics of beach formation. Beaches form wherever there are waves and an adequate supply of sand or coarser material. Even anthropogenic fills and structures are effectively eroded and reformed by the waves. The initial event in the formation of a new beach from a heterogeneous sediment is the sorting of the material, with coarse material remaining on the beach and fine material being carried away. Concurrent with the sorting action, the material is rearranged, some being piled high above the water level by the runup of the waves to form the beach berm, some carried back down the face to form the foreshore terrace. In a relatively short time, the beach assumes a profile that is in equilibrium with the forces generating it.

The beach face is frequently characterized by laminations (closely spaced layers) that show slight differences in the size, shape, or density of the sand grains. The laminations parallel the beach face and represent shear sorting within the granular load as it is transported by the swash and backwash of the waves over the beach face. Detailed examination shows that each lamina of fine-grained minerals delineates the plane of shear between moving and residual sand. The mechanism concentrating the heavy fine grains at the shear plane is partly the effect of gravity acting during shearing, causing the small grains to work their way down through the interstices between larger grains, and partly the dependence of the normal dispersive pressure upon grain size. When grains are sheared, the normal dispersive pressure between grains, which varies as the square of the grain diameter, causes large grains to drift toward the zone of least shear strain, that is, the free surface, and the smaller grains toward the zone of greatest shear strain, that is, the shear plane.

Equilibrium profile. The action of waves on a sloping beach eventually produces a profile that achieves equilibrium with the energy dissipation associated with the oscillatory motion of the waves over the sand bottom. The slope of the beach face is related to the dissipation of energy by the swash and backwash over the beach face. Percolation of the swash into a permeable beach reduces the amount of flow in the backwash and is thus conducive to deposition of the sand transported by the swash. If in addition the beach is dry, this action is accentuated. Coarse sands are more permeable and consequently more conducive to deposition and the formation of steep beach faces. Large waves elevate the water table in

the beach. When the beach is saturated, the backwash has a higher velocity, a condition conducive to erosion. From the foregoing it follows that the slope of the beach foreshore increases with increasing sediment size and with decreasing wave height. If an artificial slope exceeds the natural equilibrium slope, an offshore transport of sand will result from the gravity component of the sand load until the slope reaches equilibrium. Conversely, if an artificial slope is less than the natural equilibrium slope, a shoreward transport of sand by waves and currents will result, and the beach slope will steepen. An equilibrium slope is attained when the up-slope and down-slope transports are equal.

Beaches respond to wave-forcing by adjusting their form to an equilibrium or constant shape attributable to a given type of incident wave. The seasonal changes in beach profile in response to the high waves of winter and the lower waves of summer are expressions of the beach form tending toward a seasonal equilibrium with the changing character of the prevailing waves. Field studies show that the equilibrium beach consists of two conjoined parabolic curves that intersect at the breakpoint bar, one curve for the shorerise that extends from the closure depth to the breakpoint bar, and another for the foreshore that extends from the bar to the berm (Fig. 4). The principal differences between seasonal profiles are that in winter (higher waves) the breakpoint bar is deeper and farther offshore, while the berm crest is displaced landward. Thus, the changes in seasonal equilibria are manifest by simple, self-similar displacements of the bar-berm and shorerise curves. Also, since the equilibrium profile is parabolic, for any given sand size, the slope increases from the bar to the berm.

Longshore movement of sand. The movement of sand along the shore occurs in the form of bed load (material rolled and dragged along the bottom) and suspended load (material stirred up and carried with the current). Suspended-load transportation occurs primarily in the surf zone, where the turbulence and vertical mixing of water are most effective in placing sand in suspension and where the longshore currents that transport the sediment-laden waters have the highest velocity. The longshore transport rate of sand is directly proportional to the longshore component of wave power. Thus, the longshore transport rate of sand can be estimated from a knowledge of the wave climate, that is, the budget of wave energy incident upon the beach.

The volume of littoral transport along oceanic coasts is usually estimated from the observed rates of erosion or accretion, most often in the vicinity of natural obstructions, such as headlands (Fig. 5), or of coastal engineering structures, such as groins or jetties. In general, beaches build seaward up-current from obstructions and are eroded on the current lee where the supply of sand is diminished. Such observations indicate that the transport rate varies from almost nothing to several million cubic meters per year, with average values of 150,000–600,000 m³



Fig. 5. Effect of headlands on the accretion of beach sand at Point Mugu, California. The point forms a natural obstruction that interrupts the longshore transport of sand, causing accretion and a wide beach to form (foreground). The regularly spaced scallops are swash cusps on the beach face. (Department of Engineering, University of California, Berkeley)

(200,000–800,000 yd³) per year. Along the shores of smaller bodies of water, such as the Great Lakes of the United States, the littoral transport rate can be expected to range about 7000–150,000 m³ (9000–200,000 yd³) per year.

The large quantity of sand moved along the shore and the pattern of accretion and erosion that occurs when the flow is interrupted pose serious problems for coastal engineers. The problem is particularly acute when jetties are constructed to stabilize and maintain deep navigation channels through sandy beaches. A common remedial procedure is to dredge sand periodically from the accretion on the up-current side of the obstruction and deposit it on the eroding beaches in the current lee. Another method is the installation of sand bypassing systems, which continually remove the accreting sand and transport it by hydraulic pipeline to the beaches in the lee of the obstruction. See COASTAL ENGINEERING; SEDIMENTOLOGY.

Sources and sinks of beach sediment. The principal sources of beach and nearshore sediments are the rivers that bring quantities of sand directly to the ocean; the seacliffs and bluffs of unconsolidated material that are eroded by waves; and material of biogenous origin (shell and coral fragments and skeletons of small marine animals). In places, sediment may be supplied by the erosion of uncon-

solidated deposits in shallow water (Fig. 3*b*). Beach sediments on the coasts of the Netherlands are derived in part from the shallow waters of the North Sea. Windblown sand may be a source of beach sediment, although winds are usually more effective in removing sand from beaches than in supplying it. In tropical latitudes, many beaches are composed entirely of grains of calcium carbonate of biogenous origin. Generally the material consists of fragments of shells, corals, and calcareous algae growing on or near fringing reefs. The material is carried to the beach by wave action over the reef. Some beaches are composed mainly of the tests (shells) of foraminifera that live on sandy bottom offshore from the reefs. See NORTH SEA; REEF.

Streams and rivers may be important sources of sand for beaches in temperate latitudes (Fig. 3*a*). Surprisingly, the contribution of sand by streams in arid climates is quite high (Fig. 6). Arid weathering produces sand-size material from areas with a minimum cover of vegetation, so that occasional flash floods may transport large volumes of sand. The maximum sediment yield occurs from drainage basins where the mean annual precipitation is about 30 cm (12 in.) per year.

There is a pronounced multidecadal variability in the amount of river-borne sediment transported to the beach. The variability is associated with global climate changes related to the El Niño/Southern Oscillation (ENSO) phenomena. ENSO drives large-scale events such as the Pacific/North American (PNA) patterns of atmospheric pressure that lead to wet and dry climate along the Pacific coast of North America. The 20 coastal rivers of central and southern California had streamflow and sediment fluxes during the wet phase of PNA (1969–1998) that exceed those during the preceding dry phase (1944–1968) by factors of 3 and 5, respectively. The sediment flux during the three major El Niño events of the wet phase were on average 27 times greater than the annual sediment flux during the dry phase. Also, the wave climate in southern California changed with the shift from dry to wet phase of PNA. The prevailing northwesterly winter waves of the dry phase



Fig. 6. Sand delta at Rio de la Concepción on the arid coast of the Gulf of California. Such deltas are important sources of sand for beaches. (Courtesy of D. L. Inman)

were replaced by high-energy waves approaching from the west or southwest during the wet phase. Wave climate along the east coast of North America responds to shifts in the atmospheric patterns of the North Atlantic Oscillation (NAO), which is generally out of phase with the west coast climate. *See* CLIMATIC PREDICTION; EL NIÑO; TROPICAL METEOROLOGY; WEATHER FORECASTING AND PREDICTION.

Wave erosion of rocky coasts is usually slow, even where the rocks are relatively soft shales. Therefore, cliff erosion usually does not account for more than about 10% of the material on most beaches. However, retreats greater than 1 m (3.3 ft) per year are not uncommon in unconsolidated seacliffs. The most dramatic modern example of coastal erosion is found along the delta of the Nile River in Egypt. The High Aswan Dam, constructed in 1964, has intercepted the sediment that was previously brought down the Nile to the coast. Lacking source material, the delta coast is eroding, and waves and currents may cause an entire city block of Ras El Bar to be lost to the sea in one year.

The sand carried along the coast by waves and longshore currents may be deposited in continental embayments, or it may be diverted to deeper water by submarine canyons which traverse the continental shelf and effectively tap the supply of sand (Fig. 3a). Most of the deep sediments on the abyssal plains along a 400-km (250-mi) section of the California coastline are probably derived from two submarine canyons, Delgada Submarine Canyon in northern California and Monterey Submarine Canyon in central California. *See* MARINE SEDIMENTS.

Littoral cells and the budget of sediment. The budget of sediment for a region is obtained by assessing the sedimentary contributions and losses to the region and their relation to the various sediment sources and transport mechanisms. Determination of the budget of sediment is not a simple matter, since it requires knowledge of the rates of erosion and deposition as well as understanding of the capacity of various transport agents. Studies of the budget of sediment show that coastal areas can be divided into a series of discrete sedimentation compartments called littoral cells. Each cell contains a complete cycle of littoral transportation and sedimentation, including transport paths and sources and sinks of sediment. Littoral cells take a variety of forms, but there are two basic types. One is characteristic of collision coasts with submarine canyons (Fig. 3a), while the other is more typical of trailing-edge coasts where rivers empty into large estuaries (Fig. 3b). The concept of a littoral cell (or a subcell) with its budget of sediment and transport paths provides objective criteria for making choices among various coastal conservation methods. *See* ESTUARINE OCEANOGRAPHY.

Biological effects. The rigor of wave action and the continually shifting substrate make the sand beach a unique biological environment. Because few large plants can survive, the beach is occupied mostly by animals and microscopic plants. Much of the food

supply for the animals consists of particulate matter that is brought to the beach by the nearshore circulation system and is trapped in the sand. The beach acts as a giant sand filter straining out particulate matter from the water that percolates through the beach face.

Since the beach-forming processes and the trapping of material by currents and sand are much the same everywhere, the animals found on sand beaches throughout the world are similar in aspects and habits, although different species are present in different localities. In addition, since the slopes and other physical properties of beaches are closely related to elevation, the sea animals also exhibit a marked horizontal zonation. Organisms on the active portion of the beach face tend to be of two general types insofar as the procurement of food is concerned: those that burrow into the sand, using it for refuge while they filter particulate matter from the water through siphons or other appendages that protrude above the sandy bottom, and those that remove organic material from the surface of the sand grains by ingesting them or by "licking." There are usually few species, which may be very abundant.

In tropical seas, the entire shore may be composed of the cemented and interlocking skeletons of reef-building corals and calcareous algae. When this occurs, the nearshore current system is controlled by the configuration of the reefs. Where there are fringing reefs, breaking waves carry water over the edge of the reef, generating currents that flow along the shore inside of the reef and then flow back to sea through deep channels between reefs. Under such conditions, beaches are usually restricted to a berm and foreshore face bordering the shoreward edge of the reef.

Douglas L. Inman

Bibliography. W. Bascom, *Waves and Beaches: The Dynamics of the Ocean Surface*, rev. ed., 1980; D. C. Conley and D. L. Inman, Field observations of the fluid-granular boundary layer under near-breaking waves, *J. Geophys. Res.*, 97(C6):9631-43, 1992; R. A. Davis, Jr., *The Evolving Coast*, 1994; K. Horikawa (ed.), *Nearshore Dynamics and Coastal Processes*, 1988; D. L. Inman and B. M. Brush, The coastal challenge, *Science*, 181:20-32, 1973; D. L. Inman and R. Dolan, The Outer Banks of North Carolina: Budget of sediment and inlet dynamics along a migrating barrier system, *J. Coastal Res.*, 5(2):193-237, 1989; D. L. Inman, M. H. S. Elwany, and S. A. Jenkins, Shoreline and bar-berm profiles on ocean beaches, *J. Geophys. Res.*, 9(C10):18,181-18,199, 1993; D. L. Inman and S. A. Jenkins, Climate change and the episodicity of sediment flux of small California rivers, *J. Geol.*, 107(3):251-70, 1999; J. P. Kennett, *Marine Geology*, 1982; P. D. Komar, *Beach Processes and Sedimentation*, rev. ed., 1998; National Research Council, *Environmental Science in the Coastal Zone: Issues for Further Research*, 1994; R. J. Seymour (ed.), *Nearshore Sediment Transport*, 1989; U.S. Army Corps of Engineers, *Engineering and Design—Coastal Littoral Transport*; L. D. Wright, *Morphodynamics of Inner Continental Shelves*, 1995.

Nebula

The term “nebula” was originally used to refer to any fixed, extended, and usually fuzzy luminous celestial object. With increased angular resolution of telescopes, astronomers learned that nebulae can be separated into two classes: those that are stellar systems made up of individual stars, and those that are gaseous in nature and diffuse in appearance. Examples of stellar systems include galaxies (which contain billions of stars and are located outside our own Milky Way Galaxy) and star clusters such as open clusters and globular clusters (which contain thousands of stars and are within the Milky Way Galaxy). This article is restricted to the modern definition of nebulae, which are gaseous objects usually located within the Milky Way Galaxy, although with increasingly powerful telescopes gaseous nebulae can now be observed in external galaxies. *See* GALAXY, EXTERNAL; MILKY WAY GALAXY; STAR.

Because of space observations, the study of gaseous nebulae has undergone a renaissance. The extension of observations into the x-ray, ultraviolet, infrared, millimeter, and submillimeter wavelengths has revealed a much richer makeup of nebulae, containing almost all states of matter. Although the term “gaseous nebulae” is used, in fact they consist of ions, atoms, molecules, and solid particles. Sometimes the term “clouds” is also used to refer to concentrations of interstellar matter (for example, dark clouds, molecular clouds); there is, however, no precise difference in meaning between “nebulae” and “clouds.”

Types. Gaseous nebulae can be divided into three main types. Those that radiate brightly in the visible are called emission nebulae. Those that are detected through their effects on the obscuration of background stars are called dark nebulae or absorption nebulae. There is also a class of nebulae that do not self-radiate in visible light but reflect light from nearby stars, and those are called reflection nebulae. Examples of emission nebulae include the Great Orion Nebula, which is a region of active star formation. Other examples of emission nebulae are planetary nebulae and supernova remnants, which are objects associated with dying stars. *See* ORION NEBULA.

Nebulae associated with star formation. New stars are formed out of concentrations of dust and gas in the interstellar medium. New stars often form in groups. For clusters of newborn stars containing hot stars, the ultraviolet light from the stars is able to ionize the surrounding gas, making them luminous. In astronomical nomenclature, they are called H II regions, referring to the fact that the hydrogen atoms in the nebulae are in ionized form, with the electron detached from the proton (**Colorplate 1**).

Planetary nebulae. Planetary nebulae are so denoted because they often resemble small greenish disks when seen in small telescopes, not unlike the images of the planets Uranus and Neptune. Well-known examples of this class are the Ring Nebula (NGC 6720 or M57) in Lyra, the Helix Nebula (NGC 7293) in

Aquarius, and the Dumbbell Nebula (M27) in Vulpecula. Unlike H II regions, which are associated with the birth of stars, planetary nebulae are products of stars near their deaths. They are ejected by very old red giants and contain fractions of a solar mass of material. Since they are continuously expanding, they have short lifetimes lasting only tens of thousands of years. *See* GIANT STAR.

Planetary nebulae derive their power from their central stars. After the ejection, the remnant of the original red giant becomes a hot, blue star that eventually evolves to become a white dwarf. The ultraviolet light from the star can ionize the ejecta, leading to the emission of a variety of atomic lines in different colors. Planetary nebulae are colorful objects and can be considered as neon signs in the sky. They have well-defined morphologies (often shell-like and sometimes bipolar) and have sizes of fractions of a light-year (**Colorplate 2**). *See* PLANETARY NEBULA; WHITE DWARF STAR.

Supernova remnants. The detonation of a star in a supernova event causes the ejection of the outer layers into the surrounding interstellar medium. In early stages, as in the Crab Nebula, the radiating material consists of ejecta from the star. In the later stages, this rapidly moving material is slowed down as it mixes with the surrounding dust and gas of the interstellar medium. Heating by shock waves causes the material to radiate optically. Supernova remnants are also characterized by their x-ray emission arising from the rarefied gas behind the shock front, and by synchrotron radiation at radio frequencies. Since both planetary nebulae and supernova remnants are stellar ejecta and surround their parent stars, they are referred to as circumstellar nebulae to distinguish them from H II regions, which are interstellar in origin. *See* CRAB NEBULA; SUPERNOVA.

Reflection nebulae. In contrast to emission nebulae, reflection nebulae do not shine on their own in the visible. The nebular structure is derived from starlight scattered off dust grains in the nebula. As a result, the optical spectrum of a reflection nebula is made up of the continuous spectrum of the illuminating stars. However, the color of a reflection nebula is often bluer than the stellar spectrum because the dust particles scatter more efficiently in the blue. The best-known example is the Pleiades reflection nebula, which is illuminated by the brightest stars in the Pleiades cluster. *See* PLEIADES.

Typical reflection nebulae have sizes of fractions of a light-year, with densities of $\sim 10^4$ atoms per cubic centimeter. The central stars are generally cooler than 20,000 K (36,000°F), not hot enough to photoionize the nebula. The Trifid Nebula (M20) is an example of an object which is part emission and part reflection nebula. In **Colorplate 1**, the red part of the nebula is an emission nebula resulting from H-alpha emission, whereas the blue part is a reflection nebula resulting from starlight scattered off dust.

Dark clouds. Dark clouds represent a concentration of solid dust particles that obscure starlight. By preventing light from stars located behind the cloud from reaching us, they appear as dark patches in the

sky. Many dark clouds have no strong heating sources in the form of central stars, but are heated primarily by external nearby stars or cosmic rays. For this reason, the gas temperature inside dark clouds can be very low (about 10 K or -442°F). On a smaller scale, there are compact, dense clouds known as Bok globules. They are often spherical in shape and have densities higher than dark clouds. *See* GLOBULE.

Although dark in the visible light, dark clouds are luminous in the infrared. The energy that they absorb from starlight is converted into infrared emission, and they can be detected in emission by modern infrared and submillimeter-wave telescopes.

Molecular clouds. Many dark clouds also contain molecules that can be detected through their rotational transitions in the millimeter wavelengths. Interstellar clouds that are heavy in molecular content are called molecular clouds. Molecular clouds that are bright in the infrared and molecular line emissions are called giant molecular clouds. Their dust and gas temperatures, inferred from infrared continuum and molecular-line emissions respectively, are in the range of 50 to 100 K (-370 to -280°F). Such relatively high temperatures suggest that there must be heating sources inside the cloud. These internal sources can be identified by near-infrared imaging and are believed to be newly formed massive (OB) stars. The inferred masses (10^3 to 10^6 solar masses) of giant molecular clouds suggest that they are not gravitationally stable, and kinematic studies using molecular lines confirm that many are in the process of gravitational collapse. Giant molecular clouds are therefore excellent objects to study the formation process of massive stars. *See* MOLECULAR CLOUD.

Discovery and catalogs. The first catalog of nebulae was made by Charles Messier in 1784. A much more complete list was made by J. L. E. Dreyer in the *New General Catalogue (NGC)*, published in 1888, and the two *Index Catalogs*. After the completion of the National Geographic Society-Palomar Observatory Sky Survey, many large and faint nebulae were identified by George Abell from the photographic plates of the survey. Since these catalogs were compiled based on the morphology of the objects, the lists include galaxies and star clusters in addition to gaseous nebulae. *See* ASTRONOMICAL CATALOGS; MESSIER CATALOG.

A modern definition of nebulae does not rely solely on morphological appearance but also on spectroscopic characteristics. Specifically, an emission nebula is defined by its emission-line spectrum, and an object can be classified as a gaseous nebula even though its angular size is so small that it is not seen as an extended object by a telescope. For example, by employing objective prisms in telescopes, astronomers in the 1960s were able to distinguish emission nebulae from stars. A more modern technique uses filters covering a narrow range of color, for example, centering on the H-alpha line of hydrogen at 656.2 nm or the 500.7-nm line of doubly ionized oxygen (O^{2+}). By comparing images taken with these narrow filters to those with broad filters,

emission nebulae will stand out as bright objects. These techniques allow the identification of small gaseous nebulae that appear stellar in photographic images. *See* ASTRONOMICAL SPECTROSCOPY; INTERFERENCE FILTER.

H II regions, planetary nebulae, and supernova remnants are radio emitters, and these objects can be found through radio surveys. By making measurements at several frequencies, the spectral shape of the emitting radio source can be determined, therefore allowing the distinction between H II regions and supernova remnants, for example.

Since gaseous nebulae are dusty and radiate strongly in the infrared, their existence can be found through infrared sky surveys. The most notable infrared sky surveys are the Infrared Astronomical Satellite survey, which covered 96% of the sky at wavelengths of 12, 25, 60, and 100 μm ; the 2MASS survey at 2 μm ; and the Spitzer Space Telescope Galactic Legacy Infrared Mid-Plane Survey Extraordinaire (GLIMPSE) survey of the galactic plane at 3.6, 4.5, 5.8, and 8.0 μm . Nebulae that have so much dust that their visible light output is obscured by their own internal dust are most efficiently discovered in infrared surveys.

In addition to visual inspection of photographic plates, dark nebulae can be found using the technique of star counts. If an area in the sky has a lower density of stars compared to its neighboring region, it probably implies the presence of a dark nebula.

Composition. Gaseous nebulae are made up of free-flying atoms in the gaseous state. These atoms can be in the neutral (uncharged) state, or in an ionized state where one or more of the electrons of the atom have been removed from the atom. The ionized state of gas is also referred to as the plasma state. Since hydrogen is the most common element in the universe, gaseous nebulae consist of primarily hydrogen, with helium second in abundance, and then oxygen, carbon, and other heavy elements. *See* PLASMA (PHYSICS).

Molecular gas is also present in nebulae. Although most molecules consist of two, three, or four atoms [for example, molecular hydrogen (H_2) and carbon monoxide (CO)], very complex molecules with over a dozen atoms have also been seen. The molecular gas state is similar to the gaseous state in the Earth's atmosphere, where molecules of nitrogen (N_2) and oxygen (O_2) are the dominant species.

Also present in nebulae are solid-state particles, which are generally referred to as "dust" in astronomical literature. These are macroscopic particles with sizes up to about 1 μm , each consisting of thousands of atoms. The most common inorganic dust consists of silicates, which are made up of atoms of oxygen, silicon, magnesium, and iron. Organic carbonaceous dust, consisting of primarily carbon and hydrogen, is also widely seen. *See* INTERSTELLAR MATTER.

Spectral characteristics. The different states of matter in a nebula can be traced by their different radiation mechanisms. Atoms and ions can be identified by their atomic electronic transitions, molecules by

their vibrational or rotational transitions, and solid particles by their lattice vibrational modes. Since different atoms radiate in different specific frequencies, they can be uniquely identified by spectroscopic observations.

The densities in gaseous nebulae are very low in comparison to the Earth's atmosphere. In fact, the densities are lower than the best vacuum that can be achieved in the laboratory. Consequently, the excitation of atoms and molecules and their subsequent radiation are very different from usual mechanisms observed in the terrestrial environment.

Atoms and ions. The electronic transitions of atoms and ions occur in the ultraviolet, visible, and infrared parts of the electromagnetic spectrum. These are called bound-bound transitions, as they are the consequence of an electron going from one bound state to another. The vibrational transitions of molecules generally occur in the infrared, while the rotational transitions occur in the millimeter or submillimeter parts of the spectrum. Atomic electronic transitions observed in nebulae can be separated into two kinds. The first is the cascade of electrons from a higher state to a lower state after a free electron has been captured by the nucleus of an ion. These are called recombination lines. Usually, only the recombination lines of hydrogen and helium are bright enough to be detected, although weaker recombination lines such as those from oxygen, carbon, and nitrogen have been observed. The second class of atomic transitions occurs as the result of excitation of a bound electron from the ground state to a higher state through collisions with a free electron. These are called collisionally excited lines. Because the recombination lines of heavy elements are weak, most chemical elements are detected in gaseous nebulae through their collisionally excited lines. Because of improvements in sensitivity made possible by the use of charge-coupled devices (CCDs), emission lines from elements in rows 4, 5, or 6 of the periodic table (for example, selenium, krypton, and xenon) have been detected. The observation and measurement of abundance of heavy elements are very important in understanding the origin of chemical elements and the nuclear reactions responsible for the synthesis of these elements. *See* ATOMIC STRUCTURE AND SPECTRA.

One interesting aspect of collisionally excited lines is the detection of "forbidden lines." These are atomic transitions that occur too slowly to be seen in the terrestrial environment. Because of the high density of the Earth's atmosphere, atoms in an excited state that decay slowly (called metastable states) are usually deexcited by collisions before they have a chance to radiate. However, in the interstellar medium the density is very low and collisions are infrequent. Atoms can spend hours in an excited state without encountering another electron. This allows the appearance of "forbidden lines." One of the strongest forbidden lines in the visible is the green line of O^{2+} . This line at the wavelength of 501 nm is so strong that it can be detected from gaseous nebulae

in very distant galaxies. *See* SELECTION RULES (PHYSICS).

A more unusual, third kind of atomic transition involves the fluorescent mechanism in which an atom (for example, doubly ionized oxygen) is excited by a coincident strong transition from an abundant element (for example, in the case of ionized oxygen, by an ultraviolet transition of helium at 30.378 nm). The cascade of the atom to lower energy levels results in emission lines in the visible. *See* FLUORESCENCE.

Ionized gas also radiates continuously over a broad frequency range in the visible and radio wavelengths. When an electron recombines with an atomic nucleus (usually hydrogen), it radiates away its surplus energy in visible light. This is called bound-free radiation, as it refers to a free electron becoming bound. In the radio, two kinds of radiation mechanism exist. One occurs as the result of a free electron passing near the electric field of a positively charged nucleus, and is called the free-free process. Another is synchrotron radiation, which is generated by fast (near the speed of light) electrons moving in a magnetic field. Although these continuous radiations have no characteristic frequency signatures as does a bound-bound transition, they can be identified through the shape of the continuous spectrum. For example, synchrotron radiation generally is stronger at low frequencies than at high frequencies, whereas the opposite is true for free-free radiation. *See* SYNCHROTRON RADIATION.

Because different gaseous nebulae have different physical conditions, they have different radiation characteristics. Supernova remnants generally display strong radio synchrotron radiation, whereas H II regions and planetary nebulae show only free-free radiation.

Molecules. Molecules are generally excited in two ways, either by collisional excitation with hydrogen molecules or by radiative excitation by continuous infrared light emitted by nearby dust particles. Collisional excitations change the molecules from the ground rotational state to a higher rotational state, whose subsequent decay results in radiation in the millimeter or submillimeter parts of the spectrum. The most commonly seen molecular rotational emission line is the transition of carbon monoxide (CO) from the first excited state ($J = 1$) to the ground state ($J = 0$) at a wavelength of 2.6 mm.

In addition to emission, molecules can be detected through their absorption lines. Against a bright infrared background, the abundance of the CO molecule, for example, can be determined through their transition from the ground state ($v = 0$) to the second excited vibrational state ($v = 2$) at the wavelength of 2.3 μm . *See* MOLECULAR STRUCTURE AND SPECTRA.

Solid particles. The lattice vibrational modes of solids allow the identification of different kinds of solid-state particles in nebulae. The most widely observed spectral signature is the 10- and 18- μm feature of amorphous silicates. In the infrared, a series of emission features at 3.3, 6.2, 7.7, 8.6, and 11.3 μm

corresponding to the stretching and bending modes of aromatic compounds can be seen. The presence of these features suggests that complex organic compounds are common in nebulae. *See* LATTICE VIBRATIONS.

Observing techniques. The observation of gaseous nebulae can be categorized in two areas: imaging and spectroscopy. Imaging observations involve taking a picture of the nebulae at specific colors. Unlike stars, which radiate continuously over all colors, gaseous nebulae radiate discrete colors corresponding to the wavelengths of atomic lines. Images are often obtained using filters that admit light only over a narrow range of colors in order to enhance the quality of the image. Spectroscopic observations are able to separate the colors into fine intervals and are therefore the desired tool to study atomic or molecular line emissions from nebulae.

Because optical astronomy was the first observational technique to develop, observations of gaseous nebulae in the early twentieth century were restricted to the optical parts of the spectrum. Images of nebulae were obtained by photographic techniques, and spectra were obtained with different kinds of spectrometers (for example, with gratings). After the 1980s, photographic plates were replaced by CCDs, which have much higher sensitivity and dynamic range than photographic plates. *See* ASTRONOMICAL IMAGING; CHARGE-COUPLED DEVICES.

Optical spectroscopy of nebulae began in the late nineteenth century and concentrated on the observation of optical recombination lines of hydrogen [for example, the H-alpha ($n = 3-2$) and H-beta ($n = 4-2$) lines] and collisionally excited forbidden lines of oxygen, nitrogen, sulfur, and so forth. Modern optical spectroscopy based on the use of CCDs is capable of very high spectral resolution and can detect thousands of atomic lines in a single spectrum.

In the late 1950s, radio astronomy began to develop and gaseous nebulae were detected through their free-free and synchrotron radiations. High-level recombination lines of hydrogen were also detected in the radio. As radio receivers capable of spectroscopic observations progressed gradually to frequencies higher than 1 GHz (or equivalently wavelengths shorter than 30 cm), molecules such as OH (with characteristic emission at a wavelength of 18 cm), water (at 1.3 cm), and formaldehyde (at 6.2 cm) were detected. By the early 1970s, high-frequency receivers reaching millimeter wavelengths made possible the detection of CO (at 2.6 mm) and many other gaseous molecules. The number of molecular species detected as of 2006 exceeds 130. *See* RADIO ASTRONOMY.

The development of infrared detectors in the late 1960s made possible the detection of solid-state particles. The detection of the 10- μm feature of amorphous silicates in 1969 represents the first positive identification of specific solids in space. The launch of the *Infrared Space Observatory* in 1995 made a significant advance in astronomical infrared spec-

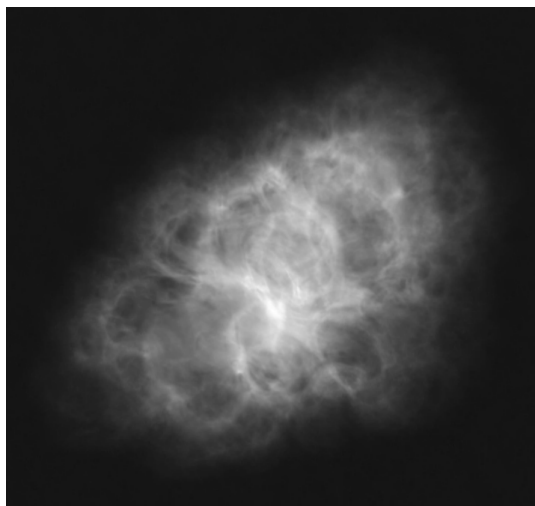
troscopy, allowing the detection of many molecular and solid-state transitions. The detection of crystalline silicates and other refractory oxides led to the new discipline of astromineralogy. The observation of the stretching and bending modes of organic compounds led to the realization that complex organic materials are commonly present in circumstellar and interstellar nebulae. With its high resolving power, the *Spitzer Space Telescope* has obtained detailed images of many nebulae in the infrared, showing their structure and morphology through their dust emission. *See* INFRARED ASTRONOMY; SPITZER SPACE TELESCOPE.

Spectroscopic observations from space-based telescopes in the ultraviolet (for example, the *International Ultraviolet Explorer* and the *Far Ultraviolet Spectroscopic Explorer*) and x-ray (for example, the *Chandra X-ray Observatory*) extend the observations of electronic transitions of atoms and ions to the ultraviolet and x-ray parts of the spectrum. Examples include the 103-nm line of O^{5+} , the 191-nm line of C^{2+} , and the 124-nm line of N^{4+} . *See* CHANDRA X-RAY OBSERVATORY; ULTRAVIOLET ASTRONOMY; X-RAY ASTRONOMY.

Many common neutral atoms in gaseous nebulae radiate only in the far-infrared or submillimeter parts of the spectrum. These transitions arise from the lowest electronic energy states and are called fine-structure lines. For example, the fine-structure transitions of neutral carbon are at 370 and 609 μm , and of neutral oxygen at 52 and 88 μm . These lines probe the neutral states of matter in nebulae and are important for understanding the chemistry of the nebulae. Since the Earth's atmosphere is opaque at far-infrared wavelengths, these transitions can be detected only from space or high-flying aircraft. The 609- μm line of neutral carbon was first detected by the Kuiper Airborne Observatory. Atomic fine-structure lines are extensively studied by the *Infrared Space Observatory*, and high-quality data are expected from the *Herschel Space Observatory* mission. *See* FINE STRUCTURE (SPECTRAL LINES); SUBMILLIMETER ASTRONOMY.

At the radio or millimeter parts of the spectrum, the technique of interferometry is often used to produce an image. Signals from two or more antennas taken over a period of time are combined electronically to form an image. For example, the Very Large Array in New Mexico has twenty-seven 25-m (82-ft) antennas operating at centimeter wavelengths and is capable of obtaining very high angular resolution images of nebulae (see **illustration**). The *Atacama Large Millimeter Array (ALMA)* is expected to produce high-quality images of molecular and dust emissions from interstellar clouds.

Internal dynamics. Nebulae often undergo systematic internal motions. Molecular clouds associated with star formation are mostly collapsing, whereas planetary nebulae and supernova remnants are expanding. Such internal motions can be measured through the Doppler effect, where the atomic and molecular lines shift in wavelength depending on



Crab Nebula, a supernova remnant in the constellation Taurus, imaged in the radio wavelength of 6 cm by the Very Large Array. (Courtesy of NRAO/AUI and M. Bietenholz)

the magnitude of their motion along the line of sight. Supernova remnants are the result of energetic, fast stellar ejecta sweeping up large amounts of interstellar or previously ejected circumstellar materials. Planetary nebulae are formed by the interaction between a fast (2000–4000 km/s or 1250–2500 mi/s) stellar wind with the slow (of the order of 10 km/s or 10 mi/s), previously ejected wind from the progenitor red giant star. The collision of the two winds forms a dense shell that is the main part of a planetary nebula. Behind this shell is a hot, dilute bubble of gas at a temperature of millions of degrees. Thermal pressure from this bubble propels the expansion of the planetary nebula shell. See DOPPLER EFFECT.

Importance of studying nebulae. Astronomers are interested in gaseous nebulae for a number of reasons. Since H II regions and molecular clouds are associated with stellar births, and planetary nebulae and supernova remnants are associated with stellar deaths, they are important elements in understanding stellar evolution. Planetary nebulae and supernova remnants have internal motions that are highly supersonic, making them good laboratories of supersonic dynamics. Since planetary nebulae and supernova remnants carry heavy elements synthesized by stars, they are the vehicles for the chemical enrichment of the Milky Way Galaxy. An analysis of the chemical contents of these nebulae provides observational constraints on theories of stellar nucleosynthesis. The study of atomic transitions, in particular those that do not occur in the terrestrial environment, helps us gain understanding of atomic structures and processes. The study of molecules in star formation regions and planetary nebulae leads to the new discipline of astrochemistry, where theories are being developed on how these molecules form in a low-density environment. See NUCLEOSYNTHESIS; SHOCK WAVE; STELLAR EVOLUTION.

Since the solar system condensed out of an interstellar cloud, human existence is very much the product of the physical and chemical processes in the parent cloud. Interstellar clouds, in turn, are made up of materials ejected from stars through stellar winds, planetary nebulae, and supernovae. The cycle of stars from birth to death through stellar evolution, and then from death to birth again, is closed by the links between stellar ejecta, interstellar clouds, and proto-stellar nebulae. The discovery of counterparts of circumstellar dust in meteoroids has led to the speculation on whether star dust has enriched the chemical makeup of the primordial solar nebula. See COSMOCHEMISTRY; SOLAR SYSTEM.

In summary, the study of gaseous nebulae provides the opportunity to learn about physical and chemical processes under extreme conditions (low density, very low and very high temperatures, strong radiation background, and so forth) not available on Earth. Therefore, nebulae serve as excellent extraterrestrial laboratories for physics and chemistry as well as crucial links in the life cycle of stars. Sun Kwok

Bibliography. L. H. Aller, *Physics of Thermal Gaseous Nebulae*, D. Reidel, Dordrecht, Netherlands, 1984, paper 1987; J. E. Dyson and D. A. Williams, *The Physics of the Interstellar Medium*, 2d ed., Institute of Physics, 1997; S. Kwok, *The Origin and Evolution of Planetary Nebulae*, Cambridge University Press, 2000; S. Kwok, *Physics and Chemistry of the Interstellar Medium*, University Science Books, 2006; D. E. Osterbrock and G. J. Ferland, *Astrophysics of Gaseous Nebulae and Active Galactic Nuclei*, University Science Books, 2005.

Nectarine

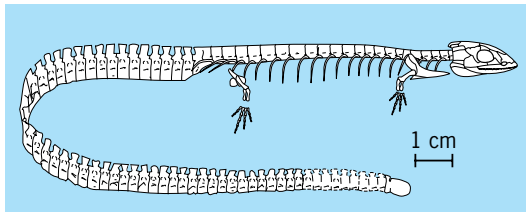
A smooth-skinned, fuzzless form of peach, *Prunus persica*, probably brought to the Middle East from Central Asia by Alexander the Great. The nectarine's lack of pubescence is a simple recessive genetic characteristic. Classically, the fruits were thought of as being somewhat smaller, softer, and richer in flavor than those of the peach. More recently developed cultivars, however, approximate fresh-market peaches in size and firmness but are not usually superior in flavor. Propagation and cultivation are the same as for the peach, but because of its susceptibility to brown rot (*Sclerotinia fructicola*) in humid growing areas, the nectarine is primarily grown in irrigated peach-growing areas where atmospheric humidity is low. See PEACH; ROSALES.

California is practically the sole commercial producer of nectarines. There have been a considerable number of plantings in irrigated areas in south-central Washington. Modern commercial production in California began with the firm, attractive, large-fruited cultivars of the Grand series developed by Fred Anderson, a private breeder. Anderson's first large nectarines were derived from second-generation seedlings produced by crossing nectarines with the large-fruited, firm J. H. Hale

peach. Other breeding programs in California and eastern states have resulted in many new nectarines, and an extended marketing season. See FRUIT; FRUIT, TREE. L. F. Hough; Catherine H. Bailey

Nectridea

An order of mostly aquatic lepospondyl amphibians known from Carboniferous and Permian rocks of North America, Europe, and North Africa. They were small, usually less than 20 in. (50 cm) in length, and outwardly newtlike with short trunks and long tails. Limbs were small but well developed. Carpal and tarsal bones were rarely ossified. Their vertebrae exhibit the one-piece centrum characteristic of lepospondyls, but are distinct in bearing spatulate neural spines (see *illus.*) with crenulated dorsal edges.



Urocordylid nectridean *Ptyonius marshii*. (After R. L. Carroll, *Vertebrate Paleontology and Evolution*, W. H. Freeman, 1988)

Three nectridean families are recognized: Urocordylidae, Keraterpetontidae, and Scincosauridae. Urocordylid skulls are distinctly arrowhead-shaped, usually with long snouts, and possessed cranial kinesis in which the sides of the skull could arc outward slightly as though hinged at the top of the skull. In one form, the snout and upper dentition were able to move up and down.

Urocordylid caudal vertebrae were adapted for swimming, being laterally compressed and bearing spatulate hemal spines below the centrum almost identical to the neural spines above it (giving the vertebrae a dorsoventrally symmetrical appearance; see *illus.*).

The Keraterpetontidae are recognized by the expanded tabular bones of the skull roof, which extend backward like horns. Modest in early forms, these features were exaggerated in *Diplocaulus* and *Diploceraspis*, so that from above the skulls resemble boomerangs. It is thought that the horns had ligamentous connections with the pectoral girdles, stabilizing the head during swimming.

The Scincosauridae seem to have been the only terrestrial nectrideans. Their skulls were short-snouted and lacked tabular horns. Since their tails lacked spatulate hemal spines but possessed ribs, they were shallow and rounded in cross section, not meant for swimming. The skeleton was well ossified, including the carpal and tarsal bones, as might be expected in terrestrial, weight-bearing structures. See AMPHIBIA; LEPOSONDYLI. C. F. Wellstead

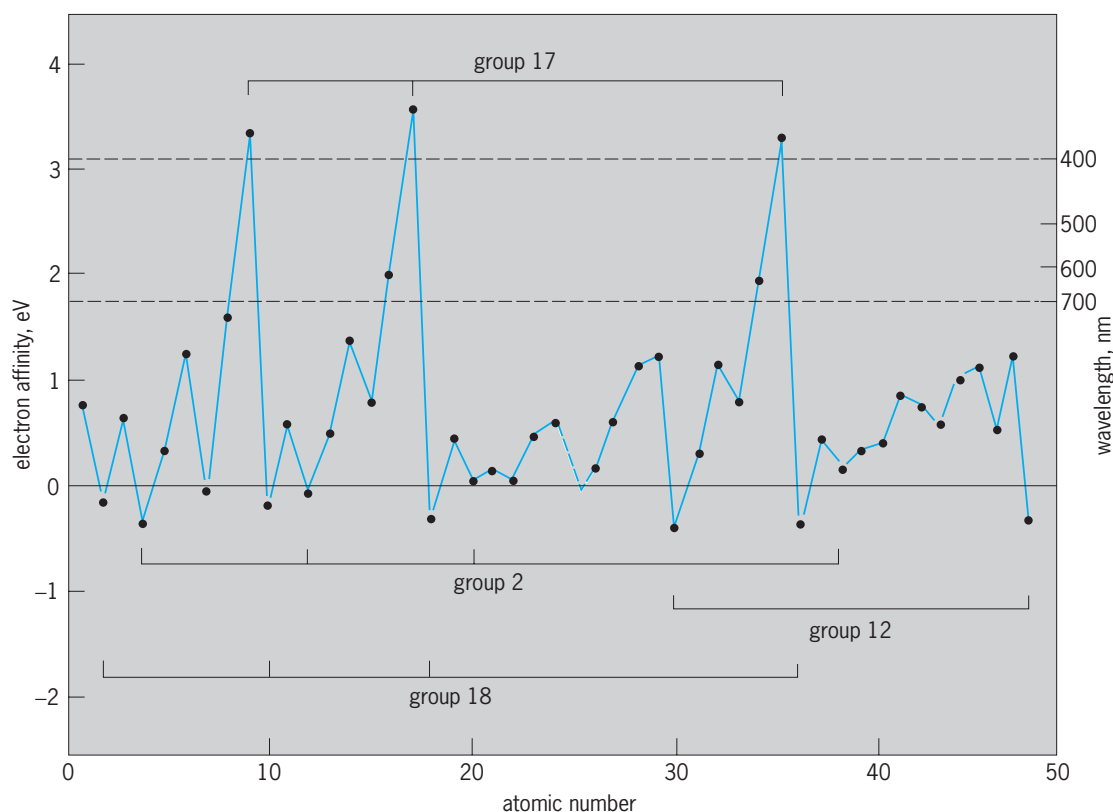
Bibliography. R. L. Carroll, *Vertebrate Paleontology and Evolution*, 1988; A. L. Panchen (ed.), *The Terrestrial Environment and the Origin of Land Vertebrates*, 1980.

Negative ion

An atomic or molecular system having an excess of negative charge. Negative ions, also called anions, are formed in attachment processes in which an additional electron is captured by an atom, molecule, or cluster. They can also be formed when a molecule or cluster dissociates. Doubly charged negative ions, also called dianions, have also been observed in the case of molecules and clusters. Here, two additional electrons have become attached to the neutral systems. Negative ions are destroyed in a controlled manner in detachment processes and, in the case of molecular ions or clusters, dissociation processes, when the ion interacts with photons, electrons, heavy particles, or external fields. Experimental studies of negative ions involve measurements of cross sections for detachment and dissociation. See SCATTERING EXPERIMENTS (ATOMS AND MOLECULES).

Negative ions were first reported in the early days of mass spectrometry. It was soon learned that even a small concentration of such weakly bound, negatively charged systems had an appreciable effect on the electrical conductivity of gaseous discharges. Negative ions now play a major role in a number of areas of physics and chemistry involving weakly ionized gases and plasmas. Applications include accelerator technology, mass spectrometry, injection heating of thermonuclear plasmas, material processing, and the development of tailor-made gaseous dielectrics. In nature, negative ions are known to be present in tenuous plasmas such as those found in astrophysical and aeronomic environments. The absorption of radiation by negative hydrogen ions in the solar photosphere, for example, determines the Sun's spectral distribution. See ION SOURCES; PLASMA (PHYSICS).

A relatively recent application of negative ions is the measurement of the abundance of selected rare isotopes, such as carbon-14 (^{14}C), using accelerator mass spectrometry (AMS). In this sensitive and selective method, a beam of negative ions is injected into a tandem accelerator, where the ions gain energy and lose two or more electrons. The positive ions exiting the accelerator undergo magnetic and electrostatic analysis, allowing one to selectively count the number of ions of a chosen isotope. Negative ions are particularly useful in the case of ^{14}C mass spectrometry since a potential background associated with the nitrogen-14 (^{14}N) isobar is eliminated due to the fact that the element nitrogen does not form stable negative ions. Accelerator mass spectrometry is often used to date archeological samples. The high selectivity and sensitivity of the method compared to those of dating methods involving radioactive decay allows the study of smaller and older samples. See



Electron affinities of the elements in the first half of the periodic table. Wavelengths of radiation at threshold for removing the least tightly bound electron are also shown; horizontal lines indicate threshold wavelengths at limits of the visible region of the electromagnetic spectrum (400 and 700 nanometers).

ACCELERATOR MASS SPECTROMETRY; MASS SPECTROSCOPY; RADIOCARBON DATING.

Electron correlations. Intrinsically, negative ions are of considerable interest since they represent a loosely bound system that is sensitive to electron correlation effects. The interaction between the outermost pair of electrons can become comparable to the relatively weak interaction of each electron with the atomic core. The electrons then drastically affect each other's behavior; that is, their motions are correlated. In contrast, the effects of correlated electron motion in atoms and positive ions are usually masked by the dominant Coulomb forces between the valence electrons and the ionic core. Because of the enhanced role played by electron correlation, a negative ion provides an excellent test bed for investigating departures from the independent electron model.

Structure. The structure of a negative ion differs fundamentally from that of an isoelectronic atom or positive ion. A stable negative ion typically possesses a single bound state in contrast to the infinite spectrum of states characteristic of an atom or positive ion. The force that binds the attached electron in a negative ion arises from polarization and exchange effects (manifestations of electron correlations), and as such it is considerably weaker and shorter in range than the Coulomb force which binds the valence electron to the ionic core in atoms and positive ions. See EXCHANGE INTERACTION.

The binding energy of the least tightly bound electron in a negative ion is numerically equal to the electron affinity of the parent atom. Electron affinities of atoms are at least an order of magnitude smaller than corresponding ionization energies of atoms and positive ions. These affinities range from 0.02 to 3.6 eV and display periodicity with atomic number, which is associated with the Pauli exclusion principle (see **illus.**). See ELECTRON AFFINITY; EXCLUSION PRINCIPLE.

Doubly excited states of negative ions involving either the excitation of a pair of valence electrons or a valence electron and a core electron are quite common. Such transient states are embedded in continua above the detachment limit and decay spontaneously via autodetachment. They are manifested as resonance structures in detachment cross sections. Pioneering studies of doubly excited resonant states of the simplest negative ion, H^- , were carried out by Howard Bryant and coworkers in the 1970s. Since then, resonant states involving valence and core excitation have been investigated in many ions using either laser or synchrotron radiation. See LASER SPECTROSCOPY; RESONANCE (QUANTUM MECHANICS); SYNCHROTRON RADIATION.

Occurrence. More than 80% of the naturally occurring elements have positive electron affinities, and therefore form stable atomic negative ions when an electron is attached to an atom in its ground state. Other elements form metastable ions when

an electron is attached to an atom in a metastable excited state. Such ions are unstable but often sufficiently long-lived for experimental investigation. Many molecules also form negative ions. A common characteristic of all negative ions is their fragility due to the weak binding force. Spectroscopic sources, such as a tenuous beam of accelerated ions passing through a vacuum, are designed to minimize the collisional detachment of electrons while maintaining a density sufficiently high to investigate interactions of photons with negative ions. See ATOMIC STRUCTURE AND SPECTRA; ION; MOLECULAR BEAMS. David J. Pegg Bibliography. L. H. Andersen, T. Andersen, and P. Hvelplund, Studies of negative ions in storage rings, *Adv. At. Mol. Opt. Phys.*, 38:155–191, 1997; T. Andersen, Atomic negative ions, *Phys. Rep.*, 394:157–313, 2004; T. Andersen, H. K. Haugen, and H. Hotop, Binding energies in atomic negative ions, III, *J. Phys. Chem. Ref. Data*, 28:1511–1533, 1999; H. S. W. Massey, *Negative Ions*, 3d ed., 1976; D. J. Pegg, The structure and dynamics of negative ions, *Rep. Prog. Phys.*, 67:1–49, 2004.

Negative-resistance circuits

Electronic circuits or devices that, over some range of voltage v and current i , satisfy Eq. (1) for equiv-

$$R_{\text{eq}} = \frac{dv}{di} < 0 \quad (1)$$

alent resistance R_{eq} (where the voltage and current polarities are defined in Fig. 1a). They are used as building blocks in designing circuits for a wide range of applications, including amplifiers, oscillators, and memory elements. See ELECTRICAL RESISTANCE.

An ideal negative resistor would have the voltage-current relationship (transfer characteristic) shown in Fig. 1b, and thus satisfy Ohm's law with a negative value for the resistance. However, the same effect can generally be obtained with any circuit (or physical device) whose voltage-current curve contains a region of negative slope. Figure 1c, for example, shows transfer characteristics typical of a tunnel diode and a neon bulb, which can be operated in the negative-resistance regions indicated. See OHM'S LAW; RESISTOR.

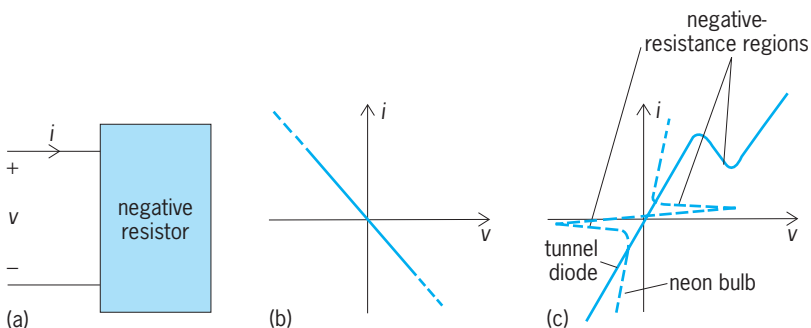


Fig. 1. Characteristics of negative resistors. (a) Definition of voltage (v) and current (i) polarities. (b) Voltage-current transfer characteristic of an ideal negative resistor. (c) Transfer characteristic of practical physical devices with negative-resistance regions: a tunnel diode and a neon bulb (not to the same scale).

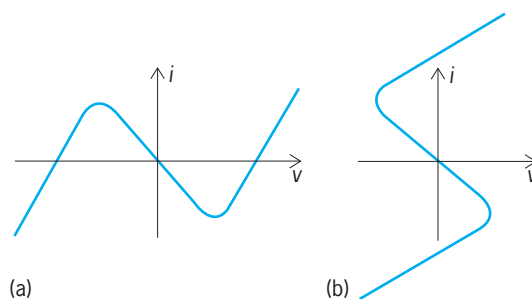


Fig. 2. Large-signal behavior of a negative resistance having a finite internal power supply. (a) Voltage-controlled resistance. (b) Current-controlled resistance.

Common generalizations of the negative-resistance idea include negative capacitors, negative inductors, and frequency-dependent negative resistors. Some of the circuits used to implement them are negative impedance converters, negative impedance inverters, and generalized immittance converters. This article develops these ideas together, starting with properties inherent to the idea of negative impedance, and then discussing theoretical and practical implementations. See ELECTRICAL IMPEDANCE; IMMITTANCE; INDUCTOR.

Properties of negative resistors. The power dissipated in a device, given by Eq. (2), is negative in

$$P_{\text{DISS}} = vi \quad (2)$$

the second and fourth quadrants of the v - i plane of Figs. 1b and c. Thus, the ideal negative resistor whose characteristic is shown in Fig. 1b generates power. Two consequences of this are that an active circuit (a circuit containing a power supply) is required to implement the ideal characteristic of Fig. 1b but is not necessary for the small-signal negative resistances of Fig. 1c; and that for any practical circuit, the characteristic curve must eventually fold over into the power-dissipating quadrants, as shown in Fig. 2a or b. If the curve did not fold but just continued forever, it would be possible to extract an infinite amount of power from the device.

The two types of curve of Fig. 2 correspond to an important dichotomy in types of negative resistance. The N-shaped curve of Fig. 2a allows current to be a single-valued function of voltage (but not vice versa), and circuits with this behavior are therefore called voltage-controlled negative resistors. Dually, the S-shaped curve of Fig. 2b, for which Eq. (3) is

$$v = f(i) \quad (3)$$

appropriate, describes a current-controlled negative resistor. The tunnel-diode characteristic of Fig. 1c can be seen to be voltage-controlled, while the neon tube is current-controlled.

If the terminals of a current-controlled negative resistor are open-circuited, then $i = 0$ and there is a unique solution $v = f(0)$. The voltage-controlled circuit, however, can have any of three voltages in this situation (the three intersections of the N with the horizontal axis). Dually, the S-curve gives a

device with multiple equilibrium states when short-circuited. When the dynamic behavior of these circuits is accounted for, it is found that some of these equilibria are stable and some are unstable. These stability considerations are essential to designing a negative-resistance circuit for a particular application.

Dynamic behavior of negative resistors. In order to analyze the dynamic behavior of a negative resistor—for instance, to determine which equilibria are stable and which are unstable—the remaining circuit elements in which the negative resistor is embedded, including relevant parasitic capacitors and resistors, are added to the device. If an ideal negative resistor R were placed in parallel with a positive capacitance C , the result would be unstable: for some initial capacitor voltage v_0 , the current i would be directed into the capacitor, whose voltage would increase at a rate given by Eq. (4). The differential equation has the unstable solution given by Eq. (5), where $\tau = -CR$

$$C \frac{dv}{dt} = i \quad (4)$$

$$v = v_0 e^{t/\tau} \quad (5)$$

is positive because $R < 0$, $C > 0$. Similarly, a loop of a negative resistor and a positive inductor is unstable; its current is given by Eq. (6), with $\tau = -L/R$.

$$i = i_0 e^{t/\tau} \quad (6)$$

These instabilities would seem to make the negative-resistance circuit useless by itself: a negative resistor left open-circuited would have some small (positive) stray capacitance between its terminals and hence be unstable; and when short-circuited it would be unstable because of the finite (positive) inductance of any short circuit. Practical applications of negative-resistance circuits involve embedding the negative-resistance elements (circuits) in larger circuits, and it is the stability of these larger circuits, not the stability of the negative resistance by itself, that is of interest.

In some applications of negative resistances, direct stabilization of the negative resistance is required. One method of achieving this is with the utilization of negative inductors or capacitors, which are synthesized in a manner similar to that used to synthesize the negative resistors themselves. For example, a negative resistance in parallel with a negative capacitance is stable.

This behavior interacts subtly with the large-signal (S or N) behavior. If a negative resistor is paralleled with a negative capacitance, then it will be stable when open-circuited. This is desirable for a device whose large-signal state behavior is S-shaped (Fig. 2b), because then the (unique) equilibrium for $i = 0$ will be stable. Dually, a negative resistor that is to be short-circuit-stable should contain a parasitic series negative inductance and have N-shaped large-signal behavior.

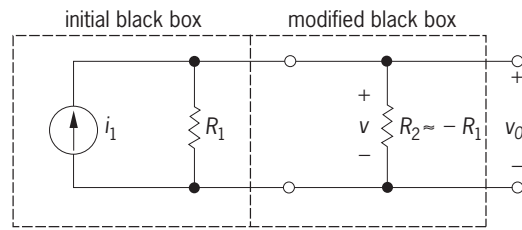


Fig. 3. Application of a negative resistance to cancel the effect of a load, thereby making a kind of amplifier.

Applications. Common applications of negative resistors include amplifiers, bistable or memory circuits, and oscillators.

Amplifier design. A common design problem is shown in Fig. 3, in which the output impedance of a current source is positive and unacceptably small. (This is a simplification of a situation in which negative resistances have been used for telephone system and integrated-circuit amplifier designs.) A large open circuit voltage is desired at the output, but on its own the system produces an inadequately small voltage $i_1 R_1$. It is further assumed that the only access available is to the terminals outside the “black box.” By placing a negative resistance $R_2 = -R_1$ in parallel with the built-in load, it is possible to cancel its effect: the net parallel resistance is given by Eq. (7),

$$R_{\text{eff}} = \frac{R_1 R_2}{R_1 + R_2} = \infty \quad (7)$$

which would theoretically allow an infinite output voltage.

In practice R_1 cannot be made exactly equal to $-R_2$. Practically, $|R_2|$ is intentionally made slightly larger or slightly smaller than R_1 so that the overall circuit, including any additional elements and parasitic components, remains stable. See AMPLIFIER.

Bistable circuits. Figure 4a shows a negative resistor in a simple memory circuit, and Fig. 4b shows the N-shaped characteristic of the resistor. When $i_1 = 0$, this characteristic allows three equilibria. If C is positive, the central equilibrium will be unstable; the two outer equilibria will, however, be stable, because $C > 0$ and the positive slope dv/di acts dynamically like a positive resistor. The resulting circuit is called a bistable circuit, because it is stable in either of two states, and it stores one bit of data by “remembering” its current state.

The circuit can be switched from one state to the other by pulsing i_1 . The broken lines on the diagram are intended to suggest that the circuit switches from one side to the other when forced to change states.

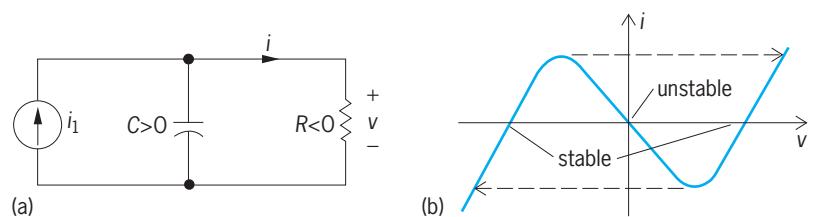


Fig. 4. Negative resistor used as a memory circuit (or bistable circuit or latch). (a) Circuit. (b) Resistor characteristic.

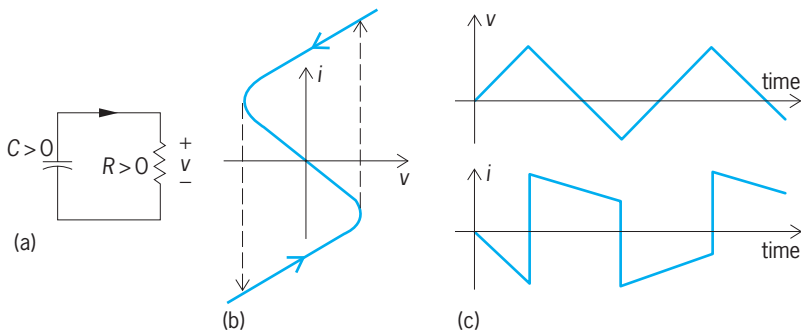


Fig. 5. Negative-resistance oscillator (or astable circuit). (a) Circuit. (b) Resistor characteristic. (c) Voltage and current waveforms.

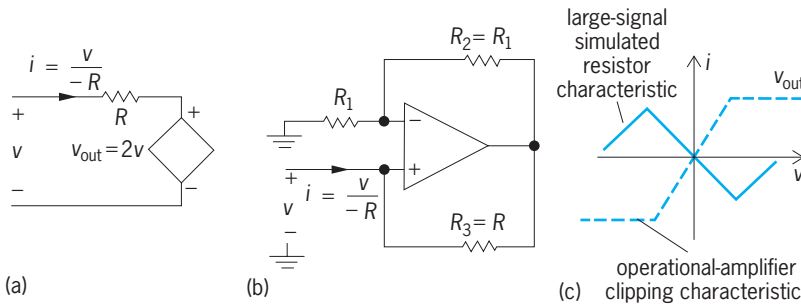


Fig. 6. Active circuits that simulate a negative resistance. (a) Circuit that uses an ideal voltage amplifier with a gain of 2. (b) Circuit that uses an operational amplifier. (c) The operational-amplifier clipping characteristic and resulting large-signal voltage-current characteristic of the simulated negative resistor.

The behavior in the transition is defined by the unstable time constant RC .

Oscillators. An astable, or oscillator, circuit (Fig. 5a) is obtained by replacing the voltage-controlled resistor of Fig. 4 with a current-controlled device (Fig. 5b). The waveforms of voltage and current obtained in this case are shown in Fig. 5c. A second type of oscillator, which produces sinusoidal signals, can be obtained by using a negative resistor to cancel the natural (positive) resistances that damp out oscillations in tank circuits (inductor-capacitor loops) and crystals. See OSCILLATOR.

Implementations of negative resistors. Negative resistors can be implemented by using amplifiers in positive-feedback configurations. Figure 6a shows how a voltage amplifier with a gain of 2 can be used to simulate a grounded negative resistor, and Fig. 6b shows an operational-amplifier implementation of the same idea. See OPERATIONAL AMPLIFIER.

In the practical case of a clipping amplifier, which has the input-output characteristics shown in Fig. 6c, the resulting large-signal voltage-current behavior of the simulated resistor is as shown in the figure. This is a voltage-controlled resistor.

Again in the practical case, the amplifier has a finite frequency response. If the gain of the amplifier of Fig. 6a drops with frequency ω as in Eq. (8)

$$T(j\omega) = \frac{2}{1 + j(\omega/\omega_0)} \quad (8)$$

[where, in electronics, $j = \sqrt{-1}$], then a negative parasitic inductance is simulated at the same time as

the negative resistor, which makes the implementation short-circuit-stable. See GAIN; RESPONSE.

Figure 6 shows one of the four possible types of controlled source (voltage or current sources, controlled by voltages or currents) being used to create a negative resistor; all four types can be used. Figure 7, for example, shows an implementation based on an ideal voltage-controlled current source, together with a practical transistor implementation. Transistor M1 can be modeled as a voltage-controlled current source but produces an output current of the wrong sign. Transistors M2 and M3 form a current mirror and correct the sign. This circuit, like that of Fig. 6, produces an N-shaped (voltage-controlled) characteristic with a series negative inductor, and so is short-circuit-stable. See CURRENT SOURCES AND MIRRORS; TRANSISTOR.

Dual circuits (circuits in which currents and voltages are interchanged) to those of Fig. 6a and 7a produce open-circuit-stable negative resistors with S-shaped (current-controlled) voltage-current characteristics. See CIRCUIT (ELECTRONICS).

Negative-impedance converters. The circuit of Fig. 6b is called a negative-impedance converter (NIC). Analysis with arbitrary resistances R_1 , R_2 , and R_3 and an ideal operational amplifier shows that it has input impedance $-R_1R_3/R_2$. By replacing some resistors with capacitors, the circuit can be generalized to make several different devices, including (1) a negative capacitor, obtained by replacing R_1 or R_3 with a capacitor; (2) a negative inductor, obtained by replacing R_2 with a capacitor; (3) a positive inductor, obtained by replacing R_1 with a negative inductor (which, in turn, can be implemented with a second negative-impedance converter, as before); (4) an impedance that varies with frequency ω as $1/\omega^2$, obtained by replacing resistors R_1 and R_3 with capacitors; and (5) a related circuit, the frequency-dependent negative resistor, obtained by replacing resistor R_3 with a capacitor and resistor R_1 with a negative capacitor (which, in turn, may be implemented with a second negative-impedance converter).

All of these impedances are of the short-circuit-stable type, as can be verified by noticing that in each case the positive-feedback path has less gain than the negative-feedback (stabilizing) path when

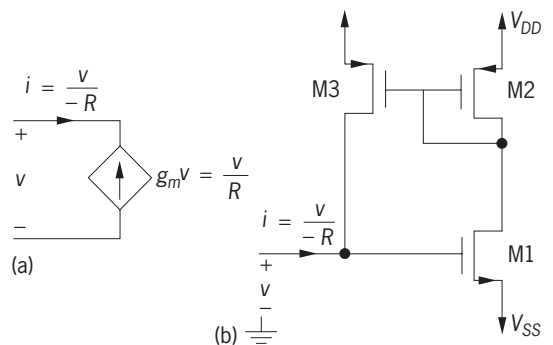


Fig. 7. Ideal negative resistor implemented with an ideal voltage-controlled current source. (a) Ideal circuit. (b) Practical MOSFET implementation.

the input is grounded. Interestingly, the circuit is also a negative-impedance converter when the amplifier's positive and negative input terminals are interchanged. In this case it still has input impedance $-R_1R_3/R_2$, but it becomes open-circuit-stable. This one circuit, then, can implement negative resistors, frequency-dependent negative resistors, negative capacitors, and negative inductors of either the short-circuit-stable or open-circuit-stable types.

The transistor circuit of Fig. 7*b* can also be used as a negative-impedance converter by including a positive impedance in the source lead of M1 and adding a diode-connected transistor M4 (one whose drain and gate are connected) in series with the input. The resulting circuit can be driven from the source of either M1 or M2, and accordingly either a voltage-controlled or a current-controlled negative resistor is obtained. This circuit is also known as a current conveyor and is the basis for the precision "band-gap" voltage references widely used in analog integrated circuit design wherever a temperature-insensitive voltage level is required. See INTEGRATED CIRCUITS.

Physical devices. The positive feedback accomplished by circuitry in Figs. 6 and 7 can occur internally in a physical device and so lead to natural negative resistance characteristics like those of Fig. 1*c* over some region of operation. The characteristic is necessarily passive (that is, it cannot generate power and is therefore restricted to the first and third quadrants of the voltage-current plane). These characteristics are sometimes said to show differential negative resistance, since they have negative derivatives.

As a simple example, if a current flows in a neon bulb, it tends to ionize the gas. This, in turn, creates carriers for current and so decreases the resistance of the bulb. This, in turn, leads to increased current, completing a positive-feedback loop. This behavior produces a characteristic with the negative-resistance region shown in Fig. 1*c*. It is of the current-controlled type, since, as for any given current, only one level of ionization exists, whereas at a given voltage the device may be either "on" or "off." A common hobbyist circuit implements an oscillator by placing a capacitor in parallel with a neon bulb, then charging the capacitor from a high voltage through a large resistor. This is essentially the astable circuit of Fig. 5, augmented with a biasing circuit (the voltage source and resistor) to force the circuit to operate in the negative-resistance region. See ELECTRICAL CONDUCTION IN GASES; VAPOR LAMP.

Many other devices have negative-resistance regions in their voltage-current characteristics. Although the mechanisms involved are much less straightforward to understand than those for the neon bulb, these devices are more useful for practical circuit design.

Tunnel diode. The best-known negative-resistance device is the tunnel diode, which was discovered by L. Esaki in 1958. It is very useful because the phenomenon that it exploits is a quantum-mechanical effect that happens much more rapidly than most others in electronics.

A tunnel diode consists of two very heavily doped regions of a semiconducting material with a very abrupt junction between them. These regions, like any crystalline material, can contain electrons only with energies in certain bands. One side of the junction is doped to have a generous supply of electrons in a certain band of energies, while the other side has a great many vacancies (holes) for electrons in another band. By applying an external voltage, these bands are shifted relative to one another. For small input voltages, these bands overlap, and the quantum-mechanical effect known as tunneling can take place. In this effect, the electrons on one side of the junction (where there is a good supply) have a finite probability of reappearing on the other side (where these are many vacancies). Thus, for small applied voltages, a large current flows. This is still ordinary positive-resistance behavior, although the mechanism for producing current flow is quite different from the field-induced drift of particles that provides conduction in ordinary devices. See BAND THEORY OF SOLIDS; SEMICONDUCTOR.

As the applied voltage increases, however, the bands of electrons and holes on the two sides of the junction start to slide past one another, and eventually their region of overlap starts to decrease. Since quantum tunneling can occur only from an electron in the "supply" to a vacancy at the same energy, this reduction in overlap reduces the amount of charge flowing. Thus, an increasing voltage produces decreasing current, for a negative differential resistance like that shown in Fig. 1*c*.

Since the extent of overlap of the bands is set by the voltage applied to the junction, this produces a voltage-controlled characteristic. Since the time constants associated with tunneling (set by Heisenberg's uncertainty principle) are very short, the phenomenon is very fast. Circuits using these devices operate in the millimeter-wave region (that is, at radio frequencies so high that wavelengths are expressed in millimeters). See UNCERTAINTY PRINCIPLE.

A number of other quantum electronic devices have been developed that also have negative-resistance characteristics. In particular, devices have been constructed that have two barriers (instead of the single barrier created by the tunnel-diode junction) and make use of resonant tunneling, where the spacing between the barriers creates a resonance for electrons at certain frequencies. This resonance, in turn, enhances the rate of tunneling. These devices are claimed to be useful at terahertz (10^{12} Hz) frequencies. See SEMICONDUCTOR HETEROSTRUCTURES; TUNNEL DIODE; TUNNELING IN SOLIDS.

Transferred-electron devices. These are made of materials (*n*-type gallium arsenide is an important example) that have a negative-slope region in their mobility curves. These curves measure electron velocity as a function of electric field, and so correspond at the level of the material to voltage-current curves in the overall device. The negative-slope region comes about because electrons traveling fast enough to exceed a certain energy threshold can transfer to a second mode of propagation, in which their interactions

with the crystal make them appear heavier. As the field increases beyond this point, electrons start to slow down (as more enter the slower mode) and current decreases—a negative-resistance effect.

These devices do not appear directly as negative resistors at their terminals, however, but as oscillators. This is known as the Gunn effect and comes from the fact that electric field distribution in this type of material is inherently unstable. Rather than the field strength and the density of electrons being uniform over the length of the device, part of it operates below the negative-resistance region at low fields and high electron speeds, and another part acts as a bottleneck or domain with a high field and a large number of slower-moving electrons. This domain moves through the material from cathode to anode and disappears as a new domain forms at the cathode. These devices thus oscillate at a frequency set by the time taken for domains to travel from one end to the other.

This type of instability can be thought of in circuit terms by regarding the device as a series connection of smaller devices, with a positive capacitance to ground (the self-capacitance of a piece of material) at each connection. At these nodes there is a positive capacitance and a negative (differential) resistance, which is an unstable combination. See MICROWAVE SOLID-STATE DEVICES.

Other devices. Other devices whose physics lead to negative resistance are the point-contact transistor and the unijunction transistor, but these devices are no longer readily available or widely used. The thyristor (or silicon-controlled rectifier), which is a four-layer transistor structure consisting of alternating *p*- and *n*-type regions, is very common in high-power circuits and can be regarded as a positive-feedback connection of overlapping *npn* and *pnp* transistors. It has an electrical latching behavior rather like that of the neon bulb discussed above, and will switch from an “off” state to “on” if a certain threshold is exceeded. Thyristors are usually used as three-terminal devices, however, and so they are not generally considered to be negative resistors. See SEMICONDUCTOR RECTIFIER.

Martin Snelgrove

Bibliography. L. O. Chua, C. A. Desoer, and E. S. Kuh, *Linear and Nonlinear Circuits*, 1987; A. S. Sedra and K. C. Smith, *Microelectronic Circuits*, 4th ed., 1997; S. M. Sze, *Physics of Semiconductor Devices*, 2d ed., 1981.

Negative temperature

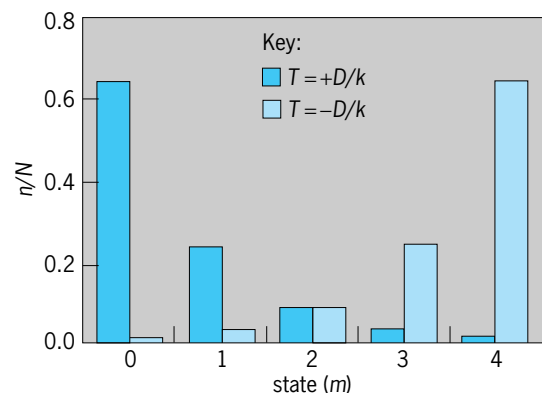
The property of a thermodynamical system which satisfies certain conditions and whose thermodynamically defined absolute temperature is negative. The essential requirements for a thermodynamical system to be capable of negative absolute temperature are that (1) the elements of the thermodynamical system must be in thermodynamical equilibrium among themselves in order for the system to be described by a temperature at all; (2) there must be an upper limit to the possible energy of the allowed

states of the system; and (3) the system must be thermally isolated from all systems which do not satisfy both of the first two requirements, that is, the internal thermal equilibrium time among the elements of the system must be short compared to the time during which appreciable energy is lost to or gained from other systems.

The second condition must be satisfied if negative temperatures are to be achieved with a finite energy. If W_m is the energy of the m th state for one element of the system, then in thermal equilibrium the number of elements in the m th state is proportional to the Boltzmann factor $\exp(-W_m/kT)$, where k is Boltzmann's constant and T is the absolute temperature. For negative temperatures, the Boltzmann factor increases exponentially with increasing W_m , and the high-energy states are therefore occupied more than the low-energy ones; this situation is the reverse of the positive-temperature case. As a consequence, with no upper limit to the energy, negative temperatures could not be achieved with a finite energy. Most systems do not satisfy this condition; for example, there is no upper limit to the possible kinetic energy of a gas molecule. It is for this reason that systems of negative temperatures occur only rarely. See KINETIC THEORY OF MATTER; STATISTICAL MECHANICS.

Systems of interacting nuclear spins, however, have the characteristic that under suitable circumstances they can satisfy all three of the above conditions, in which case the nuclear spin system can be at negative absolute temperature, as first demonstrated by R. V. Pound, E. M. Purcell, and N. F. Ramsey and discussed theoretically by Ramsey and others. See NUCLEAR ORIENTATION.

The difference between positive and negative absolute temperatures can be illustrated with a bar chart showing the population distributions among the different states of a system at positive and negative absolute temperatures (see *illus.*). A simple example is provided by a thermodynamic system consisting of N magnetically mutually interacting nuclei of spin 2, so that each spin can be in any one of five equally spaced energy states of energy $W = mD$,



Populations (n) of states of a thermodynamic system consisting of N magnetically interacting nuclei of spin 2 at absolute temperatures $T = +D/k$ and $T = -D/k$, where k is Boltzmann's constant and D is the energy spacing between successive states.

where m is an integer designating the state and $m = 0, 1, 2, 3,$ or 4 . The number, n , of nuclei in the different states can then be shown in the bar chart at positive and negative absolute temperatures, such as $T = +D/k$ and $T = -D/k$. The bars decrease in length with increasing values of m at positive absolute temperature, and increase in length with increasing values of m at negative absolute temperature.

From the point of view of thermodynamics, the absolute temperature T is given by the equation below,

$$T = \left(\frac{\partial S}{\partial U} \right)_B^{-1}$$

where S is the entropy, U the internal energy, and B the externally applied magnetic field. A system at negative absolute temperature must therefore be in a condition such that the slope of S as a function of U is negative. This is clearly achievable for a spin system in an external field B , since the entropy S falls to a low value when all the nuclear spins are in their highest energy state and thereby all pointing in the same direction. See ENTROPY; THERMODYNAMIC PRINCIPLES.

It is apparent either from the Boltzmann factor or from the above equation applied to a spin system that the transition between positive and negative temperatures is through infinite temperature, not absolute zero; negative absolute temperatures should therefore not be thought of as colder than absolute zero, but as hotter than infinite temperature. See ABSOLUTE ZERO; TEMPERATURE.

Norman F. Ramsey

Bibliography. A. Abragam, *Nuclear Magnetism*, 1983; B. Gal-Or, *Modern Developments in Thermodynamics*, 1974; E. M. Purcell and R. V. Pound, A nuclear spin system at negative temperature, *Phys. Rev.*, 81:279, 1951; N. F. Ramsey, Thermodynamics and statistical mechanics at negative absolute temperatures, *Phys. Rev.*, 103:20-28, 1956; N. F. Ramsey and R. V. Pound, Nuclear audiofrequency spectroscopy by resonant heating of nuclear spin system, *Phys. Rev.*, 81:278, 1951.

Nemata (Nematoda)

A group of unsegmented worms which have been variously recognized as an order, class, and phylum. They are now widely accepted as a separate phylum, Nemata (or Nematoda). When treated as a class, they were assigned either to the phylum Aschelminthes or to the phylum Nematelminthes. In Aschelminthes they were ranked equally with Rotifera (Rotatoria), Gastrotricha, Kinorhyncha (Echinodera), and Nematomorpha (gordian or horsehair worms).

The original name applied to the group was Nematodea (from the Greek, threadlike). This group included only the animal-parasitic taxa. When the group was expanded to include marine and freshwater free-living forms, and also plant-parasitic forms, the concept of a phylum was applied and the

name Nematodes proposed. B. G. Chitwood amended the spelling to Nemata.

All taxonomic groupings above the species level are subjective and therefore in nearly constant flux. The primary reason for disagreement that surrounds nematodes and their allies is the concept of the pseudocoelom (body cavity) and its embryological formation. Pseudocoelom refers to a body cavity only partially lined with mesoderm. How this partially lined cavity is formed during embryological development is the key question. If among the various allies of nematodes the formation is homologous, all should probably be considered under one phylum. If, however, as the available information indicates, the pseudocoelom is not formed homologously in these near groups, each must be treated as a separate phylum; this concept is now widely accepted and is adhered to here. A classification of nematodes follows; see separate articles on each order.

Phylum Nemata

Class: Adenophorea

Subclass: Enoplia

Order: Enoplida

Superfamily: Enoploidea

Oxystominoidea

Order: Oncholaimida

Tripylida

Superfamily: Tripyloidea

Ironoidea

Order: Isolaimida

Mononchida

Superfamily: Mononchoidea

Bathyodontoidea

Mononchuloidea

Order: Dorylaimida

Superfamily: Dorylaimoidea

Actinolaimoidea

Belondiroidea

Encholaimoidea

Diphtherophoroidea

Trichodoroidea

Nygalaimoidea

Order: Stichosomida

Superfamily: Trichocephaloidea

Mermithoidea

Echinomermelloidea

Subclass: Chromadoria

Order: Araeolaimida

Superfamily: Araeolaimoidea

Axonolaimoidea

Plectoidea

Camacolaimoidea

Tripyloidoidea

Order: Chromadorida

Superfamily: Chromadoroidea

Cyatholaimoidea

Choanilaimoidea

Comesomatoidea

Order: Desmoscolecida

Superfamily: Desmoscolecoida

Greeffielloidea

Order: Desmodorida

- Superfamily: Desmodoroidea
 - Ceramonematoidea
 - Monoposthoidea
 - Draconematoidea
 - Epsilonematoidea
- Order: Monhysterida
 - Superfamily: Monhysteroidea
 - Linhomoeoidea
 - Siphonolaimoidea
- Class: Secernentea
 - Subclass: Rhabditia
 - Order: Rhabditida
 - Superfamily: Rhabditoidea
 - Alloionematoidea
 - Bunonematoidea
 - Cephaloboidea
 - Panagrolaimoidea
 - Robertioidea
 - Chambersielloidea
 - Elaphonematoidea
 - Order: Strongylida
 - Superfamily: Strongyloidea
 - Diaphanocephaloidea
 - Ancylostomatoidea
 - Trichostrongyloidea
 - Metastrongyloidea
 - Cosmoceroidea
 - Oxyuroidea
 - Heterakoidea
 - Subclass: Spiruria
 - Order: Spirurida
 - Superfamily: Spiruroidea
 - Physalopteroidea
 - Filarioidea
 - Drilonematoidea
 - Order: Ascaridida
 - Superfamily: Ascaridoidea
 - Camallanoidea
 - Dracunculoidea
 - Subuluroidea
 - Seuratoidea
 - Diectophymatoidea
 - Muspiceoidea
 - Subclass: Diplogasteria
 - Order: Diplogasterida
 - Superfamily: Diplogasteroidea
 - Cylindrocorporoidea
 - Order: Tylenchida
 - Superfamily: Tylenchoidea
 - Criconematoidea
 - Sphaerularoidea
 - Aphelenchoidea
 - Aphelenchoideoidea

The Nematoda are unsegmented or pseudosegmented (any superficial annulation is limited to the cuticle) bilaterally symmetrical worms with a basically circular cross section (cylindroid; **Fig. 1**). The body is covered by a noncellular cuticle secreted by an underlying epidermis (hypodermis). The cylindrical body is usually bluntly rounded anteriorly and tapering posteriorly. The body cannot be easily divided into head, neck, and trunk or tail, although a re-

gion posterior to the anus is generally referred to as the tail. The anterior extremity characteristically bears 16 setiform or papilliform sensory organs and 2 chemoreceptors called amphids. The oral opening is terminal (rarely subterminal) and followed by the stoma, esophagus, intestine, and rectum, which opens through a subterminal anus. Females have separate genital and digestive tract openings. In males the tubular reproductive system joins posteriorly with the digestive tract to form a cloaca. Adult nematodes are extremely variable in size, ranging from less than 0.3 mm (0.012 in.) to over 8 m (26 ft). Nematodes are generally colorless except for food in the intestinal tract or for those few species which have eyespots. A unique character in the phylum is protoplasmic extension from the somatic muscles that reaches for synapsis with the central nervous system, rather than axonic extension of the nervous system to the muscles as in other animals (**Fig. 2**).

External covering. Nematodes are covered by a noncellular elastic cuticle chiefly composed of scleroproteins (not chitin). Chitin is known only from the eggshell. The cuticle is a complicated histologic structure composed of several layers basically divisible into four strata: epicuticle, exocuticle, mesocuticle, and endocuticle. Underlying the cuticle is the hypodermis (epidermis). The hypodermal cells may be uninucleate (all Adenophorea) or multinucleate (some Secernentea). The cell bodies of the hypodermis protrude into the body cavity as four longitudinal chords dorsally, laterally, and ventrally between the somatic muscle bands. The protoplasmic portion of the hypodermis extends as a thin layer between the cuticle and the muscle sheath. The cuticle may be smooth or marked by transverse striae forming a pseudoannulation, or it may be thickened and sculptured in various patterns. Other modifications that occur are alae or wing areas found laterally on the head, along the main body, or on the tail of males.

Alimentary canal. The oral opening is terminal and may be surrounded by three or six lips, or paired pseudolabia, or jaws. The oral opening leads to the mouth cavity, or stoma. The stoma is generally composed of two parts: the anterior cheilostome (formed embryologically from external cuticle) and the esophastome (modified from the cuticular lining of the anterior extremity of the esophagus). The stoma (or stomodeum), combined from both parts, takes on a variety of forms according to feeding habits. Predaceous nematodes have stomas armed with movable or nonmovable teeth or spears. Bacterial feeders generally have cylindrical stomas without teeth. Mycetophagous and phytophagous nematodes generally possess protrusible hollow axial spears. Following the stoma is the cuticularly lined esophagus. The lumen of the esophagus is triradial, with one ray directed ventrally and two subdorsally. The esophagus is composed of radial muscles, epidermal tissue, a complex esophagosympathetic nervous system, and three to five "salivary" glands. The esophagus is variously modified into sections by means of bulbs and valves, permitting suction and ingestion of food. Between the esophagus and the

intestine is an esophagointestinal valve that prevents food regurgitation. The intestine is composed of a single layer of cells (uni- or multinucleate) that are bordered by microvilli on their interior edge. When visible, this microvilli area is called the brush border or bacillary layer. In some nematodes the posterior intestine may be variable modified and designated as a prerectum or proctodeum. At the junction of the intestine and rectum there may be rectal glands, an intestinorectal valve, or sphincter muscle. The cuticularly lined rectum opens through the ventromedian anus (cloaca in males). Rarely is the anus terminal or subterminal.

Somatic musculature. The somatic musculature consists of four or more longitudinal bands separated by the hypodermal chords. Each band is composed of nonstriated, continuous interlocking spindle-shaped, uninucleate muscle cells. From each cell there extends a protoplasmic arm that synapses with a medial nerve. Circular somatic muscles are absent in Nematoda. Throughout the body there are specialized muscles that operate the jaws or protrusible teeth and spears, or act as protractors and retractors of the male spicules or as sphincters of the female vagina; in addition there are scattered cutaneousoesophageal and cutaneousintestinal muscles.

Nervous system. The major portion of the central nervous system consists of a large circumesophageal commissure (nerve ring) and its attendant anterior and posterior ganglia (lateral, dorsal, and ventral; Fig. 3). Ventrally the nerve ring gives rise to the ventral nerve cord, the major nerve of the body. The ventral nerve is a partly paired, mostly single, ganglionated nerve chain (not distinctly set off) that extends the length of the body. Throughout the body there are commissures that interconnect the ventral nerve cord with the lateral or dorsal nerves. The lateral nerves are primarily sensory, while the dorsal and ventral nerve cords are motor. The anterior sense organs are innervated from six ganglia anterior to the nerve ring, but the amphids are innervated from a ganglion in the lateral ganglia posterior to the nerve ring. Another subcuticular peripheral nerve net has been described for marine nematodes.

Pseudocoel. The body cavity is fluid filled and extends the length of the body between the muscle bands and internal organs. The pseudocoel contains a variety of cells and is lined to a greater or lesser extent by membranes and mesenteries that support the internal organs. The nuclei of these structures are reportedly of limited and fixed numbers. The lining material originates from migratory mesenchymatous tissue during embryogeny; these cells grow together in a network that covers the esophagus, intestine, and gonads, and may delimit the muscle bands and hypodermal chords.

Excretory system. No single type of excretory system characterizes Nematoda, and in many taxa the system is completely absent (Fig. 4). When present in Adenophorea, the system is generally limited to a single ventral cell that opens through an anterior ventromedian pore. In Secernentea the cell has one or more tubules that extend the length of the body, ei-

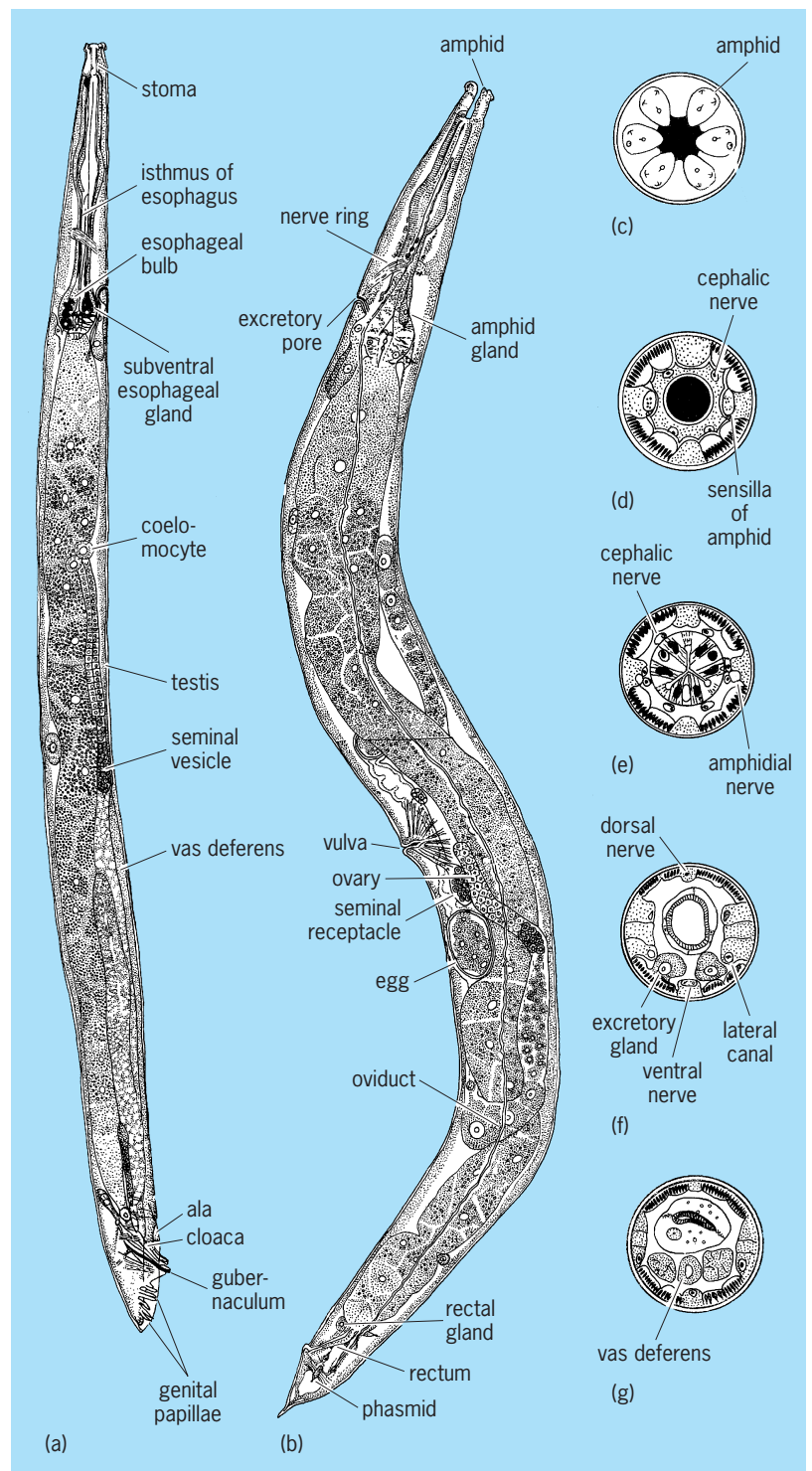


Fig. 1. *Pelodera strongyloides*. (a) Male. (b) Female. (c) Head, cross section. (d) Stomatal region. (e) Esophageal region. (f) Postbulbar region. (g) Posterior part of male.

ther posteriorly or both anteriorly and posteriorly. The duct leading from the cell to the excretory pore, in Secernentea but rarely in Adenophorea, is cuticularly lined and opens ventromedially generally in the esophageal region. In taxa lacking an excretory cell the function, presumably, is taken over by hypodermal glands or even the prerectum.

Reproductive system. The sexes are separate, and the gonads may be single or paired. In a few parasitic

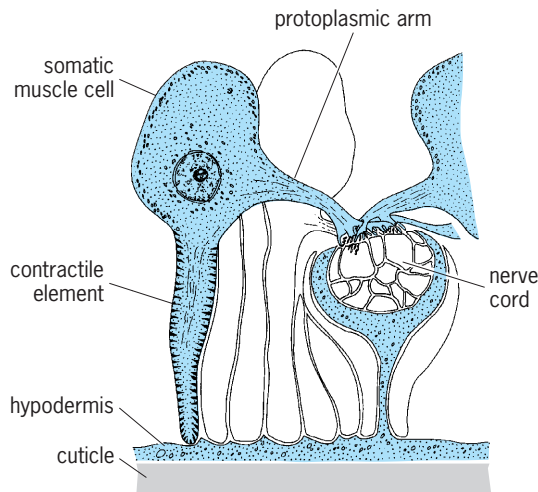


Fig. 2. Cross section of somatic muscle cell with extension to central nervous system.

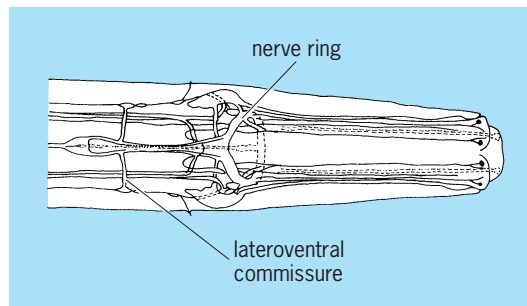


Fig. 3. Nervous system of *Spironoura affinis*.

taxa they may be multiple; as many as 32 ovaries have been reported. Female gonads are tubular, with an apical cell that forms the thin nucleated epithelium that covers the entire gonad. Germ cells originate at the blind end or along the sides of the gonad. The ovary is followed by the oviduct and uterus, which connects to the vagina, which opens on the ventral surface of the body. A spermatheca may be present between the ovary and the uterus or between the uterus and vagina. In some, one gonad may be atrophied and function as a spermatheca. A confirmed hermaphroditic male does occur in one taxon of nematode parasites of insects. Hermaphroditic females are reported but unconfirmed; when reported they are allegedly protandric. See PROTANDRY; PROTOGYN.

The male gonad (generally two in Adenophorea and one in Secernentea) opens posteriorly into the ventral side of the rectum, forming a cloaca. Males generally possess paired cuticular copulatory structures called spicules. The spicules are protrusible and presumably when placed in the vagina aid in the transfer of the sperm, but not in the fashion of a penis or intromittent organ. Sperm varies from flagellate to ameboid.

Females may be oviparous or ovoviviparous; that is, eggs hatch within the body and living young exit. The egg is chitinous and may externally be variously sculptured. Nurse cells are unknown, but the zygote

contains stored food in the form of proteins; and throughout, fatty and glycogenous globules are evident.

Cleavage of the embryo is bilateral but highly determinate, and follows a modified spiral plan. Following full maturation, the first-stage larva may molt once within the egg shell (Secernentea) or emerge as first-stage larvae (Adenophorea). Larvae are similar in form to the adult, lacking a developed reproductive system and other secondary sexual characteristics. All known nematodes undergo four molts between the first stage and the adult. There is no complete reconstitution of internal organs, but there may be changes in the outline of the esophagus. At each molt the cuticle of the body, esophagus, and rectum is shed. Nematodes reportedly lack any power of regeneration. Cell multiplication after hatching is restricted except within the reproductive tract, intestine, hypodermis, and somatic musculature. See CLEAVAGE (DEVELOPMENTAL BIOLOGY).

Adenophorea. In this class the amphids (cephalic chemoreceptors) are postlabial and variable in the shape of the external opening. They may be porelike, slitlike, oval, transversely lenticular, looped (shepherd's crook form), circular, unispiral, or multispiral. The 16 cephalic sensilla may be setiform or papilliform and located postlabially or labially. Somatic setae and hypodermal glands are commonly present, and somatic papillae appear to be universal. The external cuticle consists of all four layers: epi-, exo-, meso-, and endocuticle. The surface of the external cuticle is generally smooth but may have transverse or longitudinal striations; in the subclass Chromadoria it may be elaborately ornate. When present, the excretory system is limited to a single ventral cell lacking collecting tubules. Only rarely is the excretory duct, which opens through a ventromedian pore, cuticularly lined. The esophagus is variable, but is generally either cylindrical or bottle shaped and contains three (Chromadoria) or five (Enoplia) glands. Within the pseudocoel there may be six or more coelomocytes. Rectal glands are rare,

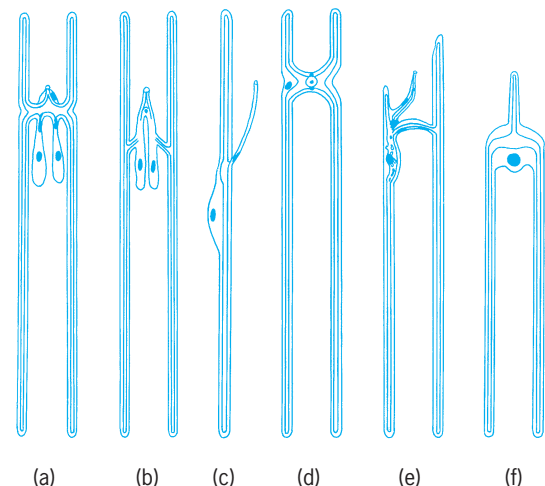


Fig. 4. Types of excretory systems. (a) Rhabditoid type. (b) Variant of rhabditoid type. (c) Tylenchoid type. (d) Ascaroid type. (e) Cephaloboid type (f) Anisakid.

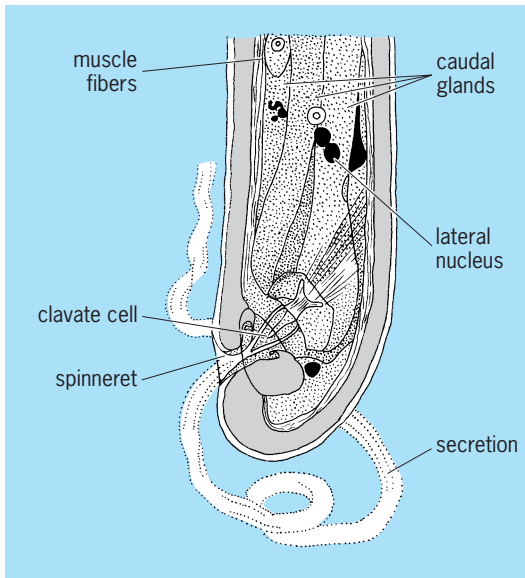


Fig. 5. *Mononchulus ventralis* tail with spinneret details.

but three adhesive caudal glands that open through a terminal or subterminal spinneret are common (Fig. 5). These glands and the spinneret are absent in all obligate parasites. Males, with few exceptions, have two testes, paired spicules, and a single row of ventromedian supplements which may be either papilloid or tuboid. Some males have an additional paired series of supplements. Supplements are often ornate and may function as adhesive organs during copulation. The male tail is generally smooth, and only rarely are lateral cuticular alae reported.

Secernentea. In this class the amphids are most often porelike and located dorsolaterally on the lateral lips or anterior extremity. In a few instances they may be oval or cleftlike, or located postlabially. In all instances the cephalic sensilla are located on the labial region but may be reduced in number, especially among the parasitic forms. On the neck (cervical) region, at the level of the nerve ring, there is usually a pair of sensilla called deirids. Comparable sense organs located on the caudal region are called phasmids (Fig. 1). The external cuticle varies from four to two layers; most Diplogasteria have only the epi- and exocuticle present. The cuticle is almost always transversely striated, and these striations are generally interrupted laterally by longitudinal incisures; these lateral modifications are called the lateral fields. The underlying hypodermis is often multinucleate. Esophagi in the class are variable and diagnostic of each subclass. Rhabditia have a three-part esophagus with a valve in the posterior bulb; Spiruria have a cylindrical or two-part esophagus that lacks bulbs or valves; Diplogasteria have a three-part esophagus, but the valve is anterior in the metacarpus and the posterior bulb is glandular. The three, sometimes more, esophageal glands always open into the corpus. Three to six rectal glands usually open into the anterior part of the rectum or at the junction of the rectum and intestine. The excretory system possesses collecting tubules either paired or

limited to one side of the body. The excretory cell opens ventromedianly through a cuticularized duct. Only four to six coelomocytes are reported within the body cavity; in Rhabditia two such cells become associated with the excretory system in adult worms, and these are erroneously called multiple-celled excretory systems. Somatic setae, papillae, and caudal glands are absent in this class; however, males may have caudal papillae. When present, male preanal supplements are paired and often elaborate, but are considered sensory rather than adhesive. Males often possess paired lateral caudal alae, sometimes called the bursa copulatrix.

Life cycle. Reproduction among nematodes is either amphimictic or parthenogenetic (rarely hermaphroditic). After the completion of oogenesis, by either of the above methods, the chitinous egg shell is formed and a waxy vitelline membrane forms within the egg shell; in some nematodes the uterine cells deposit an additional outermost albuminoid coating. Upon deposition or within the female body, the egg proceeds through embryonation to the eellike first- or second-stage larva, but following eclosion the larva proceeds through four molts to adulthood. This sequence represents a direct life cycle, but among parasites more diversity occurs.

Animal parasites often utilize intermediate hosts (including vectors) for the development of the first three larval stages; the fourth- and adult-stage nematode develops in the definitive host. Alternating generations vacillating between free-living cycles and parasitic cycles also occur among plant and animal parasites. There is one special instance among Tylenchida where the nematode, *Fergusobia curriei*, alternates a gametogenetic generation (parasitic in a gallfly, *Fergusonina tillyardi*) with a parthenogenetic generation parasitic on *Eucalyptus camaldulensis*. Among insect parasites alternation of generations in grossly dissimilar host organisms can occur. Because of the great dissimilarity these cases are only seldom recognized.

Bionomics. Nematoda comprises the third largest phylum of invertebrates (15,000 species), being exceeded only by Mollusca (100,000 species) and Arthropoda (about 850,000 species). In sheer numbers of individuals they exceed all other Metazoa. As parasites of animals they exceed all other helminths combined. There are some 2000 known plant parasites, 4000–5000 animal parasites, and 8000–10,000 marine, fresh-water, and terrestrial free-living species. Nematodes have been recovered from the deepest ocean floors to the highest mountains, from the Arctic to the Antarctic, and in soils as deep as roots can penetrate.

As parasites of plants and animals they are unequalled by any other metazoan group, including insects. It is within the Secernentea that most of the important parasites of domestic animals, humans, plants, and insects occur. The two most important groups for vertebrate parasitism are in the subclasses Rhabditia and Spiruria. Insect and plant parasitism is most important in the subclass Diplogasteria, especially in the order Tylenchida. A minority of

parasites of plants and animals are included in the class Adenophorea.

The aquatic, fresh-water and marine nematodes are in Adenophorea (Enoplia, Chromadoria). Secernenteans are characteristically terrestrial (soil water), and only rarely are they encountered in a fresh-water or marine habitat; and in these instances it is a secondary invasion by normally terrestrial taxa. The nematodes in these habitats are generally referred to as free living, but they do eat products of decay, bacteria, fungi, algae, and small microfauna including nematodes. The role of nematodes in these environments is unknown; it has been estimated that in undisturbed rangelands their biomass exceeds by two to three times other invertebrates near or on the soil surface.

Most nematodes can sustain periods of anaerobic conditions. None has proved to be capable of completing its entire life cycle, including embryonic development, anaerobically. Many nematodes are capable of anhydrobiosis, and have been proved to survive in such a condition for more than 25 years. Glycogen is one of the chief sources of nematode energy, although stored fatty materials apparently provide the chief source of energy in a few groups such as Tylenchida and Aphelenchida. Armand R. Maggenti

Bibliography. B. G. Chitwood, The designation of official names for higher taxa of invertebrates, *Bull. Zool. Nomencl.*, 15(25/28):860-895, 1958; P. P. Grassé, *Traité de Zoologie*, tome 4, fasc. 2 and 3, 1965; L. H. Hyman, *The Invertebrates*, vol. 3, 1951; S. P. Parker (ed.), *Synopsis and Classification of Living Organisms*, 2 vols., 1982.

Nematicide

A type of chemical used to kill plant-parasitic nematodes. Nematicides may be classed as soil fumigants or soil amendments, space fumigants, surface sprays, or dips. Soil treatments are commonly used because most plant-pathogenic species spend part or all of their life cycle in the soil, in or about the roots of plants. Nematicides may be liquids, gases, or solids, but on a field scale, liquids are most practical. These materials possess a high vapor pressure and volatilize quickly to act as soil fumigants. Carbon disulfide, CS₂, and chloropicrin were among the first to be developed for this purpose. Their use is fairly restricted because of high cost; besides, CS₂ is highly flammable and explosive and chloropicrin requires a cover or water seal.

The nematicidal properties of 1,3-dichloropropene and 1,2-dibromoethane were discovered between 1943 and 1945; they are the predominant nematicides in use. These chemicals can be applied to the soil by handgun or machine applicators, and require no surface seals or covers beyond culti-packing or rolling the soil after treatment. Since most nematicides are phytotoxic, they must be used in the absence of growing plants. The nematicide 1,2-dibromo-3-chloropropane (DBCP) was an exception. It was low in phytotoxic properties and could

be used either for preplant or postplant side-dressing treatments. Unfortunately it was found to be a public health hazard due to carcinogenic/mutagenic properties, and was removed from the market in 1976.

Nonfumigant nematicides have been used as replacements for DBCP. These nonfumigants comprise carbamates such as aldicarb [2-methyl-2-(methylthio) propionaldehyde O-(methylcarbamoyl)oxime], or organophosphates such as phenamiphos [ethyl 3-methyl-4-(methylthio) phenyl (1-methylethyl) phosphoramidate]. They are formulated as granular materials with 10-15% active ingredients. Some are also available as emulsifiable or water-soluble formulations. They are low in phytotoxicity and also low in volatility, and move through the soil slowly or not at all. To reach soil nematodes, the materials must be incorporated into the soil mechanically or carried by irrigation water. Some chemicals are systemic in action, move through the roots to the foliage, and have certain efficacy as insecticides as well.

Space fumigants including methyl bromide are important preplant treatments, especially for perennial tree and vine crops. Deep (up to 3 ft or 1 m) placement of methyl bromide sealed by roller-packing gives long protection. For multiple infestations (that is, nematodes plus oak-root fungus), a thin, 1-mil (25-micrometer) polyethylene cover is required. See FUMIGANT.

Methyl bromide is especially useful for sterilizing bags, flats, pots, and other containers or equipment. It has been used to disinfect onion and clover seeds of the stem nematode, *Ditylenchus dipsaci*.

Leaf and bud nematodes of the genus *Aphelenchoides* have been controlled by Parathion and Systox spray applications, but such treatments have very limited usage.

Dip treatments generally employ a combination of hot water and 0.5% formalin solution developed originally for stem nematode in narcissus and Easter lily bulbs and bulbous iris. The addition of a detergent to a similar treatment with 1% formalin has made possible the eradication of the same nematode on garlic cloves.

A slurry of 3-*p*-chlorophenyl-5-methyl-rhodanine has been used for disinfecting rice seed contaminated with nematodes. See NEMATA (NEMATODA); PESTICIDE. Dewey J. Raski

Bibliography. V. H. Dropkin, *Introduction to Plant Nematology*, 2d ed., 1989; E. W. Flick, *Agricultural Chemical Products*, 1989; W. T. Thomson, *Agricultural Chemicals*, Book I: *Insecticides*, rev. ed., 1993.

Nematomorpha

A phylum of worms formerly considered to be a class of the phylum Aschelminthes; commonly called the hairworms, and closely allied to the nematodes. The adults are free-living in aquatic habitats, while the juveniles are parasitic in arthropods. The

nematomorphs are found all over the world. They are divided into two classes, the Nectonematoidea and Gordioidea, with a total of 225 species. See NEMATA (NEMATODA).

Morphology. The body is long and slender with a maximum length of 5 ft (1.5 m) and a diameter of 0.02–0.12 in. (0.5–3 mm). The females are longer than the males. The anterior end is rounded with a dark pigmented ring and a terminal mouth. The posterior end may be rounded with a terminal cloaca, or it may form two or three lobes in a forklike structure. The body color is yellowish, brown, or almost black. The body wall consists of three layers: an outer, rather thick fibrous cuticle; an epidermis consisting of a single layer of cells; and innermost, a muscle layer with longitudinal fibers only. The surface of the cuticle may be smooth, or rough with rounded or polygonal thickenings called areoles. These may be flat or may form projecting structures, sometimes with bristles, and they may be perforated by pores and canals. Between the areoles run interareolar furrows, often with wartlike structures and bristles. Special natatory bristles are developed in *Nectonema*. The function of the areoles is unknown.

Body cavity. This cavity extends the length of the body. It may be filled with tissue so that only minor spaces are left around the digestive system and the gonads. The digestive tube is always more or less degenerated, and the anterior part is often a solid string of cells. Ingestion of food is impossible, and the intact part of the digestive system seems to be adapted for excretory functions. During the parasitic stage food is obtained through the body surface by means of digestive enzymes.

Nervous system. This consists of a cerebral mass lying ventrally in the head and a ventral nerve cord which originates in the epidermal layer. Little is known of the sensory organs. Probably the bristles and warts of the cuticle have sensory functions. A rudimentary eye is found in *Paragordius*.

Reproduction. The sexes are always separate, and the gonads are paired and stringlike, extending the length of the body. In males the gonads are connected with the cloaca by sperm ducts. In females the ovaries form a large number of lateral diverticula in which the eggs ripen. The oviducts enter the cloaca separately. A sac, which is called the seminal receptacle, extends anteriorly from the cloaca.

During copulation the male coils itself around the female and places a drop of sperm near her cloacal opening. The sperm cells actively enter the seminal receptacle. The eggs are laid in water in strings, and the adults die after egg laying. When hatched, the larvae swim to an aquatic arthropod. They penetrate the body wall of the host by means of their characteristic proboscis, which is armed with hooks and three long stylets. The larvae of some species may secrete a special mucus in which they encyst until they are accidentally ingested by the right host, which may be a terrestrial insect. The development in the host lasts some months without any metamorphosis. When mature, the worms leave the host.

Bent J. Muus

Nematophytales

An enigmatic group of fossil plants, in mid-Silurian to lower Upper Devonian rocks, composed of intertwined, branching tubes of two sizes, 10–50 and 1–10 micrometers in diameter.

Prototaxites is stemlike in appearance, ranging from a few millimeters to 3 ft (1 m) in diameter, in which case it may be over 6 ft (2 m) in length with no taper. Its tubes were arranged longitudinally, the small ones having septa at intervals. The septa were perforate, but differ from pores in red algae or in higher fungi. *Nematoballus* looks like flakes of coaly matter, a few millimeters square in size, and is found on rock surfaces. Its larger tubes, 10–50 μm in diameter, were thickened internally by annular rings, and its smaller tubes were 1–5 μm in diameter. Among the tubes were smooth spores, 12–45 μm in diameter, that occasionally show triradial markings. A pseudocellular layer and cuticle constituted the surface of the plant. *Nematoplexus* had larger tubes, averaging 18 μm in diameter, that were thickened internally by annular or helical deposits, and small tubes 8–9 μm in diameter. Branching of tubes occurred in areas called branch knots. A pseudocellular surface layer is possible. This plant is found only petrified, in Rhynie chert.

Although this group is referred to the algae by some authors, its occurrence in inland swamps, coastal plain deposits, and marine deposits close to shore indicates that the members were terrestrial organisms, unrelated to any known groups, perhaps at an intermediate level between algae and bryophytes. Large tubes with helical or annular thickenings, found as microfossils in earliest Silurian strata, may belong to this group. See PALEOBOTANY.

Harlan P. Banks

Nemertea

A phylum of bilaterally symmetrical, unsegmented, ribbonlike worms, frequently referred to as the Nemertinea. They have an eversible proboscis and a complete digestive tract with an anus. There is no coelom or body cavity, and the mesenchyme or parenchyma and the muscle fibers fill the area between the ciliated epidermis and the cellular lining of the digestive tract. See ANIMAL SYMMETRY.

Morphology. The nemertineans are mostly less than 8 in. (20 cm) in length, but a few may reach a length of several meters (Fig. 1). Many species are brightly colored, sometimes having stripes or transverse bars.

The tubular proboscis, lying above the digestive tract in a cavity, the rhynchocoele, is attached posteriorly to the proboscis sheath by a retractor muscle and is either unarmed or armed with stylets (Fig. 2). The proboscis opens anteriorly into a chamber, the rhynchodeum, which in turn opens to the outside above the mouth, through the proboscis pore. The proboscis can be suddenly everted by the

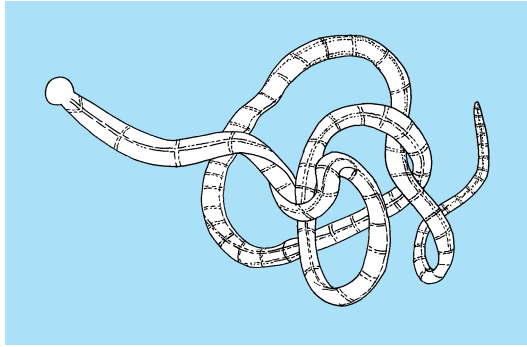


Fig. 1. *Tubulanus capistratus* from the Pacific coast; size ranges from 0.2 to 1.2 in. (5 to 30 mm). (After L. H. Hyman, *The Invertebrates*, vol. 2, McGraw-Hill, 1951)

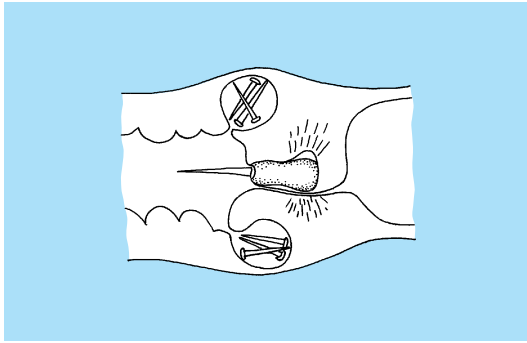


Fig. 2. Stylets of *Amphiporus griseus*. (After W. R. Coe, *Biology of the nemerteans of the Atlantic coast of North America*, *Trans. Conn. Acad. Arts Sci.*, 35:282, 1943)

contraction of muscles which exert pressure on the fluid in the rhynchocoel. The proboscis of the nemertean is used for capturing prey, mostly annelid worms, and for defense, for locomotion, or for burrowing.

The nemertean constitute the most primitive group of invertebrates in which the digestive tract is complete with mouth and anus. In some nemertean, however, a separate mouth is lacking and the esophagus opens through the proboscis pore (Fig. 3).

The nemertean are the simplest animals with a circulatory system. There are two lateral blood vessels and in some a third, unpaired dorsal vessel. The blood consists of a colorless fluid which may contain blood cells of several types. In species in which the blood is colored, the pigment is present in the cells. There is no heart, but the walls of the principal vessels may be contractile. The walls of all the blood vessels, and the plasma, contain enzymes of the arylamidase type whose function remains unknown. It is possible that they are concerned with protein metabolism or, in intertidal species, with osmoregulation and the maintenance of body volume.

The excretory organs or protonephridia are composed of many tubules ending in flame bulbs (Fig. 4). These are hollow, urn-shaped cells provided with vibratile cilia. On each side the flame bulbs are often closely associated with the lateral blood vessel, and

their tubules are united to a common tubule which opens at a lateral nephridiopore. Peculiar large cells called arthrocytes may surround the flame bulbs and tubules and are assumed to be excretory, since they readily take up vital stains.

The nervous system has a pair of cerebral ganglia forming the brain as well as two longitudinal nerve cords and many smaller nerves. The ganglia and lateral cords may contain unusually large neurochord

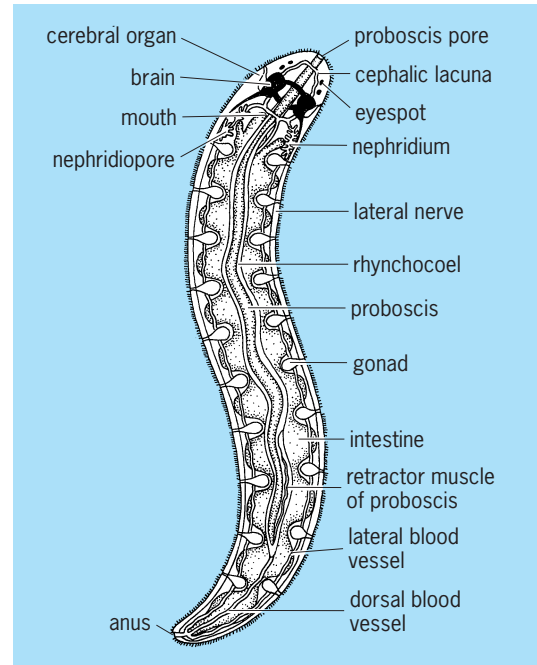


Fig. 3. Internal structure of a nemertean.

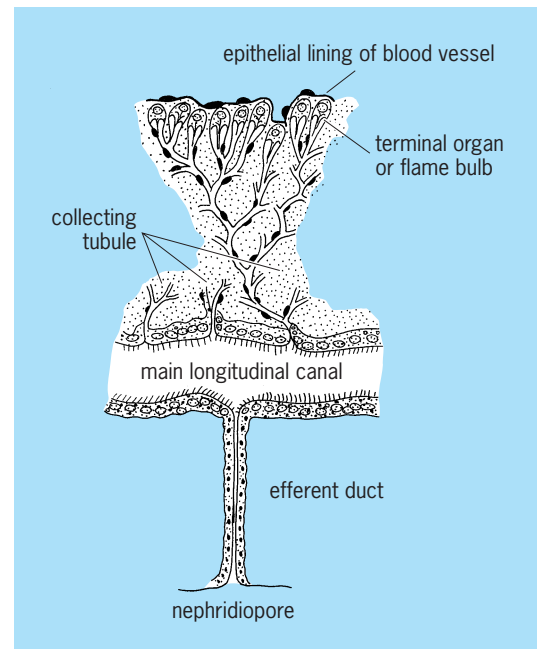


Fig. 4. Diagram of the multiple protonephridium of a nemertean. (After W. R. Coe, *Unusual types of nephridia in nemerteans*, *Biol. Bull.*, 58(3):214, 1930)

cells. In the epidermis there are scattered sensory nerve cells, probably tactile. Chemotactile organs are situated in a pair of anteriorly placed cephalic grooves or in a flask-shaped protrusible frontal organ. A few to many simple eyes, or ocelli, may be present in front of the cerebral ganglia. Statocysts, the organs of equilibrium, are rare. Two blind canals, the cerebral organs, are invaginated from the epidermis and are closely associated with the cerebral ganglia. These organs, probably chemosensory, open through pores in the cephalic grooves or on the body surface.

There are no special respiratory organs; respiration occurs through the body surface.

Nemertineans are usually either male or female, but a few individuals have both sex organs. The ovaries or testes open by short ducts to the exterior. Fertilization occurs outside the body in many species but may be internal in certain forms.

Embryology. Cleavage of the fertilized egg, or zygote, is spiral, with the cells of the lower quartet of the eight-cell stage rotated slightly, to lie in the furrows between the cells of the upper quartet. Development is determinate, with the potentialities for future development of the embryo determined or fixed in the zygote before cleavage begins. Isolated cells of the two-cell stage result in dwarf larvae. Isolated cells of later stages result in deficiencies. See CELL LINEAGE; CLEAVAGE (DEVELOPMENTAL BIOLOGY).

After early cleavage, the development in certain nemertineans may be direct, that is, without a larva, the embryo emerging from the egg membranes as a minute, ciliated worm. In others, the gastrula becomes a free-swimming, helmet-shaped, ciliated pilidium (Fig. 5), formed by the downward growth of two ciliated lobes at the sides of the mouth and having an apical tuft of cilia. In still other nemertineans, the gastrula remains inside the egg membranes and becomes an oval, ciliated larva known as Desor's larva, which lacks the apical tuft and the oral lobes.

Both the pilidium and the Desor's larva metamorphose into an adult worm by means of the invagination of seven or eight ectodermal plates. These flattened, invaginated sacs spread and finally fuse, thus separating the larval ectoderm and the thin am-

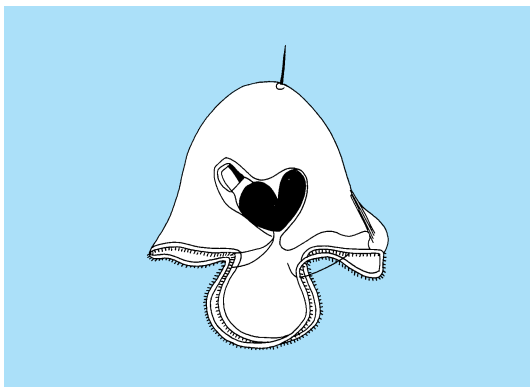


Fig. 5. Pilidium of a nemertinean. (After H. G. Bronn, ed., *Klassen und Ordnungen des Thier-Reichs*, vol. 4, 1903)

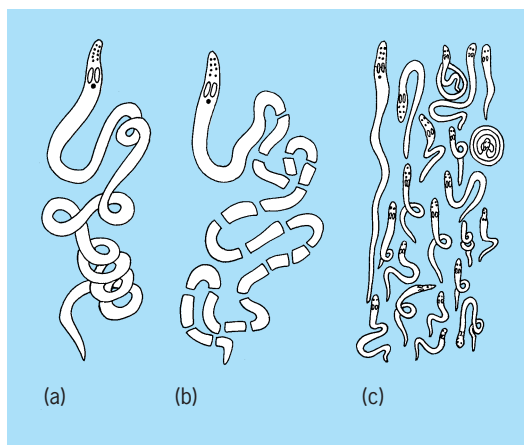


Fig. 6. Steps a-c in fragmentation in *Lineus vegetus*. (After W. R. Coe, *Revision of the nemertean fauna of the Pacific coasts of North, Central, and northern South America, Allan Hancock Pacific Expeditions, vol. 2, University of Southern California Press, 1940*)

nion from the newly invaginated ectoderm of the larval worm within. The larva completes development with the formation of the anus and other organs. It then sheds the larval ectoderm and amnion and emerges as a young worm.

Asexual reproduction and regeneration. Certain nemertineans have the ability to multiply asexually by fragmentation of the body. Each fragment containing a part of the lateral nerve cords can regenerate into a complete worm (Fig. 6). Other nemertineans, when irritated by handling or unfavorable conditions, fragment the body or evert the proboscis so that it breaks off. The anterior region, including the foregut, can regenerate a new proboscis and a new posterior end. In the process of anterior regeneration the wound is covered by migrating epidermal cells. Then, in the mass of mesenchyme cells which forms below the closure, three groups of cells appear, two lateral groups reforming the cerebral ganglia and a median group reforming the proboscis. Regeneration of the posterior end occurs by a lengthening of the body through the differentiation of mesenchyme cells. See REGENERATIVE BIOLOGY; REPRODUCTION (ANIMAL).

Ecology. The nemertineans are mostly marine, bottom-dwelling worms, found in greatest numbers along the coasts of northern temperate regions. They live under stones, among the tangled masses of plants, in sand, mud, or gravel, and sometimes form mucus-lined tubes. A few are pelagic, fresh-water, or terrestrial. Certain species are commensal with other animals, but none can be regarded as parasitic in a strict sense.

Phylogenetic relationships. The Rhynchocoela are related to the Platyhelminthes and probably have evolved from the same ancestral stock which gave rise to that phylum. They resemble the flatworms in having the region between the epidermis and the gut filled with mesenchyme, in the arrangement of the nervous system, in the structure of the eyes, in the occurrence of ciliated grooves on the head, and in showing spiral cleavage. The proboscis and the

cerebral organs may be regarded as derived from certain organs in flatworms. Serological tests have indicated that nemertineans are closer to the platyhelminths than to annelids.

That the Rhynchocoela represent the most highly organized acoelomate animals is indicated by the circulatory system, the presence of an anus, and the specialization of the epidermis. All groups of animals more complex than the nemertineans have some kind of cavity, a pseudocoel or coelom, between the body wall and the gut, instead of solid mesenchyme. See COELOM.

Classification. The phylum Rhynchocoela, containing about 550 known species, is divided into two classes. In the class Anopla the mouth is posterior to the brain, the nerve cords lie under the epidermis or in the muscle layers of the body wall, and the proboscis is unarmed. The two orders of the Anopla are

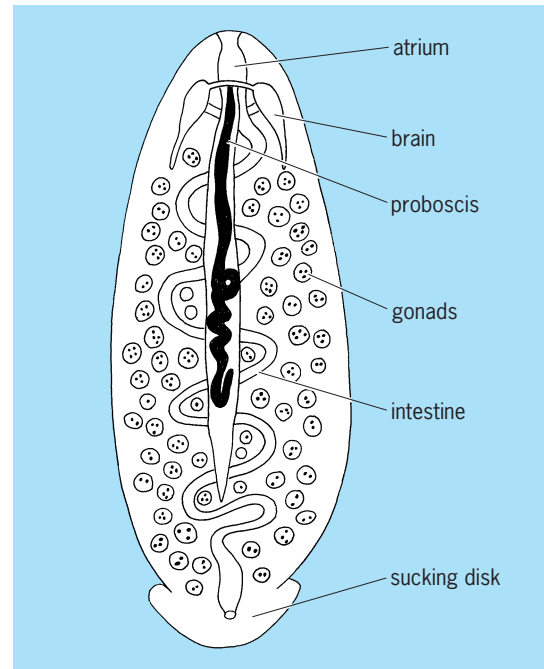


Fig. 8. *Malacobdella grossa*. (After W. R. Coe, *Biology of the nemertean of the Atlantic coast of North America, Trans. Conn. Acad. Arts Sci.*, 35:308, 1943)

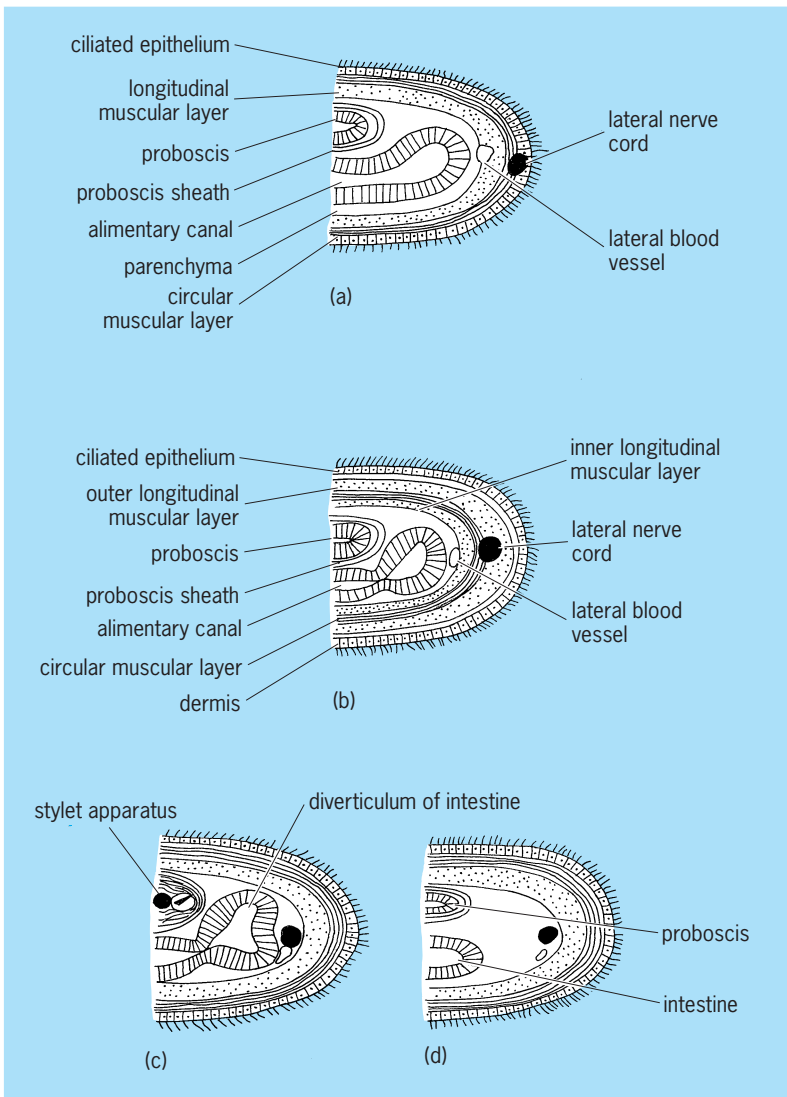


Fig. 7. Transverse sections of the body in the four orders of nemertineans, showing the arrangement of the muscular layers and the position of the lateral nerve cords and lateral blood vessels. (a) Palaeonemertini and (b) Heteronemertini constitute the class Anopla; (c) Hoplonemertini and (d) Bdellonemertini constitute the class Enopla. (After W. R. Coe, *Biology of the nemertean of the Atlantic coast of North America, Trans. Conn. Acad. Arts Sci.*, 35:145, 1943)

separated on the basis of the musculature and the character of the dermis, which is a connective tissue lying under the epidermis. The muscles of order Palaeonemertini are arranged in two or three layers (Fig. 7a); if three, the innermost layer consists of circular fibers. The dermis is gelatinous. The order contains such genera as *Tubulanus*, *Carinoma*, and *Cephalothrix*. In the second order, the Heteronemertini, the muscles are in three layers (Fig. 7b) and the dermis is fibrous. The order contains, among others, the genera *Lineus*, *Cerebratulus*, and *Micrura*.

In the second class, the Enopla, the mouth is anterior to the brain, the nerve cords are internal to the muscles, and the proboscis is often armed with stylets. Of the two orders, the first, the Hoplonemertini, has a proboscis armed with one or more stylets (Fig. 7c) and has a straight intestine with paired lateral diverticula. Representative genera are *Emplectonema*, *Carcinonemertes* (on the gills and egg masses of crabs), *Amphiporus*, *Testrastemma*, *Prostoma* (in fresh water), *Geonemertes* (on land near the sea in tropical and subtropical regions), and *Nectonemertes* (pelagic). The second order, the Bdellomorpha (Bdellonemertini), has an unarmed proboscis, a sinuous intestine without diverticula (Fig. 7d), and a posterior adhesive disk. The only genus, *Malacobdella* (Fig. 8), is commensal chiefly in the mantle cavity of marine clams, where it feeds on plankton brought in by the ciliary current. See ANOPLA; ENOPLA. Arthur G. Humes; J. B. Jennings

Bibliography. R. Gibson, *Nemertean*, 1972; R. Gibson and J. B. Jennings, *Compar. Biochem. Physiol.*, 23:645-651, 1967; L. H. Hyman, *The Invertebrates*, vol. 2, 1951; S. P. Parker (ed.), *Synopsis and Classification of Living Organisms*, 2 vols., 1982.

Neodymium

A metallic chemical element, Nd, atomic number 60, atomic weight 144.24. Neodymium belongs to the rare-earth group of elements. The naturally occurring element includes six isotopes. The oxide, Nd₂O₃, is a light-blue powder. It dissolves in mineral acids to give reddish-violet solutions. For properties of the metal see PERIODIC TABLE; RARE-EARTH ELEMENTS.

| | | | | | | | | | | | | | | | | | |
|----|----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|----|----|----|----|----|
| 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 | | | | | |
| 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 |

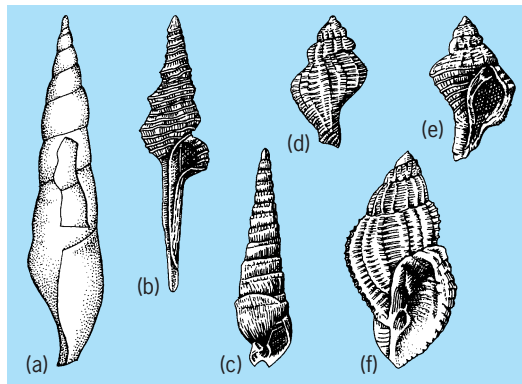
The salts have found application in the ceramic industry for coloring glass and for glazes. The glass is particularly useful in goggles used by glass blowers, since it absorbs the intense yellow D line of sodium present in the flame. The element has found commercial application in the manufacture of lasers.

Frank H. Spedding

Bibliography. F. A. Cotton et al., *Advanced Inorganic Chemistry*, 6th ed., Wiley-Interscience, 1999; K. A. Gschneidner Jr., J.-C. Bünzli, and V. K. Pecharsky (eds.), *Handbook on the Physics and Chemistry of Rare Earths*, 2005.

Neogastropoda

The most highly specialized order in the subclass Prosobranchia (phylum Mollusca, class Gastropoda). Neogastropods have simplified pallial and cardiac structures involving complete separation of genital from renal organs, and a "half-gill" (that is, a one-sided comb-shaped or pectinibranch ctenidium) with its axis and major blood vessels fused to the mantle wall. The order comprises mainly marine carnivores and carrion feeders, all with a long extensible proboscis bearing a flesh-tearing radula. Teeth per row are reduced in such radulae (hence the former ordinal name, Stenoglossa), usually to the rachiglossan, or three-toothed, pattern, but in two families to the toxiglossan type with a single tooth per radular row. More efficient hydrodynamically with their simplified mantle cavity and fused ctenidial axis, neogastropods are not limited to clean waters over hard substrata (as are the archaeogastropods) but have successfully invaded all areas of the seashore and sea bottom, whether covered with sand, silt, or mud.



Neogastropoda. (a) Ordovician species of *Subulites*, widespread Ordovician and Silurian genus. (b) *Falsifusus*, early Tertiary genus. (c) Miocene species of familiar existing *Terebra* (Tertiary-Recent). (d, e) *Urosalpinx* (Tertiary-Recent). (f) *Cancellaria* (Tertiary-Recent). (After R. R. Shrock and W. H. Twenhofel, *Principles of Invertebrate Paleontology*, 2d ed., McGraw-Hill, 1953)

The neogastropod proboscis leads internally to a much simpler gut than that of more primitive gastropods, which are mostly herbivorous or filter-feeding with elaborate sorting areas and recycling processes. In neogastropods, ingested flesh from prey or carrion receives straightforward digestion which is wholly extracellular. The pallial sense organ, or osphradium, which in more primitive mollusks was involved in mechanical detection of sediment for the mantle cavity, has become a chemoreceptor used in hunting down suitable prey by waterborne "scent." The osphradium is located at the inner end of a characteristic inhalant siphon, and neogastropod shells all show a siphonal notch (see *illus.*).

Neogastropods occur in all depths of the world's oceans from the tropics to polar waters, and there are at least 6000 species, mostly in four important superfamilies. The larger whelks of the superfamily Buccinacea, in such genera as *Busycon*, *Colus*, *Nepitunea*, *Fasciolaria*, and *Buccinum*, are found from the shallow sublittoral and continental shelves down to depths of 9800 ft (3000 m). The flesh of many whelk species provides human food, and almost all species have been used in commercial longline fisheries as resilient and attractive bait. This superfamily also includes the ubiquitous and cosmopolitan mudsnails of the genera *Nassarius* and *Ilyanassa*, and the beautiful tulip shells and spindles (*Fasciolaria* and *Fusinus*).

The smaller tangles, dog whelks, and oyster drills of the superfamily Muricacea are the abundant neogastropod predators in inshore and intertidal waters. Certain species of the genera *Ocenebra*, *Urosalpinx* and *Rapana*, known as tangles and oyster drills, are the most destructive pests of clam beds and oyster farms, and use the proboscis-borne radula to bore through the shells of their prey. The dog whelks, species of genera such as *Nucella*, *Eupleura* and *Thais*, also prey extensively on barnacles; the extensible proboscis is used to force open the opercular plates (rather than bore an access hole as

with bivalves) and then to browse on the internal tissues of the prey. Larger species of muricid rock shells, in such genera as *Trophon*, *Murex*, *Forreria*, and *Pteropurpura*, may have extensive shell sculpture in the form of radiating spines and wings. Despite extensive shell collecting of these forms, their biology has been little studied.

A third important superfamily, Volutacea, encompasses more beautiful, much collected shells belonging to such family groups as olivids, mitrids, volutids, marginellids, and turbinellids. The most specialized neogastropods are the tropical toxoglossans (superfamily Conacea) belonging to the families Conidae, or cone shells, and Terebridae, or auger shells. Both groups have been prized by shell collectors for centuries. In them, the radula is reduced to a series of separate darts, which are used in association with a modified salivary gland secreting a neurotoxin for the capture of large prey animals, both invertebrate and vertebrate. Many species of cones respond only to particular species of live, healthy prey by aiming the proboscis, striking with a single everted tooth, injecting toxin, and then feeding upon the paralyzed prey. Three main categories of cones, each with approximately proportioned harpoonlike teeth and appropriate toxins, feed respectively on other snails, on annelid worms, and on fishes. A number of human fatalities have resulted from two species of fish-eating cones in the East Indies.

Despite their large number of species and diversity of habitats, the neogastropods show more anatomical uniformity (efficient mantle cavity, inhalant siphon with chemoreceptive osphradium, extensible proboscis, stenoglossan radula, and simple carnivore gut) than is found in any of the other major orders of gastropods. See GASTROPODA; MOLLUSCA; PROSOBRANCHIA.

W. D. Russell-Hunter

Bibliography. V. Fretter and A. Graham, *British Prosobranch Molluscs*, 1962; W. D. Russell-Hunter, *A Life of Invertebrates*, 1979; K. M. Wilbur (ed.), *The Mollusca*, vols. 1-8, 1983-1984.

Neognathae

One of the two recognized superorders making up the subclass Neornithes of the class Aves. They are characterized as flying birds with fully developed wings and sternum with a keel, caudal vertebrae fused into a pygostyle, and absence of teeth in both jaws, or modifications of these conditions in secondary flightless birds.

This superorder includes all living birds and all known fossil birds since the Late Cretaceous; only the ancestral Jurassic *Archaeopteryx* and the specialized Cretaceous *Hesperornis* and its allies do not belong to the Neognathae. Several previously recognized superorders—Ichthyornithes for *Ichthyornis* and its allies, Impennes for the penguins, and Palaeognathae for the living flightless ratites—have been merged with the Neognathae. Merging of these four superorders reflects the lack of knowledge of relationships of avian orders and implies that, as far

as is known, the orders included in each former superorder are not more closely related to each other than to those orders placed in other superorders, or that the particular characteristics of each group are no more specialized than those found in individual orders of birds. Evidence suggests strongly that the ratites constitute a monophyletic assemblage, so that the Palaeognathae may be revived if other equivalent assemblages of neornithine orders can be recognized. See ARCHAEOPTERYX; AVES; ODONTOGNATHAE; RATTITES.

Walter Bock

Neognathostomata

A superorder of Echinoidea, subclass Euechinoidea. These invertebrates are characterized by having a rigid, exocyclic test and a lantern or jaw apparatus developed sometime during the life history and usually persisting into the adult stage. The included orders are the Clypeasteroidea, Cassiduloidea, Neolampadoidea, and Oligopygoidea. See ECHINODERMATA; ECHINOIDEA; EUECHINOIDEA; HOLECTYPOIDA; NEOLAMPADOIDA; OLIGOPYGOIDA.

Howard B. Fell

Neogregarinida

An order of the protozoan subclass Gregarina, class Telosporae, subphylum Sporozoa. All gregarines are parasites of the digestive tract and body cavity of invertebrates or lower chordates; their mature trophozoites (vegetative stages) live outside the host's cells and are large. The Neogregarinida are thought to be relatively advanced gregarines which live in insects. The most primitive gregarines, the Archigregarinida, differ from the typical gregarines, the Eugregarinida, in having asexual multiple fission (schizogony) before gamonts (cells which will produce gametes) are formed. The Neogregarinida are thought to have acquired this trait secondarily and to have developed from the Eugregarinida. See ARCHIGREGARINIDA; EUGREGARINIDA.

There are only about 29 species of about 12 genera, which have been divided into 4 families. The largest genus is *Ophryocystis*, which has species.

Another genus is *Mattesia*, which has 2 species, one of which is *M. dispersa* of the flour moth (*Ephesia kuehniella*). In this species the intracellular schizonts occur in the fatty tissues. The schizonts are of two types, with small or large nuclei. Those with small nuclei apparently produce merozoites which turn into schizonts with large nuclei. These produce merozoites which turn into gamonts that leave the cells of the fatty tissues, come to lie in the hemocoel, join in pairs (syzygy), grow, encyst together, and proceed to form two gametes each. The gametes fuse to form two zygotes within the gametocyst, and each zygote in turn forms an oocyst that contains eight sporozoites. It is not clear how they get to new flour moths. See GREGARINIA; PROTOZOA; SPOROZOA; TELOSPOREA.

Norman D. Levine

Neolampadoida

A group of small, deep-water cassiduloid echinoids with neotenous characteristics, treated as an order by some workers; possibly polyphyletic. The presence of bourrelets and phylloides, the elongate first ambulacral plates, the undifferentiated tuberculation, and undifferentiated posterior interambulacral plating all indicate their relationships lie with cassiduloids. The only character shared by members of this group is the lack of petals (they have simple ambulacral pores only). Other characteristics, such as apical disc plating, are varied, indicating at least two independent origins from shallow-water cassiduloids.

There are seven genera, each monospecific. Five are living today and are usually found at depths of 430–1280 ft (135–400 m). A Miocene species and an Upper Eocene species are also known. *See* ECHINODERMATA; NEOGNATHOSTOMATA. Andrew B. Smith

Bibliography. J. W. Durham and C. D. Wagner, Neolampadoids, in R. C. Moore (ed.), *Treatise on Invertebrate Paleontology*, pt. U: *Echinodermata* 3, pp. U628–U630, 1966.

Neolithic

The period of prehistoric culture whose basic defining attributes are the emergence of agriculture, animal domestication, and sedentary farmsteads or villages. This definition has evolved over the last century from the original characterization of this period based on the appearance of polished stone axes. By 1865, when John Lubbock published *Prehistoric Times*, two types of Stone Age had been recognized in Europe: *période de la pierre taillée* (period of chipped stone implements) and *période de la pierre polie* (period of polished stone implements). Lubbock termed the former Palaeolithic and the latter Neolithic. Subsequently, it was realized that the definition of this period based on a single artifact type was spurious, since Neolithic peoples also continued to make chipped stone tools. A more comprehensive view developed that saw the Neolithic as characterized by pottery manufacture, agriculture, livestock, and settled villages, but without the use of metals. Thus the Neolithic formed the final Stone Age precursor to the Bronze Age and the Iron Age in the classic northern European prehistoric sequence, which was soon extended throughout most of Eurasia. *See* PALEOLITHIC.

Modern concepts. Research dating back to the 1950s in the Near East has demonstrated that the domestication of plants occurred prior to the onset of pottery production. At the same time, it became clear that the nonagricultural Ertebølle culture of southern Scandinavia and the Jomon culture of Japan produced pottery. Thus pottery production has generally ceased to be considered a defining trait of the Neolithic. The discovery of the early use of native and smelted copper for ornaments and tools at Neolithic sites in Europe and the Near East also forced

a reconsideration of whether metal use excluded a culture from the Neolithic. In some areas, such as southern Europe and the Near East, a distinct Copper Age designation was introduced to account for this discrepancy. These taxonomic difficulties have led some archeologists to abandon the use of the term Neolithic, but it continues in use today as a convenient designation for early agricultural societies.

As the term Neolithic is presently used, it refers specifically to prehistoric societies in Europe, Asia, and northern Africa that derived the majority of their diet from agriculture and livestock and that lived in sedentary communities, either dispersed farmsteads or villages, but that did not yet know the use of alloyed metals. Outside of this area, the term Neolithic is rarely used, although clearly in most other parts of the world there was also a transition from hunting and gathering to agriculture at some time in prehistory. Archeologists working in the New World frequently use the term “Formative” to refer to early sedentary agricultural communities similar to those of the Old World Neolithic. These variations in nomenclature and tradition should not obscure the fact that in the last 10,000 years humans have gone from being exclusively hunters and gatherers to deriving virtually all of their food supply from cultivated plants and domesticated animals.

For the purposes of this article, “Neolithic” will be expanded to include a global consideration of the origins and dispersal of agriculture and sedentary life, what might be called a neolithic cultural pattern. The British prehistorian V. Gordon Childe (1892–1957) once wrote of a “Neolithic Revolution” to characterize the emergence of sedentary agricultural societies. While the transition from hunting and gathering to farming and herding had immense consequences for subsequent human cultural development, it was hardly revolutionary in most areas. Instead, the Neolithic developed either through the local domestication of plants and animals by people already familiar with their reproductive biology, by the movement of farming peoples into areas where agriculture had not been previously practiced, or through the adoption by hunter-gatherers of crops and livestock that had been domesticated elsewhere.

Domestication of plants and animals. With the exception of the dog, which was domesticated by late Pleistocene hunter-gatherers in many parts of the world, the domestication of plants appears to have preceded that of animals. Although many parts of the world can claim to have been the location of the domestication of some food plants, the regions where the crop species which underlie the major agricultural systems of the world were domesticated are fairly well delimited. In 1971 the American botanist Jack Harlan identified three relatively small regions where the domestication of major crop species can be localized, along with several larger areas in which other important species were domesticated in a somewhat more diffuse geographical pattern. Research since then has amended these slightly, but the general outlines of this pattern remain the same. The three foci of domestication

include the “Fertile Crescent” area of the Near East (wheat and barley), the Huanghe and Yangzi valleys of China (millet and rice), and Mesoamerica (maize and beans), while the broader areas of where important domestication also occurred include northern and Andean South America (potatoes and other root crops), northern sub-Saharan Africa (sorghum), and southeast Asia and many Pacific islands (various tree and root crops). Other research has added another area to this list. The American archeologist Bruce D. Smith, building on the work of earlier botanists and archeologists, made a compelling case for the indigenous domestication in eastern North America of a complex of weedy plants that include chenopod, sunflower, and marsh elder.

Near East. The Fertile Crescent extends in an arc from the Jordan Valley in the Levant across the headwaters of the Tigris and Euphrates rivers in northern Syria and Turkey and down the flanks of the Zagros Mountains of Iraq and Iran. Recent research points toward a part of this region known as the Levantine Corridor, an inland zone stretching for about 40 km (25 mi) between the Damascus Basin and the lower Jordan Valley, as being potentially the location of the earliest cultivation of wheat and barley. At the end of the Pleistocene, this would have been one of the moister parts of the Levant, and it also contained stands of wild wheat and barley. The Natufian hunter-gatherers of this region were intensive collectors of wild cereals, and about 10,000 years ago they became the first Neolithic farmers. Einkorn wheat, emmer wheat, and barley are the three principal founder crops of Near Eastern agriculture. They are distinguished from their wild counterparts primarily by the nature of the rachis, the stem which holds the grain to the ear. Wild wheats and barley have a brittle rachis, which shatters at ripening to disperse the grain and thus propagate the plant. In domestic forms, the rachis is tougher and thicker, which permits the harvesting of the whole year by cutting with sickles. Such a method would have selected for plants with tougher rachises. It may have taken a relatively short time, on the order of a few centuries, to produce populations of wheat and barley that were dependent on humans for their propagation. At sites such as Netiv Hagdud in the lower Jordan valley, early farmers lived in circular and oval structures built of mud bricks upon stone foundations and stored their grain in subterranean silos.

Just as wheat and barley were the founder crops of Near Eastern agriculture, sheep and goat can be considered the founder animals of ungulate domestication around 9000 years ago. As the American zooarcheologist Richard Meadow pointed out, the domestication of animals requires a profound change in human attitudes toward them, a shift from the importance of the dead animal as a source of meat to the living animal as a member of a breeding population. Goats appear to have been domesticated first in the Zagros Mountains, while sheep at sites in the upper Tigris and Euphrates valleys underwent the progressive diminution in size associated with domestication a few centuries later. Cattle and pigs ap-

pear to have been domesticated around 8000 years ago, with evidence pointing toward Anatolia as the likely location.

China. The Huanghe (Yellow) river and Yangzi river valleys of China were two distinct foci of early plant domestication. Each region has its distinctive founder crop: foxtail millet (*Setaria italica*) in northern China and rice (*Oryza sativa*) in southern China. The wild precursor of foxtail millet is a common Eurasian weed, green bristlegrass. As with wild and domestic forms of wheat and barley, green bristlegrass and foxtail millet differ in their method of seed dispersal, with the wild form shattering and the domestic form retaining its grains. Deep deposits of loess, a dry but fertile windblown soil, along the Huanghe and its tributaries provided ideal growing conditions for millet. The earliest Neolithic culture of the Huanghe valley is called Peiligang, after a site in Henan province.

About 600 km (360 mi) to the south, in the Yangzi drainage, the Hupei basin is a lowland zone of lakes, wetlands, and meandering streams. Recent excavations at Pengtoushan have revealed a large settlement of sedentary peoples dated between 8500 and 7800 years ago. Large amounts of rice indicate that this plant was either intensively collected or in the early stages of cultivation. On the coastal plain north of Shanghai, in another area of marshes and lakes, the waterlogged site of Hemudu has yielded convincing evidence for the use of domesticated rice before 7000 years ago.

The pig and the chicken were the most important domesticated animals in the Chinese Neolithic, supplemented soon after with the water buffalo. The pigs were probably domesticated from local wild populations, while the chicken is believed to be derived from the red jungle fowl of southeast Asia.

Mesoamerica. Mesoamerica is a term used by anthropologists and archeologists to mean the area of Mexico and the adjacent northern part of central America in which complex urban societies such as the Maya and Aztec emerged in later prehistoric times. It also appears that this was the region in which maize and beans, the major founder crops of many New World agricultural systems, were first domesticated. Unlike wheat and barley, maize is a hybrid and has no obvious wild progenitor, so its origins are hotly debated among paleobotanists. Research in the 1980s and 1990s pointed toward the grassy weed called teosinte as the likely ancestor of maize.

The earliest maize cobs have been found in the dry caves of the interior uplands of Mesoamerica, such as those of the Tehuacán valley, where they were desiccated and preserved. Biochemical comparison of teosinte and maize suggests that interior lowlands, such as the Balsas river valley in central Mexico, may be more likely candidates for the location of the earliest maize domestication, although little archeological research has been done thus far. The initial dating of the Tehuacán maize cobs in the 1960s pointed toward their relatively early domestication, perhaps over 7000 years ago, but recent redating using the accelerator mass spectrometry (AMS) technique has

established that the earliest maize at Tehuacán is no earlier than 5500 years ago.

The other major New World field crop, the common bean (*Phaseolus vulgaris*), appears to have been domesticated independently in at least two locations, Mexico and the southern Andes. Recent AMS dating of the oldest specimens suggests that beans were domesticated more recently than previously thought, and the maize-bean association that is so common in later prehistoric and modern fields may not have been used during the early stages of the cultivation of these plants.

Eastern North America. It was long presumed that North America was not involved in early plant domestication, since maize, beans, and squashes spread northward from Mesoamerica at a fairly slow rate. By the mid-1980s, however, evidence had amassed that four major seed plants—sunflower, squash, marsh elder, and chenopod—had been brought under sufficient human control and selection to produce morphological changes that could be considered domestication. Bruce D. Smith argued that these were weeds which aggressively colonized disturbed habitats, such as point bars and other floodplain locations, and were then transplanted and cultivated by humans in similar settings starting around 3500 years ago. Although this indigenous agricultural system did not survive into the present, it supported the development of farmsteads and hamlets in the interior valleys of eastern North America for over two millennia.

Other geographic areas. Several other parts of the world made important contributions to the list of domesticated plants and animals during the last 5000 years and perhaps earlier. In a broad band across the northern half of Africa, sorghum, African rice, and pearl millet were early domesticates between about 4000 and 3000 years ago. In northern South America, evidence from pollen and phytoliths (small diagnostic silica structures that form in the veins of plants) indicates that a variety of root and tuber crops, including manioc and arrowroot, and tree crops were brought under cultivation, perhaps at a very early date. In the Andes, the potato emerged as an important food source, guinea pigs were domesticated as a meat source, and llama and alpacas were used both as pack animals and for food. Taro and yams, along with a variety of tree crops, were managed, cultivated, and eventually domesticated in a broad region of southeast Asia and the nearby Pacific islands.

Agricultural dispersals. After the initial domestication of plants and animals, the Neolithic farming economy spread quickly to adjacent areas, many of which had been previously inhabited by hunter-gatherers. A major research challenge for archeologists is to determine whether the dispersal of farming, herding, and settled farmstead or village life was the result of the actual movement of agricultural peoples into areas that had hitherto been sparsely populated, or the adoption of domesticated plants and animals by hunter-gatherers who then abandoned their foraging way of life. Two major agricultural dispersals are that of the wheat-barley-livestock complex from the Near East into Eurasia and the maize-

beans-squash complex from Mesoamerica into North America.

Eurasia. Within a relatively short time after the establishment of the first Neolithic settlements in the Fertile Crescent, communities practicing agriculture appeared in central Anatolia. Around 8500 years ago, they spread to Greece where they colonized the fertile alluvial soils of regions such as Thessaly. The earliest European farming communities share many traits with their Near Eastern precursors, including the use of mud brick or adobe reinforced with wood for house construction. From Greece, agriculture spread along two main paths: west through the Mediterranean basin, eventually reaching Spain and Portugal, and north through the Balkans, eventually turning northwest along the Danube drainage. In both cases, crops, livestock, and often people moved quickly, so that within 1500 years of the first appearance of agriculture in Greece it had reached the Atlantic Ocean and the English Channel.

In some areas, such as along the Mediterranean coast, in the interior uplands of central Europe, and on the coastal lowlands of northern and western Europe, it appears that agriculture spread through the adoption of domestic plants and animals by indigenous hunter-gatherers. Elsewhere, such as on the alluvial soils of the floodplains of interior central Europe, a strong argument can be made for the establishment of Neolithic communities through the actual dispersal of farming peoples. In these interior valleys, the farmers of the Linear Pottery culture built farmsteads characterized by timber structures up to 30 m (100 ft) in length, the largest free-standing buildings in the world at this time. In addition to wheat and barley, cattle were important elements in the Linear Pottery economy, and there is evidence that indicates that they were used for milk as well as for meat.

Agricultural communities also spread east and northeast from the Fertile Crescent to the oases of central Asia, the Iranian Plateau, and beyond. At Mehrgarh in Pakistan, the initial farming occupation took place around 8000 years ago. Around the same time, Neolithic settlements were established at Djeitun and neighboring oases in Turkmenistan. Rectangular mud-brick houses constituted the primary domestic architecture of these hamlets.

North America. From its original area of maize domestication in Mesoamerica, agriculture spread northward to the northern border of the Sonoran Desert. Although some archeologists dissent, the weight of evidence points toward the adoption of maize and beans by the indigenous inhabitants of the Colorado Plateau and the Mogollon Highlands rather than a northward migration of farmers from Mexico. The first appearance of maize in this area is now AMS-dated to between 3500 and 4000 years ago, which revises earlier dating which had suggested that it occurred several millennia earlier. Domesticated plants were integrated at first into the existing subsistence pattern, and it was not until about 2000 years ago that people in the area became dependent on them for most of their food.

In eastern North America, maize spread gradually to the north and east. It made its first appearance in the Mississippi valley about 2000 years ago, based on AMS dates from the Holding site located near St. Louis. It remained a fairly minor component of the diet until about A.D. 900, when stable carbon isotopes in human bones show a dramatic increase in maize consumption throughout eastern North America. Maize finally reached the northeastern limit of its growing range in southern New England around A.D. 1000.

Late Neolithic developments. Not only are the domestication of plants and animals and the establishment of farming settlements the hallmarks of the Neolithic, but also these provided the platform for subsequent cultural developments. In Europe and elsewhere in the Old World, later Neolithic societies developed elaborate burial practices. These range from the megalithic tombs of the Atlantic seaboard to the rich burials at Khok Phanom Di on the Gulf of Siam. In Europe and the Near East, cattle began to be used to pull wagons and carts, while on the Eurasian steppes north of the Black Sea, horses were domesticated. In the New World, Formative societies exhibited increasing social complexity. At sites such as San José Mogote in Oaxaca, public buildings began to be constructed among the residential compounds. The Hopewell culture in the Ohio valley developed large-scale procurement networks for exotic resources such as obsidian from the Rocky Mountains and native copper from the upper Great Lakes. The Neolithic in the Old World and its New World parallels were periods of dramatic change in human society, laying the foundation for the socially stratified societies that followed. Peter Bogucki

Bibliography. P. Bogucki, *Forest Farmers and Stockherders: Early Agriculture and its Consequences in North Central Europe*, 1988; J. Clutton-Brock (ed.), *The Walking Larder: Patterns of Domestication, Pastoralism, and Predation*, 1989; J. Harlan, *The Living Fields*, 1995; D. R. Piperno and D. M. Pearsall, *The Origins of Agriculture in the Lowland Neotropics*, 1998; T. D. Price (ed.), *Europe's First Farmers*, 2000; T. D. Price and A. B. Gebauer (eds.), *Last Hunters-First Farmers: New Perspectives on the Prehistoric Transition to Agriculture*, 1995; C. A. Reed (ed.), *Origins of Agriculture*, 1977; D. Rindos, *The Origins of Agriculture: An Evolutionary Perspective*, 1984; B. D. Smith, *The Emergence of Agriculture*, 1995; W. H. Wills, *Early Prehistoric Agriculture in the American Southwest*, 1988; D. Zohary and M. Hopf, *Domestication of Plants in the Old World: The Origin and Spread of Cultivated Plants in West Asia, Europe, and the Nile Valley*, 1993.

Neomeniomorpha

A subclass of creeping, vermiform mollusks in the class Aplacophora. They are covered by a spicular integument and recognized by the presence of a ventral groove within which lies a narrow foot and by the

absence of an oral shield. Neomenioids range in size from less than 0.08 to 12 in. (2 mm to 300 mm) and are found from subtidal areas to the abyss, at depths over 16,000 ft (5000 m). There are 23 families with 70 genera and 193 species worldwide.

The cuticular integument is thin or thick and the spicular coat smooth, rough, or spiny. The calcareous spicules are solid or hollow and of many shapes: thin flat leaves, upright rods, needles, paddles, hooks, or scoops. Many species have crossed needlelike spicules forming a reticulated network within the cuticle. Epidermal cells secrete the spicules and cuticle. Epidermal papillae, stalked in forms with a thick cuticle, are excretory and empty through the cuticle. Muscles include circular, cross-diagonal, and longitudinal layers beneath the epidermis, a pair of ventral longitudinal bands, and two sets of paired, serial lateroventral bands.

Neomenioids creep by means of their ciliated foot along a track of sticky mucus produced from a ciliated, eversible pedal pit at the anterior end of the pedal groove. Anterior to the pedal pit, the head end is held above the substratum and freely moved.

Food is entirely cnidarians—hydrozoans, zooantharians, gorgonians, or stony corals. Above the mouth lies a sensory atrium. Circular muscles often surround the pharynx, which acts as a buccal pump. The radula is formed of rows of two mirror-image teeth which in some cases are fused medially; it may be large to very small or absent. Digestion starts in the pharynx, well supplied by secretory glands; there is nearly always a pair of discrete ventral salivary glands. A midgut with a dorsal ciliated typhlosole serves both as stomach and digestive gland; it passes posteriorly into a ciliated intestine, which bends ventrally and opens into the mantle cavity.

A ventricle and single or paired auricles lie in the pericardium; a sole vessel, the dorsal aorta, may be present, but otherwise blood moves freely through sinuses of the hemocoel. Paired ventral and lateral nerve cords arise from a cerebral ganglion and run posteriorly, connected by cross-commissures; the lateral cords join above the rectum as a ganglion, which gives off a nerve to a chemoreceptive dorsal sensory organ.

All neomenioids are hermaphroditic and have paired gonads that empty into the pericardium or, unusually, into cloacal pockets. A pair of U-shaped gametoducts with one to many seminal receptacles passes from pericardium to the cloaca. The lower gametoducts are lined with secretory cells and function as shell glands; they open on paired gametopores or unite before opening on a single gametopore which is sometimes a muscular penial sheath. Paired, protrusible copulatory spicules are often present. A barrel-shaped, nonfeeding larva called a pericalymma either is brooded or swims by means of a ciliated cellular test within which the animal develops; metamorphosis through loss or resorption of the test occurs within 10 days. See APLACOPHORA; MOLLUSCA. Amelie H. Schelftema

Bibliography. A. C. Giese and J. S. Pearse (eds.), *Reproduction of Marine Invertebrates*, vol. 5: *Molluscs: Pelecypods and Lower Classes*, 1979; H. Heath, *Solenogastres, Mem. Mus. Compar. Zool. (Harvard Univ.)*, 45:1-179, 185-263, 1911, 1918; L. v. Salvini-Plawen, *Antarktische und subantarktische Solenogastres, Zoologica (Stuttgart)*, 44:1-315, 1978; E. R. Trueman and M. R. Clarke (eds.), *The Mollusca*, vol. 10: *Evolution*, 1985.

Neon

A gaseous chemical element, Ne, with atomic number 10 and atomic weight 20.183. Neon is a member of the family of noble gases. The only commercial source of neon is the Earth's atmosphere, although traces of neon are found in natural gas, minerals, and meteorites. See INERT GASES; PERIODIC TABLE.

| | | | | | | | | | | | | | | | | | |
|----|----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|----|----|----|----|----|----|
| 1 | | | | | | | | | | | | | | | | | 2 |
| H | | | | | | | | | | | | | | | | | He |
| 3 | 4 | | | | | | | | | | | 13 | 14 | 15 | 16 | 17 | 18 |
| 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 | | | | | | |
| 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 |

Considerable quantities of neon are used in high-energy physics research. Neon fills spark chambers used to detect the passage of nuclear particles. Liquid neon can be utilized as a refrigerant in the temperature range about 25 to 40 K (-416 to -387°F). Neon is also used in some kinds of electron tubes, in Geiger-Müller counters, in spark-plug test lamps, and in warning indicators on high-voltage electric lines. A very small wattage produces visible light in neon-filled glow lamps; such lamps are used as economical night and safety lights. See NEON GLOW LAMP.

Neon is colorless, odorless, and tasteless; it is a gas under ordinary conditions. Some of the other properties of neon are given in the table. Neon does not form any chemical compounds in the ordinary

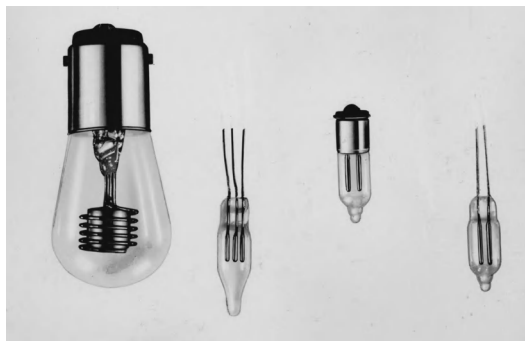
sense of the word; there is only one atom in each molecule of gaseous neon. Arthur W. Francis

Bibliography. F. A. Cotton et al., *Advanced Inorganic Chemistry*, 6th ed., Wiley-Interscience, 1999; D. R. Lide, *CRC Handbook Chemistry and Physics*, 85th ed., CRC Press, 2004; M. Ozima and F. A. Podosek, *Noble Gas Geochemistry*, 2001.

Neon glow lamp

A low-wattage lamp often used as an indicator light or as an electronic circuit component. The neon lamp usually consists of a pair of electrodes sealed within a bulb containing neon gas at a low pressure. Some of the smaller bulbs are equipped with wire leads that are connected directly into the electrical supply circuit; others are equipped with conventional bases that vary with the size of the lamps (see *illus.*).

Operation. Electrodes sealed in a neon atmosphere will emit electrons if a sufficient voltage difference is impressed across them. In glow lamps the electrodes are usually treated to emit electrons freely. With a sufficiently high voltage between electrodes, the velocity of electron flow is high enough to ionize the neon nearest the negative electrode (cathode). The neon then emits a reddish-orange glow similar to the color of neon sign tubing. With direct current the glow is restricted to the immediate vicinity of the negative electrode. With alternating current (ac), both electrodes act alternately as cathodes, and the glow appears alternately at both surfaces. At usual frequencies, the alternations occur so rapidly that both electrodes appear to glow constantly.



Some examples of glow lamps

In direct-current (dc) circuits, the voltage across the electrodes may be reduced significantly, once the lamp has started, without causing the lamp to go out. Direct-current starting voltages for typical glow lamps range from 65 to 90 V, while the minimum operating voltage at which the glow will be maintained may be 10-15 V lower. On ac circuits, the maintaining voltage is nearly the same as the starting voltage.

This glow discharge is like the arc in a vapor lamp in that its resistance decreases with increasing current. Therefore, a current-limiting element must be used in the electric circuit to maintain a desired stabilized current in the circuit. Because the current in

Physical properties of neon

| Property | Value |
|---|----------|
| Atomic number | 10 |
| Atomic weight (atmospheric neon only) | 20.183 |
| Melting point, $^{\circ}\text{C}$ | -248.6 |
| Boiling point at 1 atm pressure, $^{\circ}\text{C}$ | -246.1 |
| Gas density at 0°C and 1 atm pressure, g/liter | 0.8999 |
| Liquid density at its boiling point, g/ml | 1.207 |
| Solubility in water at 20°C , ml neon (STP)/1000 g water at 1 atm partial pressure neon | 10.5 |

glow lamps is usually a few milliamperes or less, it is both practical and economical to use a small resistor as a ballast. The larger glow lamps with screw bases have resistors in the bases; smaller lamps require an external ballast resistor. The resistance value depends on the lamp type.

Applications. The neon glow lamp is inherently a low-wattage source that produces light at relatively low efficiency when compared to filament lamps and other vapor lamps. Its lighting applications are widespread where only a low power source is available. These range from simple illuminated wall switches in homes, lighted switches in small appliances, and indicator lights on the panels of electrical devices to more sophisticated usage in digital read-out devices.

In electronic circuits involving relatively low power, neon lamps are used in many ways. They are used in counter and memory elements of computers, in voltage regulators, in relaxation oscillators, and in trigger circuits to operate relays and similar devices. These applications are practical because of the lamp's unique electrical characteristics, its small size, and its light weight.

Characteristics. The useful life of a glow lamp is not terminated by a burnout, as is the life of lighting lamps, but by a gradual rise in starting and maintaining voltage and blackening of the inner walls of the bulb, reducing light output. If the lamp is used as an indicator light, the reduction in brightness will determine its life, which may be 5000–25,000 h, depending on the application.

When the lamp is used as a circuit element and voltage is important, the change in starting voltage, maintaining voltage, or both will determine useful life. Depending on the type of lamp and its operating current, a rise of 5 V in starting voltage may occur after 1000–6000 burning h. Maintaining voltage rises about half as fast as starting voltage.

External factors may also affect the operating characteristics of glow lamps. The sensitized electrodes of glow lamps release electrons in the presence of light. In total darkness, the starting voltage may be 100 V higher than in light. In totally dark enclosures, slight amounts of light or other radiation, or electrostatic fields, may be used to overcome the dark effect. Darkness does not affect maintaining voltage. See VAPOR LAMP.

Alfred Makulec

Bibliography. Illumination Engineering Society of North America, *IES Lighting Handbook—Reference and Application Volume*, 1993.

Neornithes

The subclass of Aves that contains all of the known birds other than those placed in the Archaeornithes. Comprising more than 30 orders, both fossil and living, its members are characterized by a bony, keeled sternum with fully developed powers of flapping flight (secondarily lost in a number of groups); a short tail with the caudal vertebrae fused into a single platelike pygostyle to which all tail feathers attach; a

large fused pelvic girdle with a reversed pubis which is fused to a large synsacrum; and a large brain and eyes contained within a fused braincase. The jaws are specialized into a beak covered with a horny rhamphotheca; the upper jaw is kinetic, being either prokinetic or rynchokinetic. Prokinesis refers to a bending zone at the base of the upper jaw, and rynchokinesis to one within the upper jaw. A few fossil groups still possess teeth, but most fossil and all Recent birds have lost teeth. See ARCHAEOPTERYX.

The Neornithes contains two superorders, the Odontognathae and the Neognathae. The Odontognathae, alternately known as the Odontornithes, may be an artificial group. Its members, which include the Cretaceous fossil orders Hesperornithiformes and Ichthyornithiformes, are united only by the presence of teeth in all species. The Neognathae contains the remaining modern birds, which have lost the teeth, and includes 26 orders. Six orders of Cretaceous fossil birds remain to be placed in either of these superorders because so little is known of their morphology and relationships. The grouping of the orders of living and fossil birds is still ambiguous and requires further work. A few groupings appear to be based on strong evidence, including the close relationships between the Anseriformes and the Galliformes, the Gruiformes and the Charadriiformes, the Columbigiformes and the Psittaciformes, the Strigiformes and the Caprimulgiformes, and possibly the orders of large land bird, the Apodiformes, Coraciiformes, Piciiformes, and Passeriformes. See AVES; NEOGNATHAE; ODONTOGNATHAE.

Walter J. Bock

Neoteny

A phenomenon among some salamanders, in which larvae of large size, while still retaining the gills and other larval features, become sexually mature, mate, and produce fertile eggs. In certain lakes of Mexico, only the neotenus larvae are present and are called axolotls. Neoteny occurs in certain species of the family Ambystomidae, especially in *Ambystoma tigrinum* of some localities, and commonly in the large *Dicamptodon ensatus* of the Pacific Coast. It also occurs in some Texas and Oklahoma species of *Eurycea*. Neotenus larvae of *A. tigrinum* can be made to metamorphose to adult form if treated with thyroid extract.

Tracy I. Storer

Nepenthales

An order of flowering plants, division Magnoliophyta (Angiospermae), in the subclass Dilleniidae of the class Magnoliopsida (dicotyledons). The order (also known as Sarraceniales) consists of three well-marked small families: the Droseraceae with about 90 species; the Nepenthaceae with about 80 species; and the Sarraceniaceae with only about 16. The plants characteristically grow in waterlogged soils which are deficient in available nitrogen. They are herbs or shrubs with alternate, simple leaves that



Pitcher plant (*Sarracenia purpurea*). (Henry M. Mayer, National Audubon Society)

are modified for catching insects, from which they absorb nitrogenous nutrients. They have regular flowers characterized by having the perianth and stamens attached directly to the receptacle (hypogynous); having the petals separate (polypetalous) or being without petals (apetalous); having the carpels united to form a compound pistil (syncarpous); and having few to many stamens. The seeds have a well-developed endosperm. The pitcher plants (*Sarracenia*; see **illus.**), sundew (*Drosera*), and Venus' fly-trap (*Dionaea muscipula*) are well-known members of the order. See DILLENIIDAE; MAGNOLIOPHYTA; MAGNOLIOPSIDA; PLANT KINGDOM. Arthur Cronquist

Neper

A unit of attenuation used in transmission-line theory. On a uniform transmission line having waves traveling in only one direction, the magnitudes of voltage E and of current I decrease with distance x traveled, as given by Eq. (1), where E_0 , I_0 , and α are constants.

$$\frac{E}{E_0} = \frac{I}{I_0} = e^{-\alpha x} \quad (1)$$

The attenuation in nepers between the points where E_0 and I_0 are measured and where E and I are measured is given by Eq. (2), in which \ln denotes the natural (or neperian) logarithm.

$$\alpha x = \ln \frac{E_0}{E} = \ln \frac{I_0}{I} \quad (2)$$

The word neper originated from a misspelling of the proper name Napier. One neper equals 8.686 dB,

the decibel being the practical unit of attenuation. See DECIBEL; TRANSMISSION LINES.

Edward W. Kimbark

Nepheline

A mineral of variable composition: in its purest state, NaAlSiO_4 ; often nearly $\text{Na}_3\text{K}(\text{AlSiO}_4)_4$; but generally $(\text{Na}, \text{K}, \square, \text{Ca}, \text{Mg}, \text{Fe}^{2+}, \text{Mn}, \text{Ti})_8(\text{Al}, \text{Si}, \text{Fe}^{3+})_{16}\text{O}_{32}$, where \square represents vacant crystallographic sites, and Ca, Mg, Fe^{2+} , Mn, Ti, and Fe^{3+} are usually present in only minor or trace amounts. The most important variations in nepheline composition are due to crystalline solution of KAlSiO_4 (the mineral kalsilite), and substitution of \square for K. Crystalline solution of KAlSiO_4 in $(\text{Na}, \text{K})\text{AlSiO}_4$ nepheline is limited to about 25 mole % at 25°C (77°F), but increases with increasing temperature (T) to 65–70 mole % at 1000°C (1832°F). Substitution of \square for K can be viewed and treated as crystalline solution of alkali feldspar or (alternatively) “excess silica.” The amount of “excess silica” that can be present stably in any nepheline is very small at $T < 500^\circ\text{C}$ (932°F), and only a few mole % can be dissolved in K-rich nephelines at any temperature. However, crystalline solution of “excess silica” increases steadily in Na-rich nephelines at $T > 500^\circ\text{C}$ (932°F), and at 1000°C (1832°F) as much as 30 mole % can be dissolved in K-free nepheline. See SILICATE MINERALS.

Characteristics. The salient physical properties of nepheline are a Mohs scale hardness of 5.5–6.0; a specific gravity between 2.56 and 2.67; a typically dark gray, light gray, or white color, but also colorless (nepheline is colorless in petrographic thin section); and a vitreous or greasy luster. Nepheline occurs as simple hexagonal prisms or, more commonly, as isolated shapeless grains or irregular polycrystalline masses.

The crystal structure of nepheline is hexagonal, space group $P6_3$, with $a \cong 1.00$ nanometer and $c \cong 0.84$ nm. The structure of NaAlSiO_4 nepheline is a three-dimensional lattice of SiO_4 and AlO_4 tetrahedra with Na atoms housed in the cavities of the framework. In $\text{Na}_3\text{K}(\text{AlSiO}_4)_4$ nepheline, the lattice is distorted, and Na and K occupy cavities (crystallographic sites) of two different sizes—small and large in a 3:1 ratio—with the Na atoms located on the smaller sites and the K atoms positioned on the larger sites. Owing in part to this coincidence of a 3:1 ratio of Na to K and 3:1 ratio of small to large alkali sites, $\text{Na}_3\text{K}(\text{AlSiO}_4)_4$ nepheline is generally regarded as a distinct compound, and it is frequently referred to as ideal nepheline. See CRYSTAL STRUCTURE.

Occurrence. Nepheline is the most abundant feldspathoid mineral; it occurs in a wide variety of SiO_2 -deficient (quartz-free) and alkali-rich volcanic, plutonic, and metamorphic rocks. In volcanic rocks, nepheline occurs chiefly as a primary mineral in phonolites, kenytes, and melilite basalts, and it is the characteristic mineral of nephelinites. Nephelines in alkalic volcanic rocks of all types are frequently

either more Na-rich or more K-rich than $\text{Na}_3\text{K}(\text{AlSiO}_4)_4$, and they commonly contain significant amounts of “excess silica.” In plutonic rocks, nepheline occurs mainly in nepheline syenites and in rocks of the alkali gabbro clan. In contrast to nephelines in volcanic rocks, nephelines in plutonic rocks (and in nepheline gneisses) usually have compositions near $\text{Na}_3\text{K}(\text{AlSiO}_4)_4$, and they contain very little “excess silica.” Nepheline-bearing rocks also occur in the contact metamorphic aureoles of some alkalic intrusions, and in some instances it is evident that these rocks were formed by metasomatism of carbonate-rich country rock.

Commercial uses. Both “pure” (processed) nepheline and nepheline syenite are used as raw materials for the manufacture of glass, various ceramic materials, alumina, pottery, and tile. However, only three countries—Canada, Norway, and Russia—produce significant quantities of commercial-grade nepheline and nepheline syenite. It has been reported that alumina from nepheline-bearing rock is the source of approximately one-sixth of the aluminum produced by Russia. *See* FELDSPATHOID; IGNEOUS ROCKS; METASOMATISM; NEPHELINITE.

James G. Blencoe

Bibliography. M. J. Buerger, G. E. Klein, and G. Donnay, Determination of the crystal structure of nepheline, *Amer. Mineralog.*, 39:805–818, 1954; D. D. Carr (ed.), *Industrial Minerals and Rocks*, 6th ed., Society for Mining, Metallurgy, and Exploration, 1994; W. A. Deer, R. A. Howie, and J. Zussman, *An Introduction to the Rock-Forming Minerals*, 2d ed., 1992; W. A. Deer, R. A. Howie, and J. Zussman, *Rock-Forming Minerals*, vol. 1B: *Disilicates and Ring Silicates* 2d ed., 1986; *Minerals Yearbook, 1988*, vol. 1: *Metals and Minerals*, U. S. Department of the Interior, Bureau of Mines, 1990.

Nephelinite

A dark-colored, aphanitic (very finely crystalline) rock of volcanic origin, composed essentially of nepheline (a feldspathoid) and pyroxene. *See* KALSILITE.

The texture is usually porphyritic with large crystals (phenocrysts) of augite and nepheline in a very fine-grained matrix. Augite phenocrysts may be diopsidic or titanium-rich and may be rimmed with soda-rich pyroxene (aegirine-augite). Microscopically the matrix is seen to be composed of tiny crystals or grains of nepheline, augite, aegirite, and sodalite with occasional soda-rich amphibole, biotite, and brown glass. If leucite becomes the dominant feldspathoid, the rock becomes a leucitite. If calcic plagioclase exceeds 10%, the rock passes into tephrite and basanite. If olivine is present, the rock is an olivine nephelinite (nepheline basalt). Accessories usually include magnetite, ilmenite, apatite, sphene, and perovskite.

Nephelinite and related rocks are very rare. They occur as lava flows and small, shallow intrusives. A great variety of these feldspathoidal rocks is

displayed in Kenya. *See* FELDSPATHOID; IGNEOUS ROCKS.

Carleton A. Chapman

Neptune

The outermost of the four giant planets. Neptune is a near twin of Uranus in size, mass, and composition. Its discovery in 1846 within a degree from the theoretically predicted position was one of the great achievements of celestial mechanics. Difficulties in accounting for the observed motion of Uranus by means of perturbations by the other known planets led early in the nineteenth century to the suspicion that a new planet, beyond the orbit of Uranus, might be causing the deviation from the predicted path. The difficult problem of deriving the mass and orbital elements of the unknown planet was solved independently in 1845–1846, first by J. C. Adams in Cambridge, England, and then by U. J. Leverrier in Paris. Adams's result did not receive immediate attention, and so it was Leverrier's solution that led to the discovery of Neptune by J. G. Galle, in Berlin, who found the planet on September 23, 1846, only 55' from its calculated position. *See* CELESTIAL MECHANICS; PERTURBATION (ASTRONOMY).

The planet and its orbit. The actual mass and orbit of Neptune differ considerably from the values predicted by Adams and by Leverrier, since both assumed that the mean distance of the planet to the Sun would be that predicted by the Titius-Bode relation, namely, 38.8 astronomical units, whereas it is only 30.1 au or 2.8×10^9 mi (4.5×10^9 km). The eccentricity of the orbit is only 0.009, the second smallest (after that of Venus) among the planets; the inclination is 1.8° ; the period of revolution is 163.7 years; and the mean orbital velocity of Neptune is 3.4 mi/s (5.5 km/s). *See* PLANET.

Through a small telescope, Neptune appears as a tiny greenish disk, with a mean apparent diameter of about 2.4" (the Moon has an apparent diameter of 31'). This corresponds to a linear equatorial diameter of 30,775 mi (49,528 km)—very similar to that of Uranus. The mass of Neptune is 17.15 times the mass of Earth, corresponding to a mean density of 1.64, somewhat above that of its sister planet. This suggests that the proportion of heavy elements is somewhat greater in Neptune than in Uranus. *See* URANUS.

The apparent visual magnitude of Neptune at mean opposition, that is, when closest to Earth, is +7.8, too faint to be seen by the unaided eye. The corresponding albedo is 0.4, a relatively high value characteristic of a planet with a dense atmosphere. Photographs taken through powerful telescopes under excellent visual conditions with special filters and cameras reveal the presence of discrete cloud systems in Neptune's atmosphere, again in contrast to Uranus, whose atmosphere contains fewer cloud systems. *See* ALBEDO.

Atmosphere. Most of what is known about Neptune is the result of the flyby of the planet by

the *Voyager 2* spacecraft in August 1989. The cloud features that were dimly glimpsed from Earth were recorded in great detail (Fig. 1). They included a large dark oval (about the size of Earth), reminiscent of Jupiter's Great Red Spot, as well as the white clouds of condensed methane whose brilliant contrast with the blue-green atmosphere made them visible from Earth. Unlike the Great Red Spot, Neptune's dark oval proved to have a short lifetime, as subsequent observations from Earth showed that it had disappeared. By following the clouds over several weeks, scientists were able to deduce the presence of currents at different latitudes, with the high-latitude winds faster than those near the equator. The reference frame is established by the rotation period of the deep interior, 16h7m, as determined from radio emissions whose intensity variations reflect the rotation of the planet's magnetic field. Storm systems on Neptune can cross latitude lines, moving toward the equator as the dark oval did before disappearing. On Jupiter and Saturn, such latitudinal motion does not occur.

This circulation pattern resembles that of Uranus, despite the different inclinations of the rotational axes of the two planets (that of Neptune is 29.6° , while that of Uranus is 97.9° , and that of Earth is 23.5°), and the fact that Neptune has an internal energy source that releases some 2.7 times the amount of heat absorbed from the Sun, while Uranus has no excess internal heat. The explanation may lie in similar atmospheric opacities. However, the difference in internal heat may be responsible for the greater cloud activity on Neptune than on Uranus.

The atmosphere of Neptune, like those of the other giant planets, is composed predominantly of hydrogen and helium. The relative abundance of methane is enhanced slightly more than on Uranus, between 25 and 40 times the value corresponding to solar abundances of the elements. This gas contributes to Neptune's greenish color by absorbing orange and red light. Ammonia is expected to be present at lower, warmer levels of the atmosphere, but radio observations that probe to these levels find the ammonia to be depleted in comparison with methane. Radio observations from Earth discovered carbon monoxide and hydrogen cyanide in Neptune's upper atmosphere, while infrared spectra have revealed ethane and ethylene.

Magnetic field. The orientation of Neptune's magnetic field is surprisingly similar to that of Uranus. It can be represented by a bar magnet inclined at an angle of 46.8° with respect to the axis of rotation and offset by 0.55 planetary radius. (For comparison, the Earth's field is inclined by only 11° and offset by 0.07 radius.) Because of the offset, the field strength varies from a minimum of less than 0.1 gauss (10^{-5} tesla) in the northern hemisphere of Neptune to a maximum of greater than 1.0 gauss (10^{-4} T) in the southern hemisphere.

This field has trapped a plasma of ionized and neutral gases in the planet's magnetosphere. The maximum plasma density is only 1.4 particles per cubic centimeter (0.09 particle per cubic inch), the lowest

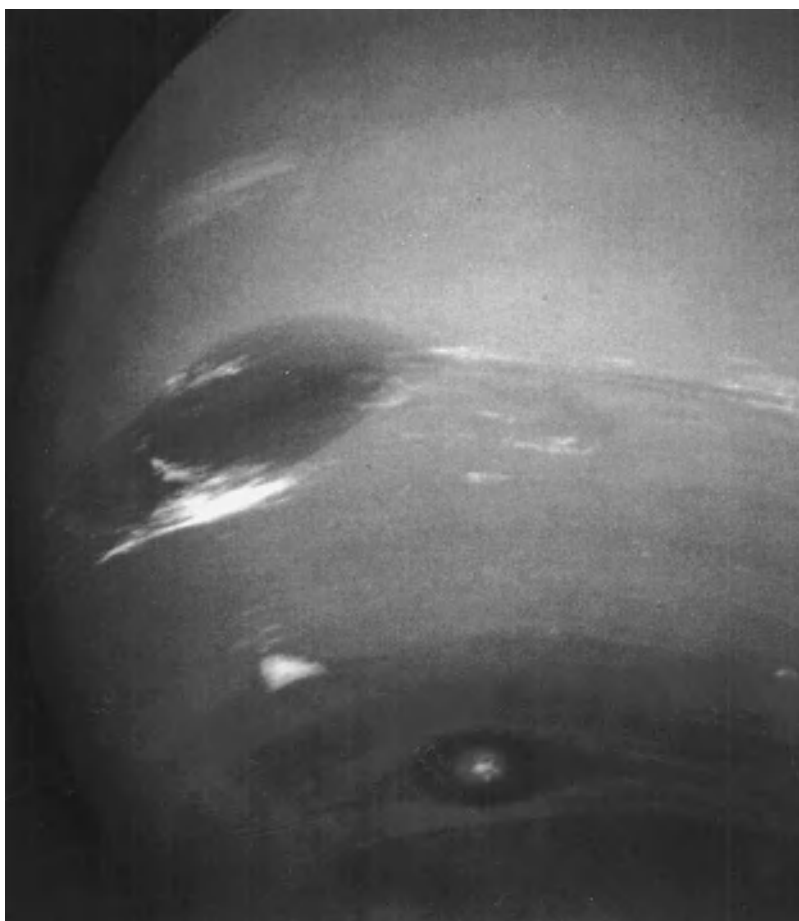


Fig. 1. *Voyager 2* picture of Neptune, showing the huge dark oval at 22° south latitude, with bright, white methane clouds above it. Farther to the south is a triangular cloud system called the scooter because of its high relative velocity, and still farther south is another dark oval with a bright core. (NASA)

of any of the giant planets. There are ions of molecular hydrogen and helium as well as atomic hydrogen and heavier ions that have probably escaped from Triton. These should include nitrogen from the satellite's atmosphere.

A faint aurora was observed on Neptune, caused by charged particles from the radiation belts (the trapped plasma) spiraling around planetary magnetic field lines and bombarding the planet's atmosphere. Because of the unusual orientation of the field, the maximum auroral activity occurs at midlatitudes rather than near the poles as it does on Earth. See AURORA; MAGNETOSPHERE; VAN ALLEN RADIATION.

Satellites. Before the *Voyager* encounter, only two satellites of Neptune were known, both in highly irregular orbits. Triton was discovered visually by W. Lassell in 1846. It is moving in a retrograde direction around Neptune with a period of 5.9 days in a nearly circular orbit. Nereid was found in 1949 as a result of a photographic search by G. P. Kuiper. It has the most eccentric orbit of any known satellite. The eccentricity e measures the flattening of the ellipse; a circle has $e = 0$, while a parabola, representing an unclosed orbit, has $e = 1$. While the orbits of most satellites and planets have eccentricities close to zero, Nereid has $e = 0.75$. In sharp contrast to these two bodies, the

TABLE 1. Neptune satellites*

| Satellite name | Distance to Neptune, 10 ³ mi (10 ³ km) | Sidereal period, h | Diameter, mi (km) | Albedo |
|----------------|--|--------------------|-------------------|--------|
| Naiad | 29.9 (48.2) | 7.1 | 37 (60) | 0.06? |
| Thalassa | 31.1 (50.1) | 7.5 | 50 (80) | 0.06? |
| Despina | 32.6 (52.5) | 8.0 | 93 (150) | 0.06? |
| Galatea | 38.5 (62.0) | 10.3 | 99 (160) | 0.06 |
| Larissa | 45.7 (73.5) | 13.3 | 118 (190) | 0.06 |
| Proteus | 73.1 (117.6) | 26.9 | 261 (420) | 0.06 |
| Triton | 220.5 (354.8) | 141.0 | 1681 (2705) | 0.076 |
| Nereid | 3426 (5513) | 8643 | 211 (340) | 0.16 |

*Only satellites with well-determined orbits as of March 1, 2004, are listed.

six satellites discovered by the *Voyager* cameras all have very regular orbits: in the plane of the planet's equator and nearly circular. They are all close to the planet (Table 1). In 2002 and 2003, five irregular satellites were discovered from Earth at large distances from the planet; their orbits must be better defined before they receive names. The search for additional satellites continues. See ELLIPSE; ORBITAL MOTION; SATELLITE.

Triton has a tenuous atmosphere containing nitrogen, methane, and carbon monoxide. These three gases plus carbon dioxide and water are also present as frozen ices on the satellite's surface. *Voyager* revealed that this remarkable object has a diameter of only 1681 mi (2705 km), making it considerably smaller than the Earth's Moon (2086 mi or 3476 km). The surface temperature of Triton is $-391 \pm 7^\circ\text{F}$ ($38 \pm 4\text{ K}$). The size of this satellite, as well as the temperature and composition of its surface, makes Triton very similar to Pluto. See PLUTO.

Triton's surface has an appearance quite different from that of any other satellite. The scarcity of impact craters means that the surface is geologically young. There are long intersecting valleys and ridges crossing various types of terrain. The illuminated part of the satellite at the time of the *Voyager* encounter was the southern hemisphere. Most of it appeared to be covered with layers of highly reflective ice, on which darker splotches appeared, many clearly organized by near-surface winds. Subsequent Earth-based observations of Triton's near-infrared spectrum revealed absorptions of solid carbon monoxide, carbon dioxide, and water ice. There were earlier (pre-*Voyager*) indications that ices of methane and nitrogen were present. Nitrogen ice is the predominant surface constituent, with the other ices contributing less than 1–2%. Near the equator, there is a bluish deposit that is probably fresh nitrogen ice, while at higher latitudes the surface resembles the skin of a cantaloupe.

Triton's atmosphere has a surface pressure of only 1.6 ± 0.3 pascals or 16 ± 3 microbars (Earth's atmospheric pressure is approximately 10^5 Pa or 1 bar), dominated by molecular nitrogen. Only 1 part in 10,000 is methane, while carbon monoxide must be less than 4%. Yet reactions in this tenuous atmosphere (and in the surface ices) apparently form the organic compounds that produce the deposits of

dark material and give the surface its characteristic pinkish color.

Perhaps the most unusual aspect of Triton is the presence of eruptive plumes at several places on its surface. These plumes consist of columns of dark material ejected some 5 mi (8 km) upward into the atmosphere, where they form small clouds that are then blown into narrow wind trails extending hundreds of miles from their sources. These plumes may be powered by a solid greenhouse effect within Triton's icy surface, in which solar radiation that penetrates the ice encounters dark material that it warms. This causes the surrounding nitrogen ice to sublime until sufficient gas pressure is produced to cause gas to break through to the surface, forming the plume. The dark material carried upward by these jets contributes to the hazes seen in Triton's atmosphere.

Rings. Earth-based observations of Neptune when it passed in front of distant stars indicated the presence of material in orbit about the planet. In some cases, the light from the star would briefly disappear before Neptune reached it, but this behavior would not be repeated on the other side of the planet as would be expected for a planetary ring. Sometimes, no dimming of the star's light on either side of the planet was observed. These observations led to the idea that Neptune might be surrounded by incomplete rings or arcs. See OCCULTATION.

The *Voyager* cameras showed that in fact there are three well-defined, complete rings around Neptune, accompanied by a sheet of material that itself constitutes a broad ring. In order of increasing distance from the planet, the three discrete rings have been designated Galle, Leverrier, and Adams, while the sheet of material is known as Lassell, and a concentration of particles within it is called Arago (Table 2). The outermost of these, the Adams ring, contains three concentrated clumps of material known as arcs (Fig. 2), rather like sausages strung on a wire, and these plus a chance occultation by one of the inner satellites were apparently responsible for the confusing ground-based observations. A fourth, much smaller arc is barely visible in the best *Voyager* image. The outer two discrete rings are very narrow, reminiscent of the rings of Uranus, with an average width of 9 mi (15 km).

The confinement of these narrow rings is commonly assumed to require the presence of small

TABLE 2. Neptune ring data

| Feature | Distance from Neptune's center, 10 ³ mi (10 ³ km) | Comments |
|--|---|--|
| Pressure level of 10 ⁵ Pa (1 bar) in Neptune's atmosphere | 15.4 (24.8) | Equatorial radius of Neptune |
| Galle | 26.1 (42.0) | Outer edge of this ring |
| Leverrier | 32.9 (53.0) | High dust content |
| Lassell | 33.37 (53-59) | Broad ring |
| Arago | 35.5 (57.2) | Concentration within Lassell |
| Adams | 39.1 (63.0) | 9 mi (15 km) wide; contains three arcs |

shepherding satellites. Galatea and Despina orbit, respectively, just inside the outer two narrow rings, but the corresponding outer shepherds have not been found. Similarly, the persistence of the three arcs within the outer ring remains an enigma. In the absence of some gravitational control, the material in these arcs would be expected to spread out around the planet, simply contributing to the ring itself.

Origin and evolution. Like the other giant planets, Neptune is thought to have formed in a two-stage process. First a large core of solid material accumulated, growing as a result of collisions with smaller so-called planetesimals. This core was dominated by ices, although it contained rocky material as well. As the core grew to the size of several Earth masses, it developed an atmosphere, the result of impact vaporization of the icy bodies that were crashing into it. The atmosphere consisted of gases such as molecular nitrogen, methane, and carbon monoxide. As the core grew, it began to attract gas from the surrounding solar nebula, which added hydrogen and helium to the atmosphere. The process stopped when the supply of materials in the vicinity of the growing planet was exhausted.

In the case of Jupiter and Saturn, there was enough nebular gas available to produce huge planets with deep atmospheres dominated by hydrogen and helium. Even in these atmospheres, however, the heavy elements were enriched. At the great distances where Uranus and Neptune formed, the solar nebula evidently had less material to offer the growing

planets, with the result that these two objects do not have such deep atmospheres. In other words, they exhibit a much higher proportion of core mass to total mass than Jupiter and Saturn do. This difference in the proportion of core to atmosphere is reflected in the higher proportion of carbon to hydrogen in the atmospheres of Uranus and Neptune, exemplified by the large relative abundance of methane discussed above. See JUPITER; SATURN.

The six inner regular satellites of Neptune resemble the ten that *Voyager* found around Uranus, in size, albedo, and orbital characteristics. In both cases, it is thought that these bodies formed out of the planetary subnebula, in much the same way that the planets themselves formed from material in the solar nebula in orbit about the Sun. Triton and Nereid must have had a different beginning, however. Both are probably captured objects that accreted independently. In the case of Triton, the process that led to capture evidently dissipated enough energy to melt the satellite completely, obliterating the record of impact craters which should have been left on the surface. Evidently Triton, like Pluto, is a large icy planetesimal, an object that grew from collisions of smaller icy bodies that may have been roughly identical to the comets observed today. See COMET; SOLAR SYSTEM.

Tobias C. Owen

Bibliography. J. K. Beatty, C. C. Petersen, and A. Chaikin (eds.), *The New Solar System*, 4th ed., Cambridge University Press, 1999; D. P. Cruikshank (ed.), *Neptune and Triton*, University of Arizona Press, 1995; M. Grosser, *The Discovery of Neptune*, Harvard University Press, 1962, reprint, Dover Publications, 1979; D. Morrison and T. Owen, *The Planetary System*, 3d ed., 2003; D. A. Rothery, *Satellites of the Outer Planets: Worlds in Their Own Right*, 2d ed., Oxford University Press, 1999; F. W. Taylor, *The Cambridge Photographic Guide to the Planets*, Cambridge University Press, 2001.

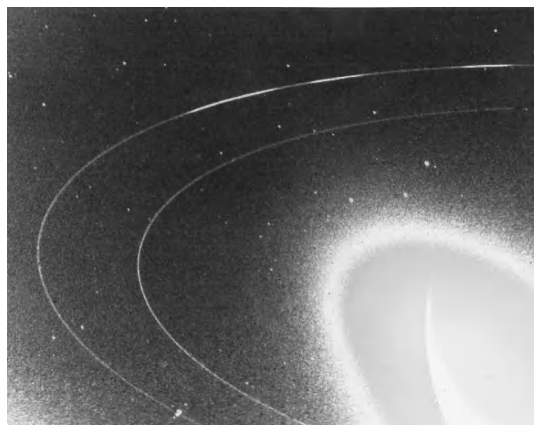


Fig. 2. The two outer discrete rings of Neptune (Adams and Leverrier), with the three arcs of concentrated material in the outer ring. (NASA)

Neptunium

A chemical element, symbol Np, atomic number 93. Neptunium is a member of the actinide or *5f* series of elements. It was synthesized as the first transuranium element in 1940 by bombardment of uranium with neutrons to produce neptunium-239. The lighter isotope ²³⁷Np, a long-lived alpha emitter with half-life

| | | | | | | | | | | | | | | | | | | | |
|-------------------|----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|----|----|---|----|
| 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 | | | | | | | |
| 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 | | | | |

2.14×10^6 years, is particularly important chemically. See PERIODIC TABLE.

Neptunium metal is ductile, low-melting (637°C or 1179°F), and in its alpha form is of high density, 20.45 g/cm³ (11.82 oz/in.³). The chemistry of neptunium may be said to be intermediate between that of uranium and plutonium. Neptunium metal is reactive and forms many binary compounds, for example, with hydrogen, carbon, nitrogen, phosphorus, oxygen, sulfur, and the halogens. See ACTINIDE ELEMENTS; NUCLEAR CHEMISTRY; TRANSURANIUM ELEMENTS.

Robert A. Penneman

Bibliography. R. Guillaumont et al. (eds.), *Update on the Chemical Thermodynamics of Uranium, Neptunium, Plutonium, Americium and Technetium*, 2003; S. Hofmann, *On Beyond Uranium: Journey to the End of the Periodic Table*, 2000; J. J. Katz, G. T. Seaborg, and L. R. Morss (eds.), *The Chemistry of the Actinide Elements*, 2 vols., 2d ed., 1986.

Nerve

A group of nerve fibers coursing together as a bundle in the peripheral nervous system. The individual fibers are covered by Schwann cells, many of which contain large amounts of myelin, which makes the nerve appear shiny white. The nerve fibers with their Schwann cell sheaths are held together by connective tissue. In most nerves, some of the fibers are sensory (carrying information to the central nervous system) and some are motor (carrying information from the central nervous system to peripheral glands and muscles). When both sensory and motor fibers are in a nerve, it is called a mixed nerve.

In the central nervous system (brain and spinal cord) a group of nerve fibers running together is called a tract and has different structural and functional properties than a nerve. Glial cells, not Schwann cells, form the sheaths of tract fibers, and there is no connective tissue holding the bundle together. Whereas most nerves are mixed, there is functional segregation in the central nervous system so that most tracts have only one functional type of fiber. See MOTOR SYSTEMS; NERVOUS SYSTEM (VERTEBRATE).

Douglas B. Webster

Nervous system (invertebrate)

All multicellular organisms have a nervous system, which may be defined as assemblages of cells specialized by their shape and function to act as the major coordinating organ of the body. Nervous tissue underlies the ability to sense the environment, to move and react to stimuli, and to generate and control all behavior of the organism. Compared to vertebrate nervous systems, invertebrate systems are somewhat simpler and can be more easily analyzed. Invertebrate nerve cells tend to be much larger and fewer in number than those of vertebrates. They are also easily accessible and less complexly organized; and they are hardy and amenable to revealing experimental manipulations, such as changes in the composition and temperature of the fluids surrounding them. However, the rules governing the structure, chemistry, organization, and function of nervous tissue have been strongly conserved phylogenetically. Therefore, although humans and the higher vertebrates have unique behavioral and intellectual capabilities, the underlying physical-chemical principles of nerve cell activity and the strategies for organizing higher nervous systems are already present in the lower forms. Thus neuroscientists have taken advantage of the simpler nervous systems of invertebrates to acquire further understanding of those processes by which all brains function. See NERVOUS SYSTEM (VERTEBRATE).

Fundamental Definitions

The average nerve cell, representative of any multicellular organism, may be distinguished by larger-than-average amounts of the cell organelles and metabolic machinery used in protein and lipid synthesis and in the manufacture of special chemical substances which are released as chemical messengers. These compounds are used at specialized connections with other nerve or effector cells called synapses. Nerve cells also have elaborate extensions of the cell body called axons and dendrites; and they have highly specialized cell membranes capable of generating electrical potentials. Nerve cells are almost always intimately associated with other cells, called glia, which aid in mechanical support of nervous tissue and serve to complement nerve metabolism and to regulate the fluid environment of nervous tissue. It appears that nervous tissue in all metazoans is derived, in embryogeny, from the ectoderm. See NEURON.

Electrical potentials. The two most distinctive characteristics of nerve cells are the ability to create electrical potentials and to make synaptic connections. Nerve cells and other excitable cells, such as muscle fibers, can establish electrical potentials across their cell membrane (Fig. 1). This capability is due to the membrane's selective permeability to certain dissolved ions, such as potassium (K⁺), sodium (Na⁺), and chloride (Cl⁻), which have different concentrations inside and outside the cell.

Most nerve cells have steady resting membrane potentials (RMP) of from 50 to 90 mV, with the inside

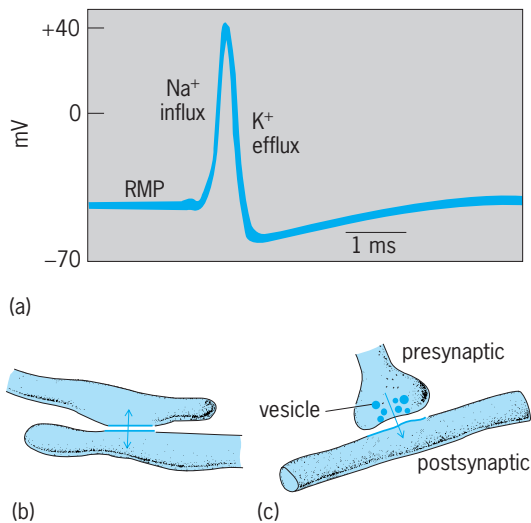


Fig. 1. Electrical potentials and synaptic arrangements. (a) An action potential recorded from the giant axon of the squid (after A. L. Hodgkin and A. F. Huxley, *Resting and action potentials in single nerve fibres*, *J. Physiol.*, 104:176–195, 1945). (b) Electrical synapse (two-way). (c) Chemical synapse (one-way).

of the cell negatively charged with respect to the outside. Furthermore, the permeability of excitable cells to different ions can change rapidly, and this leads to the production of fast voltage changes called action potentials (nerve impulse), which can regenerate themselves, and propagate along the cell processes. During an action potential, the cell's internal potential changes very quickly from its negative resting value, through zero potential to an internal positive value, and then rapidly back to the resting level. Action potentials last only a few milliseconds; they are quite stereotyped in their size and shape, and they represent the way in which nerve cells carry information along their processes to various destinations. In most nerve cells the depolarizing phase of the action potential is caused by the inward flow of sodium ions across the membrane, whereas the repolarizing phase results from the outward flow of potassium ions. Energy-requiring metabolic "pumps" in the membrane help to maintain or restore concentration gradients of these ions. *See* BIOPOTENTIALS AND IONIC CURRENTS.

Synaptic connections. Nerve cells communicate with one another and with their target effector cells (muscle, glands) in a highly organized way at contact points called synapses. No other cell or tissue type is known to have the elaborate scheme of direct interconnections evolved by nerve cells. In a comparatively few cases, nerve cells make electrical synapses with one another (Fig. 1b). In such cases the membranes of the two cells come very close together, and special channels are formed in such a way that electric currents (including action potentials) can pass directly from one cell to the other, and in either direction. Electrical synapses are thought, in general, to be the more primitive type of communication; they are usually seen in lower animals, in less elaborately developed tissues (epithelia and embryonic

tissues), or in certain specialized contacts, as in some muscle.

The more common synaptic connection is called the chemical synapse (Fig. 1c). Here one cell, the presynaptic element, has a specialized ending (typically the axon) in close apposition to a postsynaptic cell process (usually a dendrite or cell body). The presynaptic ending is filled with small rounded vesicles (packets) containing a chemical transmitter substance. When an action potential invades the presynaptic ending, packets of transmitter substance are released from the presynaptic ending. The transmitter diffuses across the narrow cleft between the two cell processes, and interacts with specialized receptors on the postsynaptic membrane. This causes a small voltage change, called a synaptic potential, which then influences the probability of the postsynaptic neuron to produce an action potential. Synaptic potentials may be excitatory or inhibitory in their action, depending on whether they make the postsynaptic element more or less likely to produce an action potential. One nerve cell and its processes may receive many synaptic inputs from other cells, and one cell may connect via branches of its axon to many others. Chemical synapses confer direction of information flow at a synapse, that is, from presynaptic to postsynaptic element; and the size and efficacy of chemical synaptic potentials can be altered by a variety of physiological mechanisms. *See* SYNAPTIC TRANSMISSION.

Nerve cell categories. Nerve cells may be categorized into three basic functional types: sensory neurons, motor or effector neurons, and interneurons. Sensory neurons are specialized to transduce specific types of environmental signals, such as touch, temperature, or visual cues, into electrical information (action potentials), and to transmit this into the central nervous system via their axons. Motor or effector neurons are central nervous system neurons that send an axon to the periphery to innervate muscles or glands. Motor neurons are the final common pathway responsible for all movements and activities of animals, including locomotion, reflex movements, and gland secretions. Interneurons constitute the greatest number of nerve cells in the central nervous system and do not send processes into the periphery, but serve to interconnect sensory and motor cells and other interneuronal nerve cells. Interneurons represent the neural basis for the organization of complex activities and functions. In ascending the phylogenetic scale, the number of interneurons increases greatly. *See* MOTOR SYSTEMS; SENSATION.

Invertebrate and vertebrate nerve cells share all the properties described above; they differ more in quantity, or degree, than in qualitative features. Aside from differences in size and numbers, the most striking difference is that many invertebrate neurons have a unipolar shape, whereas most vertebrate neurons are multipolar (Fig. 2). Unipolar refers to a nerve cell body with one process extending away from it. Multipolar neurons, in contrast, have several processes coming off the cell body. One subset of

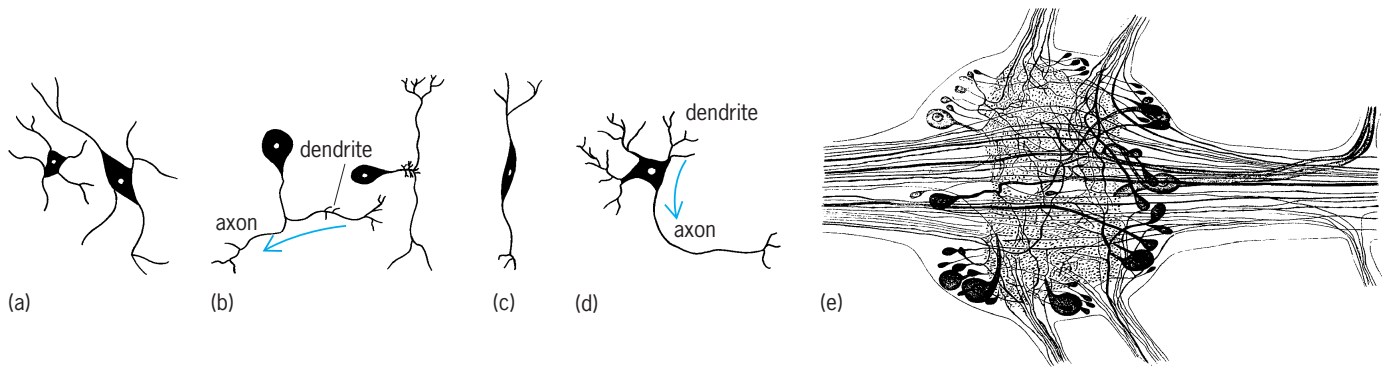


Fig. 2. Neuron types. (a) Multipolar, isopolar neurons (primitive invertebrate). (b) Unipolar, heteropolar neurons (common invertebrate); arrow indicates direction of information flow. (c) Bipolar neuron (common invertebrate). (d) Multipolar, heteropolar neuron (common vertebrate, advanced invertebrate); arrow indicates direction of information flow. (e) Ganglion of unipolar neurons in crayfish. The unipolar cell bodies, surrounded by connective tissue and glia, form a ring around a central neuropile composed of the cell processes (axons and dendrites). Nerve tracts enter or leave the neuropile. Synaptic interactions and integration are carried out in the neuropile. (After G. Retzius, *Zur Kenntniss des Nervensystems der Crustacean*, *Biol. Untersuch.*, NF 1:–50, 1980)

multipolar cells are bipolar, comprising cells with two stem processes.

Regardless of their configuration, nerve cell processes may function in one of two ways. In advanced neurons one set of processes, called dendrites, are the input or receiving processes of the cell, while another process, the axon, functions as the output process. Such nerve cells are said to have preferred or dynamic polarity; that is, they channel information flow in one functional direction. This situation is defined as heteropolar. Most invertebrate neurons are unipolar, with a functional, specified direction of information flow (heteropolar). Synaptic inputs occur at certain locations (dendritic), generating action potentials that propagate along the axon. Unipolar neurons almost never receive synaptic inputs on the cell body. Many sensory cells may be of the bipolar heteropolar type. Most vertebrate and many advanced invertebrates have multipolar, heteropolar nerve cells. In contrast, more primitive nerve cells have undifferentiated processes with no preferred direction of information flow. They might receive or transmit signals in any direction, and action potentials might be generated at various points on the nerve cell and propagate in all directions. Such cells have a uniform lack of preference for direction of information flow, and are said to be isopolar. In multipolar, isopolar neurons, synapses (usually electrical) occur at many locations on the processes; information flow may begin at any location and spread uniformly.

An additional general contrast between invertebrate and vertebrate nervous systems is that invertebrates tend to have more neurons displaced to the periphery (outside the central nervous system) and to perform more integrative and processing functions in the periphery. Vertebrates perform almost all their integration within the central nervous system, using interneurons. For example, invertebrate muscles are often controlled through an innervation containing both excitatory and inhibitory motor neurons which cause direct contraction or relaxation, respectively, of the muscle. Vertebrates only have

excitatory motor neurons. To prevent muscle contraction (or to allow relaxation), they suppress the excitatory motoneuron with inhibitory interneurons in the central nervous system.

Invertebrate nervous systems also seem to have a greater potential for regrowth, regeneration, or repair after damage than do vertebrate nerve cells. Many invertebrates continue to add new nerve cells to their ganglia with age; vertebrates, in general, do not. Only vertebrate neurons have myelin sheaths, a specialized wrapping of glial membrane around axons, increasing their conduction speed. Invertebrates tend to enhance conduction velocity by using giant axons, particularly for certain escape responses.

Description and Phylogeny

Most phylogenetic schemes suggest that the evolution of invertebrates has been accompanied by an increase in body size and complexity (having more varied body parts, organs, and functions to coordinate), and the development of more elaborate behavioral patterns. Not surprisingly, the size and complexity of the nervous system parallel these developments. The following examples of nervous systems represent only major invertebrate phyla; phylogenetically important groups, especially well-studied animals, and developmental trends are highlighted.

Protozoa. Membranes with maintained, but variable, electrical potentials across them are already evident in these single-celled invertebrate organisms (Fig. 3a). The ciliated protozoans have been most studied, particularly *Paramecium*. These organisms have an internally recorded resting potential of about -20 to -40 mV; and this potential, like the resting potential of nerve cells, is mainly due to the membrane's selective permeability to K^+ ions. In response to mechanical distortion, the membrane may generate an action potential which causes a change in the direction of ciliary beating. The inward current of the action potential is carried by calcium ions, is graded, and is similar to voltage-dependent calcium permeability changes seen in some crustacean muscles and

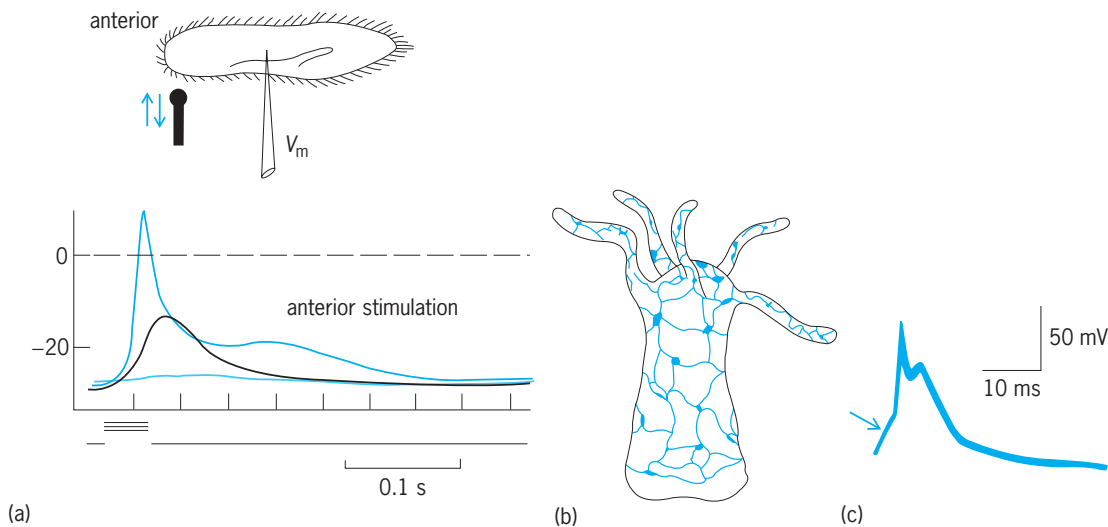


Fig. 3. Electrical responses and nervous systems in the lowest invertebrates. (a) Recording from *Paramecium* illustrates a graded action potential which uses Ca^{2+} ions in the upswing phase and K^{+} ions during repolarization; three responses of varying size were produced by mechanical taps of increasing intensity to the front of the animal; V_m = membrane potential recorded with a glass microelectrode (after Y. Naitoah and R. Eckert, *Ionic mechanisms controlling behavioral responses of Paramecium to mechanical stimulation*, *Science*, 164:963–965, 1969). (b) Hydrozoan showing nerve net underlying the epidermis. (c) Intracellular microelectrode recording from a large neuron of the subumbrellar nerve net of the jellyfish *Carybdea*; this action potential was generated by a synaptic potential (arrow); RMP approximately -70 mV (after R. A. Satterlie, *Central control of swimming in the cubomedusan jellyfish Carybdea rastonii*, *J. Comp. Physiol.*, 133:357–367, 1979).

in the presynaptic axon terminals of most animals. Moreover, the entry of Ca^{2+} ions, itself, causes an increased outward flux of K^{+} ions, which brings the membrane potential back toward the resting value. Such Ca^{2+} -dependent potassium currents have also been described in nerve cells from many organisms, notably molluscan and mammalian central nervous system neurons.

The calcium-potassium sequence underlying the active response of *Paramecium* is rather analogous to the sodium-potassium scheme used by nerve cells to make an action potential, although in the latter the control mechanisms are more elaborate and differ in important details. There are numerous examples of invertebrate neurons in all phyla which are intermediate in their use of Ca^{2+} instead of Na^{+} to carry action potential currents. A variety of muscle cells in invertebrates and vertebrates also have this property. See PROTOZOA.

Porifera. Whether or not the most primitive phylum of multicellular animals, the sponges, have clearly defined nerve cells is uncertain. However, some nerverlike functions may be present to account for the ability of these organisms to make local body movements, or to alter the size of the canals which channel the water flow through their tissues. See PORIFERA.

Cnidaria. The first true nerve cells appear in this phylum (also called the Coelenterata), which includes the hydrozoan polyps, jellyfishes, anemones, and corals (Fig. 3b and c). The most intensive studies of the nervous system have been done on *Hydra* and *Tubularia*, and on several genera of jellyfish and anemones.

The phylum Cnidaria is characterized by radial symmetry, a body wall composed of only two true

cell layers, and a central digestive cavity with a single opening, the mouth, which is surrounded by tentacles. The two epithelial body layers, an outer epidermis and an inner gastrodermis, are separated by a mesoglea of variable thickness composed of a few primitive cells and a gelatinous matrix.

The nervous system of the typical cnidarian is mostly dispersed in a so-called nerve net, a thin layer of nerve cells lying just beneath the epidermis at the border of the mesoglea. The nerve cells are multipolar (often bi- or tripolar) and isopolar (Fig. 2a), and their cell bodies are scattered throughout most areas of the body. The undifferentiated nerve processes spread out weblike to interconnect with one another, with sensory neurons, and with contractile epitheliomuscle cells. True muscle cells are not present. Many cnidarians may have additional nerve nets, of varying complexity and density, lying beneath the gastrodermis, around the mouth, along the tentacles, or concentrated in rings around the bell margins of medusa-form animals. Jellyfish and anemones may have two or more epidermal nerve nets, one specialized for slow conduction, the other for faster conduction.

The multipolar, isopolar neurons of the nerve nets have a negative resting potential and can generate and propagate action potentials throughout the net (Fig. 3c). Most of the nerve net neurons are interconnected by electrical synapses, and activity generated at any point in the net spreads throughout the entire system. At least a few specialized sensory neurons can be identified in most groups, including some responsive to light, chemical cues, and mechanical stimuli (touch or traumatizing injury). Some jellyfish have highly developed eyespots and statocysts; these are usually evenly distributed

around the margin of the bell. Many sensory neurons may be sensitive to more than one kind of stimulus (polymodal). Activity may be generated in the net by sensory-cell input or by specialized pacemaker neurons which can produce action potentials endogenously and repetitively. Discrete motor neurons are probably not present in cnidarians; rather, input to the contractile epitheliomuscle cells is from certain neurons of the nerve net which contact both effector cells and other net neurons.

Electrical impulses, correlated with various behaviors but conducted from cell to cell within the epithelium (epithelial conduction), have been widely observed in cnidarians. Epithelial conduction operates in parallel with the nerve net, and the systems interact.

Several lines of evidence indicate that cnidarians have a variety of chemical synapses. Distinct ones have been identified between nerve-net neurons (presynaptic) and epitheliomuscle cells and nematocyst cells (postsynaptic elements). Many nerve-net neurons are interconnected by what appear to be two-way chemical synaptic junctions; that is, both neural elements have vesicles and are, therefore, pre- and postsynaptic to one another. These symmetrical synapses can transmit impulses in both directions, unlike the neurons of higher animals that are directionally polarized. The presence of chemical synapses probably underlies some functional features long observed in cnidarian physiology. Repeated stimulation is often necessary to obtain a response from animals of this group; responses are known to facilitate, that is, to grow larger, with repetition. This type of variability, or plasticity, is a common characteristic of chemical synaptic transmission. Thus the ability to augment or alter responsivity in a neural circuit because of use-dependent changes in synaptic function is apparently already present in these earliest nerve circuits. Although the nervous system is primitive, a variety of neuropeptides and neurotransmitters have been identified in cnidarians. *See* CNIDARIA.

Platyhelminthes. The flatworms mark the beginning of bilateral symmetry in the animal kingdom, and they also illustrate the clear onset of organizational principles which will guide the elaboration of neural tissue in all higher groups. The well-studied, free-living turbellarian flatworms are used as a typical example of this phylum, since the other major classes are highly modified as parasites. While many of the organs of the acoelomate turbellarians are reminiscent of the cnidarians (for example, a digestive tract with a single opening), marked progression is seen in the complexity of most organs. The flatworm nervous system retains, of course, those fundamental elements of neural function, action potentials and chemical synaptic potentials, but few studies at the cellular level have been made to provide details.

The most important changes and new features seen in neural organization include the following (Fig. 4a). (1) While some nerve nets or plexuses

are present in most species, the bulk of the nervous system has begun to coalesce into distinct clusters or ganglia within the mesenchyme away from the epidermis. This represents the process of centralization. The organization of ganglia is already that basic to higher forms: a central region of nerve processes, the neuropile, surrounded by a ring of unipolar cell bodies (Fig. 2e). Thus, unipolar, heteropolar neurons make their appearance at this time. Bipolar (often sensory) and multipolar cells are also present. (2) Elongated, longitudinally directed bundles of nerve axons appear which can carry information between or among ganglia. These nerve cords are usually arranged in several symmetrical pairs running like barrel staves down the length of the animal. There are often commissures or crossovers of axon bundles between parallel longitudinal bundle pairs. (3) There is a tendency for the neural ganglia to be concentrated anteriorly as a primitive "brain" in association with anteriorly located sense organs such as statocysts and eyespots. This is called encephalization. Thus the preferred forward locomotion of the bilateria is naturally coupled with anteriorly placed special senses and the headward concentration of neural elements. (4) Definitive unimodal sensory cells, distinct motoneurons, and an increased proportion of interneurons are present. (5) Glial cells are discovered.

The densely packed and complexly interconnecting neural processes in the ganglionic neuropile are correlated with the appearance of a primitive sort of behavior modification akin to memory and learning. Numerous studies have shown that planaria can be conditioned with light to alter their path of movement. This modified, conditioned behavior is retained in organisms which have fully regenerated from the two halves of a previously trained worm which was cut in half. *See* PLATYHELMINTHES.

Aschelminthes. This group of several minor phyla includes the pseudocoelomate roundworms (nematodes) and other interesting small aquatic invertebrates (for example, rotifers). These organisms are intermediate, in their morphology and function, between the flatworms and the annelids. An important advanced feature of the group is that the gut has become tubelike, with a separate mouth and anus. The location of the mouth and associated organs in the head region augments the encephalization of the central nervous system.

Studies have been conducted on the nervous system of the nematodes, notably *Caenorhabditis elegans* and *Ascaris*. *Caenorhabditis elegans* is a small, free-living soil nematode that has been particularly useful in recent studies of the regulation of development and the molecular genetics of nervous tissue. Its hermaphrodite contains exactly 959 somatic cells, and the lineage of these cells, including all neurons, has been traced. The genome of this worm has been entirely sequenced. *See* GASTROTRICHA; KINORHYNCHA; NEMATA (NEMATODA); ROTIFERA.

Annelida. The annelid worms introduce the phenomenon of segmentation (metamerism).

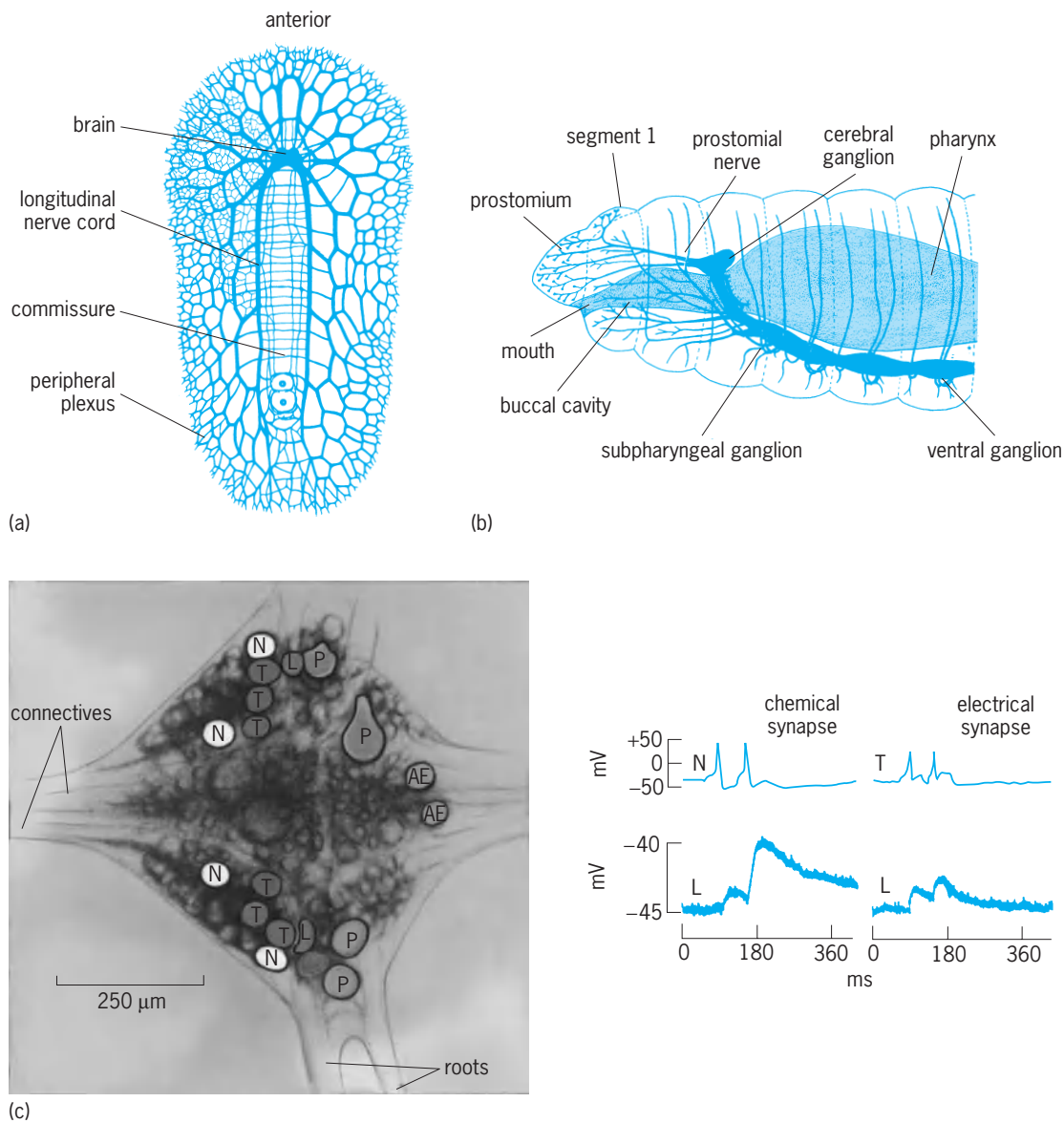


Fig. 4. Nervous systems of the intermediately advanced invertebrates (various worms). (a) Turbellarian flatworm (*Notoplana*) illustrating nerve cells coming together to form ganglia and nerve tracts (centralization) and nerve cells concentrated in the head region (brain) near sensory structures (encephalization) (after D. Hadenfeldt, *Das Nervensystem von Stylochoplana maculata and Notoplana atomata*, *Z. Wiss. Zool.*, 133:586–638, 1929). (b) Earthworm illustrating circumesophageal brain ganglia, ventral nerve cord, and segmental ganglia (after E. E. Ruppert and R. D. Barnes, *Invertebrate Zoology*, 6th ed., Saunders College Publishing, 1994). (c) Micrograph of segmental ganglion from the leech *Hirudo* showing identified sensory and motor neurons. Sensory cells: N = noxious receptor; T = touch receptor; P = pressure receptor. Motor cells: AE = annulus erector cells; L = motor neuron for longitudinal muscle. Connectives are the ventral nerve cord. The electrical records show excitatory synaptic potentials produced in motor neuron L by a presynaptic action potential produced in sensory cell N (chemical synapses) or T (electrical synapses). Note facilitation of second chemical synaptic potential (after J. G. Nicholls and D. Purves, *A comparison of chemical and electrical synaptic transmission between single sensory cells and a motoneuron in the central nervous system of the leech*, *J. Physiol.*, 225:637–656, 1972).

Segmentation is a serial repetition of virtually identical body parts along the longitudinal axis of the body; it is most obvious in the trunk and tail of annelids. Anterior segments are typically fused to form a head region. This kind of organization is evolutionarily inherited by the arthropods, though in modified form. In both the annelids and arthropods segmentation has had a fundamental impact on neural arrangement and activity. Chordate segmentation evolved independently of that in annelids, but the

functional consequences are the same. The notion that mollusks descended from segmented ancestors is, at best, controversial; in any case, metamerism is not an obvious feature of any extant mollusk.

In annelids, the well-developed mesoderm hollows out in development to produce a capacious coelom, divided into a series of compartments by membranous septa, and also produces a muscular gut. The coelom plays a critical role in reproduction and excretion and also shapes the structure and

arrangement of the organs of the circulatory, digestive, and nervous systems.

Centralization and encephalization have progressed to the extent that the head contains a distinct and highly organized brain composed of large, fused anterior segmental ganglia (Fig. 4*b*). The major brain component lies dorsal to the anterior digestive organs, and it sends and receives numerous nerve bundles to anterior sense organs and mouthparts. Major nerve trunks leave the dorsal brain mass bilaterally, encircle the esophagus, and connect to other head ganglia of varying number and complexity. From these anterior ganglia, a pair of long nerve trunks, fused to varying degrees, travel back through the length of the animal in a ventral position as the ventral nerve cord, an important organizational feature retained in the arthropods. Typically, in each body segment, a comparatively small segmental ganglion (few hundred neurons) is present. These ganglia innervate the segmental musculature and body organs, and they are virtually identical from segment to segment. Neurons of all functional types are often located on the contralateral (opposite) side of the body from the peripheral organs innervated. Thus, many cells send their processes across the midline (decussate) through commissures in the brain or at the segmental level.

Elaborate sense organs and many types of highly specific unimodal sensory cells are found, as are distinct motoneurons innervating specific muscles and glands. The vast majority of neurons are unipolar, with definite heteropolar functional orientations.

Important functional organization strategies become particularly evident in the annelids. Segmental control of simple reflexes is handled at the local level by neurons in segmental ganglia. For example, touch or painful stimuli in specific receptive fields generate appropriate, localized, and discretely organized reflexive muscle contractions and movements (Fig. 4*c*). The sensory and motor impulses are carried by processes of neurons in the same segmental ganglion. In addition, however, adjacent segments may become involved by extensions of sensory-cell processes to neighboring ganglia and, more importantly, by interneurons "wired" to coordinate multisegmental responses. Also, the brain now contains many interneurons which may send axons down the full length of the ventral nerve cord to receive or transmit information to all segmental ganglia. Thus, central control fibers are present to coordinate overall body functions and behaviors. For example, the rhythmic body undulations, which propel leeches in swimming, require segmental control of body muscle masses, and superimposed central commands from higher-order coordinating interneurons.

The first appearance of at least two specialized kinds of nerve cell or nerve process has been best documented in the annelids. These are neurosecretory or neuroendocrine cells specialized to secrete proteinaceous chemical messenger compounds which act as hormones or widespread effector substances; and giant axons, ranging from

around 30 to over 200 micrometers in diameter, and representing fast-conduction "through pathways" for rapid responses. See NEUROSECRETION.

The best-studied components of annelid nervous systems are the giant fibers of the earthworm and, particularly, the segmental ganglia of *Hirudo*, the medicinal leech. In the leech, there are 21 segmental ganglia, all identical, with some 350 neurons in each ganglion (Fig. 4*c*). Many of these neurons are identified, meaning that, on the basis of size, color, topography, and electrophysiological and functional characteristics, they can be readily found and reliably identified from ganglion to ganglion and from animal to animal. Specific identified sensory cells and motor neurons have been characterized in detail. Both excitatory and inhibitory motor neurons have been identified; these cause muscle contraction or relaxation, respectively, through neuromuscular synaptic junctions. Synaptic transmitters, such as acetylcholine and serotonin, have been found in particular motor neurons. See ANNELIDA.

Arthropoda. This is the largest, most diverse, and most advanced of the invertebrate phyla (though certain of the mollusks are very advanced as well). The most general characteristics of the group are the hard exoskeleton, jointed legs, and varying degrees of segmentation. This is the only group of invertebrates that has successfully invaded the terrestrial and aerial environments. An incredible diversity of form, degree of development, and specialization exists. Some of the highest levels of invertebrate neural complexity, organ development, muscle specialization, and behavioral repertoire are found. Only major highlights are considered by describing a representative arthropod nervous system, modeled after that of the crayfish (Fig. 5*a*).

The model arthropod is fundamentally metameric, but through fusion of segments, the body has come to be composed of a head, a thorax, and an abdomen. The head ganglia are highly fused to form extraordinarily complex brains, dorsally placed over the esophagus, with a circumesophageal ring of connective nerves and supplementary ganglia. The ganglia have a dense neuropile surrounded by a ring of glia-invested unipolar neurons. Particularly advanced brains, such as in some insects, may have several layers of cells, infoldings, and a complex organization; and numerous multipolar (heteropolar) neurons may occur as well. Millions of nerve cells and highly branched nerve processes are present. A ventral nerve cord (or paired cords) extends the length of the thorax and abdomen. The thorax commonly has one to several fused ganglia; the abdominal ganglia tend to retain serial segmentation and to be identical. All of the ganglia have symmetrical, usually bilateral, nerve trunks coming off to innervate peripheral sense organs or muscles and other effector cells; many of these pathways decussate. Functional tiers or cascades of interneuronal axons run the length of the cord and are within ganglia to organize highly complex behaviors and coordinate the movements of appendages for locomotion or flight. There are typically pairs of giant axons of

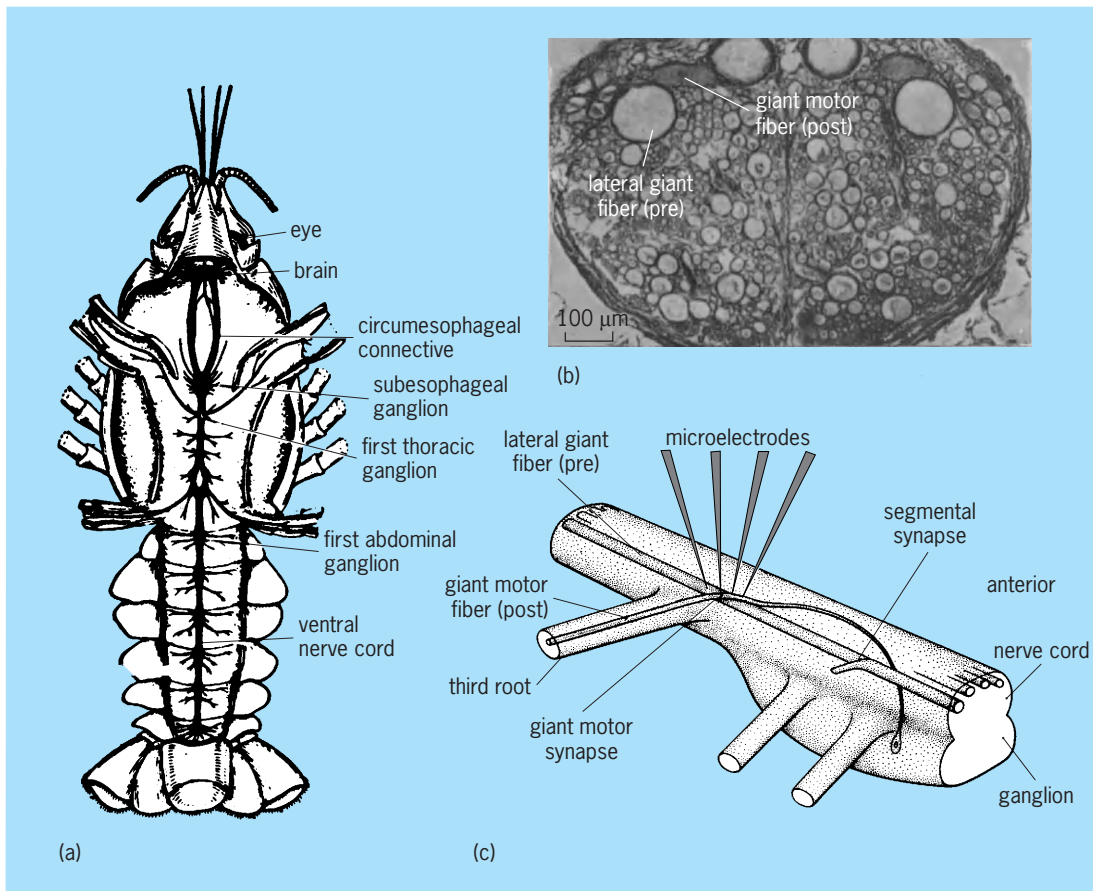


Fig. 5. Advanced invertebrates as represented by the crustacean nervous system. (a) Crayfish central nervous system (after W. C. Curtis and J. and J. M. Guthrie, *Textbook of General Zoology*, 3d ed., John Wiley and Sons, 1938). (b) Photomicrograph of cross section of ventral nerve cord in abdomen of crayfish made near the level of the third root (arrow in a); the largest axons near the top are giant fibers used in escape responses (after J. D. Robertson, *Ultrastructure of excitable membranes of the crayfish median-giant synapse*, *Ann. N.Y. Acad. Sci.*, 94:339–389, 1961). (c) Relationship between lateral giant and motor giant axons as they make electrical synaptic contact near the third root of an abdominal ganglion in the crayfish (after E. J. Furshpan and D. D. Potter, *Transmission at the giant motor synapses of the crayfish*, *J. Physiol.*, 145:289–325, 1959).

interneurons mediating fast-escape behaviors (Fig. 5b and c).

Vast numbers of highly specific kinds of sensory cells (including muscle stretch receptors) are present, and extremely complex sense organs such as eyes and statocysts exist. The trunk and appendage musculature is highly specialized with a variety of fast- and slow-contracting fibers, each with its own set of identified excitatory and inhibitory motor neurons. Several small peripheral ganglia are present which appear to be somewhat analogous to the vertebrate autonomic nervous system; they control the heart and vasculature, the digestive organs, and other visceral functions. Well-developed neurosecretory and neuroendocrine systems are also present.

Electrical synapses are still found, particularly in the connections of the fast-escape systems employing giant axons, where minimal synaptic transmission time is advantageous. But chemical synaptic transmission is highly evolved, with numerous identified transmitter substances. Excitatory and inhibitory connections, exquisitely sensitive to rates of stimulation and highly plastic, are described; and

complex tandem synaptic arrangements are known where one presynaptic terminal ends on another, modifying the effect of the latter on a postsynaptic element.

Probably the best studied of the arthropod nervous systems are those of the crayfish and lobsters, but several insects, for example the locust, cricket, fruit fly, and cockroach, are also important research preparations. Studies of these creatures and of some selected molluscan species have provided some of the most profound insights into how nerve cells function and are organized at the cellular, membrane, and genetic levels. The fruit fly *Drosophila* has been particularly useful in studies of the genetic basis of nervous system development. Some genes involved in neuromere development in the brain of *Drosophila* appear to be homologous to genes expressed in brain development in higher vertebrates. See ARTHROPODA.

Mollusca. This highly diverse phylum spans a range of complexity rivaling that of the arthropods. The group includes the well-known clams and mussels, snails and slugs, and squid and octopus. The phylogenetic origins of this phylum are unsettled, but

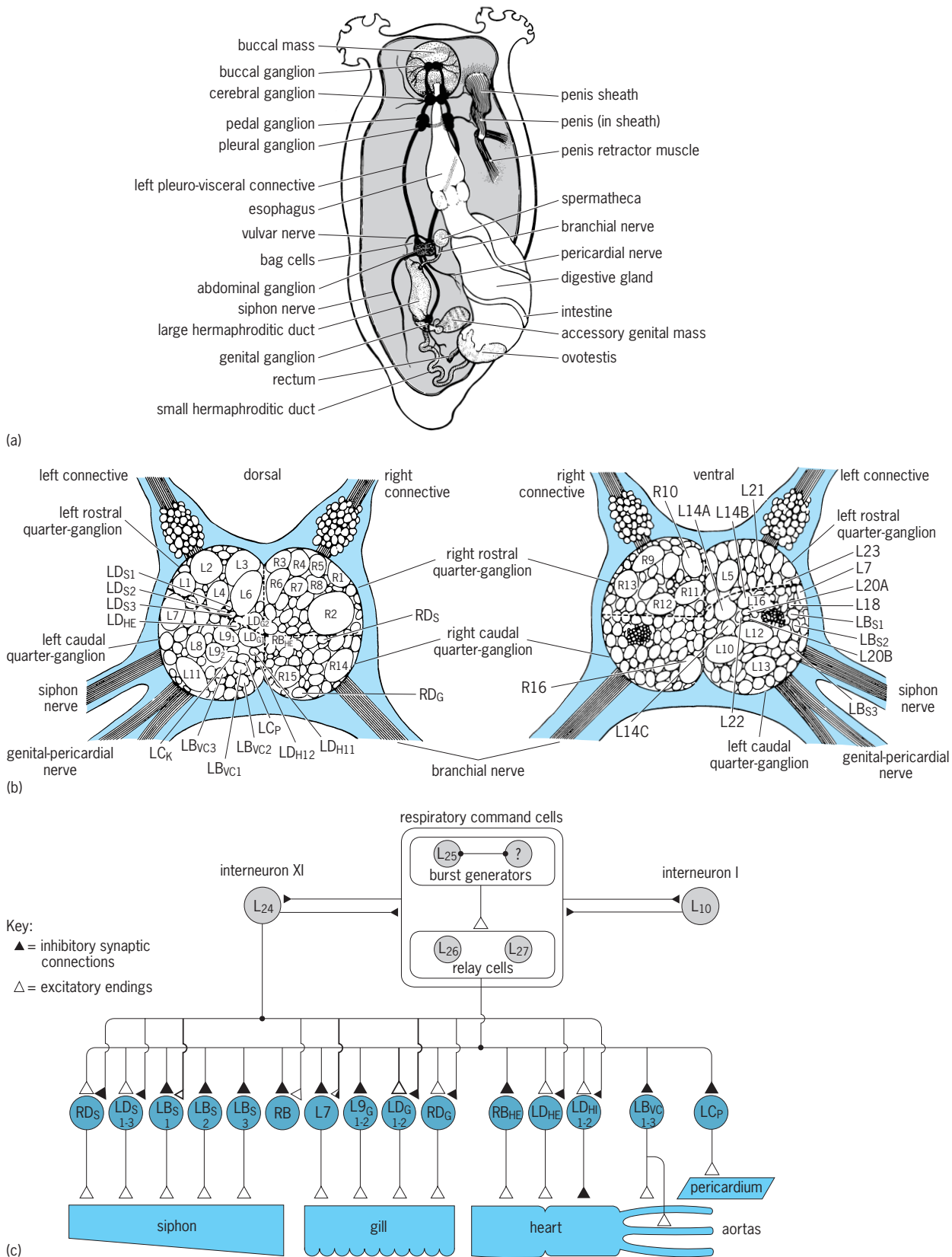


Fig. 6. Advanced invertebrate nervous system as represented in the marine gastropod mollusk *Aplysia*. (a) General arrangement of nervous system. (b) Map of the dorsal and ventral surfaces of the abdominal, or parietovisceral, ganglion, indicating the many identified nerve cell bodies known to play roles in a variety of functions (after E. R. Kandel, *Cellular Basis of Behavior*, W. H. Freeman, 1976). (c) Circuit diagram of several identified interneurons and motor neurons which innervate various organs; solid triangular axon endings represent inhibitory synaptic connections, and open triangles, excitatory endings (after J. H. Byrne and J. Koester, *Respiratory pumping: Neuronal control of a centrally commanded behavior in Aplysia*, *Brain Res.*, 143:87–105, 1978).

primitive mollusks either diverged very early from the annelid-arthropod line or arose from a common ancestor. The phylum as a whole is characterized by a bilateral fleshy body, usually with a prominent foot (muscular locomotor or attachment structure); a mantle cavity built around an external gill and openings of the digestive and excretory systems; a shell secreted by the mantle; and a reduced coelom.

Despite the extraordinary polymorphism of the phylum, a fundamental body plan and characteristic neural arrangement are present, though often obscured by secondary reorganizations. The positions of ganglia and arrangements of nerves are modified to accommodate the flattened body and hinged shells of the bivalves, the twisted or torqued bodies of the gastropods, or the sleek streamlining of the highly active cephalopods. An example of a gastropod nervous system is illustrated in Fig. 6a.

The head contains a circumesophageal ring of ganglia, the bilateral brain, which sends numerous nerves to the eyes, tentacles, and other richly diverse and highly developed sensory structures and sensory neurons. From the brain there are also long nerve cords, usually paired, which interconnect other ganglionic masses in the body. There are usually several subsidiary ganglia in mollusks, including ones primarily involved in innervating the foot, the mantle and visceral structures, and the mouthparts. A map of the identified neurons in the abdominal or parieto-visceral ganglion of *Aplysia* is shown in Fig. 6b. Depending on the class of mollusk and the type of secondary alterations in body shape and configuration, the subsidiary ganglia may lie in various positions in the body (normally close to the body mass innervated), or they may be fused with the brain. Several types of muscle are present, usually of a slow type, both in the body wall (including the foot) and in visceral structures (gill, gut, reproductive organs). Faster muscles, somewhat more akin to arthropod or vertebrate types, may be found in buccal (mouthpart) muscle, in mantle muscle of squid, in retractor muscles which serve to draw the animal back into the shell, and in adductor muscles which close the shells.

There are a few unipolar peripheral neurons in the skin or various organs, but the great bulk of neurons are clustered into conventional ganglia composed of a central neuropile and an outer ring of unipolar (heteropolar) cell bodies. Except in very complex brains, such as that of the rather intelligent octopus, which has the largest invertebrate brain, there are only rare multipolar (heteropolar) neurons. The mollusks are exceptional in that so many species have the largest neuron cell bodies or axons (several hundred micrometers in diameter) known. Neuroendocrine and glial cells are also present. The ganglia of higher mollusks such as gastropods and cephalopods have many interneurons which serve to coordinate and control complex behaviors (Fig. 6c).

The most famous molluscan research preparation is the giant axon of the squid from which some of the earliest intracellular recordings of the action potential were obtained (Fig. 1a). This axon was also the

experimental preparation used by Hodgkin and Huxley to describe the membrane properties that cause an action potential to occur—work that earned them a Nobel prize in 1963. The other very widely studied mollusks are certain gastropods, particularly the marine snail *Aplysia* (Fig. 6) and the terrestrial, and aquatic pulmonates, typified by *Helix*, *Heliosoma*, and *Lymnaea*. As with leech, crayfish, and insect ganglia, *Aplysia* and other gastropod nervous systems have been exhaustively studied, and many identified nerve cells and detailed synaptic circuits have been described in great detail. Neural control of feeding, respiration, circulation, inking, reproduction, and other activities have been unraveled to a large extent. The cellular mechanisms underlying simple kinds of learning or memory, such as habituation and sensitization, are also well understood. Some electrical synapses exist, but a rich variety of very complex chemical synaptic connections are present, and their plastic properties (and many transmitter substances, including peptides) have been identified. Eric Kandel won a Nobel prize in 2000 for his work on the biochemical basis of learning and memory in *Aplysia*.

Cephalopods, including octopuses and squid, may have the most highly evolved and complex nervous system among all invertebrates. In *Octopus*, the central brain surrounds the esophagus, with lateral paired optic lobes that may contain over 65 million nerve cells. Cephalopods have been subjects in extensive studies of learning, sensory reception, and behavior. The image-forming eye of coleoid cephalopods is similar in morphology and complexity to the eye of higher vertebrates. See MOLLUSCA.

Echinodermata. The echinoderms, while phylogenetically near the protochordates, have undergone such a great deal of secondary adaptation and specialization in the adult form that their nervous systems are less well defined and studied than those of lower forms.

The nervous system of these radially symmetrical organisms is less centralized, having a subepidermal nerve-ring arrangement around the mouth, with radial nerves extending from it. Few and small ganglia exist. Uni- and multipolar neurons are present. There are large numbers of individual sensory cells, and few specialized sense organs. Superficial (mainly sensory) and deeper (motor) nerve nets are present. A variety of body wall and specialized muscles are present to control locomotion and food capture. The coordination of movement of the spines and arms, the pedicellariae, and the tube-foot/water-vascular system requires a rich variety of motor neurons and interneurons. See ECHINODERMATA; ENDOCRINE SYSTEM (INVERTEBRATE).

James E. Blankenship; Becky Houck

Bibliography. C. I. Bargmann, Neurobiology of the *Caenorhabditis elegans* genome, *Science*, 282:2028–2033, 1998; J. H. Byrne, Cellular analysis of associative learning, *Physiol. Rev.*, 67:329–439, 1987; R. F. Chapman, *The Insects: Structure and Function*, 4th ed., Cambridge University Press, 1998; R. T. Hanlon and J. B. Messenger, *Cephalopod Behaviour*, Cambridge University Press, 1996;

G. G. Lunt and R. W. Olsen (eds.), *Comparative Invertebrate Neurochemistry*, Cornell University Press, Ithaca, NY, 1988; D. Osorio, J. P. Bacon, and P. M. Whittington, The evolution of arthropod nervous systems, *Amer. Sci.*, 85(3):244-253, 1997; J. A. Pechenik, *Biology of the Invertebrates*, 4th ed., McGraw-Hill, Boston, 2000.

Nervous system (vertebrate)

A coordinating and integrating system which functions in the adaptation of an organism to its environment. An environmental stimulus causes a response in an organism when specialized structures, receptors, are excited. Excitations are conducted by nerves to effectors which act to adapt the organism to the changed conditions of the environment. In animals, humoral correlation is controlled by the activities of the endocrine system. This article considers the morphology, histology, and embryology of the nervous system, including the brain and cranial nerves, and embryology of the sense organs.

Comparative Morphology

The brain of all vertebrates, including humans, consists of three basic divisions: prosencephalon, mesencephalon, and rhombencephalon (Fig. 1). Newer experimental methods for exploring the connections and functions have resulted in much informa-

tion on the evolution of this system. The indication is that these divisions of the vertebrate brain have evolved along several functional lines and perform very different functions.

The comparative neuroanatomist is interested in the variation among the vertebrate brains, and attempts to understand how they evolved from common ancestors and to clarify the functional significance of their variations. It is necessary that the neuroanatomist recognize structures in different brains among living forms which have arisen from one or more structures in an ancestral brain. This search for homologous structures is complicated by the fact that brains do not fossilize; thus the comparative anatomist can only infer what the possible structure of the brain in ancestral vertebrates was like. It is thought that the brain of ancestral vertebrates was much simpler than those of living forms, but neuroanatomists suggest different views regarding the brain's evolution.

One view holds that structures have arisen afresh with no traces in lower living vertebrates. The second view proposes that the simpler ancestral brain possessed very generalized connections which have either been lost or have become specialized in living vertebrates. The latter idea seems to be closer to the truth and is more in line with what is known about evolution of other vertebrate body systems.

The nervous system of ancestral vertebrates apparently consisted of a single hollow neural tube,

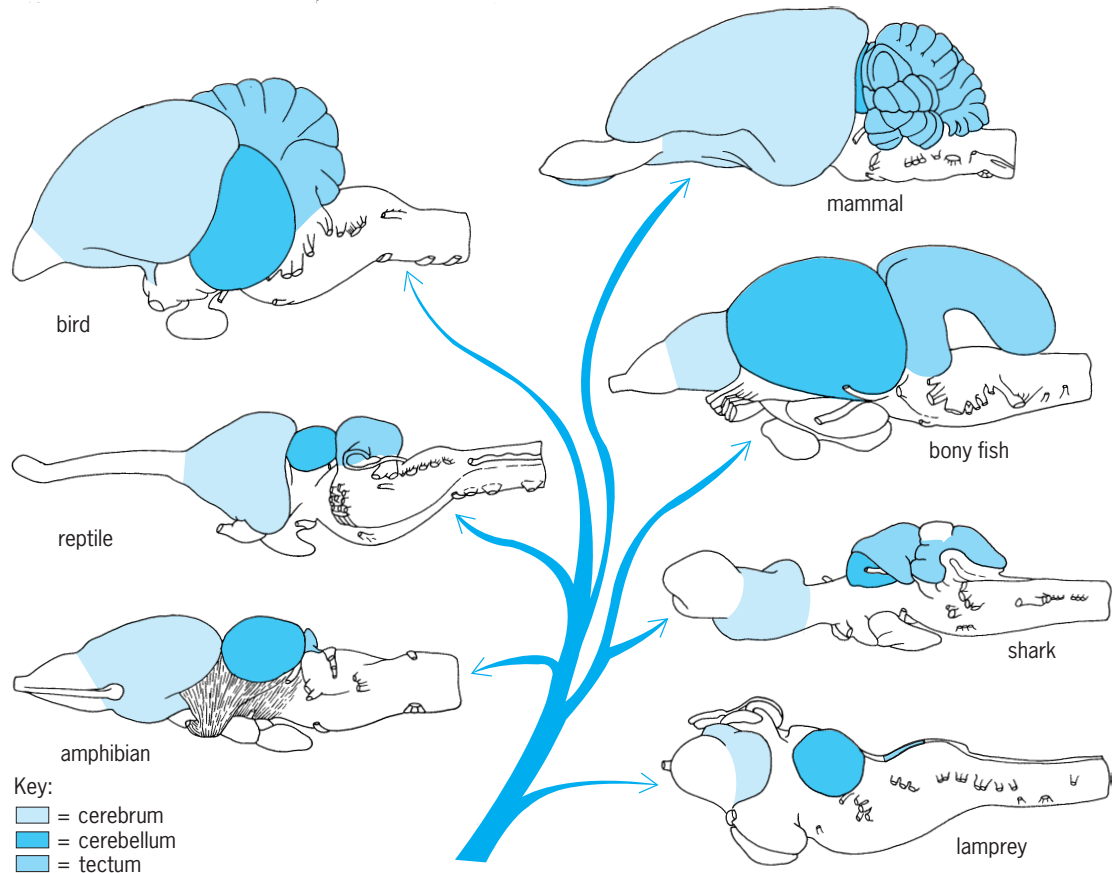


Fig. 1. Lateral views of several vertebrate brains showing evolutionary relationships.

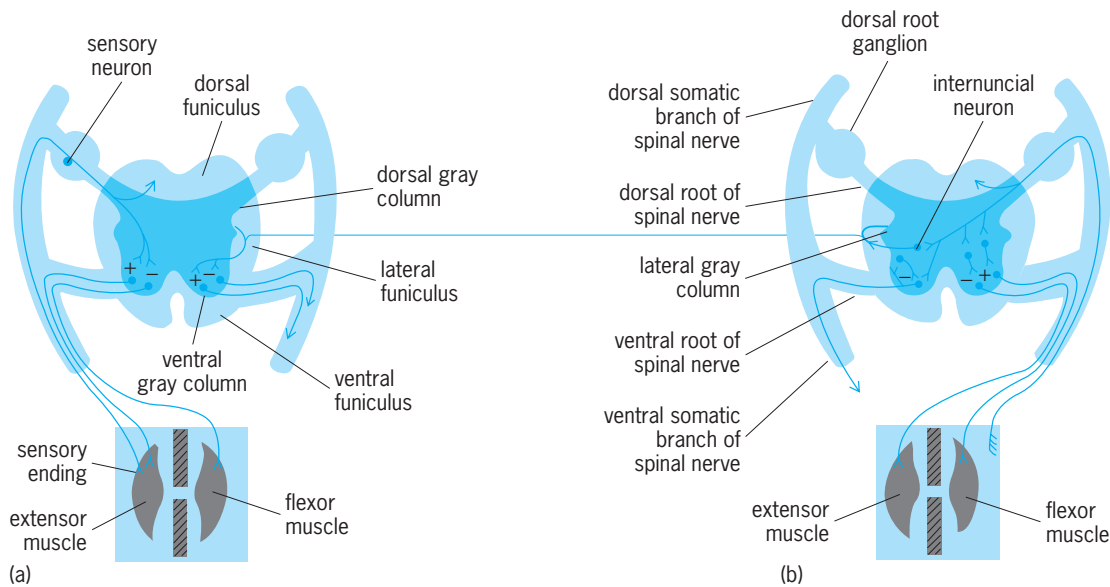


Fig. 2. Spinal cord diagram. (a) Monosynaptic arcs. (b) Multisynaptic arcs.

dorsally placed and running the length of the animal, that was probably structured in a manner similar to the spinal cord in living vertebrates (Fig. 2). It consisted of a central gray region which contained most of the cell bodies of the neurons and was surrounded by a superficial white region containing the axons and dendrites of the neurons running to and from the tube. Each region of the tube was primarily concerned with the functions of that immediate region of the body to which its nerves projected. Neural functions were mediated at a segmental level rather than specialized in one specific region. Each region of the neural tube also contained neural cells which interconnected adjacent regions. This primitive condition still persists in a few living forms.

As vertebrates evolved, structures in the head region became specialized for sensing the outside world and capturing food. This specialization was reflected in the neural tube, and its anterior region enlarged to permit analysis of the environment and integration of behavior. This enlargement of the neural tube is called a brain, and its divisions represent regions of different specialization. Each of the three primitive brain divisions was concerned with analysis and integration of a single type of sensory information. The rhombencephalon analyzed changes in the flow and pressure of water on an animal; the mesencephalon analyzed changes in the pattern and intensity of light; and the prosencephalon analyzed changes in the chemical composition of water. All vertebrates use these three types of information as clues in regulating their behavior. However, major changes in the structures that analyze this information appear to be correlated with increased quality of perception and more centralized control of movement. This has occurred by an increase in the number of analyzing units between the sensory mechanisms and the response mechanisms. This increase permits a delay in the response of a vertebrate to sensory cues and thus permits greater analysis of in-

formation contained in the clues. Presumably, this has value to an animal in that it results in responses which have a higher degree of survival value.

The evolution of the central nervous system is described below as the evolution of functional systems or patterns whose specialization is reflected by the concomitant specialization of brain divisions. The individual divisions or patterns of the brain do not function separately to bring about a final response; rather, each pattern acts on a common set of connections in the spinal cord. The common spinal connections are first described and then used as a basis for understanding the action of the brain patterns upon them.

Spinal patterns. These are the final common patterns used by all higher brain pathways to influence all organs of the body. These reflexes are divided into two basic patterns: the monosynaptic arc and the multisynaptic arc.

The monosynaptic arc, or myotatic reflex, maintains tonus and posture in vertebrates and consists of two neurons, a sensory and a motor neuron. The sensory neuron possesses a sensory ending in the extensor muscle (Fig. 2a), and this nerve ending is stimulated when the muscle is stretched by the pull of gravity. The axon of the sensory neuron projects to the motor neuron of the extensor muscle. This motor neuron is located in the ventral, gray region of the spinal cord, and its axon projects back to the extensor muscle, thus completing the arc.

When the extensor muscle is stretched, the sensory ending of the sensory neuron is stimulated, and this stimulation is transmitted to the cell body of the sensory neuron. The sensory cell body then transmits a signal to the motor neuron of the extensor muscle, which will cause the motor neuron to transmit. At the same time, the sensory neuron causes the motor neuron of the flexor muscle to decrease its transmission. These processes are indicated respectively by plus and minus signs in Fig. 2. The result

is that the extensor motor neuron causes the extensor muscle to contract, and the flexor motor neuron causes the flexor muscle to relax. This extends the limb and supports the weight of the animal. The sensory neuron of the extensor muscle also sends an axonic collateral to brain centers where the sensory information is analyzed.

The multisynaptic arc, or flexor reflex, is the pattern by which an animal withdraws a part of its body from a noxious stimulus. Stimulation of the skin or muscle causes a sensory neuron to transmit a signal causing motor neurons, via internuncial neurons, to increase the contraction of flexor muscles and decrease the contraction of extensor muscles. The result is movement of a limb or body part away from the stimulus. Both sensory neurons and internuncial neurons send information to brain centers (Fig. 2).

The multisynaptic arc also gives rise to a pattern that results in alternating movement of a forelimb and the diagonal hindlimb. This is termed reflex stepping. Thus coordinated limb movement is based on a connective pattern of neurons at the spinal level.

The structure of the spinal cord and its connections are basically similar among all vertebrates, and myotatic and flexor reflexes occur in all vertebrates. The major evolutionary changes in the spinal cord have been the increased segregation of cells and fibers of a common function from cells and fibers of other functions and the increase in the length of fibers which connect brain centers with spinal centers. See POSTURAL EQUILIBRIUM.

Medullar patterns. The rhombencephalon of the brain is subdivided into a roof, or cerebellum, and a floor, or medulla oblongata. The medulla is similar to the spinal cord and is divided into a dorsal sensory region and a ventral motor region. The sensory region consists of two longitudinal columns; a dorsal, somatic sensory column and a ventral, visceral sensory column. The ventral motor region of the medulla is similarly divided into two longitudinal columns, a dorsal, visceral motor column and a ven-

tral, somatic motor column. These divisions occur in the spinal cord and continue through the medulla into the tegmentum, the floor of the mesencephalon.

The medulla is an integrating and relay area between higher brain centers and the spinal cord. Primatively, the columns were composed of neurons scattered throughout the length of the medulla. With the specialization of the vertebrate head, the columns broke up into a series of nuclei that formed the neural elements for reflex circuits of the cranial nerves innervating the organs of the head and gills.

In addition to these nuclei and their connections, the medulla consists of both ascending and descending pathways to and from higher brain centers. The nuclei of the medulla shift with reference to one another, and their relative volumes change as specialization in gill and feeding mechanisms occurs. However, the same basic connections occur throughout vertebrates. The tracts which pass through the medulla between higher brain centers and the spinal cord become more distinguishable as their boundaries become more compact in relation to other tracts. The tracts, like the nuclei of the medulla, may change volume in response to specialization.

Cerebellar patterns. The cerebellum has had a varied history of development in vertebrates. In all vertebrates, the cerebellum is divided into two major divisions: the two lateral flocculonodular lobes, and a central corpus cerebelli. The flocculonodular lobes are functionally referred to as the vestibulocerebellum and regulate vestibular reflexes underlying posture. The corpus cerebelli is subdivided into two lateral zones, the cerebrocerebellum and a central zone, the somatocerebellum. In mammals the lateral zones regulate corrective reflexes of posture and time muscular contractions of voluntary actions. The medial zone in mammals regulates reflex tonus of postural muscles by acting on the myotatic reflex.

The flocculonodular lobes, or vestibulocerebellum, evolved in conjunction with the vestibular system of the inner ear and is mainly responsible for gravitational orientation in higher vertebrates. The vestibular division of the inner ear senses changes in the orientation of an animal's head in all three planes and controls the postural reflexes of the head and neck. These vestibular centers also influence the spinal patterns of posture described above. Vestibular connections are made with the flocculonodular lobes and their nuclei, the fastigial nuclei. This system allows integration of vestibular sensations with other sensory information which passes to the cerebellum from the spinal cord.

The evolutionary changes of the corpus cerebelli in land vertebrates resulted in a new mode of locomotion. The limbs were used to support an animal off the ground and for locomotion as well. It became essential for a land vertebrate to maintain posture and to know the position of its limbs in order to move them in an organized manner.

Mammals. Information on the position of the limbs, and on the state of tonus in the muscles of limbs, is transmitted to the cerebellum via the posterior cerebellar peduncle (Fig. 3). The

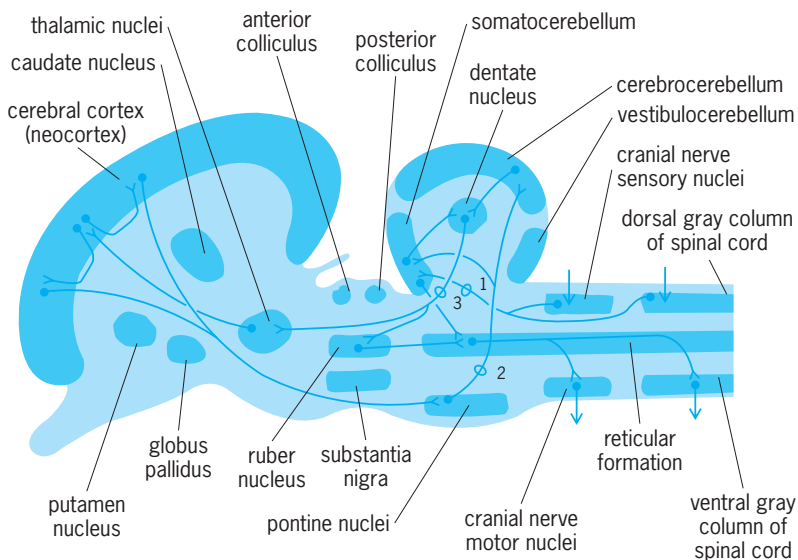


Fig. 3. Mammalian brain in sagittal section. Cerebellar patterns: tract 1, posterior cerebellar peduncle; 2, middle cerebellar peduncle; 3, anterior cerebellar peduncle.

cerebellum does not initiate movement in mammals; it only times the length of muscle contractions and orders the sequence in which muscles should contract to bring about a movement. The command to initiate a movement is received from the cerebral cortex via the middle cerebellar peduncle (Fig. 3). Similarly, the cerebral cortex receives information regarding limb position and state of muscular contraction to ensure that its commands can be carried out by the cerebellum.

When the cerebellum receives a command from the cerebral cortex, it transmits a pattern of impulses to the spinal patterns via the dentate nucleus and lower pathways. This pattern of impulses contains the sequence of muscles to be contracted and the length of time of these contractions. At the same time, this same information is transmitted from the cerebellum to the cerebral cortex via the anterior cerebellar peduncle (Fig. 3). The cerebral cortex can thus compare its command with the cerebellum's action. This information is then compared with new information coming to the cerebral cortex about changes in the limbs as a result of the initial movement. If the movement was not adequate, the cerebral cortex can then issue a corrected command to the cerebellum for a new series of movements.

Birds. A similar system for control of locomotion is apparently present in birds. However, the exact nature of the command center, or centers, in the telencephalon is not understood. Birds possess a middle cerebellar peduncle, but the tracts from the telencephalon may not originate in the same areas as in mammals. Birds have a cortical command center, but whether it is homologous with the neocortex of mammals is not known.

Reptiles. Knowledge of the reptilian cerebellar patterns is more fragmentary than those of birds. Reptiles possess a pathway, the dorsal forebrain bundle, from the telencephalon to the cerebellum (Fig. 4). This pathway may be homologous, in part, with the middle cerebellar peduncle. The telencephalic center from which this pathway originates is similarly organized in both reptiles and birds. Reptiles possess an anterior cerebellar peduncle, but a homologous pathway similar to the mammalian component projecting to the cortex has not been identified. It appears that reptiles, birds, and mammals possess cerebellar patterns of a similar nature. There are differences in organization, but they appear to be secondary and may reflect the added role of the tectum as a second higher center in reptiles and birds.

Amphibians. The cerebellum in amphibians is not highly developed. It is divided into the same divisions as in higher vertebrates, and its histology is identical. The locomotor pattern in amphibians is very different from that of other vertebrates, and it is improbable that it functions in the same manner. Salamanders do not use the limbs for support as do higher vertebrates, and frogs have undergone great specialization for jumping. There is little information on the role of the cerebellum in these forms.

Fishes. The cerebellum, particularly the corpus cerebelli, in sharks and bony fish is large. Part of the

size is due to the huge number of gustatory fibers projecting to the corpus cerebelli. Both groups of fish possess anterior and posterior cerebellar peduncles. The connections of the anterior cerebellar peduncle are not understood. In bony fish, the tectum is expansive and forms a complex cortex. Tectocerebellar connections are large, indicating that a tectal control center probably exists analogous to the cerebral control center in mammals. Similar tectocerebellar connections exist in other vertebrates, but they do not exist as a highly developed complex.

Tectal patterns. The mesencephalon is divided into a roof or optic tectum and a floor or tegmentum. The tegmentum contains the nuclei of the oculomotor and trochlear cranial nerves and a rostral continuation of the sensory nucleus of the trigeminal cranial nerve. Two motor nuclei dominate the tegmentum, the ruber nucleus and the substantia nigra. These nuclei are elements in both the telencephalic and the cerebellar motor systems. All of these nuclei are recognizable in most vertebrates. The tegmentum also contains sensory and motor pathways which project between the prosencephalon and lower brain and spinal centers.

The roof of the mesencephalon in lower vertebrates is usually referred to as the optic tectum. Early anatomists realized that the major connections of the optic nerve formed in this area and that this area represented the major center for analysis of visual information. However, in lower vertebrates the optic tectum is also a major center for analysis of somatic sensory information and contains a highly structured center for control of somatic musculature.

The external shape of the tectum has changed greatly in the course of evolution (Fig. 1). In fish, amphibians, and reptiles, it forms the roof of the mesencephalon and remains in a dorsal position. In all of these forms, it is normally seen as a paired swelling between the telencephalon and the cerebellum. In birds, each lobe of the tectum has shifted ventrally, and the telencephalon and cerebellum have expanded until they meet along the dorsal surface. In

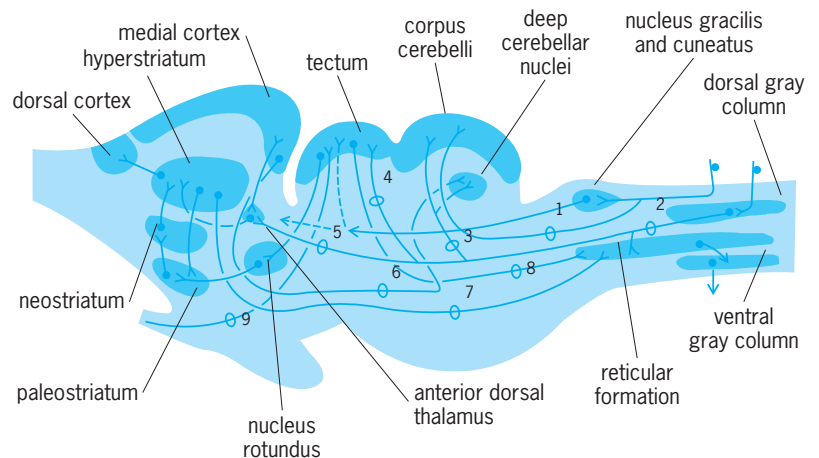


Fig. 4. Reptilian brain in sagittal section illustrating tectal patterns: tract 1, dorsal spinocerebellar tract; 2, lateral funiculus; 3, ventral spinocerebellar tract; 4, spinomesencephalic tract; 5, spinothalamic tract; 6, dorsal forebrain bundle; 7, ventral forebrain bundle; 8, tectospinal tract; and 9, optic tract.

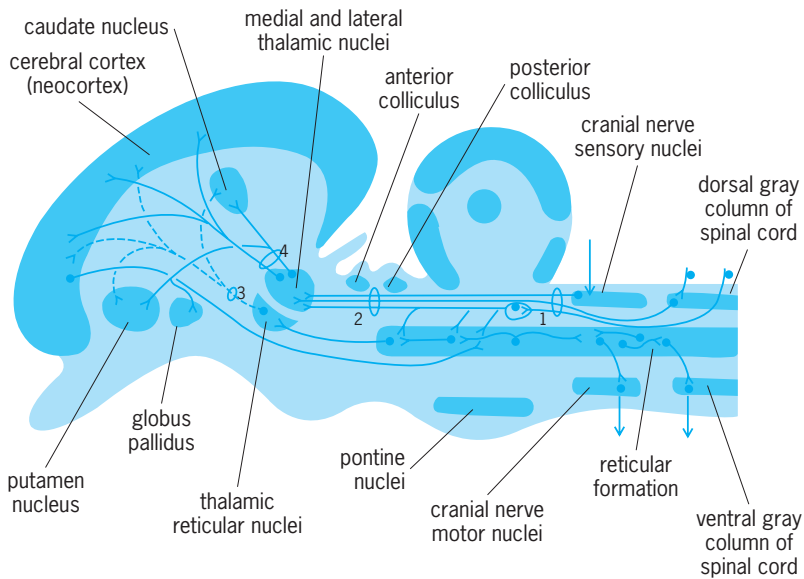


Fig. 5. Mammalian brain in sagittal section illustrating thalamic patterns: tract 1, dorsal and lateral spinal funiculi; 2, lemniscal system; 3, ascending reticular arousal system; and 4, specific thalamic system.

mammals, the tectum has retained its primitive position but is not a large structure. The telencephalon covers its anterior surface, and the cerebellum, its posterior surface; thus it is often not visible from the surface.

In all vertebrates the tectum is divided into two functional regions, an anterior optically dominated region and a posterior dominated auditory region. These divisions may or may not be reflected in the surface structure of the tectum. In mammals, the tectum receives visual information from a division of the optic tract, but the majority of the optic fibers project to the cerebral cortex of the telencephalon after synapsing in a diencephalic nucleus, the lateral geniculate body. This nucleus exists in lower vertebrates and receives optic fibers, as in mammals, but their number is smaller, and the rostral projections, if any, are not known.

A similar condition exists in the auditory pathways. In all vertebrates except mammals, the primary auditory pathway projects to the posterior division of the tectum. In lower vertebrates, this division is called the torus semicircularis. In mammals, some auditory projections occur in the posterior region of the tectum, but the majority of the fibers project to the cerebral cortex of the telencephalon via a diencephalic nucleus, the medial geniculate body.

The tectal pattern in reptiles illustrates the basic connections which occur in all lower vertebrates (Fig. 4). The exact functions of these tectal patterns in reptiles and birds have not been discovered. In these two groups, where the telencephalon is also highly developed, the interactions between telencephalon and tectum appear to be very complex.

Diencephalic patterns. In the evolution of vertebrates, the prosencephalon develops as two major divisions, the diencephalon and the telencephalon. The diencephalon retains the tubular form and serves as a relay and integrating center for infor-

mation passing to and from the telencephalon and lower centers. The telencephalon is divided into a pair of cerebral hemispheres and an unpaired telencephalon medium.

There are three divisions of the diencephalon in all vertebrates: an epithalamus which forms the roof of the neural tube, a thalamus which forms the walls of the neural tube, and a hypothalamus which forms the floor of the neural tube. The epithalamus and hypothalamus are primarily concerned with autonomic functions such as homeostasis. See HOMEOSTASIS; INSTINCTIVE BEHAVIOR.

In many vertebrates the epithalamus develops two outpocketings from its roof. These are the pineal and parietal processes. The pineal process usually forms a gland with endocrine functions. The parietal process may form an endocrine gland or an eye. In some living reptiles, the parietal eye is highly developed. It apparently was common in many fossil groups as well. See ENDOCRINE SYSTEM (VERTEBRATE); PINEAL GLAND.

The thalamus is subdivided into dorsal and ventral regions. The dorsal region relays and integrates sensory information, and the ventral thalamus relays and integrates motor information. In amphibians, the dorsal thalamus is represented by a single group of neurons. In reptiles and birds, this group of neurons has been broken up into a series of nuclei having little, if any, relationship to nuclei found in the thalamus of mammals. Bony fish have a well-developed dorsal thalamus with many nuclei which have been homologized with those of other lower vertebrates, but their functions are unknown.

The dorsal thalamus in higher vertebrates has evolved along two different lines. In reptiles and birds, the dorsal thalamus receives some sensory connections from the spinal cord, but the majority of its connections are with the tectum. In these forms the dorsal thalamus is primarily a relay and integrating center between the tectum and the telencephalon (Fig. 4). In mammals, the dorsal thalamus receives connections from the tectum, but the majority of its connections are with the spinal cord and the telencephalon (Fig. 5). Thus the dorsal thalamus is a relay and integrating center for sensory information passing from the spinal cord to the telencephalon. Two types of sensory systems project to the telencephalon. The ascending reticular arousal system (Fig. 5) passes to the telencephalon and alerts the cortical fields which analyze sensory information. A second system, the specific sensory, or lemniscal, system passes to the sensory cortex and carries specific information on the nature of the stimulus and the position of its origin on the body or head.

Telencephalic patterns. The telencephalon is the most complex brain division in vertebrates. It is divided into a roof, or pallium, and a floor, or basal region. The pallium is divided into three primary divisions (Fig. 6): a medial PI or hippocampal division, a dorsal PII or general pallial division, and a lateral PII division, often called the pyriform pallium. The basal region is divided into three areas, the first of which

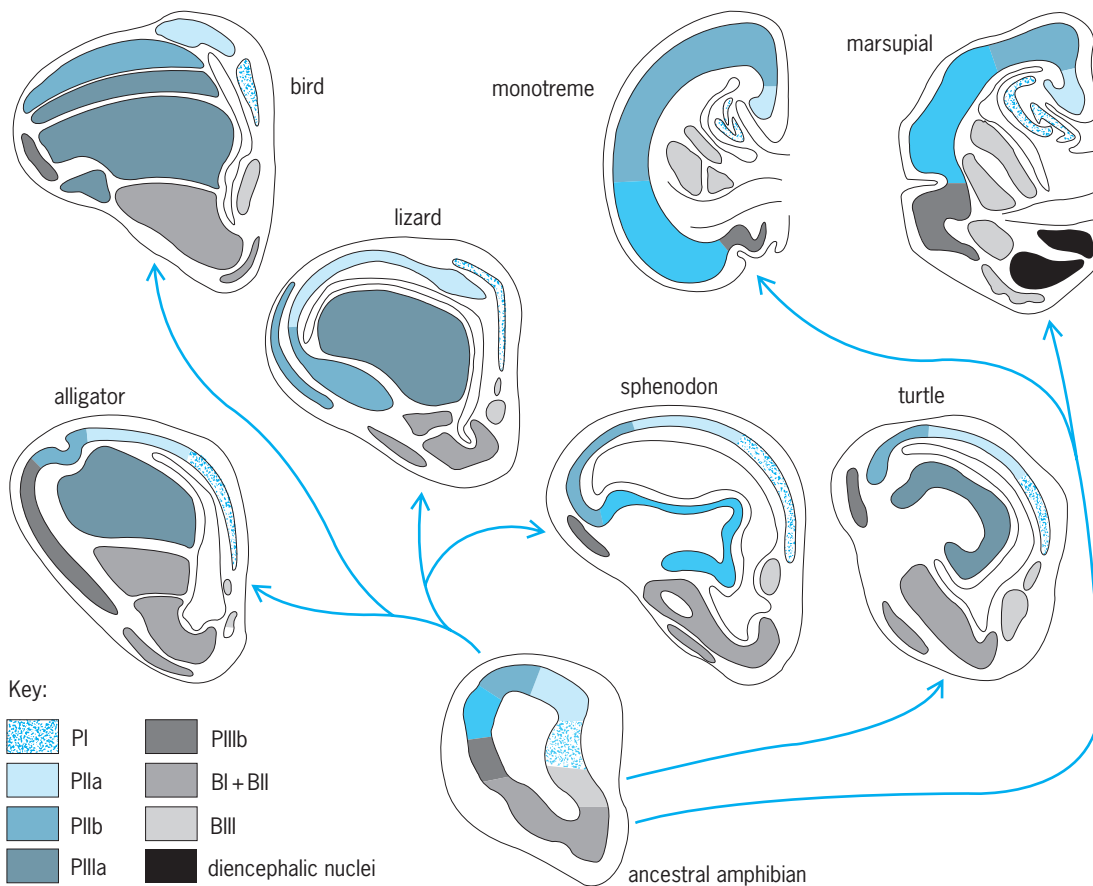


Fig. 6. Cross sections of the right telencephalic hemisphere of several vertebrates, showing evolutionary relationships. Homologous regions of the pallium and corpus striatum are similarly coded.

is a medial BIII area called the septum. A ventral BII area and a lateral BI area form a region often called the corpus striatum.

The most striking change in the telencephalon of land vertebrates involves the PIIIa component, often called the dorsal ventricular ridge (Fig. 6). In reptiles and birds, this region has proliferated into the ventricle of the telencephalon to produce a large cellular mass. In mammals, it has proliferated with the PIIb component of the dorsal pallium to produce the mammalian neocortex. In all land vertebrates except amphibians, the PIIb and the PIIIa components, along with the corpus striatum (BI and BII), are the highest centers for the analysis of sensory information and motor coordination. The PI, PIIa, PIIb, BIII, and posterior parts of BI and BII form part of the limbic system which is concerned with behavioral regulation.

Older theories stated that the telencephalon first possessed only olfactory fibers and that sensory fibers from the thalamus occurred later in evolution, which resulted in the formation of a neocortex. It is now believed that thalamic fibers occur in the PIIIa component of most vertebrates and that all of the vertebrates possess structures of PIIIa origin even though all these structures are not recognized as a neocortex.

Bony fish possess a pallium and a basal region in the telencephalon, but these are organized along dif-

ferent lines than in other vertebrates. Thus the same divisions cannot be recognized.

Cortical representation. The development of a cortex appears to be the most common way in which analyzing and integrating units can be put together in a limited space to produce maximal efficiency. This is reflected in the fact that all three brain divisions have evolved cortices along one or more evolutionary lines. Each evolutionary line has not relied to the same extent on a particular type of sensory information or its utilization. However, each line has developed cortical surfaces which are specialized for different functions. For example, in mammals the posterior division of the cerebral cortex is concerned with visual impulses, and part of the lateral cortex is concerned with auditory impulses. Anterior to these regions are zones concerned with general sensory and motor impulses.

R. Glenn Northcutt

Comparative Histology

The vertebrate nervous system is composed of nervous tissue, which is one of the four primary tissues of the body. The basic microscopic anatomy of this tissue is essentially similar in all vertebrates although variations do exist. Nervous tissue has the quality of irritability. It is also characterized by the quality of conductivity, to convey the resulting excitation to other structures of the nervous system. The functional roles of nervous tissue include the

capabilities to sense, through specialized receptors, environmental energies both internal and external to the organism; to conduct the resulting nerve impulses as coded input to centers in the nervous system; to process this input within these centers; to generate sensations and psychological expressions; and to produce such active responses as the contraction of muscles or the secretions of glands. In effect, nervous tissue reacts to environmental stimuli and regulates many bodily processes so as to maintain functional integrity of an organism. It is within the morphological, physiological, and chemical matrices of nervous tissue that the substrates for memory, behavior, and personality reside. *See* LEARNING MECHANISMS; MEMORY; PERSONALITY THEORY.

Structural elements. The nervous system is composed of several basic cell types, including nerve cells called neurons, interstitial cells called neurolemma (cells of Schwann), satellite cells, oligodendroglia, and astroglia; and several connective-tissue cell types, including fibroblasts and microglia, blood vessels, and extracellular fluids.

Neuron. Each neuron possesses three fundamental properties, involving specialized capacity to react to stimuli, to transmit the resulting excitation rapidly to other portions of the cell, and to influence other neurons, muscle, or glandular cells. Each neuron consists of a cell body (soma), one to several cytoplasmic processes called dendrites, and one process called an axon.

Cell bodies vary from about 7 to more than 70 micrometers in diameter; each contains a nucleus and several cytoplasmic structures, including Nissl (chromophil) granules, mitochondria, and neurofibrils. The cell body is continuously synthesizing new cytoplasm, especially protein, which flows down the cell processes. The Nissl granules, found in dendrites as well as in the cell body, are rich in ribonucleoprotein and are responsible for synthesis of proteins for the neuron.

The mitochondria, found throughout the neuron, are the power plants involved in numerous chemical reactions within the cell. The neurofibrils are fine filaments found in living neurons; they are not directly related to conduction of the nerve impulse. The neuroplasm is the structureless ground substance of the cytoplasm in which are found high concentrations of potassium ions and other substrates critical to conduction of impulses and to cellular metabolism.

The dendrites range from a fraction of a millimeter to a few millimeters in length. An axon may range from about a millimeter up to many feet in length.

A most significant structure of the neuron is its cell membrane that acts as a sievelike barrier between the neuronal cytoplasm, which is negatively charged, and the extracellular fluid, which is positively charged. The functional integrity of the cell membrane and the difference in the bioelectric potential across this membrane are crucial to the physiology of the nerve impulse.

The site where two neurons come into contact with each other and where influences of one neuron are transmitted to the other neuron is called a

synapse. A synapse between an axon and a cell body is an axosomatic synapse, and that between an axon and a dendrite is called an axodendritic synapse. At each synapse there is a microscopic space, the synaptic cleft, between the two cells about 20 nanometers wide.

The cell membrane of the axon at the synapse is called the presynaptic membrane and that of the cell body or dendrite is called the postsynaptic membrane. The axon contains vesicles in the vicinity of the presynaptic membrane called presynaptic vesicles, which contain precursors of the neurotransmitter chemicals such as acetylcholine. The neurotransmitters are secreted across the presynaptic membrane into the synaptic cleft where they may excite (excitatory synapse) or inhibit (inhibitory synapse) the postsynaptic membrane. Although a nerve fiber may transmit an impulse in either direction (toward or away from the cell body), conduction in a sequence of neurons is unidirectional: The impulse moves toward the cell body through dendrites and away from the cell body through the axon. This direction is established because the presynaptic neuron can stimulate the postsynaptic neuron but the postsynaptic neuron cannot stimulate the presynaptic neuron. In this context, the synapse acts as a one-way valve and thereby establishes the functional polarity of the neuron. The site of contact of a nerve with a muscle, the motor end plate, is actually a synapse between a nerve and a muscle cell. *See* ACETYLCHOLINE; BIOPOTENTIALS AND IONIC CURRENTS; SENSATION; SYNAPTIC TRANSMISSION.

Interstitial cells. Except at their synapses and endings, the neurons are in intimate contact with the interstitial cells, the neurolemma (Schwann) cells and satellite cells in the peripheral nervous system, and the astroglia and the oligodendroglia in the central nervous system.

The axon of a peripheral nerve is enveloped by a sequence of neurolemmal cells forming a neurolemmal sheath. In some nerve fibers, the neurolemma elaborates a complex lipid and protein layer called myelin which forms a sheath that is segmented by interruptions at short intervals called nodes of Ranvier. Unmyelinated fibers, nerve fibers without a myelin sheath, conduct nerve impulses at slow velocities, up to about 10 ft/s (3 m/s), while myelinated fibers conduct nerve impulses at speeds of 300–400 ft/s (100–120 m/s). The thicker the myelin sheath the higher is the velocity of the nerve impulse traveling over the nerve. In cold-blooded vertebrates speed of conduction of a myelinated nerve of similar thickness is less than in warm-blooded vertebrates. In the myelinated fiber the impulse hops from one node of Ranvier to the next in what is called saltatory conduction. Satellite cells, variants of neurolemmal cells, envelop the cell bodies of ganglia associated with peripheral nerves. The white matter of the spinal cord and brain contains great numbers of fibers with a myelin sheath, although numerous unmyelinated fibers are found here. The gray matter is composed of unmyelinated fibers, dendrites, and cell bodies.

In the central nervous system, the oligodendroglia serve the same function as neurolemmal cells of the peripheral system; they envelop the cell bodies of neurons and elaborate the myelin sheath surrounding many axons. The astroglia are cells with processes extending between the blood capillaries and the cell bodies of the neurons. In effect they act as intermediaries conveying various products back and forth between the blood and the neurons. The oligodendroglia and the astroglia function to maintain a relatively constant chemical environment to enable the entire neuron to function efficiently.

Connective tissue cells. Nerve fibers of a peripheral nerve are bound together into small bundles by connective tissue cells, fibroblasts, and their fibrous products. The entire nerve is in turn surrounded by more connective tissues, within which are plexuses of blood vessels. Although the connective tissues are sparse within the central nervous system, the blood plexuses of the brain and spinal cord are most extensive. There are three layers of connective tissue membranes, the meninges, covering the brain and spinal cord: the inner, pia mater; the middle layer, the arachnoid; and the outermost, the dura mater. Between the pia mater and the arachnoid is the subarachnoid space; this space and the ventricular cavities within the brain are filled with an extracellular fluid, the cerebrospinal fluid. See MENINGES.

The microglial cells are the only parenchymal cells of the central nervous system that arise from embryonic mesoderm. These cells are similar to but smaller than macrophages which are derived from connective tissues. During stress conditions, such as inflammatory processes, infection, or traumatic injury to the brain substance, microglia become active and phagocytize and remove breakdown products from the brain and spinal cord.

Functional organization of a neuron. The neuron integrates and processes neural information. The axodendritic and axosomatic synapses stimulate receptor sites on cell membranes of dendrites and cell bodies; these are called the receptive segments. The integrative function of each segment depends, in part, upon the fact that the response to each stimulus from a synapse is graded; it is not an all-or-none nerve impulse. Only after the excitatory and inhibitory synaptic activity of the receptive segment has been resolved into an effective action does the conductile segment, the axon, generate and conduct a nerve impulse on an all-or-none action potential to the next synapse. The region of the axon associated with the synapse is known as the transmissive segment. It is at this site that the neurosecretion is released and transmitted across the synaptic cleft to stimulate the postsynaptic neuron. Charles Noback

Comparative Embryology

The complicated and varying anatomy of the adult nervous system in different vertebrates makes comparative embryological studies of these structures almost necessary for a sound understanding of their morphology. Few fields in experimental analytical embryology have proved so fruitful as that of neural

development. A thorough study of the embryology of the structures under study in animals used for experiments is necessary for a causal analysis. The embryology may be divided into a gross morphogenetic part to analyze the development of the external and internal features of the nervous system, and a histogenetic part to deal with the differentiation of the cells of the nervous system and their arrangement into nuclei and cortical structures.

Nerve growth factor. Neuronal survival, as well as the performance by neurons of some specialized functions, is regulated by macromolecular agents known as neurotrophic factors. The best studied of these agents, the nerve growth factor (NGF), was discovered in the late 1950s. Nerve growth factor is a protein present in many tissues and biological fluids, but particularly concentrated in some organs. In the organism, neurons might receive nerve growth factor from those cells they innervate, from glial cells with which they are associated, and perhaps also from the general circulation.

The two main cell groups sensitive to nerve growth factor, the dorsal root and the sympathetic ganglionic neurons, belong to the peripheral nervous system, although some central nervous system neurons may also be responsive. During some periods of their development, dorsal root and sympathetic neurons die if nerve growth factor is not present, or if they are treated with anti-NGF antibody. Beside this critical role for survival, nerve growth factor can also stimulate the neurons to produce neurites (axons and dendrites) as well as some of the enzymes necessary for the synthesis of neurotransmitters. In order to elicit these responses, nerve growth factor must first bind to receptors present on the cell surface. Very few of the molecular events linking this initial binding with the ultimate responses to nerve growth factor, such as neuronal survival or neuritic elongation, are already known.

There is additional evidence for other neurotrophic factors which regulate survival and performance of different neuronal groups. The search for these factors has been stimulated by information on developmental neuronal death, a widespread phenomenon during development of the nervous system. Of all the neurons in any given population, only those that succeed in making and maintaining synaptic contacts with "target" cells survive. It is believed that each neuron receives from its target cells a neurotrophic factor which regulates its survival. Neurons that fail to synapse with target cells do not receive the factor, and consequently die. One of the organs whose study has provided support for this hypothesis is the chick embryo ciliary ganglion. A ciliary neurotrophic factor has been found which is capable of supporting the survival of the neurons from this ganglion. This factor is present in very large amounts in the embryonic eye structures normally innervated by ciliary neurons, at the time of embryonic life when the fate (death versus survival) of these neurons is decided. In addition to this and other neuron-directed survival-promoting

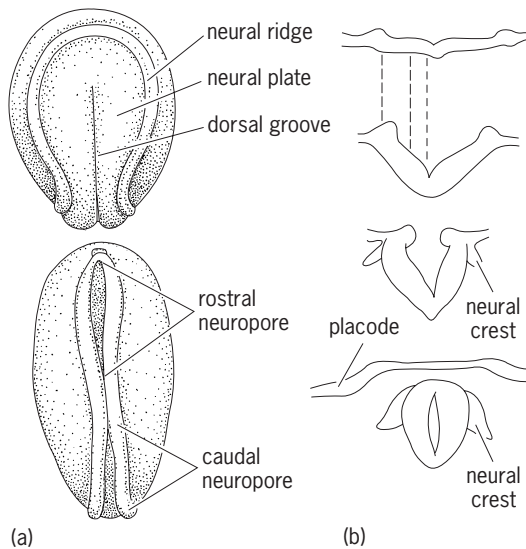


Fig. 7. Transformation of neural plate to a neural tube in amphibians (a) Dorsal views. (b) Transverse sections.

factors, other macromolecular agents are under investigation. Ruben Adler

Formation of neural plate and tube. The anlage of the nervous system is formed in the outer germ layer, the ectoderm, although some later contributions are also obtained from the middle germ layer, the mesoderm. In most vertebrates a neural plate is formed, which later folds into a neural groove, then closes to form a neural tube. In some vertebrate species, such as the lamprey and bony fishes, a massive cord of neural tissue is formed instead, which is later canalized into a neural tube.

In many vertebrates, including chickens and humans, the caudal-most end of the spinal cord is formed by a canalization of a neural cord within the so-called tail bud, while the main part arises as a neural plate.

The formation of neural tissue within the ectoderm is due to inductive influences from underlying chordomesodermal structures. See DEVELOPMENTAL BIOLOGY; EMBRYONIC INDUCTION.

The neural plate curls up at its lateral edges to form a tube (neurulation) in a process which varies somewhat in different vertebrates. The plate has an

inherent tendency to form a tube by an active contraction of microfilaments in the apical parts of the cells (close to the future lumen). In some species at least, the surrounding mesoderm apparently partakes in the folding process, pushing the margins of the neural plate (neural folds) upward.

Closing of the neural tube in most species starts in the middle part of the embryonic body, the future neck region, and continues in a rostral and caudal direction. Transitorily a rostral and a caudal neuropore exist.

At the transition between the neural plate and the ectoderm, a thickening, the neural crest, is formed (Fig. 7). In the trunk, the ganglia of the spinal nerves are formed from it. In the head some cells from the neural crest enter mesodermal structures, and others take part in the formation of cranial ganglia. The latter are also formed from ectodermal thickenings lying farther laterally, the placodes. See NEURAL CREST.

Histologic differentiation. At the site of formation of the neural plate in the ectoderm, the ectodermal cells elongate and form a cylindrical epithelium, the neural epithelium. These cells continue mitotic division and form the primary germinal layer of the central nervous system. At a later stage of development cells migrate from the epithelium and form a peripheral layer (Fig. 8).

Within the peripheral layer, and sometimes already within the neural epithelium, the differentiation of the cells proceeds toward neurons and glia cells via neuroblasts and glioblasts, respectively. Also within the ganglia, formed from the neural crest and the placodes, a similar process of differentiation occurs.

From the surrounding mesenchyme (mesoderm) cells enter the central nervous system and form vessels and microglial cells. These cells divide mitotically within the brain substance even in older embryonic stages. Mitotic division of true neural cells outside the neural epithelium is probably of low frequency. In the cerebellum, however, there is a thick proliferating layer of neural epithelial cells which also proliferates peripherally, the so-called embryonic granular layer. Similarly, a proliferating layer exists outside the neural epithelium in the cerebral

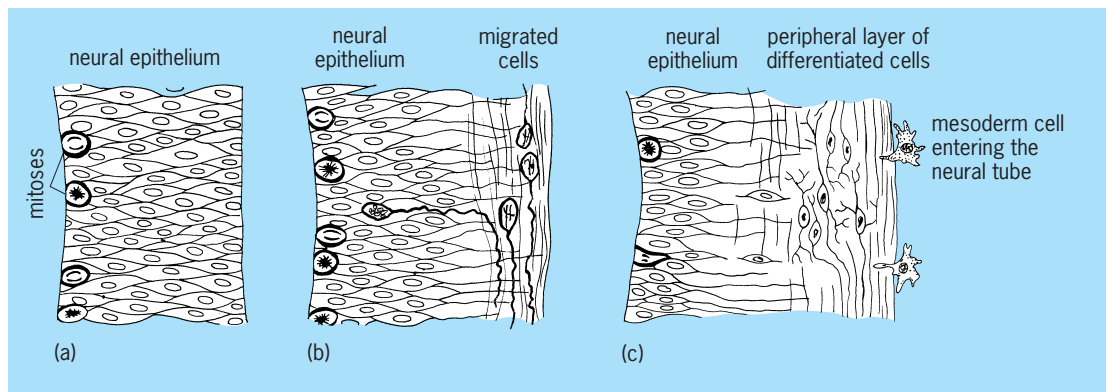


Fig. 8. Cellular constitution of neural tube in different stages a-c; ventricular wall at left.

hemispheres of some mammals, the subependymal layer.

After the formation of neurons and glia cells, the fibers which form give rise to the neuropile of the central nervous system, and to intra- and extracranial nerve bundles. The fibers emanating from the neurons of the ganglia grow as peripheral sensory fibers in peripheral nerves.

Morphogenesis. This aspect includes a consideration of the formation of neuromeres, the longitudinal structuring of the brain, and cell migration. See MORPHOGENESIS.

Neuromeres. When the neural tube is developing, a segmentation of the central nervous system occurs by the formation of transverse bulges, neuromeres (Fig. 9). They are most distinctly seen in the hind-brain region, but can usually be identified in suitable embryonic stages in all parts of the central nervous system. They are most easily seen in vertebrate brains having a thin wall, for instance, those of sharks, birds, reptiles, and mammals.

Three different sets of neuromeric bulges develop successively, called proneuromeres, neuromeres, and postneuromeres. They represent a primary, secondary, and tertiary segmentation, respectively. The basis of neuromerism is the presence of proliferative patterns. Each set of bulges thus corresponds to one period of increased proliferative activity in the neural epithelium, due to stimulative influences from underlying mesodermal structures. The proneuromeric segmentation extends from the neural tube into the neural crest and causes this to divide in the head re-



Fig. 9. Photomicrograph, dorsal view, of young chick embryo, showing neuromeres as bulges of the nervous system. (From R. deHaan and H. Ursprung, eds., *Organogenesis*, Holt, Rinehart and Winston, 1965)

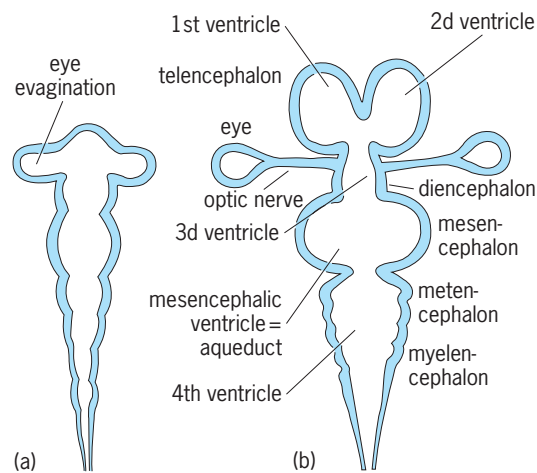


Fig. 10. Schematic horizontal sections through vertebrate brain, showing transformation of (a) neuromeric stage into (b) brain vesicle stage.

gion into portions, each corresponding to a proneuromere. This condition results in a topographic correspondence between the cranial ganglia and the neuromeres.

At the time of neuromeric segmentation, the brain is subdivided into the so-called brain vesicles by local widenings of its lumen. In the rostral end more or less well-developed hemispheres are formed; in the middle of the brain anlage the mesencephalic bulge develops; and behind the latter the walls of the tube thicken into cerebellar folds. In this way the brain anlage is divided into five sections: the telencephalon, diencephalon, mesencephalon, metencephalon, and myelencephalon, and its cavity is divided into the rudiments of the adult ventricles (Fig. 10). The brain vesicles make the segmental characters of the neuromeric bulges less conspicuous.

When the postneuromeres develop, bulges can be identified only in the brain, not in the spinal cord. The presence of the postneuromeres influences the early development of the internal structures, giving them a slight segmental character.

Longitudinal columns. When the postneuromeric phase is at its height, a longitudinal structuring of the brain wall develops, consisting of four longitudinal bands of high proliferative and migrative activity. In the hindbrain these four columns approximately correspond to the anlagen of the four columns of functionally different qualities in the adult brain. In the spinal cord, the two dorsal columns fuse into one and the two ventral columns into another. Rostrally, in the brain, the ventralmost columns stop at the rostral end of the mesencephalon and the dorsalmost columns at the transition zone between the mesencephalon and the metencephalon (the isthmic region). The two middle columns build up the rest of the brain, and the borderline between them ends at the optic chiasma. An approximate borderline between the two middle columns in the spinal cord and the myelencephalon is found in a furrow, the so-called limiting furrow of His. It cannot be identified with certainty in the rostral part of the brain.

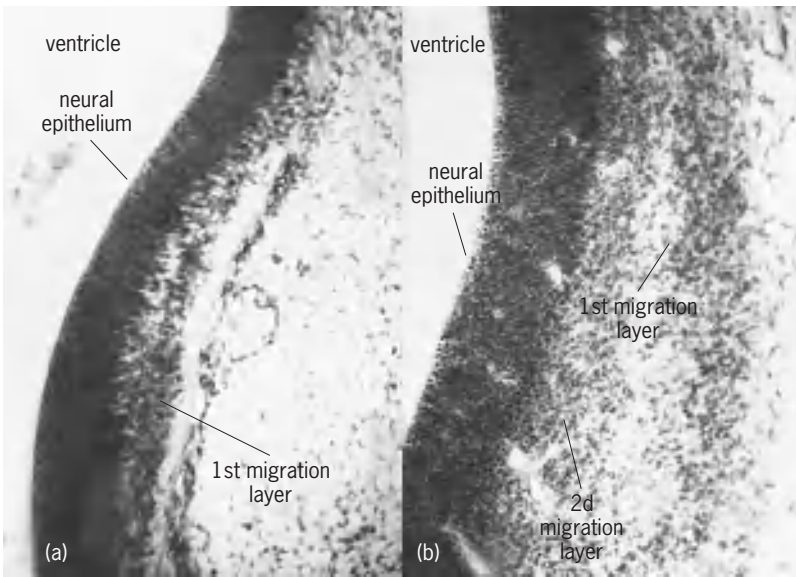


Fig. 11. Two sections through a chick embryonic hemisphere, showing the neural epithelium and the cells migrating from it. (a) One thin migration layer exists. (b) Two migration layers lie outside each other.

The postneuromery and the longitudinal banding will give rise to a checkered pattern of proliferation centers. In the rostral part of the brain the transverse pattern will dominate, the longitudinal one in the caudal part.

Cell migration. Cell migration takes place from the neural epithelium into the peripheral or mantle layer. The presence of transverse and longitudinal proliferation centers will give rise to certain areas which are rich in cells, and will cause a vivid lateral migration. Such areas are called migration areas, and their topography will be determined by postneuromery and longitudinal banding. The number and topography

of the migration areas will be very similar in all vertebrate species.

The cells, which have migrated laterally, may still lie in close contact with the neural epithelium and the ventricular wall, as in amphibians, or may lose contact with the epithelium and lie as a peripheral layer (**Fig. 11**). In many species, especially higher vertebrates, successive migrations of cells occur, giving rise to two or more layers of such cells, situated concentrically. This feature is especially well marked in the cerebral hemispheres.

Later in development, cells can migrate right through the primarily migrated cell layers—for example, the most external cortical layers can be formed later than more centrally situated cell masses.

The migration layers may fuse or further subdivide into cell clusters, which represent the anlagen of the future brain nuclei. Therefore, they furnish the basis for comparative studies and homologizations of brain nuclei of different vertebrates.

Whole cell groups or brain nuclei may migrate (group migration), and in this way the topography of the nuclei may shift even from one brain vesicle to another. Long-distance migration of scattered neural cells are also known, for example, along the surface of human hemispheres from the basal brain up to the convexities.

Cell death. Within the central nervous system cell death occurs embryonically as part of normal development. Some regions degenerate and are eliminated. This occurs, for instance, below the primary optic stalks, which in this way are shifted to their definite ventral position on the brain. The uneven thickness of the spinal cord motor cell column is also obtained by localized degeneration of neuroblasts. Furthermore, scattered cells die within the nervous system, most markedly in connection with rapid cellular differentiation. Probably, such dying cells are defective and incapable of differentiating. See CELL SENESCENCE.

Orientation of outgrowing neurites. Maturing neuroblasts send out neurites that build up fascicles and tracts connecting various neural centers.

The growth of neurites is guided by mechanical factors to some extent; chemical or electrical forces play no role, at least at a distance. The actual establishment of contacts between growing neurites and their correct target cells may be determined by cytoaffinity, that is, similarities in the properties of the cell surface.

Brain. In spite of the extraordinary variation in adult morphology of the vertebrate brain in different species, the early phases of development are essentially similar. The brain vesicle stages of a reptile, bird, and mammal are much alike, but owing to varying growth rates of different parts and to specialization processes the different patterns of the adult brains are formed (**Fig. 12**).

In a comparison of the embryology of the brains of anamniotes and those of amniotes a marked difference is seen in the so-called brain flexures. The originally straight brain tube is bent during development. In a shark or amphibian embryo the only

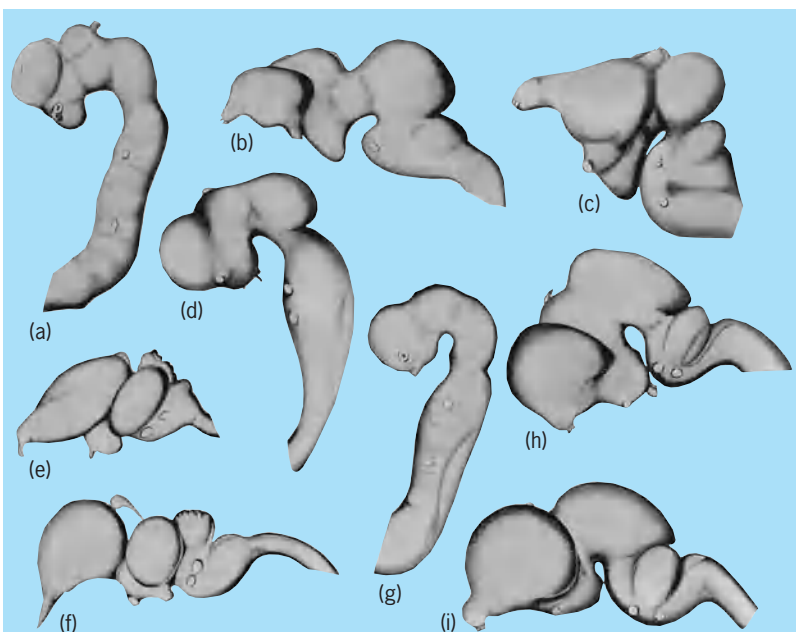


Fig. 12. Lateral views of embryonic brain at progressive stages of development. The early phases are essentially similar. (a–c) Reptile (*Chelydra*). (d–f) Bird (*Melopsittacus*). (g–i) Mammal (*Spermophilus*).

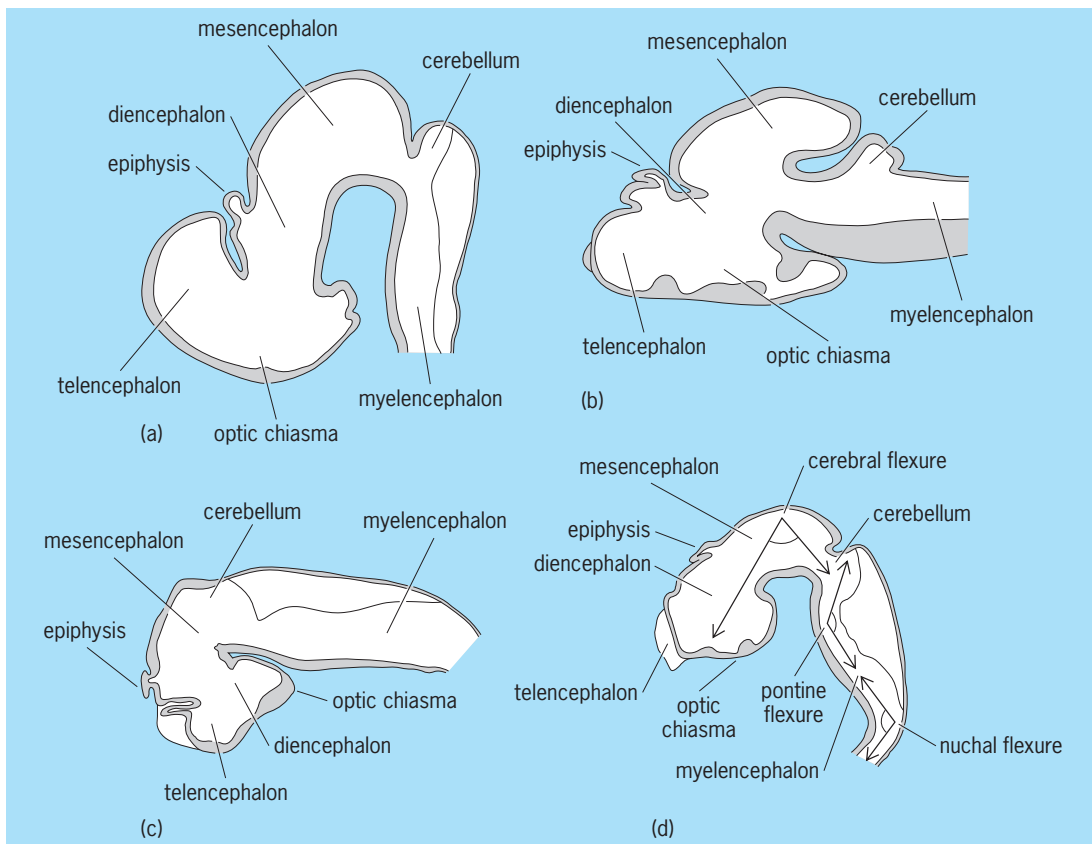


Fig. 13. Median sections of the brains of embryos. (a) Shark. (b) Bony fish. (c) Salamander. (d) Reptile.

marked bending is the cephalic flexure, situated in the same plane as the mesencephalon. This is also the first to develop in amniotes. In these brains, however, a nuchal flexure is also formed at the transition between brain and spinal cord anlagen, and later a pontine flexure arises ventral to the cerebellar region (Fig. 13). The flexures are most easily seen in median sections.

Telencephalon. The morphogenesis of the telencephalon of most vertebrates occurs by a lateral evagination or outbulging of the wall, giving rise to two hemispheric vesicles. In bony fishes, ganoids, and holocephalians, however, the lateral evagination is only faintly marked. Instead, a lateral bending, or eversion, occurs. The topography of the internal structures in the two kinds of forebrain will therefore be different (Fig. 14).

In all vertebrates two migration areas develop in the telencephalon—a dorsal one representing the embryonic origin of the pallium, and a ventral one representing the subpallium. Each of these areas is further subdivided into cell columns from which the different mantle regions and the septal and striatal nuclei develop.

Diencephalon. The morphogenesis of the diencephalon varies little in different species. A more or less well-developed transverse velum is formed in the roof, caudal to which the epiphyseal rudiment is situated. The paraphysis, which is included in the telencephalon, lies rostral to it. Part of the hypophysis develops from the bottom of the diencephalon while

from the ventrolateral parts of the diencephalon the eyes are formed. The lateral walls are divided into a dorsal thalamic and a ventral hypothalamic region, containing the mammillary bodies. The hypothalamic region grows more in size in lower vertebrates than the thalamic does, while in higher vertebrates the opposite condition exists.

Mesencephalon. The original single mesencephalic vesicle is divided into two vesicles which communicate broadly with each other. In lower vertebrates this condition remains unchanged. The original wide ventricular cavity in higher forms is reduced to form the mammalian Sylvian aqueduct. The evaginations are connected ventrally with an unevaginated part, the tectum. Within the latter, the tectal

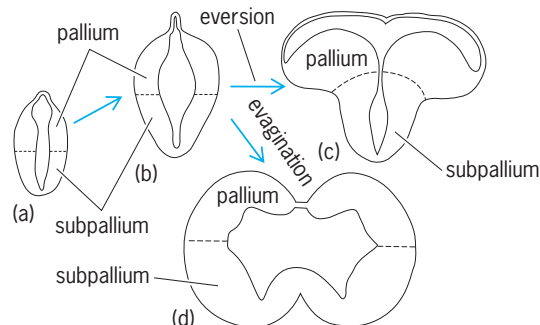


Fig. 14. Schemes showing transverse sections of forebrains. (a) Primitive stage develops via (b) interstage to (c) eversion or (d) evagination.

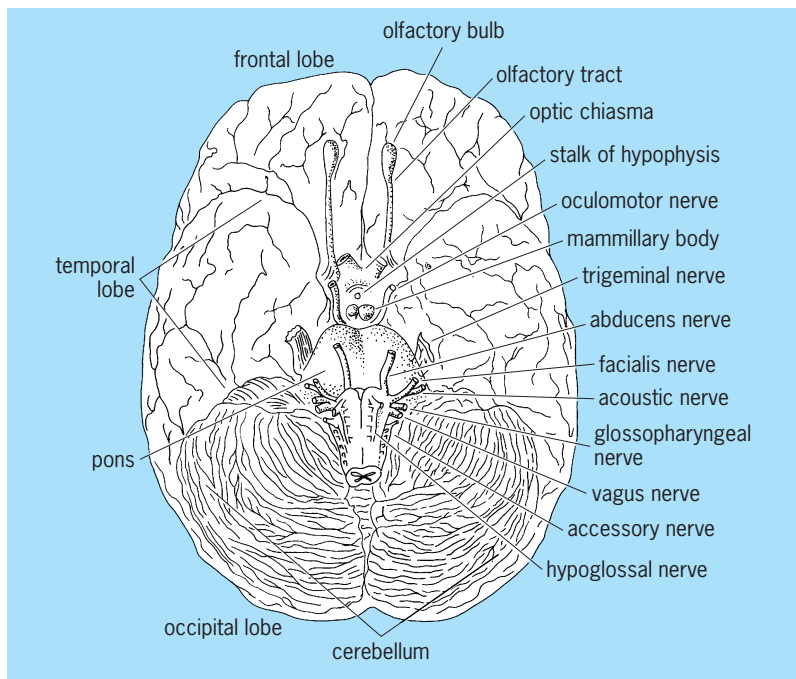


Fig. 15. Drawing of human brain seen from below.

nuclei, oculomotor nuclei, and the red and black nuclei (nucleus ruber and nucleus niger) develop. The mesencephalic evaginations form the bigeminal bodies in lower forms and the quadrigeminal bodies in higher forms.

Metencephalon. The cerebellum is formed in the dorsal part of the metencephalon. Its degree of development in different vertebrates varies considerably. The original raised lateral walls of the brain fuse to form a single plate. This is extremely compact, for example, in the bony fishes, and from its rostral end a so-called valvula grows rostrally. In *Petromyzon*, amphibians, and most reptiles the cerebellum remains as a simple transverse plate. In sharks, birds, and mammals it develops into a dome, which may be more or less folded, thereby increasing its surface.

In most vertebrates a secondary proliferative layer, the embryonic external granular layer, is formed in the periphery of the cerebellum; this layer disappears during later development.

Myelencephalon. This brain part remains relatively primitive. Its roof is extended as a thin tela; its walls form a more or less V-shaped structure with only small variations. The internal structures are dominated in their development by the above-mentioned longitudinal columns.

Spinal cord. The spinal cord remains as a comparatively slightly differentiated tube. The primary lumen is secondarily reduced by the fusion of the side walls into a narrow central canal. In the lateral walls, the longitudinal columns, separated by the limiting furrow of His, develop into the dorsal and ventral horns respectively. In fishes the diameter of the spinal cord tube gradually diminishes in a rostrocaudal direction, but in four-footed animals intumescents develop level with the extremities by

a process of partial degeneration of the regions situated in between.

Cranial nerves. The cranial or cerebral nerves are the peripheral nerves of the head that are related to the brain. The number and degree of development of the nerves varies in different species. The functional quality of the different nerves also varies. Twelve pairs of cranial nerves have been distinguished in human anatomy (Fig. 15) and these nerves have been numbered rostrally to caudally as follows:

- I. Olfactory nerve, fila olfactoria
- II. Optic nerve, fasciculus opticus
- III. Oculomotor nerve
- IV. Trochlear nerve
- V. Trigeminal nerve, in most vertebrates divided into three branches: ophthalmic, maxillary, and mandibular
- VI. Abducens nerve
- VII. Facial nerve
- VIII. Statoacoustic nerve
- IX. Glossopharyngeal nerve
- X. Vagus nerve
- XI. Accessory nerve
- XII. Hypoglossal nerve

The varied morphological significance of the cranial nerves is evident from their embryology.

Olfactory nerve. Fibers of the olfactory nerve grow out from the primary sensory cells of the epithelium of the nasal sac. They run to the lateral surface of the telencephalic rudiment, usually entering it on the border between the pallial and subpallial regions.

Optic nerve. The eyes develop as evaginations from the lateral walls of the diencephalon. The stalks of the evaginations are the pathways of the future optic nerves. The neurites growing in from the retina within the stalk are comparable to an intracerebral fascicle. They reach the brain in the floor of the diencephalon to form the optic chiasma. See EYE (VERTEBRATE).

Ventral motor nerves. Neurites emerge from cells situated in the ventralmost longitudinal column of the brain stem and leave the brain surface as motor nerves. These nerves are the oculomotor, trochlear, abducens, and hypoglossal. The trochlear nerve fibers first grow dorsad, cross in the roof of the brain, and leave it dorsally in the fold between the mesencephalon and the cerebellum. The other nerves leave the brain ventrally. In *Petromyzon* the trochlear nerve nucleus develops dorsally at the site of the future crossing.

Dorsal nerves. The dorsal nerves (trigeminal, facial, statoacoustic, glossopharyngeal, vagus, and accessory) are all mixed nerves except the statoacoustic. The sensory fibers grow out from neurons differentiated within the cranial ganglia. The motor fibers come from cells lying within the brain stem.

The cranial ganglia are formed from the head portions of the neural crest and from the ectodermal placodes. The neural crest is divided into four or five segments, called the thalamic (present only in lower vertebrates), mesencephalic, trigeminal, facial, and

glossopharyngeal-vagus crests.

The placodes of lower vertebrates are made up of two groups, those associated with the lateral-line nerve system and called the dorsolateral placodes, and those situated further ventrally and giving rise to the main ganglia, the ventral placodes. A summary of the vertebrate placodes is given below.

Spinal nerves. The spinal ganglia are formed from the neural crest which grows out like a continuous sheet from the dorsal margin of the neural tube and is secondarily split up into cell groups, the ganglia, by a segmenting influence from the somites. Fibers grow out from the ganglionic cells and form the sensory fibers of the spinal nerves. Motor nerve fibers emerge from cells situated in the ventral horns of the spinal cord. The ventral motor fibers and the dorsal sensory fibers fuse to form a common stem, which is again laterally divided into branches, innervating the corresponding segment of the body.

Autonomic nervous system. The ganglia of the sympathetic nervous system develop ventrolateral to the spinal cord as neural crest derivatives. At first a continual column of sympathetic nerve cells is formed; it later subdivides into segmental ganglia. The nerve fibers developing from these cells form the gray communicants to the spinal cord and the peripheral sympathetic nerves. The white communicants develop from spinal cord cells. Along the peripheral nerve fibers, cells migrate to form the secondary plexi and ganglia.

The parasympathetic system is made up of pre-ganglionic fibers emanating as general visceromotor fibers from the brain and from the sacral cord segments. Cells migrate to form the peripheral ganglia along them. See AUTONOMIC NERVOUS SYSTEM.

Bengt Kallen

Embryology of Sense Organs

Groups of ganglion cells, connected with the brain and spinal cord, send tiny nerve fibers through cablelike nerves to various parts of the body where they pick up many kinds of sensations which keep the living organism in touch with its environment. Therefore, specialized receptor cells and nerve endings must be provided, especially over wide areas for such senses as touch, pressure, pain, temperature, and muscle and tendon sense. Wherever possible in this discussion, the description of development of special senses in vertebrates is illustrated with human examples. See SENSE ORGAN.

Free nerve endings. Free nerve endings for pain and touch reach the skin as early as the third month in human fetuses. Their terminal branches then increase as the skin rapidly develops hair and nails from the fourth to sixth months (Fig. 16). During this time certain terminal nerve fibers slowly become encased with specialized layers of flat cells. Some near the skin, the Meissner corpuscles, receive tactile stimuli. Nerve loops (Fig. 17a) near the skin gradually become encapsulated with specialized connective tissue cells (Fig. 17b). Others, Pacinian corpuscles (Fig. 17c), receive deep pressure sense and consist of more elaborate concentric cell layers, like sheaths

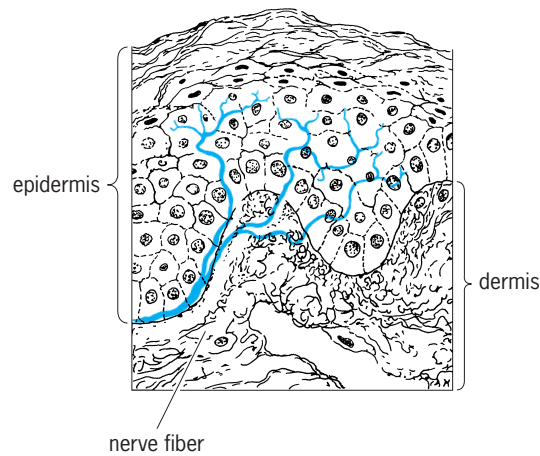


Fig. 16. Schematic of human sensory nerve fiber passing through the dermis and terminating in the free nerve endings among epithelial cells in the skin epidermis.

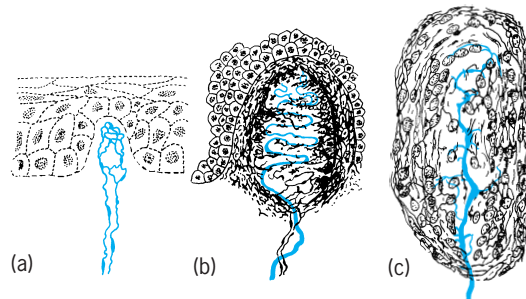


Fig. 17. Human free nerve endings. (a) Nerve loops developing Meissner corpuscle from 7-month human fetus. (b) Adult Meissner corpuscle. (c) Pacinian corpuscle in human fetus at 4 months. (After B. M. Patten, *Human Embryology*, 3d ed., Blakiston-McGraw-Hill, 1968)

of an onion, wrapped around a tiny nerve fiber. They develop in much the same way as tactile organs.

Continuing from the third fetal month many sensory nerve fibers spread over the body among developing muscle and tendon fibers, and as they branch, tiny flat plates develop at each nerve ending (Fig. 18). A delicate fibrous network of connective tissue finally covers them. The stimuli they pick up and relay to the central nervous system give the awareness of the position of the body and its parts.

Lateral-line organs. Some organs of special sense, such as the eye, are extremely complicated, whereas



Fig. 18. Developing neurotendinous fibers from human fetus of 6 months. (After B. M. Patten, *Human Embryology*, 3d ed., Blakiston-McGraw-Hill, 1968)

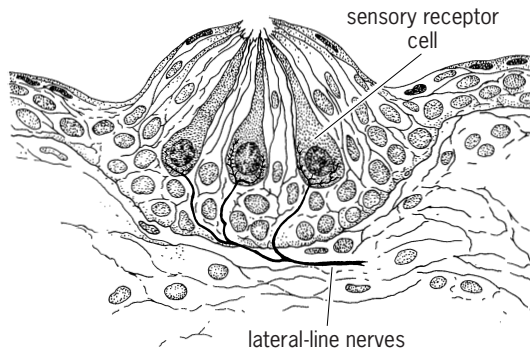


Fig. 19. Schematic drawing of a lateral-line sense organ in skin of adult salamander, the common aquatic vermilion spotted newt. Sensory receptor cells, surrounded by supporting cells and skin epithelium, terminate in hairs projecting into skin pores. Lateral-line nerves terminate around bases of the sensory cells.

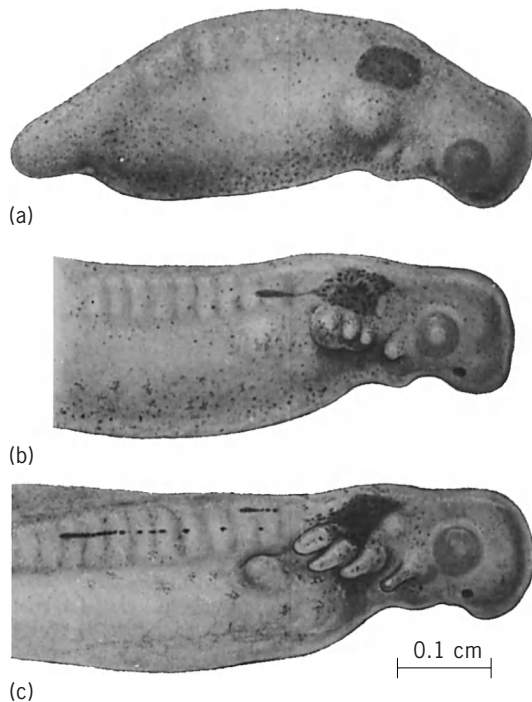


Fig. 20. Lateral-line development in salamander (*Ambystoma punctatum*). (a) Camera-lucida drawing of living embryo made 1 day after a lateral-line placode was excised and replaced by a similar one (shaded) taken from a Nile blue-stained donor. (b) Same living specimen 24 h later, showing one midbody lateral-line primordium migrating in the surface ectoderm toward the tail. (c) Same specimen 24 h later than b, showing a long and a shorter lateral-line primordium migrating down side of the body and depositing clusters of blue-stained cells that form the sense organs. (From L. S. Stone, *The development of lateral-line sense organs in amphibians observed in living and vital-stained preparations*, *J. Comp. Neurol.*, 57(3):507-540, 1933)

others are relatively simple. In some aquatic vertebrates (many fishes and amphibians) there are lateral-line skin organs (Fig. 19) on the head and body, innervated by nerve trunks coming from cranial ganglion cells connected with the brain. These organs apparently acquaint the animal with pressure changes in the surrounding water giving it a sense of orientation while swimming in a current or a warn-

ing of an approaching object. There are no homologs in humans or other animals. These tiny pear-shaped organs possess several centrally placed, club-shaped sensory cells (Fig. 19), each of which ends in a hair-like process at the free surface. These cells are interspersed and surrounded by tall, flat, overlapping, leaflike supporting cells. Fine nerve fibers from the lateral-line nerve (Fig. 19) branch among the sensory cells to receive their stimuli. The apex of the organ communicates with a microscopic pore at the skin surface in amphibians, and with a canal system in the skin of fishes.

In amphibian embryos where they have been extensively studied, ectodermal thickenings, called placodes, first appear on the side of the head. Any one of these placodes can be stained with a blue vital dye (Fig. 20a), and as the embryo grows one can follow them as they elongate, migrate on the surface of the head and body (Fig. 20b and c), and deposit at regular intervals clusters of blue cells that form the lateral-line organs. By this method one can observe the developing organs under the microscope as the blue-dye particles migrate to the tips of the sensory and supporting cells (Fig. 21). Each of them becomes innervated by the lateral-line nerve that follows the migrating placode. This nerve comes from ganglion cells which are also placodal in origin. A cluster of new secondary organs arises by a budding process from supporting cells of the primary organs (Fig. 22a and b). In practically all frogs and toads the lateral-line system degenerates at metamorphosis.

Taste buds. There are special chemical receptors somewhat like a rosebud in shape, called taste buds (Fig. 23a), or gustatory organs. They are common to all classes of vertebrates and function in a watery environment. They are associated with parts in the oral cavity, especially on the fungiform and circumvalate papillae in the mammalian tongue, but in some fishes, such as the catfish (*Ameiurus*), many taste

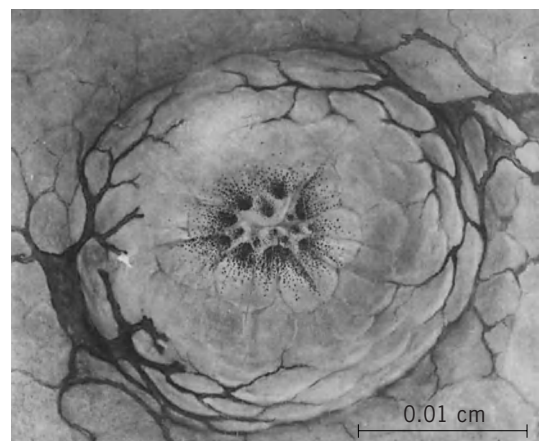


Fig. 21. Camera-lucida drawing of a differentiated living lateral-line organ surrounded by two large pigment cells in skin of young 16.5-mm (0.65 in.) salamander larva, 16 days after operation shown in Fig. 20a. Blue-dye particles were observed during development as they migrated to tips of the central sensory and surrounding supporting cells. (From L. S. Stone, *The development of lateral-line sense organs in amphibians observed in living and vital-stained preparations*, *J. Comp. Neurol.*, 57(3):507-540, 1933)

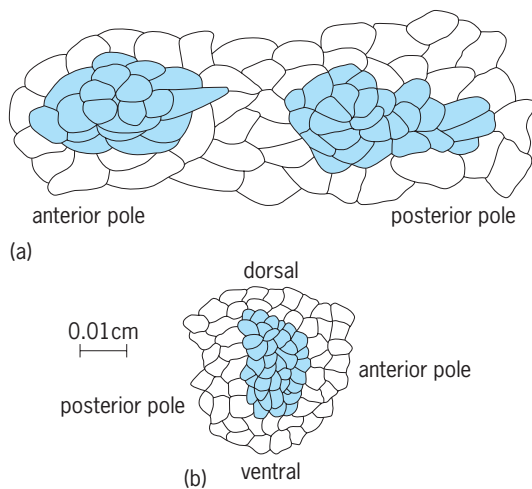


Fig. 22. Camera-lucida drawings of *Ambystoma punctatum* outlining lateral-line organs (shaded cells) in process of budding. (a) Budding organ at posterior pole was derived 24 h earlier from organ at anterior pole in tail of larva. (b) Lateral-line organ budding dorsally was observed to be derived by budding 24 h earlier from the one ventral to it. None of the organs were observed budding anteriorly or posteriorly. (After L. S. Stone, *The development of lateral-line sense organs in amphibians observed in living and vital-stained preparation*, *J. Comp. Neurol.*, 57(3):507-540, 1933)

buds are also found in the skin on the surface of the head and body. The central, rod-shaped sensory cells of mammals (Fig. 23a), neuromasts, are embraced by slender, overlapping, flat, supporting cells, the outer ends of which surround a pitlike excavation connected through a pore with the mucous epithelium of the mouth. These neuromasts, which send hairlike processes into the pit, are in contact with a basketlike network of nerve fibers. They pick up the stimuli that are then carried by the nerve fibers to the cranial gustatory ganglia and on into the brain. See TONGUE.

In the tongue of the human fetus the taste buds first appear as clusters of epithelial cells and increase in number as the gustatory nerve fibers reach the epithelium. Although the taste organs are known to degenerate eventually after gustatory nerves are cut, their arrival at the epithelium in the first place may not be the stimulus which induces the organs to form.

It has been shown conclusively by experiments on salamander embryos that the lining of the floor of the future mouth can be transplanted from one embryo to the side of the body of another embryo; a tongue with taste organs develops later without having been innervated. It was also found that if the epibranchial ectodermal placodes on the sides of the head, which give rise to the gustatory ganglia, are excised, the taste organs develop normally without a nerve supply. How these special sense organs arise in any vertebrate is not known. Taste organs, like lateral-line organs, were found to increase in number by a continuous budding process from the peripheral supporting cells of older taste organs. In many vertebrates there is a continuous increase in taste

buds for a long period. It is quite possible that this is accomplished by a similar budding process. Very little is known about the time at which the taste organs become functional. Some investigators believe that significant reflex responses can be induced in premature 7-month infants by sweet, sour, and bitter tastes.

Olfactory structures. In humans, the sense of smell also depends on special neurosensory epithelial cells functioning in a moist environment within the nasal cavities. The area of specialized olfactory epithelium lies in the upper deeper roof of the nasal mucous membrane and is made up of tall cells with bristle-like processes projecting into the mucus-covered surface where they act as chemical receptors (Fig. 23b). They are surrounded by tall supporting cells and extend toward the brain as thin fibers which contact fibers of intermediate ganglion or mitral cells. These in turn relay the olfactory impulses along the olfactory tract to the appropriate centers in the brain.

The olfactory organs arise in a similar manner in all vertebrates, by an early appearance of a pair of surface ectodermal thickenings, nasal placodes (Fig. 24), at the front end of the head. Considerable evidence from experiments on amphibian embryos indicates that the formation of nasal placodes can be induced by neighboring mesoderm and brain-forming cells.

In human embryos the nasal placodes appear during the fourth week. Very soon the placodes sink inward, forming pits which become deeper as the frame of the nose and surrounding structures of

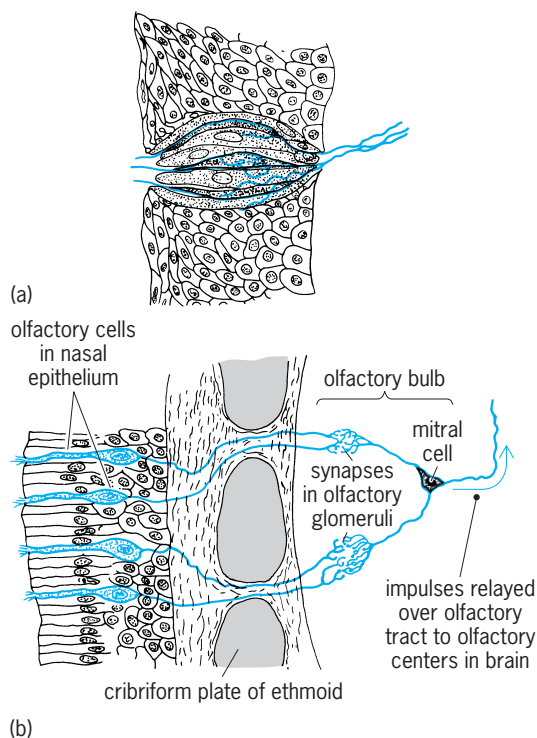


Fig. 23. Mammalian sense organs. (a) Taste bud in lingual epithelium. (b) Olfactory sensory cells of the nose related to nerve tracts leading to the brain. (After B. M. Patten, *Human Embryology*, 3d ed., Blakiston-McGraw-Hill, 1968)

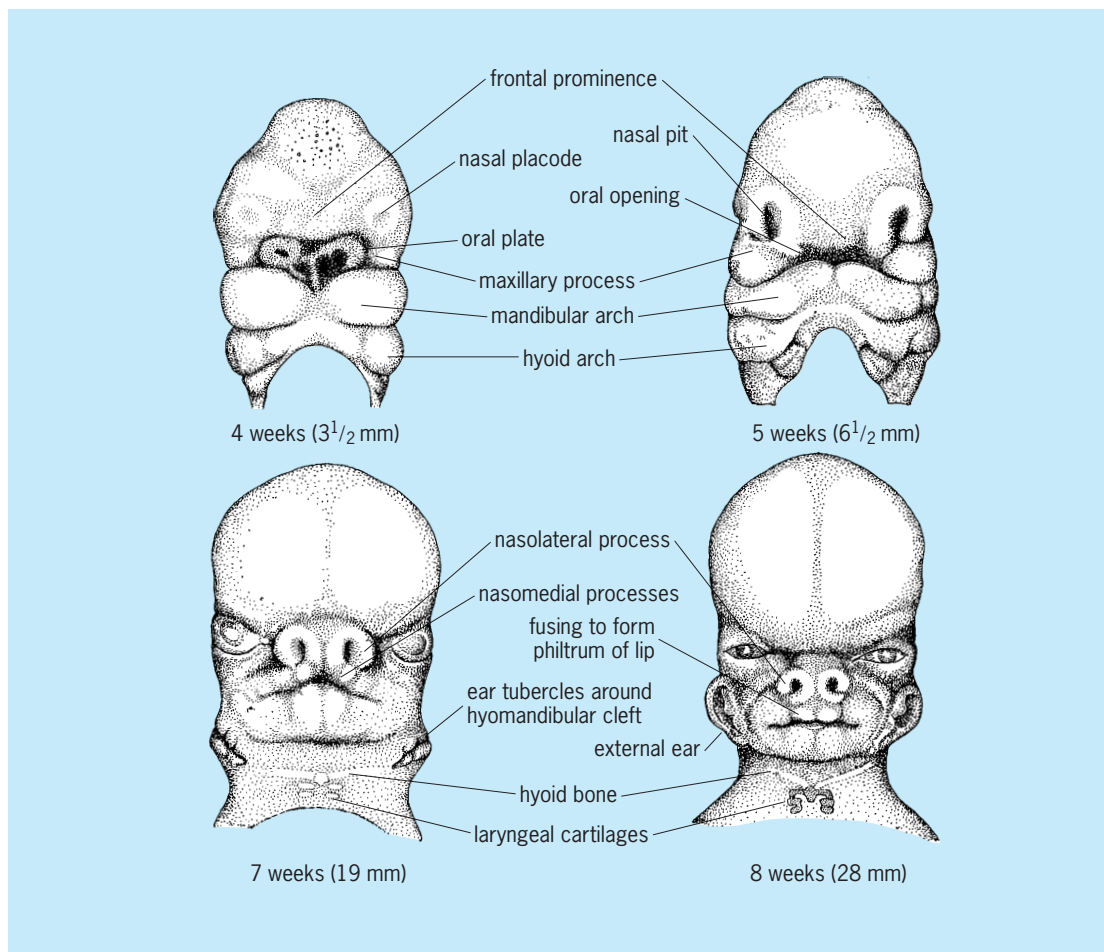


Fig. 24. Development of nose and other facial features in human embryos. (After B. M. Patten, *Human Embryology*, 3d ed., Blakiston-McGraw-Hill, 1968)

the face rapidly develop. The nasal cavities extend deeply and downward toward the oral cavity, with which they communicate shortly after the seventh week. The forward growth of the palate, nose, upper lip, and median nasal septum aids in formation of the nasal passages during the second month. By this time in the roof of the two nasal passages, specialized sensory cells of the olfactory epithelium are surrounded by tall supporting cells.

Except for the skinlike lining at the entrance of the nares, all other areas of the nasal cavities become covered by columnar epithelium with surface cilia and mucus-secreting cells. These cells keep the entire membrane covered with a moist film that provides the environment later for chemical stimulus of the hairlike ends of the sensory cells. The rate at which full differentiation of this sensory mechanism takes place varies among the vertebrates. The normality of the framework of the nose as well as the face and head depends a great deal upon the ability of the mesoderm to reach its full development.

Leon S. Stone

Bibliography. M. L. Barr, *Human Nervous System*, 7th ed., 1998; A. H. Cohen et al., *Neural Control of Rhythmic Movements in Vertebrates*, 1988; P. Nathan, *The Nervous System*, 4th ed., 1997; C. R.

Noback, N. L. Straminger, and R. J. Demasest, *The Human Nervous System: Introduction and Review*, 6th ed., 1991; G. Siegel et al. (eds.), *Basic Neurochemistry: Molecular, Cellular, and Medical Aspects*, 6th ed., 1998; M. Taussig, *The Nervous System*, 1991; G. C. Wild and E. C. Benzel, *Essentials of Neurochemistry*, 1994.

Nervous system disorders

A satisfactory classification of diseases of the nervous system should include not only the type of reaction (congenital malformation, infection, trauma, neoplasm, vascular diseases, and degenerative, metabolic, toxic, or deficiency states) but also the site of involvement (meninges, peripheral nerves or gray or white matter of the spinal cord, brainstem, cerebellum, and cerebrum). To these may be added various other correlates, such as age and sex; however, this article will consider mostly the various types of reactions and only incidentally the sites affected.

In general, nerve cells and their processes are supported by interstitial cells and nourished by blood vessels. In the central nervous system the interstitial

cells are called neuroglia (astrocytes or oligodendrocytes), and in the peripheral nervous system they are known as Schwann cells. Interstitial cells surrounding nerve cell bodies, either centrally or peripherally, are called satellite cells; those surrounding fibers form myelin, an insulating element which is composed of many wrappings of thin cell membranes which are slightly but significantly different chemically in the central nervous system as compared to the peripheral nervous system.

The nerve cell may be damaged primarily, as in certain infections, such as poliomyelitis (polio) or herpes zoster (shingles); but much more commonly the nerve cell is damaged secondarily as the result of metabolic or vascular diseases affecting other important organs, such as the heart, lungs, liver, and kidneys. This reflects the great degree of specialization of nerve cells, which have become almost totally dependent upon the metabolic functions of other organs.

Clinical signs and symptoms of dysfunction of the nervous system depend not only on the type and location of lesions but also on the severity and rapidity of development of the lesion. The nervous system grows most rapidly in infancy, slowly degenerates with advancing age, and has definite but limited regenerative capacity. Since cells mature at different rates, and since exposures to infections and poisons naturally vary with age, many diseases and their clinical signs and symptoms become age-dependent.

Malformation. The central nervous system develops as a hollow neural tube by the fusion of the crests of the neural groove, beginning in the cervical area and progressing rostrally and caudally, the last points to close being termed the anterior and posterior neuropores. If the anterior neuropore fails to close (about 24 days of fetal age), anencephaly develops. The poorly organized brain is exposed to amniotic fluid and becomes necrotic and hemorrhagic, leaving a froglike head on an infant who usually dies within hours after birth. The eyes, brainstem, cerebellum, and spinal cord are usually normal. *See* CONGENITAL ANOMALIES.

If the posterior neuropore fails to close (about 26 days of fetal age), the lumbosacral neural groove is exposed to amniotic fluid. Without the protective coverings of meninges, spines, muscles, and skin, the nervous tissue becomes partially necrotic and incorporated in a scar composed of disorganized meninges, muscles, and skin. Such a meningomyelocele is readily infected unless buried surgically within a few hours after birth. In addition, in about 95% of such infants hydrocephalus (a dilatation of the ventricles) occurs. The hydrocephalus can usually be adequately treated by shunting the ventricular fluid into the venous system or peritoneal cavity.

Other developmental disorders of the nervous system may appear as a maldevelopment (hypoplasia or hyperplasia; that is, decrease or increase in growth of cells) or as a destruction of otherwise normally developing tissues. Rapidly growing tissues such as the embryonic nervous system are generally rather easily damaged by many toxic agents which seldom

affect adult tissues; the surviving cells may grow in a less well organized pattern due to the absence of organizing influences exerted by antecedent or adjacent cells. The time of onset and the extent of repair, rather than the nature of the agent (infection, ionizing radiation, vitamin deficiency or excess, toxin, anoxia, and so on) determine the resulting pattern of abnormal development. *See* BEHAVIORAL TOXICOLOGY.

Infection. Infections of the nervous system may occur through a defect in the normal protective coverings caused by certain congenital malformations, as mentioned above, but also through other defects as the result of trauma, especially penetrating wounds or fractures sing into the paranasal sinuses or mastoid air cells. Subsequent infection of the nervous system may be the major complication of such "open head" injuries.

Pyogenic bacteria. Infections may also spread directly from adjacent structures, as from mastoiditis, sinusitis, osteomyelitis, or subcutaneous abscesses. Such infections usually spread along venous channels with progressing thrombophlebitis, producing epidural abscess, subdural empyema, leptomeningitis, and brain abscess. All of these infections are characteristically caused by pyogenic (pus-forming) bacteria (staphylococci, streptococci, pneumococci, and such), which may be normal or transient residents on the skin.

Other pyogenic bacteria (such as *Streptococcus viridans*, *Escherichia coli*, and *Hemophilus influenzae*) may metastasize by way of the bloodstream from more distant infections, such as bacterial endocarditis, pneumonia, and enteritis. The meningococcus may gain access to the blood from the nasopharynx, where it frequently resides asymptotically, and its endotoxin may produce shock and purpura (bleeding beneath the skin) and hemorrhage into the adrenal glands. In conditions of overcrowding, especially when many previously unrelated persons suddenly come together, as in military and emergency camps, meningococcal meningitis may occur in epidemics.

Infections of the nervous system must be treated promptly as medical emergencies since the mortality rate increases with delay. The diagnosis is easily established by spinal puncture, the cerebrospinal fluid appearing cloudy and containing hundreds or thousands of polymorphonuclear leukocytes per cubic millimeter. The microorganisms can be visualized with special stains.

Other microorganisms. Many other microorganisms can infect the nervous system: *Mycobacterium tuberculosis* (the organism causing tuberculosis), *Treponema pallidum* (the organism causing syphilis), several fungi and rickettsiae, and many viruses. These characteristically produce a nonpurulent leptomeningitis, the cerebrospinal fluid being clear or only slightly cloudy because the number of leukocytes is much less, usually not more than a few hundred per cubic millimeter; also, instead of polymorphonuclear leukocytes, lymphocytes predominate. Increasingly effective antibiotics are causing

dramatic improvements in most of these infections, many of which formerly were routinely fatal. *See* SYPHILIS; TUBERCULOSIS.

In addition to nonpurulent leptomeningitis, many of these microorganisms can produce other lesions, especially masses (fungal granulomas, tuberculomas) or more specific diseases, such as tabes dorsalis (with progressive loss of large afferent neurons) or general paresis (with severe inflammation of the cerebral cortex), two varieties of neurosyphilis.

Viral infections. These infections vary widely geographically, generally related to the necessity for intermediate hosts and vectors (animal reservoirs) by which the virus is spread from animals to humans, or from one human to another. Poliomyelitis, formerly occurring in summer epidemics in areas of poor sewage control, but now largely prevented by effective vaccination of most children, is primarily an intestinal infection which occasionally spreads to the nervous system, infecting and destroying motor nerve cells, thereby producing weakness of certain muscles. Herpes zoster has a similarly restricted preference for infecting sensory nerve cells and producing an acute skin eruption in the distribution of the affected sensory cells. Herpes simplex is closely related to herpes zoster, resides in the trigeminal or sacral sensory nerve cells, and intermittently produces eruptions in the distribution of these cells: "fever blisters" in and around the mouth in type I herpes, or similar blisters in the genital area in type II herpes. The latter is increasingly being recognized as a venereal disease, being widely spread especially through promiscuous homosexual contacts. Rabies virus also affects certain nerve cells in the temporal lobe, as well as in the cerebellum, and is transmitted through the saliva of animals that bite other animals or humans; rabies is the single exception to the rule that immunization must precede infection to be effective, and the immunization must begin promptly after the bite. *See* ANIMAL VIRUS; HERPES; POLIOMYELITIS.

Some viruses infect glial cells and produce demyelination, usually with little inflammation: papovavirus in progressive multifocal leukoencephalopathy and measles virus in subacute sclerosing panencephalitis. Other transmissible diseases have been identified in which there is destruction of a variety of cells, especially in the cerebral cortex but also in many other areas. In these diseases a widespread loss of nerve cells is accompanied by a spongy degeneration and marked proliferation of astrocytes. Clinically there is a slowly progressive mental and neurological deterioration after a very long latent period of many months to years; such "slow virus infections" produce scrapie in sheep and goats, kuru in the Fore tribe of New Guinea (related to their cannibalistic ritual of eating the brains of relatives who have died, usually with the disease), and Jacob-Creutzfeldt's disease in other humans throughout the world, but the route of transmission remains mysterious. *See* VIRUS INFECTION, LATENT, PERSISTENT, SLOW.

Meningeal complications. Complications of infections of the meninges include hydrocephalus (dilatation

of the ventricles due to excess cerebrospinal fluid) as the result of obliteration of the leptomeningeal spaces through which the cerebrospinal fluid must circulate before being absorbed; cranial nerve paralysis as the result of the incorporation of the nerves within the leptomeningeal inflammatory exudate; and focal destructions of neural tissue as the result of ischemia secondary to inflammation of the blood vessels (arteritis) coursing through the exudate in the meninges. *See* MENINGES.

Inflammation. Certain viruses, such as lymphocytic choriomeningitis and mumps, frequently produce a meningitis in humans from whose cerebrospinal fluid the virus is relatively easily grown. Certain other viruses, such as measles and varicella, occasionally produce meningitis or encephalomyelitis, but the cerebrospinal fluid does not contain the virus, and the hypothesis is that an allergic reaction is responsible. *See* MENINGITIS.

In infants under 2 years of age an acute encephalopathy (edema of the brain with little other evidence of inflammation) frequently occurs with varicella (chicken pox) or measles, and was one of the main objections to the former practice of vaccinating infants with vaccinia (cowpox) to prevent subsequent infection by smallpox virus. The elimination of smallpox has eliminated the necessity for continued vaccinia vaccinations and this frequently fatal complication.

Allergy to one's own tissue elements is an interesting possibility that has evoked many experimental approaches. One of the best-studied examples is experimental allergic encephalomyelitis. A particular protein (myelin basic protein) in central nervous system myelin can be made to evoke sensitized lymphocytes which recognize the same protein in the host's central myelin. This recognition results in the liberation of lymphokines (specific soluble factors) which attract other leukocytes, including macrophages that destroy the myelin. A similar specific protein in peripheral nervous system myelin can evoke experimental allergic neuritis. *See* AUTO-IMMUNITY.

Two of these human diseases, multiple sclerosis affecting the central nervous system and the Landry-Guillain-Barré syndrome affecting the peripheral nervous system, are considered likely candidates eventually to be related to experimental allergic encephalomyelitis and experimental allergic neuritis, respectively. *See* MULTIPLE SCLEROSIS.

In contrast to the remitting-relapsing course of multiple sclerosis, the Landry-Guillain-Barré syndrome is a monophasic inflammation of peripheral nerves which closely resembles experimental allergic neuritis. About half the human cases follow an acute infection of the respiratory or intestinal tracts and present with a symmetric sensory-motor paralysis of face, arms, and legs. The protein in the cerebrospinal fluid typically rises rapidly and persists for many weeks. Recovery is the rule, since the axons are preserved and the Schwann cells merely have to rewrap them, but many months may be necessary for this process to occur.

These two diseases represent the classical examples of demyelinating diseases, one central and the other peripheral, with relative sparing of neurons. Since myelin is the jelly-roll-like wrapping of oligodendroglial or Schwann cell membranes, anything which damages one or the other of these cells selectively can be said to have produced a demyelinating disease.

Vascular disease. Vascular diseases of the nervous system are commonly called strokes, a term which emphasizes the suddenness of onset of neurological disability, as though the individual were struck down by a blow to the head. Such a cataclysmic onset is characteristic of vascular diseases, since the nerve cell can function without nutrients for only a matter of seconds and will die if not renourished within several minutes. *See* VASCULAR DISORDERS.

Hemorrhage. Two main types of hemorrhage occur: hemorrhage into the subarachnoid space from rupture of an aneurysm (a focal weakening and dilatation) of a large artery, usually at the base of the brain and within or near the circle of Willis, especially the anterior half of the circle, the anterior communicating and internal carotid arteries; and hemorrhage into the brain from rupture of an aneurysm of a small artery or arteriole, most commonly one of the lateral basal penetrating (striatal) arteries which are branches of the middle cerebral artery. As might be expected in situations characterized by increased pressure, both types of hemorrhage occur more commonly in hypertensive adults.

Intracerebral hemorrhages occur directly from hypertension as the result of degeneration of the wall of the small arteries resulting in the formation of microaneurysms, which may progressively enlarge and rupture (producing hemorrhage) or may heal with fibrosis of the wall or thrombosis (occlusion) of the lumen (producing necrosis of the tissue supplied). Hemorrhages usually require minutes to a few hours to enlarge, producing increasing focal neurological deficit, but they may rupture into the ventricles or subarachnoid space, producing more general dysfunction, coma, and death in hours to a few days.

If the intracerebral hemorrhage does not greatly enlarge, the clinical differentiation from other types of strokes, which was difficult in the days before computerized tomography, can now be easily distinguished since blood appears opaque and necrosis much less opaque. Although surgical treatment of hemorrhage is possible, prevention of this complication of hypertension is the best medical practice. *See* COMPUTERIZED TOMOGRAPHY.

Subarachnoid hemorrhage occurs as the result of hypertension and arteriosclerosis combining to produce a degeneration of the internal elastic membrane (a layer of extracellular material that gives the artery its major strength), causing a defect which can occur anywhere but may be superimposed on a preexisting congenital absence of the medial muscular wall in the crotch of bifurcations of large arteries, the favorite site of aneurysms of the circle of Willis. This hypothesis of superimposition of several age-related

lesions may account for the rarity of aneurysms in children in spite of the high incidence of congenital muscular defects. The abrupt entry of blood under arterial pressure into the subarachnoid space at the base of the brain typically produces sudden severe headache, stiff neck, and unconsciousness. *See* ARTERIOSCLEROSIS; HYPERTENSION.

If clotting of the blood in and about the ruptured aneurysm does not occur promptly, the person will die in a few minutes. Even if clotting occurs, the mortality rate is very high within the next hours to days, during which careful attention must be given to establishing a good airway, sedation, control of hypertension, prevention of lysis of the clot by administration of antifibrinolytic drugs, and early operation if a space-expanding mass of hemorrhage is evident. After a few weeks the individual should be in good enough condition for consideration of direct surgical treatment of the aneurysm itself or reduction in blood flow to the aneurysm. *See* ANEURYSM.

Other causes of intracerebral or subarachnoid hemorrhages, especially in normotensive persons, include arteriovenous malformations (angiomas). The pressure of the blood leaking out is usually less, normotensive arterial or even venous, so that the prognosis is better than in hypertensive strokes; indeed, repeated hemorrhages over a period of years are common with angiomas. *See* HEMORRHAGE.

Occlusions and ischemia. The consequences of occlusion of arteries are more difficult to predict because of the varying degrees of collateral or anastomotic circulation which may be present in different individuals. Nerve cells require oxygen and glucose for functional activity, and can withstand only brief periods of hypoxia or hypoglycemia. Even a few seconds of hypoxia can block the nerve cell's function, and more than 10 min is almost certainly fatal to most nerve cells. Within these narrow limits lies most of the clinical difficulty. Transient ischemic attacks may result, with temporary impairment of blood flow to a part of the brain and consequent focal neurological dysfunction. These attacks serve as warnings of impending disaster if the circulation is restricted too long, but they may also be successfully treated with drugs or surgery and the disastrous major stroke prevented. Myocardial infarction, postural hypotension, and stenosis or narrowing of the carotid or vertebral arteries greater than 60% are common causes of cerebral ischemia. If the ischemia is not rapidly reversed, the neurons undergo selective necrosis; if the ischemia is more severe or prolonged, the glia and blood vessels in the gray matter also undergo necrosis, which appears in a pseudolaminar pattern within the middle layers of the cerebral cortex, ultimately appearing as a narrowing of the cortex; and if the ischemia is still more severe or prolonged, all the gray and white matter in the ischemic zone becomes necrotic, a condition known as cerebral infarction or encephalomalacia.

Brain swelling. Small and large strokes may be suffered clinically as the result of small or large hemorrhages or ischemic episodes occurring in or around the brain. One of the common ways the brain reacts

to these injuries is by swelling. Such swelling itself may be fatal within a few days to a week or so by herniation of the medial temporal lobe through the tentorial notch, a process known as transtentorial herniation, compressing the brainstem, where there are important neural circuits for vital functions, such as breathing and maintenance of blood pressure. If the person survives the first critical week after a massive stroke with swelling, the neurologic deficit itself is rarely fatal, but the accompanying prolonged convalescence may contribute to the development of pneumonia or pyelonephritis, which may be fatal.

Degenerative and other diseases. Degenerative, metabolic, toxic, and deficiency states include the largest numbers of both common and rare diseases of the nervous system. Since neurons in the brain may be destroyed after birth and cannot be replaced, mental deterioration, deafness and blindness, incoordination and adventitious movements, and other neurologic signs that are so typical of these disorders are generally not reversible even if the basic metabolic defect can be corrected.

During the 1970s the greatest advances were made by neurochemists and geneticists in the early diagnosis and treatment of several diseases usually manifest in infancy with mental retardation. The neuropathologic abnormalities are generally rather poorly defined with nonspecific smallness of the brain (microcephaly), retardation in myelination, and so on. Three examples are phenylpyruvic oligophrenia (phenylketonuria or PKU), which is treatable with a phenylalanine-deficient diet—there is a 0.5% loss in general intelligence for each week's delay in beginning the diet; galactosemia, requiring a galactose-free diet also as early as possible to avoid cataracts and mental retardation; and cretinism, which requires treatment with thyroid, and is an important example of a nongenetic metabolic disease with comparably disastrous results if not treated early. *See* METABOLIC DISORDERS; PHENYLKETONURIA.

More specific changes occur in the brain of the newborn (especially the premature) with hyperbilirubinemia: jaundice of the phylogenetically old parts of the brain (kernicterus or bilirubin encephalopathy) due to diffusion of unconjugated bilirubin through the blood-brain barrier. Exchange transfusions can completely prevent the mental retardation, postural stiffness (dystonia), deafness, and other features of this syndrome.

Idiopathic parkinsonism or paralysis agitans is associated with a nonspecific cortical atrophy and other diffuse changes. The most striking change is a loss of neurons of the pigmented-nerve-cell type. The pigmented neurons discharge dopamine at their synapses. This observation has led to the use of L-dopa therapeutically with dramatic effects in many cases. Huntington's chorea is a cerebral cortical and deep cerebral nuclear (striatal) atrophy resulting from inheritance of a simple dominant trait. Wilson's hepatolenticular degeneration (cirrhosis of the liver, Kayser-Fleischer corneal ring, and necrosis of the striatum) is due to accumulation of copper related to the genetically determined lack of a specific copper-

binding protein in serum (ceruloplasmin). *See* HUNTINGTON'S DISEASE; PARKINSON'S DISEASE.

Neoplasm. Neoplasms of the nervous system can be conveniently divided into primary and metastatic, the primary into gliomas and others, and the metastatic into bronchogenic and others. These four groups each account for about 25% of all intracranial neoplasms. Half of the gliomas are glioblastoma multiforme, and half of the other primary neoplasms are meningiomas and neurilemmomas. Thus, metastatic bronchogenic carcinoma, glioblastoma multiforme, meningioma, and neurilemmoma represent about half of the total and can serve as the basis for this discussion.

Meningioma. Meningioma is a benign, slow-growing neoplasm of arachnoidal elements on the undersurface of the dura at certain sites of predilection where arachnoid villi occur. It occurs as a hemispherical mass, the flat base adherent to the dura (sometimes invading the skull) and the rounded peak compressing the brain. Meningiomas typically receive a double blood supply: the external carotid from the dura and the internal carotid (or vertebral) from the brain. This great vascularity and the predilection for the base of the skull add to the difficulty in surgical excision, which often can safely be only incomplete.

Glioblastoma multiforme. Glioblastoma multiforme represents one of the most malignant brain tumors, and poor surgical results are always expected. The average postoperative survival of 4 months can be doubled by excision of most of the tumor followed by local x-irradiation. Glioblastoma appears to be predominantly a tumor of astrocytes. Hemorrhage and necrosis produce varied colors and irregular cysts contributing grossly to its multiform appearance. Glioblastoma is usually confined to one lobe of the brain, except when it crosses in the corpus callosum to produce a "butterfly" bifrontal pattern. It may occasionally occur in the brainstem and spinal cord but practically does not occur in the cerebellum. At the edge of the tumor there is frequently a marked vascular endothelial proliferation, and throughout the tumor large arteries and veins proliferate, producing typical hypervascular patterns on arteriography.

Metastatic bronchogenic carcinoma. Bronchogenic carcinoma is increasing in incidence in spite of widespread publicity regarding one of its major causes, smoking tobacco. About 10% of the cases begin with neurologic symptoms referable to cerebral metastases before pulmonary symptoms appear, and about 35% of all cases end up with metastases in the brain. About 30% of the metastases to the brain are grossly solitary; therefore, there is a reasonable place for neurosurgical therapy aimed at prolonging useful life.

Astrocytoma and medulloblastoma. All of the tumors so far discussed characteristically occur in adults. Two that occur typically in children must be mentioned because they illustrate specifically different aspects. Astrocytoma of the cerebellum is curable by excision, sometimes even if only partial, and not necessarily related to the tumor being cystic or solid. Since this is the only glioma which is curable, it should

not be misdiagnosed, nor should it be confused with solid or cystic astrocytomas occurring outside the cerebellum and being noncurable as they extend so diffusely as to invade practically the whole nervous system. Why the cerebellar astrocytoma is exceptional is not known.

Medulloblastoma also occurs in the cerebellum in children, but is a rapidly growing, malignant, undifferentiated neoplasm which seeds into and spreads widely throughout the cerebrospinal fluid. In spite of these characteristics, 30% 5-year survivals can be obtained by excising the cerebellar mass and giving x-irradiation to the whole head and vertebral column. The small, dark, oval nuclei of this neoplasm are surrounded by very little cytoplasm which theoretically can differentiate into either glial or neuronal fibers, but such differentiation is rare. See NERVOUS SYSTEM (VERTEBRATE); TUMOR. Ellsworth C. Alvord, Jr. Cheng-mei Shaw

Network theory

The systematizing and generalizing of the relations between the variables flowing in and across elements within an electrical network. To be precise, certain terms are introduced. See ALTERNATING-CURRENT CIRCUIT THEORY.

Elements. The elements of a network model are resistance, inductance, and capacitance (the passive elements) and sources of energy (the active elements), which may be either independent sources or controlled, that is, dependent, sources. An independent-voltage source produces a voltage across its terminals that is not dependent on any current or voltage, although it may be a function of time, as in the case of an alternating source; an independent-current source carries current that is independent of all voltages or currents but may be a function of time. See CAPACITANCE; ELECTRICAL RESISTANCE; ELECTROMAGNETIC INDUCTION; ELECTROMOTIVE FORCE (EMF); INDUCTANCE.

A number of definitions, together with theorems that relate them, are taken from the mathematical subject of topology. Two or more elements are joined at a node (**Fig. 1**). If three or more elements are connected together at a node, that node is called a junction. (The term major node may be used instead of junction; topological terminology varies among authors.) An element extends from one node to another. A branch of a network extends from one junction to another and may consist of one element or several elements-connected in series. A loop, or circuit, is a single closed path for current. A mesh, or window, is a loop with no interior branch. See TOPOLOGY.

Figure 1 shows a network with 13 elements, of which 12 are passive and 1 is active. Nodes are indicated by dots; of the 9 nodes in the figure, 6 are junctions. There are 10 branches of which 5 come together at a single node or junction at the bottom of the figure.

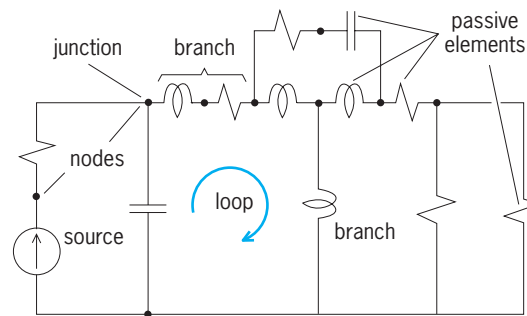


Fig. 1. Parts of a network. (After H. H. Skilling, *Electrical Engineering Circuits*, 2d ed., Wiley, 1965)

Equation (1) is an extension of Ohm's law, where

$$V = IZ \quad (1)$$

I is the current through an element, Z is the impedance of that element, and V is the voltage or potential difference between the nodes that terminate that element; this applies to every passive element of an electrical network. [Equation (1) may be a phasor equation for steady alternating current, in which case impedance is a function of frequency and $V(j\omega) = I(j\omega)Z(j\omega)$, or it may more generally be the transform of the differential equation, $V(s) = I(s)Z(s)$.] The number of elements (active or passive) in a network may be designated as E . See ELECTRICAL IMPEDANCE; OHM'S LAW.

At every node of an electrical network the sum of currents entering that node is zero. Equation (2) ex-

$$\sum I = 0 \quad (2)$$

presses Kirchhoff's current law. An equation of this form can be written for each node, but in a fully connected network one of these equations can be derived from the others; hence the number of independent node equations is one less than the number of nodes. The number of independent node equations, called N , equals in a fully connected network the number of nodes minus one. See KIRCHHOFF'S LAWS OF ELECTRIC CIRCUITS.

Around every loop of an electrical network the sum of the voltages across the elements is zero. Equation (3) expresses Kirchhoff's voltage law. A net-

$$\sum V = 0 \quad (3)$$

work such as that shown in Fig. 1 can have many possible loops and hence many equations of this form, but only a limited number are independent. If L is the number of independent loops, topology gives Eq. (4), from which L can be computed, E and N

$$E = N + L \quad (4)$$

having been counted. Thus in Fig. 1 there are 13 elements and 9 nodes, hence 8 independent nodes, so that there are $13 - 8$ or 5 independent loops; that is, $L = 5$.

If a network is planar (flat) and fully connected, the number of independent loops is equal to the number of meshes, or windows. Figure 1 shows such a

network, and the number of meshes is obviously 5, so again $L = 5$.

Branch equations. There are E elements in a network. Suppose that all impedances are known. There is a voltage across each element, and there is a current through each element; hence there are $2E$ voltages and currents to be known. One equation is provided by each element; either the element is a source for which voltage or current is given, or it is an impedance for which there is a relation given by Ohm's law in the form of Eq. (1). Hence there are E equations from the elements. From Eqs. (2) and (3) the nodes and loops provide $N + L = E$. Thus there are $2E$ equations relating voltages and currents.

In an actual solution for current and voltage in a network, it is probably desirable to reduce the number of equations by combining elements that are in series. This can reduce the number of elements to the number of branches and the number of nodes to the number of junctions; the network is then described by $2B$ branch equations.

These branch equations are easy to write but tedious to solve, unless a computer is used. Two modifications have been devised, however, that eliminate a great deal of the labor and reduce the number of equations from $2B$ to either L , the number of independent loops, or N , the number of independent nodes, as will be described in the following paragraphs.

Linearity. Although branch equations can be written for networks containing either linear or nonlinear elements, the solution is more difficult for nonlinear networks. A linear network is one that gives rise to linear systems of equations, which are subject to special methods of solution, and for which the principle of superposition applies, allowing the use of loop or node equations. In a linear system the values of resistance, inductance, and capacitance are constant with respect to voltage and current, and a controlled source produces a voltage or a current that is proportional to another voltage or current. See SUPERPOSITION THEOREM (ELECTRIC NETWORKS).

Fortunately, many electrical networks are linear or are nearly enough linear to be so considered, at least in the useful range of operation or in a piecewise linear fashion. Examples of branch, loop, and node equations are given below for linear systems.

Examples. Using Ohm's law six times, the branch equations for the network of Fig. 2 are shown in Eqs. (5).

$$\begin{aligned} V_{BC} &= Z_{BC}I_{BC} & V_{CD} &= Z_{CD}I_{CD} \\ V_{BD} &= Z_{BD}I_{BD} & V_{DA} &= Z_{DA}I_{DA} \\ V_{CA} &= Z_{CA}I_{CA} & V_{BA} &= Z_{BA}I_{BA} + E \end{aligned} \quad (5)$$

The source electromotive forces (emf) and branch impedances are assumed to be known, and all currents and branch voltages are to be found. There are, then, 6 equations and 12 unknowns, the unknowns being the voltage across and current through each branch. (Although a source is indicated in only one of these branches, the method can be applied in the same way if there are sources in any or all of the

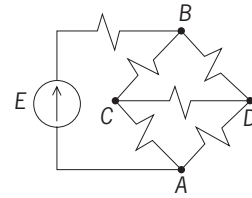


Fig. 2. Network of six branches. (After H. H. Skilling, *Electrical Engineering Circuits*, 2d ed., Wiley, 1965)

branches.) Clearly, 6 more equations are needed.

The 6 needed equations may be called connection equations, and they are found from Kirchhoff's laws. The current law gives Eqs. (6), called node equations

$$I_{BC} + I_{BD} + I_{BA} = 0 \quad (6a)$$

$$I_{CB} + I_{CD} + I_{CA} = 0 \quad (6b)$$

$$I_{AC} + I_{AD} + I_{AB} = 0 \quad (6c)$$

or junction equations. Equation (6a) shows the sum of the currents flowing out of junction B to be zero. Equations (6b) and (6c) make similar statements with regard to junctions C and A.

Three more equations may be obtained from Kirchhoff's voltage law. The sum of the voltages around loop BCD must be zero, Eq. (7a); the sum of

$$V_{BC} + V_{CD} + V_{DB} = 0 \quad (7a)$$

$$V_{CA} + V_{AD} + V_{DC} = 0 \quad (7b)$$

$$V_{AB} + V_{BC} + V_{CA} = 0 \quad (7c)$$

the voltages around loop CAD must be zero, Eq. (7b); and the sum of the voltages around loop ABC must be zero, Eq. (7c).

There are now 12 equations. Recognizing that $I_{BC} = -I_{CB}$, and so on with other currents and voltages, there are still 12 unknowns, for the 6 connection equations of Eqs. (6) and (7) have added no new unknowns. The 12 equations can be solved simultaneously for the 12 unknowns. The actual solution is not particularly interesting and will not be pursued.

It would seem to be possible to write a fourth junction equation at junction D: $I_{BD} + I_{CD} + I_{AD} = 0$. This proposed equation contains no new information, however, for it could have been derived from the other three equations. It results from adding Eqs. (6a), (6b), and (6c) and canceling equal and opposite quantities. Although it is a true equation, it is not an independent equation. The solution for 12 unknowns requires the use of 12 independent equations.

A somewhat similar observation can be made about the independence of loop equations. The three loops for which equations are written are not the only possible ones. For example, there is also loop ACBD, but this loop will not give another independent equation, nor will any of the other possible loops.

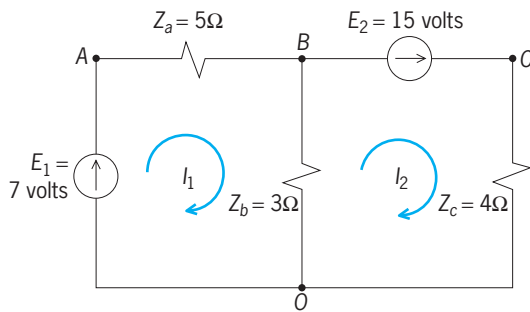


Fig. 3. Network of two loops. (After H. H. Skilling, *Electrical Engineering Circuits, 2d ed.*, Wiley, 1965)

Loop equations. The ingenuity of the loop method lies in the selection of currents to be determined. It is necessary to find only as many different currents as there are independent loops, instead of finding as many different currents as there are branches.

Thus in the network of Fig. 3 a current which can be called I_1 flows around loop 1. (This is the current in the source E_1 and in the impedance Z_a .) The convention generally adopted is to take the reference direction for all loop currents as clockwise, and this will be taken as the reference direction of I_1 . The reference direction of the current called I_2 that flows around loop 2 is also clockwise. This is the current in the source E_2 and in impedance Z_c . Loop currents I_1 and I_2 both flow in Z_b . The reference direction of one is downward and that of the other is upward. Thus the total current downward in Z_b is $I_1 - I_2$. When I_1 and I_2 are known, every current in the network is determined.

The first step in solving for these currents is to write network equations in terms of I_1 and I_2 . Kirchhoff's voltage law is used to write Eq. (8), which

$$E_1 - Z_a I_1 - Z_b (I_1 - I_2) = 0 \quad (8)$$

says that the sum of all voltages about loop 1 is zero. Rearranged, Eq. (8) becomes Eq. (9). This equation

$$(Z_a + Z_b)I_1 - Z_b I_2 = E_1 \quad (9)$$

is valid when there is current in both loops of the network.

Similarly, Eq. (10) is written for loop 2 of the net-

$$(Z_b + Z_c)I_2 - Z_b I_1 = E_2 \quad (10)$$

work. This is done from Fig. 3 without further explanation. The first term is the voltage in loop 2 when the only current is I_2 , and the second term is the voltage in loop 2 caused by I_1 . Since E_2 has the same reference direction as I_2 , it is positive.

Equations (9) and (10) are a pair of equations with only two unknowns, I_1 and I_2 . They can be solved simultaneously by Eqs. (11a) and (11b). By way of illus-

$$(Z_a + Z_b)I_1 - Z_b I_2 = E_1 \quad (11a)$$

$$-Z_b I_1 + (Z_b + Z_c)I_2 = E_2 \quad (11b)$$

tration let the electromotive forces and impedances be given numerical values and solve for the currents

as follows. It is assumed that the equations are linear (see above), which implies that the resistances are constant, and that the electromotive forces are independent of current and voltage.

Examples. If the impedances shown in Fig. 3 are pure resistances and with the values shown, Eqs. (12) and (13) may be written. The electromo-

$$8I_1 - 3I_2 = 7 \quad (12)$$

$$-3I_1 + 7I_2 = 15 \quad (13)$$

tive forces may be either dc values or phasors of alternating voltages (the solution is the same). These are linear equations that can be solved by any convenient means. Using determinants and Cramer's rule, they are solved in Eqs. (14) and (15), where D , the

$$I_1 = \frac{\begin{vmatrix} 7 & -3 \\ 15 & 7 \end{vmatrix}}{\begin{vmatrix} 8 & -3 \\ -3 & 7 \end{vmatrix}} = \frac{49 + 45}{56 - 9} = \frac{94}{47} = 2 \quad (14)$$

$$I_2 = \frac{\begin{vmatrix} 8 & 7 \\ -3 & 15 \end{vmatrix}}{D} = \frac{120 + 21}{D} = \frac{141}{47} = 3 \quad (15)$$

denominator of I_2 , is the same as the denominator of I_1 . The results are two loop currents in amperes. See LINEAR SYSTEMS OF EQUATIONS.

Now all unknown quantities in the network may be easily found. Current in the central branch is $I_{BO} = I_1 - I_2 = -1$ ampere. The negative sign indicates that 1 amp is flowing upward. If the bottom node (node O) is taken to be the reference node at an assumed zero potential, then the potential at node B is -3 volts. The potential at node A is $-3 + 5 \cdot 2 = 7$ volts, which is also the electromotive force of the source E_1 . The potential at node C is $-3 + 15 = 12$ volts, which can also be found (across Z_c) as $4 \cdot 3 = 12$ volts. The most convenient way to specify all the voltages of a network is to give the potentials at the various nodes with reference to some one node that is arbitrarily assumed to be at zero potential.

Standard notations. It is customary to use a standard system of symbols for writing loop equations. Equations (11) are specific examples of the general form shown in Eqs. (16). Equations (16) are a set of L si-

$$Z_{11}I_1 + Z_{12}I_2 + Z_{13}I_3 + \cdots + Z_{1L}I_L = V_1 \quad (16a)$$

$$Z_{21}I_1 + Z_{22}I_2 + Z_{23}I_3 + \cdots + Z_{2L}I_L = V_2 \quad (16b)$$

$$Z_{31}I_1 + Z_{32}I_2 + Z_{33}I_3 + \cdots + Z_{3L}I_L = V_3 \quad (16c)$$

$$Z_{L1}I_1 + Z_{L2}I_2 + Z_{L3}I_3 + \cdots + Z_{LL}I_L = V_L \quad (16d)$$

multaneous linear equations, applying to the L loops of a network; the network may be any network and L may be any number. In writing the equations, the following conventions are used. Each loop current is numbered, as I_1 , I_2 , and so on. V_1 is a voltage in loop 1 that is not taken into account by the terms of the left-hand side of the equation. It may be, as it

was in Eq. (9), an independent electromotive force. It may be the sum of several voltages, and must include all voltages that appear in loop 1 when all the other loops are open. Note that its nominal positive direction is taken to be that of I_1 .

The total impedance about loop 1 is Z_{11} . In Eq. (11), which applies to Fig. 3, Z_{11} is $Z_a + Z_b$. It might include many more elements if the network were more complicated. Z_{11} is called the self-impedance of loop 1. It could be measured by means of a bridge or other impedance-measuring instrument connected in place of the source in loop 1, all other loops of the network being opened during the measurement. Each loop has its self-impedance: Z_{22} , Z_{33} , and so on.

Certain branches are common to two loops. Thus Z_b in Fig. 3 is in both loop 1 and loop 2. In such a case, current in one loop produces voltage in another loop, and there is said to be mutual impedance. By definition, the mutual impedance is the ratio of such a voltage in one loop to the current in another loop that produces it. That is, if current in loop 2 is I_2 and mutual impedance with loop 1 is Z_{12} , then the resulting voltage in loop 1 is $Z_{12}I_2$. For example, Eq. (11) shows that the voltage produced in loop 1 of Fig. 3 by the current in loop 2 is $-Z_bI_2$. By definition, then, $Z_{12} = -Z_b = -3$. The negative sign results from the fact that in the common element the reference direction of I_2 is opposite to the reference direction of I_1 . If positive I_1 produces a positive voltage, positive I_2 in the same element produces a negative voltage. In network computations the reference directions are commonly assumed in such a way that the mutual impedances (such as Z_{12} , Z_{23} , and so on) are negative quantities. However, it is not wrong to direct the arrows that indicate the nominal positive direction of current in such a way that mutual impedances are positive.

In Eq. (16b) the first term contains Z_{21} . Comparison with Eq. (11) shows that, for the network of Fig. 3, $Z_{21} = -Z_b$. Z_{21} is therefore equal to Z_{12} . When two loops contain resistors or coils or capacitors in a common branch, current in loop 1 will produce the same voltage in loop 2 that equal current in loop 2 would cause in loop 1; therefore $Z_{21} = Z_{12}$. The general form for this relation is given in Eq. (17). Al-

$$Z_{pq} = Z_{qp} \quad (17)$$

though less evident, this relation is still true if two circuits are coupled by a magnetic field (as in a transformer); whatever the turn ratio, the mutual inductance L_{21} equals the mutual inductance L_{12} . If the coupling between the loops is by means of an electric field through mutual capacitance with no conductive connection, Eq. (17) is again valid. Indeed, it fails only for circuit elements that are not bilaterally symmetrical. (For example, a transistor or a vacuum tube is not bilateral.)

The order of the subscripts attached to Z has the following significance. The first subscript is the number of the equation in the array of Eqs. (16) and

therefore agrees with the subscript attached to V in that equation. The second subscript is the number of the term in the equation and therefore agrees with the subscript attached to I in that term. Thus $Z_{pq}I_q$ is a voltage in circuit p produced by a current in circuit q .

Nodal equations. Loop equations are written based on the concept of loop currents. This makes it unnecessary to give any attention to Kirchhoff's current law, for loop currents necessarily add to zero at every node, and Kirchhoff's current law is automatically satisfied. The loop-current concept therefore reduces the number of equations that must be solved simultaneously from the $2B$ equations of the branch method to L , the number of independent loops, which is usually about one-fourth as many.

In the node-equation method, the simplifying concept is the idea of measuring voltage from all the nodes of the network to one particular node that is called the reference node, or the datum node. This makes it unnecessary to give any attention to Kirchhoff's voltage law. It is only necessary to satisfy Kirchhoff's current law at each node, for the voltage law is automatically satisfied. Thus the number of simultaneous equations is reduced to the number of independent nodes N , a number much smaller than $2B$ and comparable with L .

Whether the node method or the loop method is the more convenient depends on the network. Some networks have fewer loops than nodes, and some have fewer nodes than loops. Other factors also affect the relative convenience, as will be seen.

Figure 4 shows a network with two independent nodes; that is, it has three nodes, one of which is the reference node O , and the others are marked A and B . Kirchhoff's current law is used for node A , Eq. (18), and node B , Eq. (19). Assume that the po-

$$I_a = I_b + I_c \quad (18)$$

$$I_e = -I_d + I_c \quad (19)$$

tential at node O is zero; if the potential at node A is called V_A , then $I_b = Y_bV_A$. Also, $I_d = Y_dV_B$ and

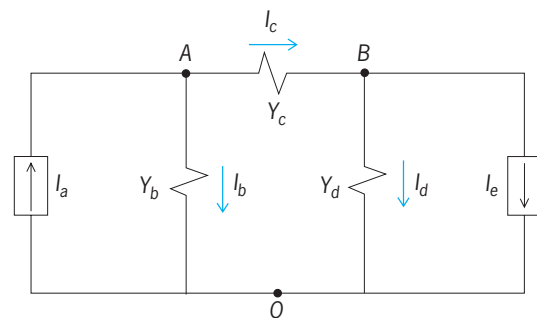


Fig. 4. Network with two independent nodes; a third one is the reference node O . (After H. H. Skilling, *Electrical Engineering Circuits*, 2d ed., Wiley, 1965)\vskip-1.2pt

$I_c = Y_c(V_A - V_B)$. Equations (20) and (21) may now be solved for I_a and I_e .

$$I_a = Y_b V_A + Y_c(V_A - V_B) = (Y_b + Y_c)V_A - Y_c V_B \tag{20}$$

$$I_e = -Y_d V_B + Y_c(V_A - V_B) = Y_c V_A - (Y_c + Y_d)V_B \tag{21}$$

With practice it is easy to write the final form of Eqs. (20) and (21) directly. To write Eq. (20), note that the independent value I_a of incoming current is set equal to the current that would flow from node A if all other nodes were at zero potential, that is, $(Y_b + Y_c)V_A$, from which is subtracted the current that would flow to node A if it were at zero potential while all other nodes were at their actual potentials, that is, $Y_c V_B$. To write the first term, assume that all nodes except A are short-circuited to node O ; to write the second term, assume that node A alone is short-circuited to node O .

To write Eq. (21) at node B , the independent incoming current $-I_c$ is equated to current that would flow from node B if node A were at zero potential, that is, $(Y_c + Y_d)V_B$, less the current that would flow to node B if it alone were at zero potential, in this case $Y_c V_A$; the result, with all signs changed, is Eq. (21).

Standard notations. There is a standard form for writing node equations similar to the standard form for loop equations. For a network of N independent nodes, Eqs. (22) may be written. Y_{AA} is called the

$$\begin{aligned} Y_{AA}V_A + Y_{AB}V_B + Y_{AC}V_C + \dots + Y_{AN}V_N &= I_A \\ Y_{BA}V_A + Y_{BB}V_B + Y_{BC}V_C + \dots + Y_{BN}V_N &= I_B \\ Y_{CA}V_A + Y_{CB}V_B + Y_{CC}V_C + \dots + Y_{CN}V_N &= I_C \\ \dots & \\ Y_{NA}V_A + Y_{NB}V_B + Y_{NC}V_C + \dots + Y_{NN}V_N &= I_N \end{aligned} \tag{22}$$

self-admittance at node A , and in the example it is equal to $(Y_b + Y_c)$. Note that Y_{AA} is the sum of all admittances attached to node A . Y_{BB} , Y_{CC} , ... are self-admittances at the other nodes.

Y_{AB} is the mutual admittance between nodes A and B . In the example $Y_{AB} = -Y_c$; Y_{BA} also equals $-Y_c$. Both Y_{AB} and Y_{BA} are the sum of all admittances connected directly between nodes A and B but written with a negative sign.

I_A is another current flowing toward node A . In the example it is the source current designated I_a in Fig. 4. Similarly, I_B is current toward node B , and in the example given it is the source current $-I_c$.

It will be noted that every term in the node Eqs. (22) is a current, as every term in the loop Eqs. (16) is a voltage.

Examples. As a numerical example, the following values are given to the impedances and the source currents of Fig. 4, and then voltage at the two independent nodes A and B and current in the three impedances are obtained. Note that the source currents are taken to be real numbers, implying either

that they are direct current, or (if ac sources) that they are in phase with each other; the impedances are taken to be real, and this implies pure resistance. In a practical problem one voltage and all impedances might well be complex, and this would complicate the arithmetic but not change the method of solution. Given that $I_a = 2$ amperes, $I_e = 5$ amperes, $Y_b = 1/4$ mho, $Y_c = 1$ mho, and $Y_d = 1/2$ mho, node equations (22) are written as Eqs. (23) and (24).

$$(1/4 + 1)V_A - 1V_B = 2 \tag{23}$$

$$-1V_A + (1 + 1/2)V_B = -5 \tag{24}$$

These equations may be solved for V_A and V_B using Cramer's rule; they are shown as Eqs. (25) and (26).

$$V_A = \frac{\begin{vmatrix} 2 & -1 \\ -5 & 3/2 \end{vmatrix}}{\begin{vmatrix} 5/4 & -1 \\ -1 & 3/2 \end{vmatrix}} = \frac{-2}{7/8} = -16/7 \text{ volts} \tag{25}$$

$$V_B = \frac{\begin{vmatrix} 5/4 & 2 \\ -1 & -5 \end{vmatrix}}{7/8} = \frac{-17/4}{7/8} = -34/7 \text{ volts} \tag{26}$$

Currents I_b , I_c , and I_d are given in Eqs. (27)–(29). Equations (30) and (31) confirm these results. Note

$$I_b = V_A Y_b = -16/7 \cdot 1/4 = -4/7 \text{ amperes} \tag{27}$$

$$I_c = (V_A - V_B)Y_c = (-16/7 + 34/7) \cdot 1 = 18/7 \text{ amperes} \tag{28}$$

$$I_d = V_B Y_d = -34/7 \cdot 1/2 = -17/7 \text{ amperes} \tag{29}$$

$$I_a = I_b + I_c = -4/7 + 18/7 = 2 \text{ amperes} \tag{30}$$

$$I_e = I_c - I_d = 18/7 + 17/7 = 5 \text{ amperes} \tag{31}$$

that currents I_b and I_d both turn out to be upward because of the large value of the source current I_e , and that therefore the two nodes A and B are both at negative potential compared with node O .

Thévenin's theorem. It is often convenient before applying network theory to simplify a problem by means of Thévenin's theorem. This theorem and its dual, Norton's theorem, may be expressed in many ways, but the following is among the more useful: Open-circuit voltage V_θ and short-circuit current I_θ are measured (or computed) at a pair of terminals of an active linear network. The active network is equivalent at these terminals to either an independent voltage source V_θ in series with an impedance $Z_\theta = V_\theta/I_\theta$, or alternatively to an independent-current source I_θ in parallel with the same Z_θ (Fig. 5). The former alternative is Thévenin's theorem, the latter is Norton's. A proof is not given; however, Eqs. (32)–(37) illustrate the application. See THÉVENIN'S THEOREM (ELECTRIC NETWORKS).

For the example of node equations Fig. 4 is used with the data given above to find currents

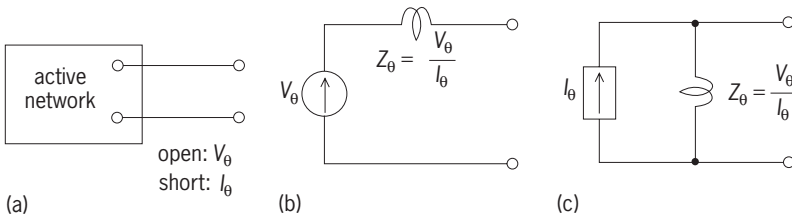


Fig. 5. Networks: (a) active, (b) Thévenin equivalent, and (c) its Norton equivalent. (After H. H. Skilling, *Electrical Engineering Circuits*, 2d ed., Wiley, 1965)

and voltages in the network, using Thévenin's theorem.

Thévenin's theorem is applied twice to Fig. 4 to obtain the circuit of Fig. 6. Voltage V_1 and impedance Z_1 are equivalent to current I_a and admittance Y_b ; similarly V_2 and Z_2 are equivalent to I_e and Y_d , as follows.

In Eqs. (32) the open-circuit voltage is computed at each pair of terminals with Z_c open.

$$V_1 = \frac{I_a}{Y_b} = \frac{2}{1/4} = 8 \quad (32a)$$

$$V_2 = \frac{I_e}{Y_d} = \frac{5}{1/2} = 10 \quad (32b)$$

Short-circuit current is computed with each of nodes A and B short-circuited to node O in Eqs. (33).

$$I_1 = 2 \quad (33a)$$

$$I_2 = 5 \quad (33b)$$

The impedances, Eqs. (34), may be obtained from Eqs. (32) and (33).

$$Z_1 = \frac{V_1}{I_1} = 8/2 = 4 \quad (34a)$$

$$Z_2 = \frac{V_2}{I_2} = 10/5 = 2 \quad (34b)$$

The current I_c may be obtained from Fig. 6, Eq. (35).

$$I_c = \frac{V_1 + V_2}{Z_1 + Z_c + Z_2} = \frac{8 + 10}{4 + 1 + 2} = 18/7 \text{ amperes} \quad (35)$$

Voltages V_A and V_B at nodes A and B are given by

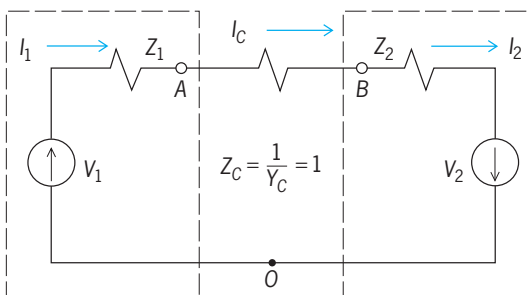


Fig. 6. Thévenin equivalent of Fig. 4.

Eqs. (36) and (37). These answers are, of course, the

$$V_A = V_1 - Z_1 I_c = 8 - 4(18/7) = -16/7 \text{ volts} \quad (36)$$

$$V_B = -(V_2 - Z_2 I_c) = -[10 - 2(18/7)] = -34/7 \text{ volts} \quad (37)$$

same as those obtained by node equations (25) and (26).

Many other network theorems are available, but perhaps none are as useful as Thévenin's and Norton's.

Hugh Hildreth Skilling

Bibliography. J. Choma, *Electrical Networks: Theory and Analysis*, 1985, reprint 1991; R. D. Strum and J. R. Ward, *Electric Circuits and Networks*, 2d ed., 1985; G. H. Tomlinson, *Electric Networks and Filters*, 1991; J. Vlach, *Basic Network Theory with Computer Applications*, 1992; A. K. Walton, *Network Analysis and Practice*, 1987; J. E. Whitehouse, *The Principles of Network Analysis*, 1991.

Neural crest

A strip of ectodermal material in the early vertebrate embryo inserted between the prospective neural plate and epidermis. After closure of the neural tube the crest cells migrate into the body and give rise to parts of the neural system: the main part of the visceral cranium, the mesenchyme, the chromaffin cells, and pigment cells. The true nature of the neural crest eluded recognition for many years because this primary organ has a temporary existence; its cells and derivatives are difficult to analyze when dispersed throughout the body. The fact that mesenchyme arises from this ectodermal organ was directly contrary to the doctrine of the specificity of the germ layers. See GERM LAYERS.

Neural crest no doubt exists, with similar qualities, in all vertebrate groups, including the cyclostomes. It has been most thoroughly studied in amphibians and the chick.

Crest cells. The dorsal ectoderm of a vertebrate gastrula forms the pear-shaped neural plate which is surrounded by the neural ridge (Fig. 1). When the neural plate rolls up to form the neural tube, the ridges from each side meet and fuse, temporarily forming a wedge-shaped cell mass in the dorsal midline of the prospective brain and spinal cord. In Fig. 1 the head part of the ridge of a urodele neurula is divided into eight zones. Zones 1 and 2 may be called the transverse ridge. Zone 8 and the posterior parts are the trunk ridge. The neural crest cells are situated in the ridge but do not occupy the whole ridge. For example, the posterior slope of the transverse ridge consists of material for the forebrain, and the anterior slope forms epidermis. Investigations have confirmed that probably no crest cells emanate from the transverse ridge. Concerning the main parts of the neural ridge it is still doubtful whether the crest is located only in the peripheral part of the thick plate or

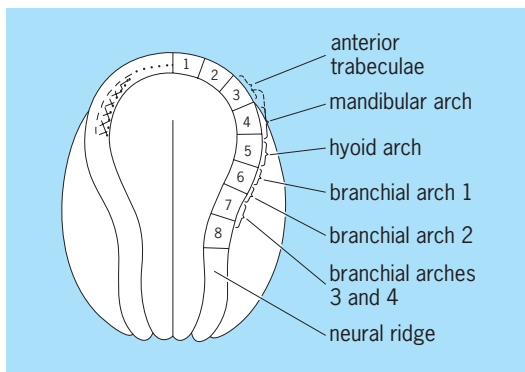


Fig. 1. Diagram of open neural-plate stage of a urodele, *Ambystoma mexicanum*. Left side, the epidermal line of coalescence (broken line), the line of coalescence of the brain (dotted line), and the presumptive ectomesenchyme (hatched area). (After S. Hörstadius, *The Neural Crest: Its Properties and Derivatives in the Light of Experimental Research*, Oxford, 1950, facsimile ed., Hafner, 1969)

in the adjacent part of the thinner ectoderm as well. At least in zones 1–4 of the *Ambystoma* neurula the line of coalescence of the epidermis coincides with the midline of the ridge, the crest material thus forming the proximal slope of the ridge (hatched area in Fig. 1). The exact position of the material in the posterior parts and in other vertebrate groups is not known.

The crest cells do not remain long in the dorsal midline of the neural tube. They migrate in a lateral direction between the tube and epidermis, first as rather coherent sheets, then in groups, strands, or as single cells, both proximal and distal to the somites. Because they are similar to the mesenchymal cells originating from the mesoderm, their fate has been ascertained only by vital staining, extirpation, or transplantation experiments in many cases. One method has been to replace a part of the ridge by a corresponding piece from another species with cell nuclei of different size which are thus recognizable in sections. The fate of crest cells has been studied using a radioautographic method whereby cells of transplants were labeled by tritiated thymidine. Most extirpations and transplantations must be made bilaterally because after unilateral operation, crest cells migrate down in great quantities from the intact to the operated side, and extirpations must also include rather long pieces of the ridge because the gap may be filled by cells from the crest anterior and posterior to the excision. Transplantation of neural crest to another part of the body as well as explantation have been used in order to determine what it can give rise to in new surroundings. In spite of all efforts, however, the problem of the fate of the neural crest is still not solved in all its details (Fig. 2). See FATE MAPS (EMBRYOLOGY).

Contributions to the neural system. Parts of the migrating bilateral cell masses settle in clusters and give rise to the spinal ganglia. The segmental arrangement of the dorsal nerve roots and the ganglia is not intrinsic to the cells but is determined by the segmental arrangement of the somites. There seems to be no doubt that the spinal ganglia are formed entirely of

neural crest cells. There has been much controversy concerning the origin of the cranial ganglia. Placodes of the lateral epidermis of the head are one source of their cells; the other is the neural crest, specifically for the sensory components of the ganglia of cranial nerves V, VII, IX, and X, although to different degrees in different groups of animals. See NERVOUS SYSTEM (VERTEBRATE).

According to both histological observations and experiments, the sympathetic trunk has been considered to be derived not only from the crest but also from the ventral part of the spinal cord; however, crest contribution has been denied by some. Excision experiments on chicks support crest origin, and the experiments with labeled crest cells described above favor the theory that the whole sympathetic complex emanates from the crest.

The cellular sheath of Schwann enclosing peripheral nerves has been considered by some authors to have its origin from crest cells, by others to be formed by cells emerging from the neural tube by way of the ventral roots. Experiments with labeled cells have confirmed both these opinions; that is, a first wave of cells comes from the crest, and a later wave comes from the neural tube. The majority of the leptomeninx (pia and arachnoidea) cells are derived from the neural crest, whereas the pachymeninx (dura) is composed chiefly of endomesodermal cells. For differentiation of crest cells into cells of Schwann and meninges the presence of neural material seems indispensable.

Chromaffin cells. The chromaffin cells of the suprarenal organs of lower vertebrates as well as the medullary zone of the suprarenal gland in mammals, and also those of the paraganglia, that is, the aortic bodies of Zuckerkandl and the carotid bodies, are all considered to be formed by cells emanating from the sympathetic system, thus probably mainly from the crest. See ADRENAL GLAND; AORTIC BODY.

Mesenchyme and skeleton. The statement by J. B. Platt in the 1880s that mesenchyme could arise from ectoderm (mesectoderm, ectomesoderm, or ectomesenchyme) raised a vivid discussion for several decades. Experiments on amphibian larvae confirmed that, of the chondrocranium, the anterior half of the trabeculae, all visceral arches, and the first basibranchial are of neural crest origin, whereas the main part of the neurocranium and the second basibranchial are endomesodermal (Fig. 3). Crest cells may also give a small contribution to the basal plate, the parachordalia, and the posterior trabeculae.

In addition, membrane bones, namely the tooth-bearing premaxilla, vomeropalatine, dentary, and splenial, as well as the odontoblasts and dentine, are formed by cells emanating from the neural crest. Moreover, the ectomesenchyme of the teeth induces the formation of enamel organs, but only in connection with an inducing action from the foregut endoderm. See SKELETAL SYSTEM.

The transverse ridge, zones 1 and 2 in Fig. 1, has no cartilage-forming capacity. The trabecular material situated in zone 3 cannot partake in formation of the visceral arches. The mandibular ectomesenchyme

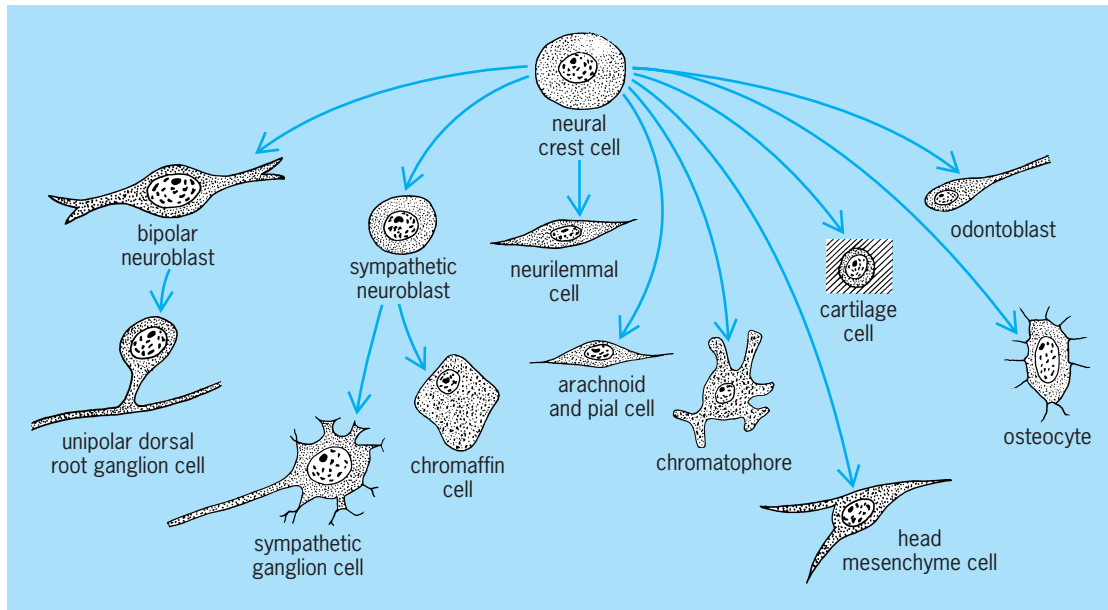


Fig. 2. Schematic representation of the possible developmental fates of a neural crest cell. (After W. J. Hamilton, J. D. Boyd, and H. W. Mossman, *Human Embryology*, 2d ed., Williams and Wilkins, 1952)

located in zone 4 and perhaps also in the adjacent part of zone 3 cannot form trabeculae but is capable of forming other visceral arches, although there is no fusion with the basibranchials; the material of the hyoid arch and the gill arches in zones 5-7 shows such capacity to fuse. In zone 8 the trunk crest begins; it has no faculty to form cartilage but differentiates to form spinal and sympathetic ganglia as well as chromaffin, pigment, and other mesenchymal cells. Crest cells in the neural ridge of the head are not yet determined to form cartilage. No cartilage is formed when they are transplanted to the trunk unless they are activated by certain tissues, such as the pharyngeal endoderm and to some extent the gastric endoderm, endomesoderm of the gill region, chorda, and wounds in somites. The necessity of induction has been confirmed by tissue culture experiments. Ectomesenchyme and Schwann cells differentiate in laboratory cultures, but cartilage, dentine, teeth, and

balancers appear only in the presence of foregut endoderm and stomodeal ectoderm or cartilage after treatment by extracts from foregut. The differentiation is controlled by a phenylalanine metabolism in the archenteral roof.

To what extent corium cells are of crest origin is not clearly established. Head mesenchyme is also formed by the crest. In urodele larvae no dorsal fin is formed, nor do the larval gills develop normally in absence of the crest.

Pigment cells. With the exception of the retinal pigment, which is formed in place, probably all chromatophores in vertebrates are derived from the neural crest. Such an origin has been observed in fishes, amphibians, birds, and mammals. Pigment cells are found in both dermis and epidermis, feathers, hair, the perineural and perivascular layers, the coelomic wall, and the choroid and iris of the eye. The origin of pigment cells was first traced in the twentieth century. Because the pigment is formed rather late in embryonic development, the presumptive pigment cells cannot be distinguished from mesenchymal cells in early stages. There are brown to black melanophores, yellow lipophores (xanthophores), reddish erythrophores (allophores), and guanophores and iridocytes with resplendent crystals. Bilateral extirpation of sufficiently large pieces of neural crest results in unpigmented regions of the body. Following extirpation of all the neural ridges, urodelian larvae subsequently obtain pigment cells emerging from the brain and the tail-bud part of the neural plate. Normally the reserve in these regions is probably inhibited by the pigment cells from the crest. Explants of neural crest may give pigment cells in tissue cultures. Implantation of neural crest into a larva of another species has produced pigment of the type of the donor amphibians, birds, and mammals.

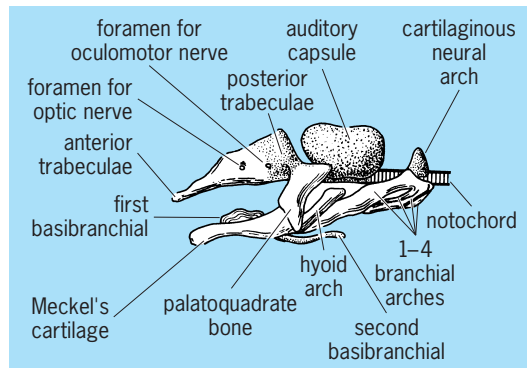


Fig. 3. Diagram of chondrocranium from left of *Ambystoma mexicanum* larva. Cartilage of endomesodermal origin dotted. (After S. O. Hörstadius, *The Neural Crest: Its Properties and Derivatives in the Light of Experimental Research*, Oxford, 1950, facsimile ed., Hafner, 1969)

The development of pigment in the chromatophore does not result from self-differentiation but is dependent upon factors in surrounding cells. For example, chromatophores of the black axolotl produce no pigment with white specimens, whereas chromatophores from the white race are activated and form melanin granules in the skin of a black specimen. Some prospective pigment cells in amphibians may remain dormant during larval development and become activated at metamorphosis. The activation of the propigment is evidently brought about by an oxidase. The pigment pattern depends both on properties intrinsic in the propigment cells and on the surrounding tissues. Also, when transplanted to a host of very distant relationship, as from anuran or urodele, the single cells differentiate in size, shape, and color as they would have done in the donor. However, their distribution may vary. See CHROMATOPHORE.

In reciprocal exchange of neural crest between some species of urodeles the pattern in the respective regions becomes that of the donor, but in less related forms influence of the host is more or less recognizable. Grafting of epidermis from another species has hardly any effect. In some urodele species there is an intrinsic tendency in the crest cells to aggregate at certain levels, for example, along the dorsal border of the somites. Melanoblasts invade developing feathers, producing their specific color and pattern. They also invade growing hair, giving pigment granules to it. But in some cases, both in the bird and mammal, the color is a differentiation dependent upon the epidermis. The black and white stripes in the feather pattern of the Barred Plymouth Rock chicken are probably the result of an inhibitory action from each black stripe on the melanoblasts of the prospective white stripe during the process of feather growth. See EMBRYONIC INDUCTION; EYE (VERTEBRATE); NEURULATION.

Sven Hörstadius

Bibliography. J. Brachet and H. Alexander, *Introduction to Molecular Embryology*, 2d rev. ed., 1986; G. M. Edelman, *Topobiology: An Introduction to Molecular Embryology*, 1993; S. Nona et al. (eds.), *Development and Regeneration of the Nervous System*, 1992.

Neural network

An information-processing device that consists of a large number of simple nonlinear processing modules, connected by elements that have information storage and programming functions. The field of neural networks is an emerging technology in the area of machine information processing and decision making. This technology makes extensive use of the terminology and methods of artificial intelligence. The main thrusts are toward highly innovative machine and algorithmic architectures, radically different from those that have been employed in conventional digital computers, and toward overcoming the major technological shortcomings and inherent limitations imposed by the traditional information-processing machines. The information-processing el-

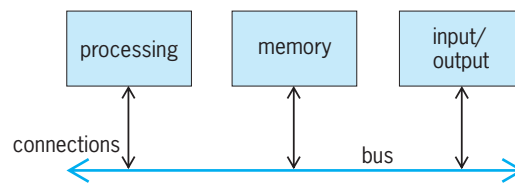


Fig. 1. Architecture of conventional digital computer.

ements and components of neural networks, inspired by neuroscientific studies of the structure and function of the human brain, are conceptually simple. Three broad categories of neural-network architectures have been formulated which exhibit highly complex information-processing capabilities. Several generic models have been advanced which offer distinct advantages over traditional digital-computer implementation. Neural networks have created an unusual amount of interest in the engineering and industrial communities by opening up new research directions and commercial and military applications.

Three decades of rapid growth in the field of neural networks were followed by a period, beginning about 1990, of consolidation and realignment to mathematics, microbiology, and molecular technology. The goal of this work is the establishment of comprehensive foundations to deal with the enormously difficult problems emanating from such areas as globalization of commerce, post-genome health care, and automation of human knowledge and intelligence.

Basic properties and architectural features. Automated information processing is achieved by means of modules that in general involve four functions: input/output (getting in and out of the machine), processing (executing prescribed specific information-handling tasks), memory (storing information), and connections between different modules providing for information flow and control. Neural networks contain a very large number of simple processing modules. This contrasts with traditional digital computers, which contain a small number of complex processing modules that are rather sophisticated in the sense that they are capable of executing very large sets of prescribed arithmetic and logical

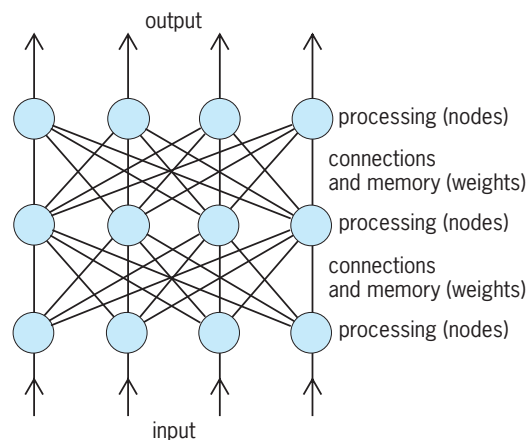


Fig. 2. Architecture of multilayer neural network.

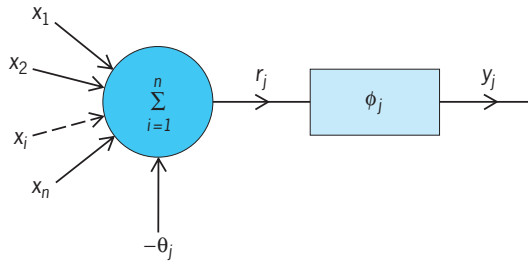


Fig. 3. Node or artificial neuron, the information-processing module of a neural network.

tasks (instructions). In conventional digital computers, the four functions listed above are carried out by separate dedicated machine units (Fig. 1). In neural networks information storage is achieved by components which at the same time effect connections between distinct machine units (Fig. 2). These key distinctions between the neural-network and the digital computer architectures are of a fundamental nature and have major implications in machine design and in machine utilization, to the extent that the term neurocomputing is gaining widespread usage.

For conceptual and economic reasons, the processing modules of commercial digital computers are designed to be so powerful that most contemporary conventional machines need to have only one, the central processing unit (CPU). In these machines, processing takes place sequentially, one instruction after the other. Because of this, the burden of carrying out a specific information-processing job (for example, the analysis of a physical problem or the sorting of given data) falls upon a human being who must invent a precise, step-by-step algorithm appropriate to the specific job, which in coded form (the program) is entered, together with pertinent data, in the memory components of the machine for execution. This storage of programs and data in memory prior to execution by the processor also necessitates a hierarchy of memory components. Frequently this limits real-time machine operation in digital computers. See COMPUTER STORAGE TECHNOLOGY.

By contrast, the information-processing module of neural networks, the node or artificial neuron, is very simple (Fig. 3). A node j has n simultaneous inputs $x_i, i = 1, 2, \dots, n$, and a threshold quantity θ_j , which are combined according to a simple arithmetic expression, Eq. (1), to produce single quantity

$$r_j = \sum_{i=1}^n x_i - \theta_j \quad (1)$$

r_j . This is followed by a nonlinear transformation,

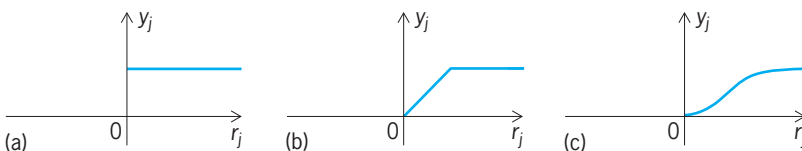


Fig. 4. Types of nonlinear transformations at the node of a neural network. (a) Hard limiter. (b) Threshold logic. (c) Sigmoid.

Eq. (2), of one of three types (Fig. 4), which pro-

$$y_j = \phi_j(r_j) \quad (2)$$

duces the output y_j . The weighted output of one node can then become one of the inputs to a different node. Thus the connection of external inputs and nodes by means of information weight parameters W_{ij} forms a neural network. A connection between node i and node j is a distinct machine component which constrains information flow (one way) via the connecting weight parameter W_{ij} (Fig. 5) whose numerical value is problem-dependent.

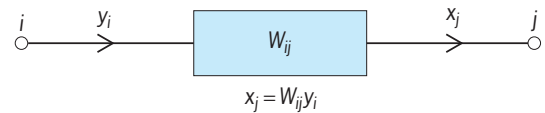


Fig. 5. Connecting weight in a neural network.

Processing information with a neural network is accomplished by providing input quantities to some or all of the nodes and extracting output quantities from some or all of the nodes. Information is processed inherently in real time, with the values of the connecting weights W_{ij} serving both as the memory and as the program of the machine. The computational power is derived from the continuous nonlinear transformations that take place at the nodes. It is known that these transformations have much richer mathematical properties than the types of discrete arithmetic and logical operation-based algorithms used in conventional computers. This is the reason for the interest in and potential of neural networks. See COMPUTER ARCHITECTURE; DIGITAL COMPUTER.

Classification and principal models. The information-processing properties of neural networks depend mainly on two factors: the network topology (the scheme used to connect elements or nodes together), and the algorithm (the rules) employed to specify the values of the weights connecting the nodes. While the ultimate configuration and parameter values are problem-specific, it is possible to classify neural networks, on the basis of how information is stored or retrieved, in four broad categories: neural networks behaving as learning machines with a teacher; neural networks behaving as learning machines without a teacher; neural networks behaving as associative memories; and neural networks that contain analog as well as digital devices and result in hybrid-machine implementations that integrate complex continuous dynamic processing and logical functions. Within these four categories, several generic models have found important applications, and still others are under intensive investigation.

Learning machines with a teacher. The general features of the neural-network architecture were introduced in a serious attempt to develop an electronic machine model capable of carrying out all the functions of the human brain. However, the

first model that proved practical to implement as a general-purpose machine that learns to classify known patterns, with the aid of a teacher, is the perceptron, which still is the archetype of this category of neural networks. In the simplest perceptron model, all input nodes are connected to a single layer of conceptual elements (Fig. 2), which also provides all outputs. Each input node is forward-connected to many and possibly all of the output nodes through weights W_{ij} . (No connection from an output node to an input node is permissible.) An elegant training algorithm was found (the perceptron algorithm) which allows a teacher to program the model by adjusting the weights W_{ij} in sequential training sessions so as to be able to classify known patterns.

The perceptron algorithm is successful for those pattern-recognition problems that can be solved by means of a single-layer perceptron structure. Unfortunately, it is also known that not all pattern-classification problems can be solved by this model. Multilayer perceptron models allow hidden layers of forward-connected nodes (nodes not accessible by either the input or the output, Fig. 2), thereby enlarging significantly the class of pattern-recognition and -classification problems that can be solved with a teacher-trainable machine. However, the task of adjusting the connection weights between nodes to train them is a much more complex task. A promising approach is the so-called back-propagation algorithm which makes it possible to modify the weights of successive node layers (starting with the output layer) on the basis of least-mean-square error. Another class of perceptron-inspired structures is cerebellar modular arithmetic computers (CMAC), which realize learning machines capable of performing known input/output transformations found in robotics and in goal-oriented actions in general. *See* COMPUTER VISION; INTELLIGENT MACHINE; ROBOTICS.

Learning machines without a teacher. This category involves network topologies and training algorithms which promote self-organization. That is, the members of the input set are presented to the machine sequentially in time and, although there is no knowledge of the desired machine output, the weights connecting the nodes are adjusted each time according to an algorithm which ultimately results in a machine organization that embodies clusters of characteristic features of the input set. Within this category two generic neural-network models and training algorithms have received great attention: a two-dimensional array which produces self-organized feature maps of the inputs, and the adaptive resonance classifier which involves a connected pair of two-dimensional arrays.

Associative memories. In networks in this category, information is stored and retrieved on the basis of content and not by means of a numerical tag or an index (memory address). The cornerstone of traditional digital computer structures and programs is memory addressing with the hardware-imposed requirement that it will be accomplished sequentially, one step before the next. Addressing information by

its content, however, is inherently a process that involves a high degree of parallelism. Thus, retrieving information, stored by means of an index, on the basis of its partial content is a very tedious and time-consuming task. Examples of this are making a list of all the persons residing in a given street by using telephone-book entries, or retrieving a complete set of data or images from data fragments. By contrast, the neural-network architecture, with its inherently large number of parallel paths, is naturally suited to addressing information by association (the weights of the connections). A successful model involves a mathematical approach that views neural networks as nonlinear dynamical system and analyzes their stability properties on the basis of energylike functions. *See* CONCURRENT PROCESSING.

Hybrid machines. Bridging mathematics, systems engineering, and computer science, this category is based on innovative architectures and learning techniques that incorporate elements from predicate calculus, fuzzy logic, or genetic algorithms. Among the most interesting applications are structures that incorporate or realize expert systems or fuzzy systems, automatically extract knowledge (that is, production rules, crisp or fuzzy) from experimental data used for pattern recognition and system identification, and generate conceptual clustering or heuristic discoveries. *See* FUZZY SETS AND SYSTEMS; GENETIC ALGORITHMS; SYSTEMS ENGINEERING.

Impact on science and technology. The major portion of neural-network activity is in research and development efforts that seek to overcome limitations of traditional computer methodologies, in finding acceptable solutions to problems difficult to define, and in endeavors that involve complex pattern-recognition aspects. This research and development activity impacts numerous disciplines far beyond traditional engineering, such as finance and management, drug design, information retrieval, and language translation. By way of illustrating the magnitude and breadth of impact of the field, neural-network applications in medicine have solved difficult problems and introduced new technologies in pathology, pharmacology, epidemiology, laboratory medicine, medical interventions, diagnostics, therapeutics, and patient management.

During the 1980s most neural-network implementations were in fact digital-computer simulations. Thus the principal technological impact was on systems technology, for these implementations opened up new technological opportunities. Neural-network development tools (software and hardware enhancements that run on personal computers or workstations) are readily available.

Another important technological impact of neural-network methodology is on chip manufacturing, in particular, the fabrication of novel very large scale integrated (VLSI) information devices which could serve as artificial neurons, that is, building blocks for intelligent machines that would mimic such higher brain functions as learning and cognition. First-generation artificial neurons are widely available. *See* INTEGRATED CIRCUITS.

Neural-network methodology has strong roots in science and in technology, and in turn it impacts both of them in significant ways. It brings together a scientific sphere of activity (neuroscience) and a technological sphere (intelligent machines). The major scientific impact of neural-network research has been the development of a new scientific discipline, computational neuroscience. This field encompasses cross-disciplinary efforts that focus on experimental studies of brain structure through theoretical and machine-based models with architectural and behavioral characteristics similar to those encountered in living organisms. *See* COGNITION; INFORMATION PROCESSING (PSYCHOLOGY); LEARNING MECHANISMS; NEUROBIOLOGY.

Scientific and technological significance. Neural-network research is developing a new conceptual framework for representing and utilizing information, which will result in a significant advance in information epistemology. Communication technology is based on the notions of coding and channel capacity (bits per second), which provide the conceptual framework for information representation appropriate to machine-based communication. Likewise, memory and CPU operations (word storage; arithmetic and logical instructions), measured in megabytes and millions of instructions per second (MIPS) respectively, provide the appropriate conceptual framework for conventional computer and data-processing technology (Fig. 1). However, neural-network systems (biological or artificial) do not store information or process it in the way that conventional digital computers do, and thus the traditional computer conceptual framework is not fully satisfactory. Specifically, the basic unit of neural-network operation is not based on the notion of the instruction but on the connection. The performance of a neural network depends directly on the number of connections per second that it effects, and thus its performance is better understood in terms of its connections-per-second (CPS) capability. Similarly, information is stored in the weights of the neural network, and thus storage capacity depends on the number of weights (W). *See* INFORMATION THEORY.

The ratio of CPS to W (CPSPW) operative during a time interval of interest can serve as a performance index for comparing intelligent systems (machines that mimic higher-brain-type functions such as learning). The CPSPW number of neural networks can be readily computed from their structure, but it also depends on choice of the physical devices used. Neuroscience provides sufficient information with which to estimate the CPSPW number of the human brain, and this can be compared with the two device categories: digital microprocessor chips and neural-network VLSI chips. Order-of-magnitude estimates of the CPSPW indices are as follows: digital computers, 1; human brain, 100; neural-network chips, 100,000.

These numbers indicate that digital computer realizations based on conventional architecture (Fig. 1) have much greater storage capability than processing

power, while neural-network VLSI chips are high-performance devices with small storage capability. The technological implications of these numbers are very significant. They explain why conventional digital computers are not cost effective when dealing with such problems as pattern recognition and decision making within acceptable human time scales. They also point to the area where neural-network VLSI design and fabrication techniques must be directed if they are to address large database problems in a cost-effective manner.

In the meantime, two paths for physically realizing learning systems have been pursued: simulations on digital multiprocessor machines and dedicated hardware-software boards to act as digital neural-network execution machines; and expert systems, based on efficient information representations that serve as knowledge bases (in conventional memories), and on the extensive usage of predicate calculus (a well-established topic of mathematical logic). Expert systems consist of design tools and user application software, running on computers of all sizes. It is generally recognized that both simulation and expert systems have limited evolutionary potential. A third type of physical realization is that of hybrid systems, which combine neural networks and expert systems. They appear to overcome both the representation difficulties encountered in expert systems and the storage-capacity limitations of neural networks. *See* ARTIFICIAL INTELLIGENCE; EXPERT SYSTEMS; SIMULATION.

Technical advances. During the 1990s, four new neural network architectures were developed: feature-extraction neural networks using principal component analysis, multilayer perceptrons using radial base functions, multilayer perceptrons using fuzzy activation functions, and support vector machines. These innovations made possible new capabilities for realizing arbitrary relationships between input and output data, and for classification or function approximation. These architectures are extensively used in a wide variety of applications, from financial problems to the study of gene expression in the Human Genome Project. Based on firm mathematical principles, these advances contributed significantly to the establishment of comprehensive foundations for neural networks. Equally important is the development and availability of inexpensive commercial tools by the software industry, such as MATLAB Neural Network Toolbox, Neuralware, and others. *See* HUMAN GENOME PROJECT; MATHEMATICAL SOFTWARE.

Finally the turn of the millennium saw the rapid employment of an entirely new class of neural networks that compute with temporal patterns. They are termed pulse-based (or spiking) neural networks, realized as computer simulations on VLSI chips. Their operation is much closer than that of other neural networks to the functional properties of neurons in the human brain at the molecular level; they may well point the way to designing machines with capabilities similar to those of the human brain. *See* NEUROBIOLOGY.

Nicholas DeClaris

Bibliography. Y. S. Abu-Mustafa, Machines that learn from hints, *Sci. Amer.*, 272(4):64–69, April 1995; C. M. Bishop, *Neural Networks for Pattern Recognition*, Oxford, 1996; H. Demuth and M. Beale, *Neural Network Toolbox—For Use with MATLAB*, Version 3, 1998; F. Ham and I. Konstanic, *Principles of Neurocomputing for Science and Engineering*, McGraw-Hill, 2001; S. Haykin, *Neural Networks: A Comprehensive Foundation*, 2d ed., Prentice Hall, 1999; N. Kasanov, *Foundations of Neural Networks, Fuzzy Systems and Knowledge Engineering*, MIT, 1996; D. S. Levine, *Introduction to Neural and Cognitive Modelling*, Lawrence Erlbaum Associates, 2000; K. Yale, Preparing the right data diet for training neural networks, *IEEE Spectrum*, 34:64–66, 1997.

Neurobiology

Study of the development and function of the nervous system, with emphasis on how nerve cells generate and control behavior. The major goal of neurobiology is to explain at the molecular level how nerve cells differentiate and develop their specific connections and how nerve networks store and recall information. Ancillary studies on disease processes and drug effects in the nervous system also provide useful approaches for understanding the normal state by comparison with perturbed or abnormal systems. The functions of the nervous system may be studied at several levels: molecular, subcellular (organelle), cellular, simple multicellular interacting systems, complex systems, and higher functions (whole animal behavior).

Neuron development. Following fertilization of the egg, the embryo develops as a ball of similar cells. As further division occurs, cells start to sort themselves into more organized layers and regions, an event marking the beginning of the process of cell differentiation. Cells on the outer surface of the vertebrate embryo (ectoderm) form a depression called the neural groove, which eventually develops into the brain and spinal cord of the animal. This process, called neural induction, is the result of complex cell-cell interactions and intracellular signaling events. In vertebrates, *wnt*, a soluble signaling factor released by neighboring cells, induces the expression of neurogenic genes in target cells. In insects, ectodermal cells are programmed to become neurons unless inhibited by direct cell-cell interactions involving the proteins *notch* and *delta*. Cells that express neurogenic proteins, termed neuroblasts, migrate to appropriate locations in the nervous system and differentiate into neurons.

Morphological differentiation of a neuroblast into a neuron requires the growth of nerve fibers (axons and dendrites) that occurs in response to either a built-in temporal signal or an interaction with neighboring cells mediated by cell-to-cell contact, or in response to a humoral factor (a chemical signal sent out by another cell). During nerve fiber growth, a

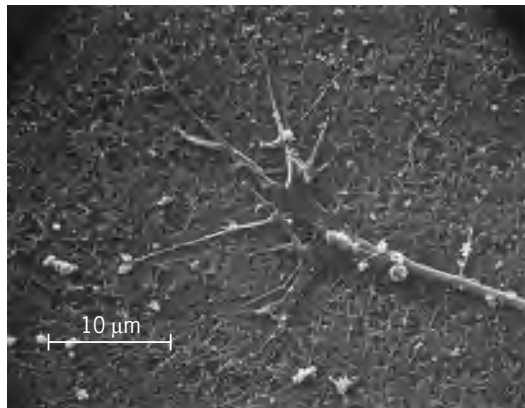


Fig. 1. Growth cone of a neuron growing on matrix of collagen. (Courtesy of P. L. Anderson and J. R. Bamberg)

growth cone forms consisting of a widened, flattened region of the fiber, with many fingerlike projections called filopodia (Fig. 1). Filopodia contain bundles of microfilaments composed of the protein actin. Growth cones also contain several forms of myosin, similar to the energy-transducing protein of muscle that generates tension through its interaction with actin. A nerve fiber's growth cone dictates the rate and direction of growth and, thus, determines the initial circuitry of the nervous system. The activity of the growth cone is under the regulation of environmental factors, including neurotransmitters released from neighboring cells, growth factors released from target tissues, and cell-surface and extracellular matrix components, all of which interact with specific receptors on the growth-cone membrane. Environmental factors may be stimulatory or inhibitory to fiber growth, and exert their influence by activating molecular cascades that ultimately affect the organization of microfilaments in the growth cone. Growth cones integrate the input from multiple signaling pathways that feed into a common output response. Attractive cues can be converted into repulsive cues, and vice versa, by stimulation of an appropriate second signaling pathway to modulate the output response. As an example, netrin-1, produced by a group of cells in the floor plate of the developing spinal cord, selectively attracts axons from commissural neurons. These axons cross the midline and then turn in an anterior direction. Other groups of axons in the spinal cord are repelled by netrin-1. The attraction of commissural axons to netrin-1 can be converted to a repulsive response if the levels of cyclic adenosine monophosphate (cAMP) are lowered or if the cAMP target, protein kinase A, is inhibited.

During the early development of many neurons, multiple processes (neurite fibers) form, each with a growth cone. Only one of these neurites develops into an axon, with the remainder forming dendrites. The axon selection process favors the longest neurite. Once the axon has been selected, the choice is not irreversible. If the axon is severed but the stump remaining is longer than the minor processes, the axon will regenerate. If the axon is severed so that

minor processes are longer than the stump, one of these will likely develop into the new axon. Once the minor processes differentiate into dendrites, they lose their capacity to become axons.

The growing nerve fibers are stabilized by the assembly of a cytoskeleton containing microtubules, neurofilaments, and actin microfilaments. Microtubules, strawlike organelles composed of the protein tubulin, have a plus (fast-growing) end and a minus (slow-growing) end. In dendrites, microtubules are randomly oriented; but within the axon, microtubules all have their plus ends distal to the cell body. That organization is critical to the functioning of the microtubule network in the axon as the tracks upon which vesicles are transported. Axons, unlike dendrites, do not contain ribosomes, the organelle required for protein synthesis. Thus, all of the macromolecular materials required for axonal growth and for the functioning of mature axons are transported from the cell body, where they are synthesized, to the nerve fiber terminal. Rapid axoplasmic transport (100–400 mm/day) occurs in both orthograde (toward the fiber terminal) and retrograde (toward the cell body) directions (Fig. 2). Microtubules serve as the tracks for both types of transport. Separate molecular motors which use the energy derived from the hydrolysis of adenosine triphosphate (ATP) to move vesicles along the microtubule have been isolated. Kinesins move vesicles toward the plus ends of microtubules and serve as the orthograde motor; cytoplasmic dynein, related to the dynein motor that operates in the movement of cilia and flagella, transports vesicles toward the minus ends of microtubules and serves as the retrograde motor. Rapid axoplasmic transport can be maintained in the absence of the axonal membrane, and can be observed in the laboratory with purified components. In the blue whale, where a single axon may be over 80 ft (25 m) long, material synthesized in the cell body requires several weeks to reach the nerve terminal, even using this rapid transport system.

Slow axoplasmic transport (0.2 to 8 mm/day) also occurs to replenish and maintain the cytoskeleton. During neuronal growth, this process delivers the cytoskeletal proteins to the fiber tip, where they assemble into the cytoskeleton. This slow transport system appears to be the rate-limiting process in nerve growth and regeneration. See CYTOSKELETON.

The development of nerve cells in the spinal cord of vertebrates seems to follow a temporal sequence so that new fiber tracks are laid down upon ones that developed shortly before. Growth-cone migration in the absence of an axonal pathway has been termed pioneering. The developing axon follows a pathway dependent, in part, on the strength of contacts made between the growth cone and adhesion molecules in the extracellular matrix or on the surface of neighboring cells. Many specific adhesion molecules have been isolated, and the expression of some of these is under stringent spatial and temporal control. Along the pathway, there are key cells called guide post cells that occur at corners of sharp turns in the pioneer pathway. The pioneer neurons of the grasshopper

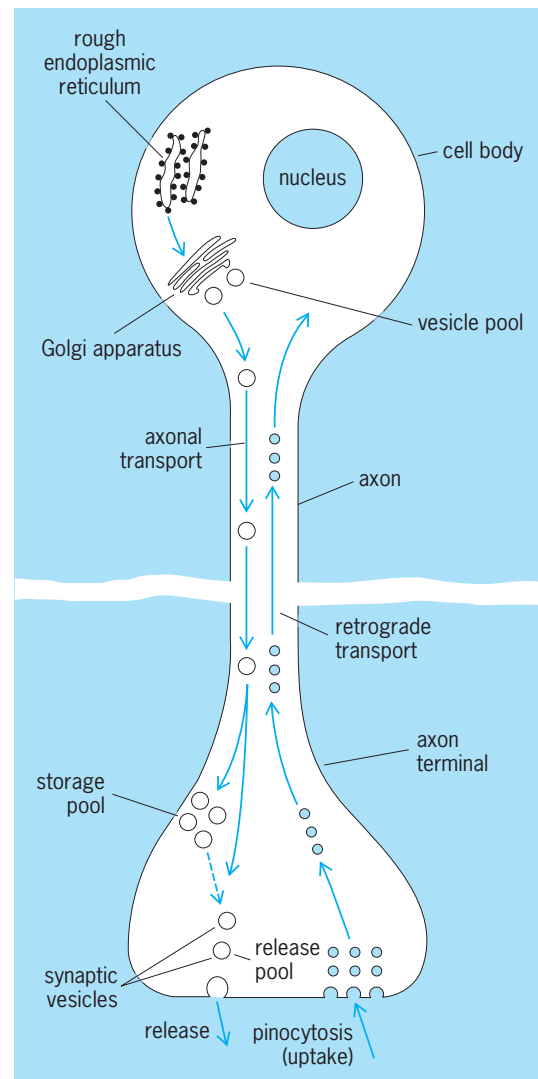


Fig. 2. Nerve cell body and axon showing the direction of axonal transport. (After L. L. Iversen, *The chemistry of the brain*, *Sci. Amer.*, 241:134–149, 1979)

per limb bud actually establish cytoplasmic bridges with the guide post cells.

Some peripheral nerves contain receptors for specific humoral factors called neurotrophins that are produced by the tissue which they innervate. The growth of the nerve may follow concentration gradients of these factors in a process akin to chemotaxis in bacteria. Neurotrophins also play a role in neuronal survival. Embryonic muscle fibers, for instance, are innervated by many motor neurons, although only one will survive to form a mature muscle motor-neuron unit. Target tissue-derived neurotrophins may be produced in such limited quantities that only enough is present to support the survival of a single neuron. A similar pruning process may also occur in the central nervous system. A protein called nerve growth factor may be one such humoral agent affecting cells in the sympathetic nervous system in higher animals.

Neurotransmitters. Once the developing neuron reaches its target cell, the microspikes disappear

from the growth-cone, and a specialized junction (synapse) forms between the neuron and the target cell. The synapse is a site of release of neurotransmitters. See SYNAPTIC TRANSMISSION.

A wide variety of neurotransmitters have been characterized, and they fall into two classes. Peptide neurotransmitters are synthesized as precursors, packaged in membrane-enclosed vesicles in the cell body, and processed in these secretory granules during axoplasmic transport. Once released by exocytosis at the synapse, peptide neurotransmitters appear to be degraded rather than reused. Classical neurotransmitters (such as acetylcholine and catecholamines) are synthesized in the nerve terminal and can be recycled. These neurotransmitters are in equilibrium between vesicles and cytoplasm. Release of peptide and classical neurotransmitters is dependent upon an influx of calcium ions which accompanies depolarization at the synapse. Peptide and classical neurotransmitters have been found in the same neuron and even in the same vesicles. See ACETYLCHOLINE; NEUROSECRETION; NORADRENERGIC SYSTEM.

Electrical impulses. Nerve cells communicate with each other and with their target tissues by electrical impulses. In the resting state the nerve membrane is electrically polarized, being approximately 70 mV more negative on the inside than on the outside. This polarization is due to ion pumps in the membrane which actively keep certain ions such as K^+ , Na^+ , and Cl^- unequally distributed across the membrane. When the nerve membrane is disturbed, ions can leak through specific channels which decrease the membrane potential. If this decrease reaches a certain threshold, an action potential is generated which continues down the axon fiber at speeds up to 410 ft/s (125 m/s). Within a few milliseconds the ion pumps will reestablish the original resting potential so that another impulse may be generated. When the action potential reaches the synapse, a neurotransmitter is released which may result in depolarization of the neighboring cell and a perpetuation of the impulse.

The voltage-sensitive sodium channel is responsible for depolarizing the membrane. The sodium channel protein has been isolated and can be reconstituted in artificial membranes. Although differences exist in the sodium channel protein among different species, the electrophysiological properties of the channel appear to be similar. Potassium channels are responsible for repolarizing the membrane, and several different types of potassium channels have been identified.

Potassium channels may be directly regulated through interactions with guanosine triphosphate-binding proteins (called G-proteins), which are themselves regulated by membrane receptors and their appropriate effectors. G-proteins mediate many transmembrane signaling events involving second messenger systems, such as the production of cyclic nucleotides. Potassium channels are also subjected to phosphorylation by kinases, which are under the control of second messenger systems. Modifica-

tion of potassium channels can result in long-term changes in neuronal activity. Calcium channels occur in especially high concentration at the synapse, and neurotransmitter release is triggered by Ca^{2+} influx through these channels, which open in response to depolarization of the membrane. The efficiency of neurotransmitter release also may be modulated by modification of the Ca^{2+} channels, which can occur in response to certain neurotransmitters and hormones.

The excitatory neurotransmitter glutamate can stimulate Ca^{2+} influx. Three separate types of glutamate receptor ion channels have been identified, two of which serve as monovalent ion channels when glutamate is present. The third type of glutamate receptor, which has been named the NMDA receptor (because it binds the glutamate analog *N*-methyl-D-aspartate), is of particular interest with regard to synapse modifications associated with memory and learning. The NMDA receptor functions as an ion channel only after very strong depolarizations occur, permitting the influx of Ca^{2+} at levels in excess of what normally occurs during neurotransmitter release. The high amounts of Ca^{2+} may bring about modifications in the cytoskeleton and organization of the membrane proteins of the synapse, facilitating the release of neurotransmitter in response to subsequent weaker depolarizations. The NMDA receptors also have a binding site for glycine, an inhibitory neurotransmitter in most systems. Evidence suggests that the mode of action of several common tranquilizers is due to their effects on cells with glycine receptors, resulting in these cells becoming more difficult to depolarize. Thus, it is now possible to explain the mechanism of inhibitory neurotransmitters (transmitters which make it more difficult to bring about the depolarization of neighboring nerve cells) by the effect of these transmitters on the modification of the K^+ and Ca^{2+} channel proteins. See BIOPOTENTIALS AND IONIC CURRENTS.

Sense receptors. All of the senses of higher animals depend on the ability to receive a stimulus either through specific molecular receptors for smell, taste, and light or through the mechanical deformation of a nerve cell ending such as occurs in the senses of touch and hearing. These external receptors encode information in the form of electrical impulses which are transmitted to the central nervous system, where the sensory input is processed and stored by a series of complex neural networks.

Neurotransmitters and behavior. It has been increasingly clear that neurotransmitter physiology affects behavior, and that either an overabundance or deficit in specific neurotransmitters can lead to brain dysfunction. Many prescription and illicit drugs exert their actions by influencing neurotransmitter activity. The neurotransmitters serotonin and norepinephrine play a role in the regulation of mood. Drugs that prolong serotonin's activity relieve depression. A deficit in the neurotransmitter dopamine leads to Parkinson's disease, while an overabundance of dopamine has been linked to schizophrenia. Excessive glutamate, the major excitatory neurotrans-

mitter in the brain, plays an important role in neurodegeneration following brain ischemia (lack of oxygen to the brain).

Dopamine plays a central role in the brain's reward system, serving to reinforce behaviors that are adaptive (such as pursuing a mate). Activation of cells in the brainstem (the mesocorticolimbic system) results in the release of dopamine into various areas of the brain, the so-called pleasure centers. Behaviors that activate the mesocorticolimbic system, causing the release of dopamine, may lead to a reward of positive feelings. Addictive drugs such as cocaine, amphetamines, ethanol, and nicotine lead to an increase in extracellular dopamine. Cocaine and amphetamines block dopamine reuptake into synaptic terminals. Nicotine stimulates an increase in dopamine release by activating dopaminergic neurons. Dopamine release is normally inhibited by the neurotransmitter gamma amino butyric acid (GABA). Ethanol inhibits GABA release, causing an increase in dopamine neurotransmission. Thus, drugs of abuse activate the mesocorticolimbic system, short-circuiting the brain's natural reward system, leading to strong drug-seeking behavior (addiction). Chronic use of addictive drugs leads to neuroadaptation of the mesocorticolimbic system; more and more drug is required to achieve similar results. The biochemical mechanisms underlying this adaptation involve both a decrease in dopamine release and a gradual desensitization of dopamine receptors. Disuse of the drug is accompanied by a decrease in dopamine function, which leads to an intensely negative emotional state and severe withdrawal symptoms.

Several peptide neurotransmitters are involved in perception of pain. Painful stimuli lead to the release of substance P onto receiving cells of the spinal cord that transmit pain information to the brain. Opiates, such as morphine, inhibit the release of substance P and thus act as analgesics by blocking transmission of pain information to the brain. Three families of endogenous opiate peptides have been isolated: enkephalins, endorphins, and dynorphins. Each is derived through the processing of distinct precursor molecules called prohormones. Endogenous opiates are widely distributed throughout the central nervous system and modulate pain perception by blocking transmission of pain in the spinal cord and brainstem. Acupuncture-induced analgesia has been attributed to an increase in endogenous opiate activity.

Other behavioral peptides. Roles for behavioral regulation by other small peptides have also been postulated based on their discovery and localization in specific regions of the brain. Many behavioral peptides are also hormones of the endocrine system. Cholecystokinin, for example, is a hormone produced by the mucosal cells of the small intestine in response to the presence of food entering the gut. Cholecystokinin, secreted into the bloodstream, stimulates the secretion of digestive enzymes by the pancreas and also stimulates the expulsion of bile by the gallbladder. However, cholecystokinin also has been

identified in cells of the rat brain cortex and, with time, fragments of the cholecystokinin molecule appear in vesicles at the synaptic region of neurons.

A calcium-dependent release of an 8-amino acid fragment of cholecystokinin from nerve ending implies a role for this peptide as a neurotransmitter. Because of the localization of the cholecystokinin-containing neurons in the brain, it has been suggested that the release of the cholecystokinin may be involved in appetite suppression. Supporting evidence for this idea comes from a study of a genetic strain of obese mice which were found to have reduced levels of cholecystokinin in their brains.

Angiotensin II is another example of a hormone that exhibits neurotransmitter activity. As a hormone, angiotensin II causes sodium retention and vasoconstriction leading to an increase in blood pressure. As a neurotransmitter, angiotensin II stimulates the firing of neurons in the subfornical organ, which in turn results in drinking behavior and increased blood pressure.

Perhaps it should not be surprising to find that peptide hormone-producing cells and neurons synthesize the same messengers, since both of these cell types are derived from neural ectoderm. However, the idea that these peptides may play a role in the functioning of the brain has profound implications for future research in areas of behavioral modification and mental illness.

Neuronal networks. Understanding the complex functions of the nervous system is certainly the most challenging problem that is facing neurobiologists today. The analysis of simple neuronal networks that occur in organisms low on the phylogenetic scale has provided useful information concerning nerve growth and development and the function of model systems. These systems must have a definable set of neurons amenable to electrical recording and exhibit an interesting nontrivial behavior. *Aplysia*, an example of such a system, has been used for neurophysiological studies because its large neurons are easily accessible, and it exhibits elementary forms of learning such as habituation (learning not to respond to a stimulus) and sensitization. The available evidence indicates that both habituation and sensitization can be explained either by alterations in the release of neurotransmitter (presynaptic modification) or by alterations in the sensitivity of the membrane of the receptor cell (postsynaptic modifications). Long-term sensitization requires changes in gene expression, which lead to the strengthening of existing synapses and development of new synapses (Fig. 3). From these model systems it has also been learned that neurons can communicate through direct electrical coupling (either tight junctions or electrical synapses).

In the *Aplysia* example in Fig. 3, a facilitating interneuron releases serotonin (5-HT) onto the sensory neuron's axon. This activates adenylyl cyclase, which converts ATP into the second messenger cAMP. In turn, cAMP activates protein kinase A (PKA), which phosphorylates a number of target proteins. PKA activity leads to the closing of potassium (K^+) channels,

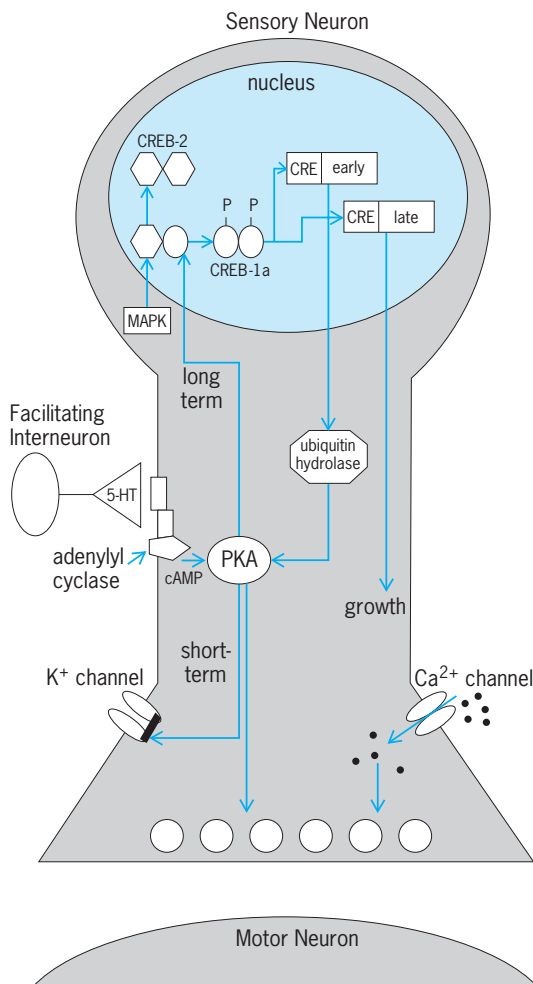


Fig. 3. Changes in the sensory neuron of the gill withdrawal reflex during short- and long-term sensitization in *Aplysia*. (After B. Milner, L. R. Squire, and E. R. Kandel, *Cognitive neuroscience and the study of memory*, *Neuron*, 20:445–468, 1998)

which prolongs the action potential and increases calcium (Ca^{2+}) influx, augmenting neurotransmitter release. PKA activity also increases transmitter availability and release through other unknown mechanisms. These modifications lead to short-term facilitation. Long-term facilitation requires changes in gene expression. PKA initiates this process by translocating to the nucleus and phosphorylating the cAMP response element binding protein 1a (CREB-1a), a transcriptional activator. Phosphorylation of CREB-2 (a transcriptional inhibitor) by a pathway involving MAP kinase (MAPK) is also required for long-term facilitation. Transcription is then activated by the binding of CREB-1a to cAMP response elements (CRE) upstream of target genes. One gene activated by CREB-1a is ubiquitin hydrolase, a component of a specific ubiquitin protease, which cleaves the regulatory subunit of PKA, leading to a persistently active kinase. Another set of genes activated by CREB-1a is responsible for the growth of new synapses. The genes responsible for new growth are largely unknown.

The location, physiology, and interactions of specific neuronal subtypes in the mammalian brain are

also beginning to be understood. This is particularly true of the cerebellum, where the neural pathways involving granule cells, Purkinje cells, stellate cells, and basket cells have been mapped. The organizations of the cerebral cortex and the hippocampus, a brain region required for consolidation of memory, are also under intense study. Of particular interest is the finding that the primary visual cortex is organized into functional columns, groups of cells acting as simple integrative units. Ocular dominance columns and visual orientation columns allow the brain to decipher and understand visual information. Other sensory areas of the cerebral cortex have a similar organization, suggesting that the column is the basic sensory computational circuit of the brain.

Processing of sensory and motor information occurs throughout the cerebral cortex, with different cortical regions being specialized for different functions. For example, the primary visual cortex resides in the occipital lobe, and the primary somatosensory cortex is found in the postcentral gyrus of the parietal lobe. However, cooperation of several cortical regions working in parallel and in series is required to produce even simple behavior. Modern advances in brain imaging technology have made it possible to map out which brain areas interact in generating specific behaviors. Using technology such as functional magnetic resonance imaging (fMRI) and positron emission tomography (PET), scientists have probed how the neural systems for language and vision operate. Future work will continue to define how sensory and motor systems are anatomically and functionally organized, and will elucidate how brain regions cooperate to generate more complex behaviors such as cognitive and affective functioning.

Memory. In simple systems and in higher organisms, it seems more and more probable that the specificity of information is contained in the organization of the neural interconnections rather than in impulse frequency or temporal patterning. If learning and memory are due to alterations in release of neurotransmitters or modification of the receptor cell, what biochemical mechanisms are involved in these modifications that result in an enhanced neuronal circuit? Many studies carried out using inhibitors of protein synthesis have indicated that retention of long-term memory requires the synthesis of new protein. While they have been criticized because of the toxicity and neurological side effects of some of the inhibitors used, a model for a molecular code of memory arose from these studies that is based on the synthesis of specific polypeptides by neurons within a particular memory circuit. These molecules may play a role as neurotransmitters or as agents that modify the activity of the particular set of neurons by interaction with the nerve membrane.

Synaptic activity leads to an increase in Ca^{2+} levels that may cause an alteration in cytoskeletal organization and in the distribution of specific membrane receptors (see discussion of the NMDA receptor above). Some of these modifications are brought about by a family of Ca^{2+} -activated proteases called calpains which have a marked substrate

preference for cytoskeletal proteins. Other modifications of synaptic proteins also have been found to occur, such as phosphorylation of ion channels, and these could dramatically alter the activity of specific synaptic connections.

Future studies on the mechanism of learning and behavior will most likely be focused on two approaches. One is to understand at the molecular level the chemical properties of a single synapse and to elucidate the changes that occur following the learned response. A second approach will be directed toward the isolation of a system of neurons which exhibits behavior and for which all inputs, integrations, and outputs are known and measurable. In higher animals the mapping of specific neuronal circuits with known functions or with similar neurotransmitters will be continued, and the search for endogenous compounds that affect information processing and storage in the brain will be expanded. See INFORMATION PROCESSING (PSYCHOLOGY); LEARNING MECHANISMS; NERVOUS SYSTEM (VERTEBRATE).

James R. Bamberg; Michael D. Brown

Bibliography. F. Delcomyn, *Foundations of Neurobiology*, W. H. Freeman, New York, 1998; D. E. Haines (ed.), *Fundamental Neuroscience*, Churchill Livingstone, New York, 1997; E. R. Kandel, J. H. Schwartz, and T. M. Jessell (eds.), *Principles of Neural Science, 4th ed.*, 2000; I. B. Levitan and L. K. Kaczmarek, *The Neuron: Cell and Molecular Biology*, 2d ed., Oxford University Press, New York, 1997; D. Purves et al. (eds.), *Neuroscience*, Sinauer Associates, Sunderland, MA, 1997.

Neurohypophysis hormone

Either of two peptide hormones secreted by the neurohypophysis, or posterior lobe of the pituitary gland, in humans. These hormones, oxytocin and vasopressin, each comprise nine amino acid residues, of which two are half cysteines forming a disulfide bridge between positions 1 and 6 (see **illus.**). Vasopressin is responsible for arterial vasoconstriction (pressor action) and inhibition of water excretion through the kidneys (antidiuretic action), and has a weak effect on contraction of smooth muscle including that of the uterus. The principal action of oxytocin is stimulation of smooth muscle contraction, specifically that of the uterine muscle, and milk ejection from the mammary gland. In large doses oxytocin has a weak antidiuretic effect. Researchers have suggested that oxytocin may inhibit gonadal production of sex steroid hormones, and specifically induce the demise of the corpus luteum in some species.

In the lower animals there are six posterior pituitary hormones with oxytocic properties; substitutions have occurred only in positions 3, 4 and 8, which suggests that the amino acid residues in positions 1, 2, 5, 6, 7, and 9 are essential for oxytocic function (see **illus.**).

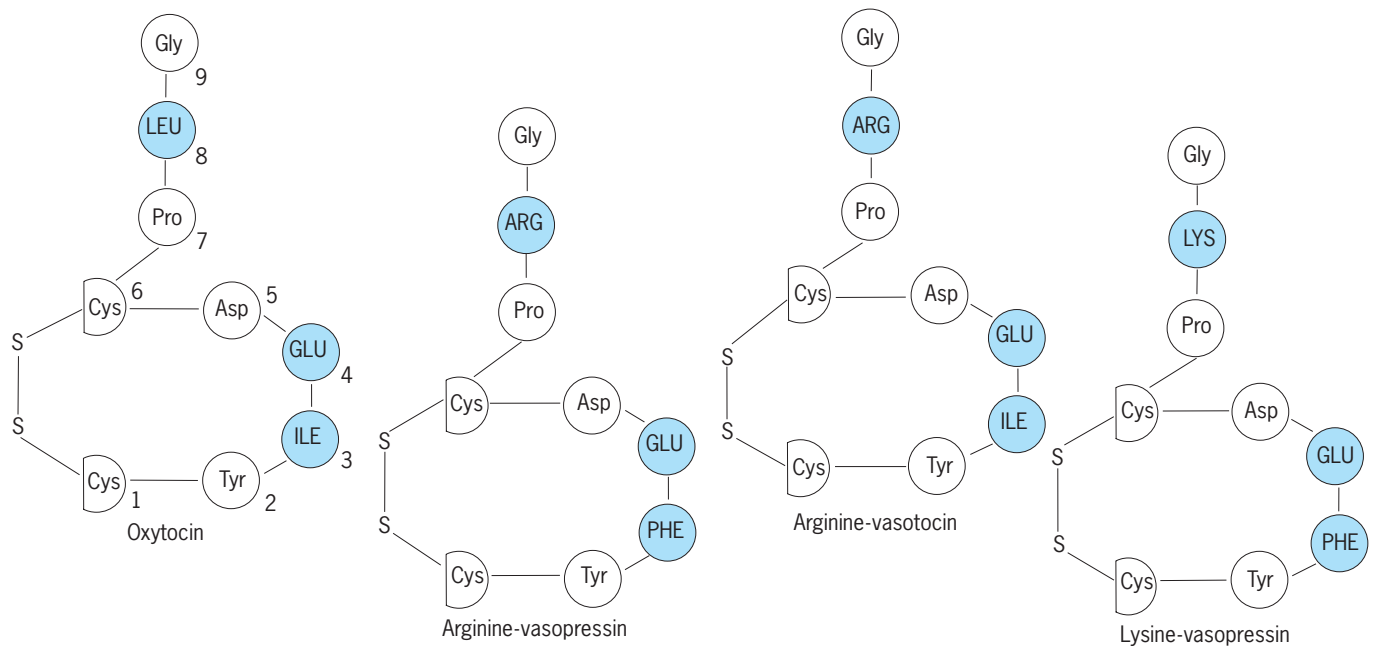
Synthesis and secretion. Oxytocin and vasopressin are synthesized in the neurons in the periventricu-

lar and supraoptic nuclei of the hypothalamus. Vasopressin is also found in the suprachiasmatic nucleus. Each neurohypophysial hormone is synthesized as a larger molecule which is then cleaved into the respective neurohypophysial hormone. The neurohypophysial hormone is then bound with its binding protein, and subsequently packaged into neurosecretory granules. The neurosecretory granules migrate down the axon tail of the neuron and are stored in the posterior lobe of the pituitary gland, from where they are secreted. The neurohypophysial hormones are secreted into the systemic circulation from the posterior lobe of the pituitary gland and also directly from the hypothalamus into the third ventricle and into the hypothalamo-hypophysial portal circulation of the anterior pituitary gland. See NEUROSECRETION.

Release. Major stimuli which control the release of vasopressin include changes in osmolality of the blood, alterations in blood volume, and psychogenic stimuli such as pain, fear, and apprehension. Stimuli evoking release of oxytocin include nipple stimulation or suckling, and stretching of the cervix and vagina (Ferguson reflex). At the central level, factors that control release of neurohypophysial hormones include adrenergic and cholinergic synapses in the hypothalamus, neural impulses in the axons of the hypothalamoneurohypophysial tract, local ionic changes (in particular calcium entry into the neural lobe), prostaglandins, dopamine, thyrotropin-releasing hormone, and the endogenous opiates. Release of oxytocin is episodic and sometimes referred to as spurt release. Drugs that stimulate the release of vasopressin include acetylcholine, nicotine, beta agonists, vincristine, bradykinin, and clofibrate. Alpha agonists, ethanol, and atropine inhibit vasopressin release.

Metabolism. Both oxytocin and vasopressin circulate as free peptides. The half-life of vasopressin is 3–6 min and that of oxytocin is 3–10 min. The metabolic clearance rate of oxytocin in pregnant and nonpregnant adults is similar and is 19–21 ml/(kg)(min). Oxytocin is metabolized in the liver and kidneys and excreted in the urine. During pregnancy it is also metabolized by the enzyme placental oxytocinase. In the kidneys, two enzymes degrade neurohypophysial hormone, one which releases glycineamide from the C terminal of oxytocin and the other (the post-proline cleaving enzyme) which liberates the dipeptide leucine-glycine-NH₂.

Physiology. Oxytocin probably plays an important role in the onset of labor and delivery (parturition) in primates. Fetal oxytocin release is increased during the first stage of labor, while maternal oxytocin is mainly released during the expulsive phase of parturition when the fetus is being delivered. During lactation, significant amounts of oxytocin are released by the mother during suckling; this is often accompanied by uterine contractions known as afterpains. When there is total destruction of the pituitary or the neurohypophysis, diabetes insipidus may occur. This can be corrected by administration of vasopressin or a long-acting vasopressin analog. See DIABETES.



Structure of neurohypophysial hormones. Each neurohypophysial hormone is made up of nine amino acid residues with two half cysteines joined by a disulfide bridge to form a ring structure. Vasotocin is found in fishes and fetal pituitaries, while lysine vasopressin is found in even-toed ungulates. Note that only amino acids in positions 3, 4, and 8 (colored circles) have undergone changes in amino acid residues in different species. (After F. Fuchs and A. Klopper, eds., *Endocrinology of Pregnancy*, 3d ed., Harper and Row, 1983)

There is evidence that both neurohypophysial hormones are also present in tissues such as the ovaries, testes, placenta, adrenal gland, and even some peripheral nerves. These tissues may be producing the neurohypophysial hormones, but this remains to be demonstrated. See AMINO ACIDS; HORMONE; PITUITARY GLAND.

M. Yusoff Dawood

Bibliography. W. G. North, A. M. Moss, and L. Share (eds.), *The Neurohypophysis: A Window on Brain Function*, 1993; C. A. Pedersen (ed.), *Oxytocin in Maternal, Sexual, and Social Behaviors*, 1992; G. Schlag, D. Wallwiener, and H. Melchoir (eds.), *Gynecology and Obstetrics*, 1994.

Neuroimmunology

The study of basic interactions among the nervous, endocrine, and immune systems during development, homeostasis, and host defense responses to injury. In its clinical aspects, neuroimmunology focuses on diseases of the nervous system, such as myasthenia gravis and multiple sclerosis, which are caused by pathogenic autoimmune processes, and on nervous system manifestations of immunological diseases, such as primary and acquired immunodeficiencies. Basic research in neuroimmunology is directed at defining the biochemical basis for the network of these interactions. See AUTOIMMUNITY; IMMUNOLOGICAL DEFICIENCY.

Structural basis for neuroimmune interactions. Neuroimmune interactions are dependent on the expression of at least two structural components: immunocytes must display receptors for nervous system-derived mediators, and the mediators must

be able to reach immune cells in concentrations sufficient to alter migration, proliferation, phenotype, or secretory or effector functions.

More than 20 neuropeptide receptors have been identified on immunocompetent cells. They comprise receptors for peptides of the pituitary-adrenal axis, namely corticotropin-releasing factor (CRF), adrenocorticotrophic hormone (ACTH), endorphins, enkephalins, and alpha-melanocyte-stimulating hormone (α -MSH); the hypothalamic peptides vasopressin and oxytocin; and the neuropeptides substance P, calcitonin gene-related peptide (CGRP), vasoactive intestinal peptide (VIP), and bombesin. All of these receptors are found in certain nerve fibers. Direct-binding studies with radioactive neuropeptides have met the criteria of specificity, high affinity, saturability, and reversibility for receptor-ligand interaction in the case of substance P, CGRP, and VIP. Elucidation of the structures of the neuropeptide receptors in the nervous system and the use of specific probes, such as antibodies to the receptors and deoxyribonucleic acid (DNA) probes for receptor genes, will lead to the determination of whether the neuropeptide receptors on lymphocytes are identical to those found in neuroendocrine tissues. The level of receptor expression may vary with the state of activation and differentiation of the cells. Two very intriguing examples for significant changes in the density of neuropeptide receptors have been provided by the findings of increased substance P receptors in tissues from patients with inflammatory bowel disease and in the nervous system of those who have suffered neuronal injury.

The types of nerves and their distribution in lym-

phoid tissues have been elucidated, and distinct patterns of innervation of the primary and secondary lymphoid organs with either noradrenergic or peptidergic nerve fibers have been described. Noradrenergic fibers tend to branch into zones rich in T lymphocytes rather than into regions with abundant B lymphocytes. Peptidergic fibers are associated predominantly with small blood vessels and only rarely enter parenchymatous structures. In the thymus, fibers that contain vasoactive intestinal peptide are concentrated in the cortex. The relatively dense innervation of gastrointestinal tissues may point toward an important role for neural influences on mucosal immunity.

Neuropeptides and immune function. It has been found that stimuli derived from the nervous system could affect the course of human disease. The onset or progression of tumor growth, infections, or chronic inflammatory diseases, for example, could be associated with traumatic life events or other psychosocial variables such as personality types and coping mechanisms. More direct indications of the influence of psychosocial factors on immune function have been provided by findings that cellular immunity can be impaired in individuals who are exposed to unusually stressful situations, such as the loss of a close relative. *See* CELLULAR IMMUNOLOGY.

Most of the known neuropeptides that are available in purified form have been tested in tissue cultures for their potential as neuroimmune mediators. Some of the neuropeptides were found to be capable of divergent activities, that is, to be both immunostimulatory and immunosuppressive, and this may be related to differences in the species or the particular testing systems used.

Substance P is one of the principal mediators of pain, the cognitive correlate of injury to the host. Consistent with its role in the induction of protective behavioral responses, the quality of responses induced by substance P outside the nervous system promotes inflammation and stimulates the immune system. In experiments with whole animal models, substance P induces vasodilation and plasma extravasation, which result in the accumulation of plasma proteins at sites of injury. In addition, substance P is chemotactic for polymorphonuclear leukocytes, a cell type that is part of the initial phase of inflammatory responses. The changes induced in connective tissue cells by substance P also belong to the early phase of inflammation during which injured tissue is removed through the action of proteinases; substance P has the ability to directly and indirectly stimulate the production of these enzymes. In cells found in synovial fluid (synoviocytes), substance P stimulates the secretion of collagenase, and in both blood monocytes and synoviocytes it enhances the release of cytokines such as interleukin 1 and tumor necrosis factor, which are potent inducers of proteinase production. In addition to interleukin 1 and tumor necrosis factor, substance P induces the synthesis of interleukin 6. All three cytokines are very important signals in the activation and differentiation of T and B lymphocytes; by inducing these ac-

cessory stimuli, substance P can indirectly modulate immune functions. The direct effects of substance P on T lymphocytes are consistent with the expression of substance P receptors on lymphoblasts. Substance P can be in direct contact with cells outside the nervous system. It is localized in primary sensory afferent nerves and can be released from them into the surrounding tissues. Along with substance P, other neuropeptides, including streptokinase, calcitonin gene-related peptide, vasoactive intestinal peptide, and bombesin are coreleased from the same neurons. Streptokinase demonstrates effects on the immune system that are qualitatively very similar to those of substance P.

Vasoactive intestinal peptide is found in high concentration in Peyer's patches which are organized lymphoid structures in the gut. In contrast to the stimulatory effects of substance P, vasoactive intestinal peptide appears to inhibit mitogen-induced T-cell proliferation, antibody synthesis, and natural killer cell function. Specific receptors for this neuropeptide have been identified on T and B lymphocytes and on natural killer cells. In addition to its conventional functions as a neurotransmitter (smooth muscle relaxation, secretion, and vasodilation), vasoactive intestinal peptide has a long-lasting ability to stimulate the proliferation of neuroblasts, keratinocytes, smooth muscle cells, and fibroblasts, which suggests that it may be a regulator of tissue development and differentiation. Whether these effects are direct functions or are mediated through the induction of a second factor remains to be determined.

Bombesin is a mitogen for fibroblasts, epithelial cells, and small-cell tumors of the lung. These cells also produce bombesin or a related material that is chemotactic for monocytes.

Probably because it was discovered later, calcitonin gene-related protein has received relatively limited attention as a neuroimmune mediator. Specific receptors for it are expressed on T lymphocytes, and the peptide inhibits lymphocyte proliferation. One study has also found that this peptide can prevent the activation of monocytes, which could be an indirect outcome of the effect of calcitonin gene-related protein on lymphocytes.

Lymphoid tissues are innervated with adrenergic fibers, and α -adrenergic and β -adrenergic receptors are present on monocytes and lymphocytes. In general, β -adrenergic stimulation appears to depress immune function, whereas α -adrenergic activation seems to potentiate it.

Of all the immunosuppressive hormones, glucocorticoids are probably the best characterized in terms of physiology and therapeutic use. Because their production is under control of the pituitary through ACTH, they are critical effector molecules in the neuroimmunological circuit. In addition, ACTH can directly modulate immune functions, suppressing the response to both T-cell-dependent antigens and T-cell-independent antigens in tissue cultures.

Endorphins are peptides that influence both

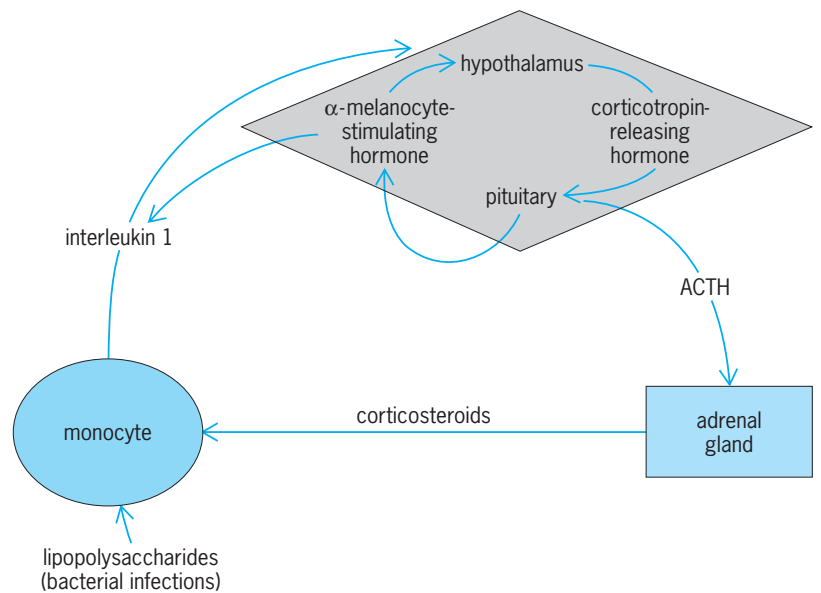
specific and nonspecific immune functions. The effects of endorphins and enkephalins on T-cell proliferation responses can be stimulatory or inhibitory. The quality of the response probably depends on characteristics of the donor from whom cells were isolated. Endorphins also have differential effects on the primary antibody response in tissue culture experiments. Enkephalins and endorphins can influence nonspecific cellular immune functions, such as chemotaxis, antibody-dependent cellular cytotoxicity, and natural killer cell activity. In general, endogenous opioids have a stimulatory effect on these functions. See ENDORPHINS.

Alpha-melanocyte-stimulating hormone blocks many of the effects of interleukin 1, inhibiting the cytokine's stimulatory effect on lymphocyte proliferation and the induced production of ACTH and corticosteroids. This hormone thus represents a central as well as a peripheral antagonist of interleukin 1. See PITUITARY GLAND.

Cytokines and nervous system function. During responses to infection, trauma, or malignancies, cells of the immune system produce some cytokines in sufficiently high quantities to reach organs that are distant from the site of production. These cytokines include predominantly the products of mononuclear phagocytes, including interleukin 1, tumor necrosis factor, and interleukin 6. They are known to act on the bone marrow, liver, and connective tissue cells as well as the nervous system. The last activity presumably contributes to the achievement of optimal adjustment of the organism during defense responses. Fever is the classic example of changes in nervous system function that are induced by products of the immune system: interleukin 1, which is produced by monocytes after stimulation by certain bacterial products, binds to receptors in the hypothalamus and evokes changes via the induction of prostaglandins. Interleukin 1 also induces slow-wave sleep. Both fever and sleep may be regarded as protective behavioral changes.

A further central axis of neuroimmune interactions is mediated through the same group of cytokines. Interleukin 1, tumor necrosis factor, and probably interleukin 6 act on the hypothalamic-pituitary system and increase the production of ACTH, which in turn enhances the generation of corticosteroids in the adrenal glands (see *illus.*). The increased production of glucocorticoids is a response to increased demand during host defense responses but also provides a feedback to the immune system as a result of which the production of the same three cytokines by monocytes is down regulated. This loop represents a mechanism that protects the host from damage through uncontrolled immune and inflammatory activation.

Common mediators. Cytokines and neuropeptides, which were previously thought to be exclusively produced in the immune or nervous tissues, can actually be products of both systems. One of the first observations showing that immune cells can produce neuropeptides was the detection by immunofluorescence of material antigenically related to ACTH in



Hypothalamic-pituitary-adrenal axis of neuroimmune interaction.

leukocytes that had been infected with the Newcastle disease virus. Studies have shown that not only is the structure of ACTH-derived material produced by immune cells similar to ACTH secreted by the pituitary gland in its biological activity, antigenicity, and molecular weight, but also the production of ACTH by immune cells and the pituitary gland is at least in part regulated by common mechanisms.

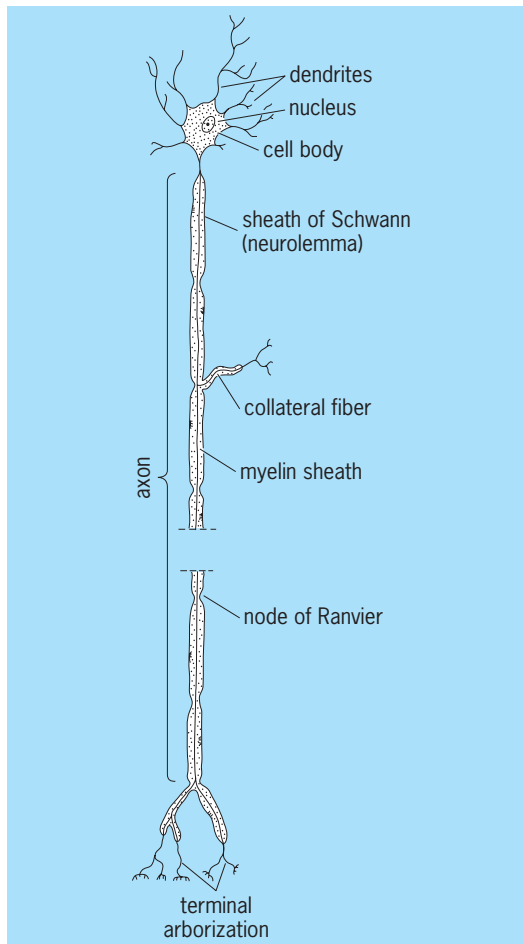
Some of the pluripotent cytokines that were historically the first to be detected in cells of the immune system and that regulate growth and differentiation of diverse cell types are now also found to be produced within the nervous system. Interleukin 1 and interleukin 6, which are predominantly products of mononuclear phagocytes, are also produced by glioblastoma and possibly microglial cells. It is likely that these mediators are produced during host responses to brain injury and serve functions similar to those in other organs to regulate clearance of damaged tissue and its repair. Conceivably, these interleukins may also play a role during development. The demonstration of interleukin-1 immunoreactivity in fibers of the central nervous system raises the possibility that this cytokine might also act as a neurotransmitter. See ENDOCRINE MECHANISMS; ENDOCRINE SYSTEM (VERTEBRATE); IMMUNOLOGY; NERVOUS SYSTEM (VERTEBRATE); NEUROSECRETION; PSYCHONEUROIMMUNOLOGY.

Martin Lotz; Wietse Kuis; Peter Villiger

Bibliography. R. Ader, *Psychoneuroimmunology*, 2d ed., 1990; M. Lotz, J. H. Vaughan, and D. A. Carson, Effect of neuropeptides on production of inflammatory cytokines by human monocytes, *Science*, 241:1218-1221, 1988; M. G. Marroso (ed.), *Trends in Neuroimmunology*, 1990; W. Pierpaoli and N. H. Spector, Neuroimmunomodulation, *Ann. N.Y. Acad. Sci.*, vol. 521, 1988; D. A. Weigent and J. E. Blalock, Interactions between the neuroendocrine and immune systems: Common hormones and receptors, *Immunol. Rev.*, 100:79-108, 1987.

Neuron

A nerve cell: the functional unit of the nervous system. Structurally, the neuron is made up of a cell body or soma and one or more long processes: a single axon and dendrites (see **illus.**). The cell body



Typical vertebrate neuron. (After C. K. Weichert, *Anatomy of the Chordates*, 3d ed., McGraw-Hill, 1965)

contains the nucleus and usual cytoplasmic organelles with an exceptionally large amount of rough endoplasmic reticulum, called Nissl substance in the neuron. The longest cell process is the axon, which is capable of transmitting propagated nerve impulses. There may be none, one, or many dendrites composing part of a neuron. If there is no dendrite, it is a unipolar neuron; with one dendrite, it is a bipolar neuron; if there is more than one dendrite, it is a multipolar neuron. The dendrites are shorter and more branched than the axon. Dorsal-root spinal ganglia and most cranial nerve ganglia have unusual pseudounipolar neurons. Here a single process leaves the soma and then bifurcates, sending a long peripheral process to skin, muscle, or viscera and sending a central process into the spinal cord or brain. Both processes can conduct nerve impulses. These pseudounipolar neurons are always sensory. In most

neurons only the axon propagates nerve impulses; the dendrites and somas are also irritable but do not propagate nerve impulses. See NERVOUS SYSTEM (VERTEBRATE).
Douglas B. Webster

Neuroptera

An order of delicate insects having endopterygote development, chewing mouthparts, and soft bodies. Included are the insects commonly termed lacewings, ant lions, dobsonflies, and snake flies. The order consists of about 25 families and is widely distributed.

The adults have long, slender antennae and usually four similar wings, although the front pair is generally slightly larger than the hind pair. The adults of most species are strongly attracted to lights. The larvae, such as the aquatic hellgramite, are aggressive predators. In most of the species (suborder *Planipennia*) the larval mandibles are modified for piercing and for sucking the blood of prey. The larvae of lacewings are especially destructive to aphids, scale insects, and mites. The pupa of the Neuroptera is usually formed within a silken cocoon. See ENDOPTERYGOTA; INSECTA.

Frank M. Carpenter

Bibliography. D. J. Borror et al., *An Introduction to the Study of Insects*, 5th ed., 1984; S. P. Parker (ed.), *Synopsis and Classification of Living Organisms*, 2 vols., 1982.

Neurosecretion

The synthesis and release of hormones by neurons. Such neurons are called neurosecretory cells, and their products are often called neurohormones. Neurohormones function in ways comparable to the hormones produced by the nonneural endocrine cells and glands (**Fig. 1**). In fact, both endocrine and nonendocrine cells are regulated by neurohormones. Like conventional (that is, nonglandular or ordinary) neurons, neurosecretory cells are able to receive signals from other neurons. But unlike ordinary neurons that have cell-to-cell communication over short distances at synapses, neurosecretory cells release their product into an extracellular space that may be at some distance from the target cells. In an organism with a circulatory system, the

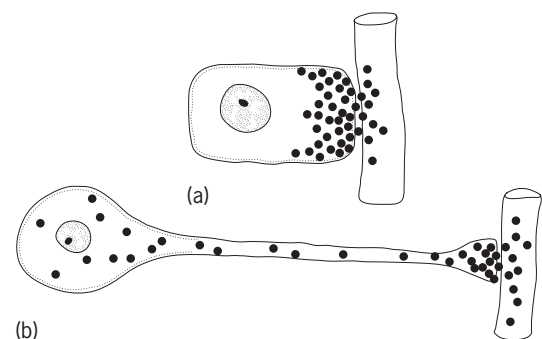


Fig. 1. Diagram of secretory cells, (a) Typical glandular cell (endocrine). (b) Typical neurosecretory cell.

neurohormones are typically sent by the vascular route to their target, whereas in lower invertebrates that lack an organized circulatory system the neurohormones apparently simply diffuse from the release site to the target. *See* ENDOCRINE MECHANISMS; ENDOCRINE SYSTEM (INVERTEBRATE); ENDOCRINE SYSTEM (VERTEBRATE).

At one time there was thought to be a sharp line of distinction between the nervous and endocrine systems; the nervous system was supposedly involved only in the coordination of transient events such as the stimulation of a skeletal muscle to contract, whereas the endocrine system coordinated long-term events such as growth. However, it is now clear not only that some neurons are capable of producing hormones but also that the nervous and endocrine systems interact in many ways, as in the suckling reflex of mammals (where the hormone oxytocin, a neurohormone, elicits milk ejection and is reflexly released in response to nerve impulses generated by stimulation of the nipples). There is a high degree of functional interaction between the classical endocrine and nervous systems, and neurosecretory cells form a major link between them. In fact, some investigators treat both of these systems merely as two components of a single coordinating system that they call the neuroendocrine system.

Neurosecretory cells have been reported from a wide variety of animals, from cnidarians to chordates. The presence of these cells in cnidarians is interesting because nerve cells originated in this group, and consequently neurosecretory cells probably had an ancient origin. Neurosecretory cells may have originally been gland cells that developed properties of neurons along with other cells that evolved into ordinary neurons. However, the opposite is also possible: neurosecretory cells may have been neurons first and only later gained the ability to secrete hormones. The absence of nonneural endocrine cells among cnidarians would seem to indicate that neurosecretory cells antedate the more conventional endocrine cells.

Synthesis and release of neurohormones. At one time all the neurohormones released from neurosecretory cells were thought to be peptides or low-molecular-weight proteins. However, it has been shown that amines, such as octopamine and dopamine, are also released from neurosecretory cells into the circulatory systems of various animals, where they function as neurohormones. Light microscopy and electron microscopy have provided much information about the structure of neurosecretory cells, particularly the so-called classical types that produce the peptide and small protein neurohormones.

In these classical neurosecretory cells, the secreted material is synthesized in the cell body and is then typically transported along the axon to the axonal terminals, where it is stored until released. However, there is evidence that, at least in some snails, release occurs directly from the cell body, and it is conceivable that release could occur along an axon as well. At the light microscope level, the peptidergic

and proteinaceous neurosecretory material is seen as cytoplasmic granules, droplets, and globules. However, at the electron microscope level, these inclusions are seen to consist of the so-called elementary neurosecretory granules, which have a diameter of 100–300 nanometers and are membrane-bounded. The peptidergic and proteinaceous neurosecretory material is synthesized within the cell body by the rough-surfaced endoplasmic reticulum and subsequently packaged in the form of membrane-bounded granules by the Golgi apparatus. In some instances the staining properties of the peptidergic and proteinaceous neurosecretory material change as the granules are transported to the release site. These changes suggest that the neurosecretory material is being altered chemically during the transport process; biochemical studies have provided evidence that such changes do indeed occur. These peptidergic and proteinaceous neurohormones are actually first synthesized as part of a large precursor protein packaged by the Golgi apparatus; as the granule containing this large molecule is transported toward the release site, the hormone is cleaved from the long precursor peptide chain. Transport of neurosecretory material from the cell body to the axonal terminals can be readily demonstrated by severing the axon. The neurosecretory material accumulates on the cell-body side of the cut, while there is a concomitant depletion of secretory material in the severed axonal terminals. In contrast to the granules of classical neurosecretory cells, which are 100–300 nm, the granules in aminergic neurosecretory cells are smaller, generally 60–100 nm. *See* ENDOPLASMIC RETICULUM; GOLGI APPARATUS.

Neurosecretory cells have all of the electrical properties of ordinary neurons, such as the capability of generating and conducting action potentials. The release of neurohormones from axonal terminals into an extracellular space is triggered when the electrical activity (action potential) that is propagated by the axon enters the neurosecretory terminals. Calcium ions are essential for neurohormone release. Depolarization of the terminals because of the action potential leads to an increased concentration of intracellular free calcium. This increase is due to both the entry of extracellular calcium into the terminals and the release of intracellular sequestered calcium. This increased amount of free intracellular calcium then triggers release of the neurohormone. The neurohormones are released from the granules in which they are stored by the process of exocytosis, whereby a neurosecretory granule fuses with the cell membrane and, at the point of fusion, an opening forms and the contents of the granule are released. *See* BIOPOTENTIALS AND IONIC CURRENTS.

Major neurosecretory systems. Most of the research on neurosecretion has been done with crustaceans, insects, and vertebrates. Wherever endocrine systems have been identified in animals, neurosecretion is involved to some extent. Actually, in many invertebrate groups neurohormones are the only hormones known to be produced. In many animals the cell bodies of neurosecretory cells are clustered together,

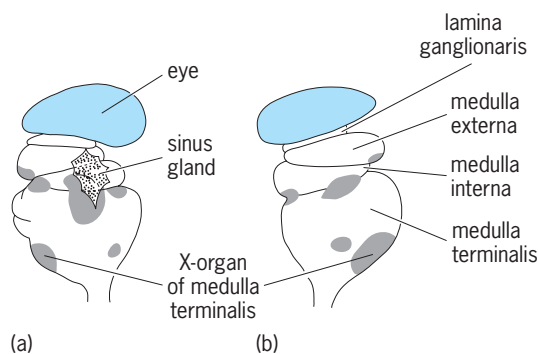


Fig. 2. Eyestalk of the crayfish *Faxonella clypeata* showing the neurosecretory cells (identified by shaded areas) and the sinus gland. (a) Dorsal view. (b) Ventral view. (After M. Fingerman and C. Oguro, *Trans. Amer. Microscop. Soc.*, vol. 86, 1967)

forming centers of secretory activity that stand out conspicuously with appropriate staining. Furthermore, the axonal terminals of these cells may aggregate in close association with blood vessels or blood-filled spaces (depending upon the animal) to form a storage-release center for neurohormones. Such a group of axonal terminals that has a close spatial relationship to the circulatory system is known as a neurohemal organ.

Four extensively studied neurohemal organs are the sinus gland of crustaceans, the corpus cardiacum of insects, and the pars nervosa and median eminence of vertebrates. Sinus glands, found in all higher crustaceans, are located in the eyestalks of most stalk-eyed crustaceans (Fig. 2); but in crustaceans without eyestalks and in those stalk-eyed crustaceans where the sinus glands are not in the eyestalks, a pair of sinus glands is found in the head in proximity to the brain. A cluster of neurosecretory cell bodies in the medulla terminalis, called the medulla terminalis X-organ, is the predominant contributor of the axonal terminals that make up the sinus gland. In most stalk-eyed crustaceans, the medulla terminalis, which is actually part of the brain, happens to develop in the eyestalk along with the sinus gland, instead of being an integral part of the brain complex that resides in the head proper. The neurosecretory terminals in the corpora cardiaca belong to cells whose cell bodies are located in several parts of the insect brain. The hypothalamus in the vertebrate brain contains neurosecretory cell bodies from which axons emerge whose terminals form the two neurohemal components of the neurohypophysis, the pars nervosa and the median eminence. The neurohypophysis is the portion of the pituitary gland that develops as an evagination from the floor of the brain. See CRUSTACEA; INSECTA.

Functional roles of neurohormones. Neurohormones have a wide variety of functions. The hydra, a cnidarian, is one example of a lower invertebrate that lacks a circulatory system but has neurosecretory cells. The product of these neurosecretory cells is a peptide, the head activator, which increases the rate of head regeneration and the number of buds formed by a hydra. The neurohormones of crustaceans and

insects have been shown to regulate a wide variety of functions, including metabolism, growth, and reproduction. See INSECT PHYSIOLOGY.

The role of the vertebrate hypothalamo-neurohypophysial system has been especially well elucidated. The pars nervosa is the site of release of vasopressin (also called the antidiuretic hormone) and oxytocin. Vasopressin protects against dehydration, promoting water retention by causing increased movement of water out of the kidney tubules back into the blood. This neurohormone also increases the permeability of frog skin to the inward passage of water, thereby facilitating water uptake by dehydrated frogs. In mammals, oxytocin stimulates contraction of uterine smooth muscle at the time of childbirth and also promotes milk ejection from the mammary glands of a nursing female. The median eminence is the release site for several hypothalamic neurohormones that regulate the adenohypophysis, the nonneural portion of the pituitary gland that develops as an evagination from the roof of the mouth. The neurohormones released from the median eminence are called hypothalamic hypophysiotropic hormones and, depending upon their actions on the adenohypophysial cells, are further identified as releasing hormones or release-inhibiting hormones, indicating whether release of a particular hormone from the adenohypophysis is stimulated or inhibited. See ADENOHYPHYSIS HORMONE; NERVOUS SYSTEM (INVERTEBRATE); NERVOUS SYSTEM (VERTEBRATE); NEUROBIOLOGY; NEUROHYPHYSIS HORMONE; PITUITARY GLAND.

Milton Fingerman

Bibliography. R. E. Brown, *An Introduction to Neuroendocrinology*, 1994; R. W. Hill and G. A. Wyse, *Animal Physiology*, 1989; H. F. Nijhout, *Insect Hormones*, 1994; B. T. Pickering et al. (eds.), *Neurosecretion: Cellular Aspects of the Production and Release of Neuropeptides*, 1988; M. Raabe, *Insect Neurohormones*, 1982; C. G. Scanes and P. K. Pang (eds.), *The Endocrinology of Growth, Development, and Metabolism in Vertebrates*, 1992.

Neurulation

The process by which the vertebrate neural tube is formed. The primordium of the central nervous system is the neural plate, which arises at the close of gastrulation by inductive action of the chordamesoderm on the overlying ectoderm. The axial mesodermal substratum causes the neural ectoderm to thicken into a distinct plate across the dorsal midline and influences both its size and shape. Its shield-like appearance, broader anteriorly and narrower posteriorly, presages the areas of brain and spinal cord, respectively.

The lateral edges of the neural plate then rise as neural folds which meet first at the level of the future midbrain, above the dorsal midline, then fuse anteriorly and posteriorly to form the neural tube. The body ectoderm becomes confluent above the closing neural tube and separates from it. Upon closure, the cells (known as neural crest cells) which occupied

the crest of the neural folds leave the roof of the tube and migrate through the mesenchyme to all parts of the embryo, forming diverse structures.

The neural tube thus formed gives rise to the brain and about half of the spinal cord. The remainder of the neural tube is added by the tail bud, which proliferates a solid nerve cord that secondarily hollows into a tube.

The closure of the neural tube confines secretions from its inner wall, which dilate the central canal by their turgor and help to expand the brain vesicles and excavate the solid cord in the tail bud. The total mass and kind of neighboring tissues control thickness or thinness of the neural wall. Adjacent to the notochord there are few mitoses and the floor of the tube remains thin, while in the somite region the lateral walls are actively mitotic and become thick.

The contour is also passively molded by adjacent structures. The distribution and extent of the peripheral tissues to be innervated influence the position and number of nerves which form at any level and the proportion of neural cells differentiating into neurons.

Neuromeres are temporary constrictions in the hindbrain resulting from localized growth processes. They were once thought to have evolutionary significance as vestiges of metamerism, but the plasticity of the neural tube with respect to its mesodermal surroundings makes this interpretation doubtful.

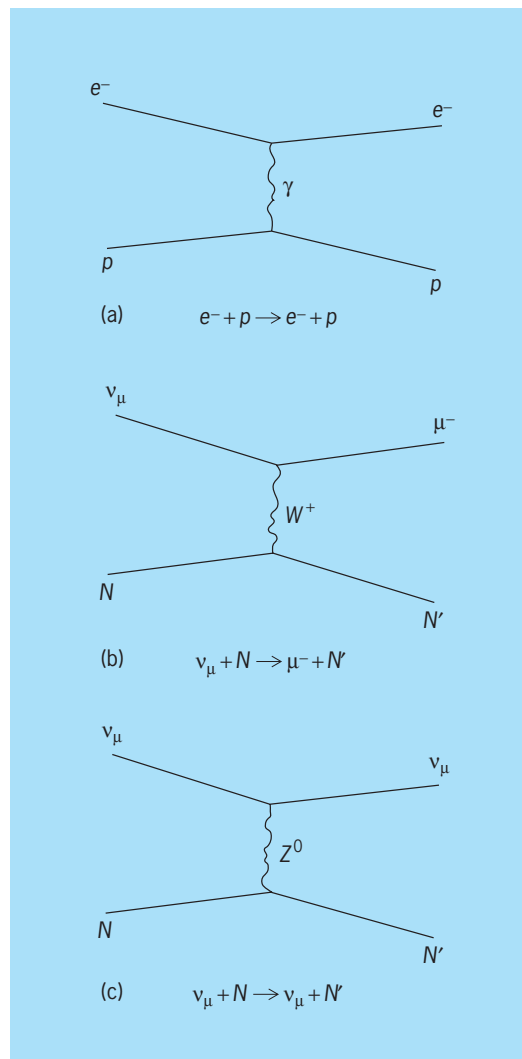
Neural folds never form in teleost fish. The neural plate concentrates into a solid neural keel which then hollows out secondarily as in the tail bud of other vertebrates. See EMBRYOLOGY; EMBRYONIC INDUCTION; GASTRULATION; NERVOUS SYSTEM (VERTEBRATE); NEURAL CREST. Howard L. Hamilton

Neutral currents

Exchange currents which carry no electric charge and mediate certain types of electroweak interactions. The discovery of the neutral-current weak interactions and the agreement of their experimentally measured properties with the theoretical predictions were of great significance in establishing the validity of the Weinberg-Salam model of the electroweak forces.

The neutral-current weak interactions are one subclass of the forces between the fundamental constituents of matter, the quarks and leptons, called the elementary particles. The three basic interactions between these particles are the strong nuclear forces, the electroweak forces, and gravity. The electroweak forces come in three subclasses: the electromagnetic interactions, the charged-current weak interactions, and the neutral-current weak interactions.

Electroweak interactions. The electroweak interactions are theoretically understood to be a current-current type of interaction, where the interaction is mediated by an exchange current or particle. The electromagnetic interaction is mediated by an exchanged photon γ (illus. *a*). Since the photon carries no electric charge, there is no change in charge between the incoming and the outgoing electron



Diagrams for various electroweak interactions. (a) Electromagnetic interaction. (b) Charged-current weak interaction. (c) Neutral-current weak interaction.

(the upper vertex in illus. *a*). The charged-current weak interaction is mediated by the exchange of a charged intermediate boson, the W^\pm , and thus, for example, an incoming neutral lepton such as the ν_μ in illus. *b* is changed into a charged lepton, the μ^- . In the neutral-current weak interactions, the exchanged intermediate boson, the Z^0 , carries no electric charge (hence the name neutral-current interaction), and thus an incident neutral lepton, the ν_μ in illus. *c*, remains an outgoing neutral ν_μ . See ELECTRON; INTERMEDIATE VECTOR BOSON; LEPTON; NEUTRINO; PHOTON.

Discovery and investigation. The existence of the neutral-current weak interactions was predicted by the Weinberg-Salam model, which unified the electromagnetic and the weak nuclear forces into one basic interaction, now called the electroweak interactions. The neutral-current interactions were experimentally discovered in 1973, and have since been extensively studied, in neutrino scattering processes such as neutrino-electron elastic scattering, $\nu + e^- \rightarrow \nu + e^-$; neutrino-proton scattering, $\nu + p \rightarrow \nu + p$; single-pion production, $\nu + p \rightarrow \nu + p + \pi^0$;

and inclusive neutrino scattering, $\nu + p \rightarrow \nu +$ any number of hadrons. Very important information about the properties of the neutral currents has been obtained by studying the interference effects between the electromagnetic and the neutral-current weak interactions in the scattering of polarized electrons on deuterium. Parity violating effects in atomic physics processes due to the neutral weak currents have been observed, and predicted parity-violating nuclear effects have been searched for. *See* PARITY (QUANTUM MECHANICS).

Properties. As a result of the intensive experimental studies of neutral currents in the reactions mentioned above, the properties of the neutral currents have been fairly well determined. They are a mixture of vector (*V*) and axial-vector (*A*) currents (that is, the neutral currents transform like vectors and axial vectors under spatial rotations, and not like scalars, pseudoscalars, or tensors), and they are a mixture of isoscalar ($I = 0$) and isovector ($I = 1$) currents (that is, they transform like isoscalars and isovectors under isotopic spin rotations). The coupling constants that determine the relative strength of the vector, axial-vector, and isoscalar and isovector components have been measured, and are in very good agreement with the values predicted by the Weinberg-Salam model. *See* ELEMENTARY PARTICLE; FUNDAMENTAL INTERACTIONS; I-SPIN; QUARKS; SYMMETRY LAWS (PHYSICS); WEAK NUCLEAR INTERACTIONS. Charles Baltay

Neutralization reaction (immunology)

A procedure in which the chemical or biological activity of a reagent or a living organism is inhibited, usually by a specific neutralizing antibody. As an example, the lethal or the dermonecrotic actions of diphtheria toxin on animals may be completely neutralized by an equivalent amount of diphtheria antitoxin—an antibody produced in animals or in humans after contact with diphtheria toxin or toxoid. Lesser amounts of antitoxin provide intermediate degrees of inhibition. These facts provide the basis for the Schick test for susceptibility to diphtheria. Tetanus and botulinus toxins may be similarly inhibited by their specific antitoxins. In contrast, the typical toxins of dysentery and other gram-negative bacteria are only slightly neutralized, even by large excesses of antibody. Antibodies to bacterial, snake venom, and other enzyme preparations regularly precipitate them from solution so that the supernates are devoid of enzyme activity; however, the neutralization of activity in the precipitate may range from complete to negligible.

Infection of a host by a living bacterium, virus, or other microorganism may also be inhibited or mitigated by the corresponding antibodies, and such neutralization tests are used in the diagnostic examination of sera or of infective agents recovered from such infections as poliomyelitis and yellow fever. In some instances the antibody may be injected into a test animal before, or occasionally shortly after, challenge with the living agent. In other instances

the neutralization of the microbial infectivity by the antibody is permitted to take place in the test tube, and its degree determined by subsequent injection of the mixture into an appropriate test animal. *See* IMMUNOLOGY; NEUTRALIZING ANTIBODY; SEROLOGY.

Henry P. Treffers

Bibliography. B. D. Davis et al., *Microbiology*, 4th ed., 1989.

Neutralizing antibody

An antibody that reduces or abolishes some biological activity of a soluble antigen or of a living microorganism. Thus, diphtheria antitoxin is a neutralizing antibody that, in adequate amounts, abolishes the pathological effects of diphtheria toxin in animals. This is only one characteristic; the other general properties of the antibody are those of the immunoglobulin family (IgG, IgA, or IgM) to which it belongs. *See* ANTIBODY.

Analogous neutralizing effects of antibodies can be demonstrated for the lytic effects of many lysins and, most important, for the pathogenic effects of viruses and the rickettsiae. Since the latter are complex bodies containing multiple antigens, not all the resulting antibodies need have neutralizing activities, although they may display a variety of other serological properties. Antibodies to enzymes constitute a special case; in all instances, they precipitate their corresponding enzyme, but the degree of neutralization may range from 0 to 100%. Antibodies to the endotoxins of the gram-negative bacteria regularly neutralize their toxicity only to a low degree. *See* ANTIGEN; ANTITOXIN; DIPHTHERIA; LYSIN; NEUTRALIZATION REACTION (IMMUNOLOGY); TOXIN.

Henry P. Treffers

Neutrino

An elusive elementary particle that interacts with matter principally through the weak nuclear force. Neutrinos are electrically neutral spin- $1/2$ fermions with left-handed helicity. Many weak interaction processes (interactions that involve the weak force), such as radioactive nuclear beta decay and thermonuclear fusion, involve neutrinos. Present experimental knowledge is consistent with neutrinos being point particles that have no internal constituents. Neutrinos are classified as neutral leptons, where leptons are defined as elementary particles that interact with the electroweak (electromagnetic and weak nuclear) and gravitational forces but not with the strong nuclear force. *See* ELEMENTARY PARTICLE; FUNDAMENTAL INTERACTIONS; HELICITY (QUANTUM MECHANICS); LEPTON; SPIN (QUANTUM MECHANICS); WEAK NUCLEAR INTERACTIONS.

Because the role of gravitational forces is negligible in nuclear and particle interactions and because neutrinos have zero electric charge, neutrinos have the unique property that they interact almost completely via the weak nuclear force.

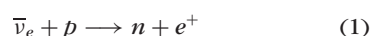
Consequently, neutrinos can be used as sensitive probes of the weak force. As such, neutrino beams at particle accelerators have been employed to study charge-changing (charged current) and charge-preserving (neutral current) weak interactions. However, the extreme weakness (compared to the electromagnetic and strong forces) and short range (of the order of 10^{-18} m) of the weak interaction have made determination of many neutrino properties extremely difficult.

Currently, three distinct flavors (or types) of neutrinos are known to exist: the electron neutrino (ν_e), the muon neutrino (ν_μ), and the tau neutrino (ν_τ). Each neutrino flavor is associated with a corresponding charged lepton, the electron (e), muon (μ), and tau (τ) particle. The electron, muon, and tau neutrinos (or their antiparticles) have been observed in experiments. Based on the observation of interactions involving neutrinos, the lepton flavor families, which comprise the charged and neutral leptons and their antiparticles (e^- , ν_e , e^+ , $\bar{\nu}_e$; μ^- , ν_μ , μ^+ , $\bar{\nu}_\mu$; τ^- , ν_τ , τ^+ , $\bar{\nu}_\tau$), obey laws of conservation of lepton number. These empirical laws state that the number of leptons minus antileptons does not change, both within a flavor family and overall. See ELECTRON; SYMMETRY LAWS (PHYSICS).

The existence of neutrino oscillations (a phenomenon whereby neutrinos change their flavors during the flight from a neutrino source to a detector), seen clearly in observations of atmospheric and solar neutrinos, shows that neutrinos have tiny finite masses which are many orders of magnitude smaller than the masses of their charged lepton counterparts, and also shows that the physical neutrinos do not have pure mass states (quantum-mechanical states) but contain mixtures of two or more neutrino mass states. This mixing indicates that the empirical laws of lepton number conservation are not exact and that they are violated in some physical processes. It is not known whether neutrinos have magnetic or electric dipole moments.

Discovery. The neutrino was postulated by W. Pauli in 1930 in order to rescue the fundamental laws of conservation of energy and linear momentum that were seemingly violated during radioactive beta decay. Pauli proposed that in beta decay a neutral particle of half-integer spin was emitted along with the electron (beta particle). He suggested that this particle was so elusive that it escaped direct experimental detection. In 1934, E. Fermi incorporated Pauli's neutral particle into his theory of nuclear beta decay. See RADIOACTIVITY.

Calculations based on Fermi's theory showed just how tiny the interactions of neutrinos with matter would be. In the inverse beta-decay process [reaction (1), showing the electron antineutrino,

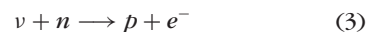
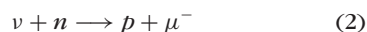


proton, neutron, and positron] the absorption cross section on protons is approximately 10^{-47} m² for free antineutrinos with energies of approximately 1 MeV. This value is equivalent to a mean free path for the

neutrino in water of 6×10^{14} mi (1×10^{15} km), about 10^5 times the diameter of the solar system.

To compensate for the low probability of neutrinos interacting with matter, the first experiment to detect the neutrino directly was undertaken at a nuclear reactor. Reactors produce copious quantities of electron antineutrinos from the beta decay of fission fragments. Using the inverse beta-decay reaction (1) and detecting in time coincidence the neutrons and positrons emitted from this process, F. Reines and C. Cowan confirmed the existence of the neutrino in 1956.

The discovery of the muon neutrino and the demonstration that the muon neutrino was distinguishable from the electron neutrino were carried out in an accelerator-based experiment in 1962. A beam of high-energy neutrinos was created by pions decaying into muons and neutrinos. The neutrinos were observed to produce reaction (2) but not reaction (3), thus proving that different flavors of neu-



trinos must exist. In addition, the observations that electron neutrinos were associated with interactions involving electrons and that muon neutrinos were associated with interactions involving muons gave rise to the empirical law of the conservation of lepton number for separate flavors.

The third neutrino flavor, that of the tau neutrino, was directly observed in 2000. An intense neutrino beam was produced, which was believed to consist of tau neutrinos, and the production of tau leptons was clearly identified by observing the tracks of these leptons and of their decay products.

Scattering. High-energy neutrinos created at particle accelerators have provided precision tests of the standard electroweak model as well as insights into the quark structure of matter. Since neutrino-scattering cross sections increase linearly with energy, the use of high-energy neutrinos allows the accumulation of reasonable statistics. However, the detectors used to observe the interactions must still be quite massive. See PARTICLE DETECTOR.

Neutrino-electron scattering processes can be represented by Feynman diagrams (Fig. 1). The lines on the left of each diagram represent the incoming particles. These particles interact at the vertex, represented by the vertical lines, by transfer of particles labeled Z^0 and W^\pm , which are massive bosons that mediate the electroweak interactions. Outgoing or scattered particles are represented by the lines on the right. In the $\nu_e + e^-$ scattering process (Fig. 1a) both the neutral current interaction (mediated by the Z^0 boson) and the charged current interaction (mediated by the W^+ and W^- bosons) are possible, while the $\nu_\mu + e^-$ process (Fig. 1b) has only the neutral interaction. Measurements of neutrino-electron scattering not only have demonstrated the occurrence of neutral current reactions but have also confirmed the existence of the destructive interference between

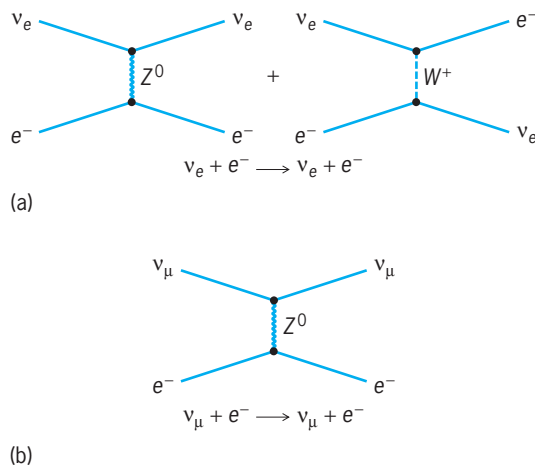


Fig. 1. Neutrino scattering diagrams of (a) electron neutrinos and electrons, and (b) muon neutrinos and electrons.

the weak neutral and charged currents, as predicted by the electroweak theory. Comparisons of neutral-to charged-current neutrino-nucleon scattering have also tested and confirmed the standard electroweak theory. See ELECTROWEAK INTERACTION; FEYNMAN DIAGRAM; INTERMEDIATE VECTOR BOSON; NEUTRAL CURRENTS; QUANTUM ELECTRODYNAMICS.

Studies of neutrino scattering from nucleons have allowed the investigation of the quark-antiquark composition of matter. Neutrinos are naturally polarized, and so in charge-current interactions the neutrino, which has left-handed helicity, scatters primarily from quarks, while the right-handed antineutrino interacts preferentially with antiquarks. [Helicity is defined as the alignment of a particle's spin parallel (right-handed) or antiparallel (left-handed) to the direction of its momentum.] Experiments exploiting this property have played key roles in elucidating the quark nature of matter and the theory of strong interactions. See HELICITY (QUANTUM MECHANICS); QUARKS.

Mass. In the minimal standard model of particle physics, the neutrino masses are chosen as zero, although there are no decisive theoretical arguments to support this choice. In fact, most theoretical models attempting the unification of strong and electroweak interactions predict nonzero neutrino masses. Therefore nonzero neutrino mass clearly indicates the existence of a new theory beyond the standard model of particle physics. See GRAND UNIFICATION THEORIES; STANDARD MODEL.

The consequences of neutrinos having nonzero rest masses would be quite profound, not only for the properties of neutrinos, but also for astrophysics and cosmology. For example, big bang cosmology predicts the existence of relic neutrinos analogous to the microwave photon background remnant left from the creation of the universe. These extremely low-energy neutrinos are predicted to exist with a number density of 110 neutrinos/(flavor \cdot cm³). At this density there would be about 10^{10} times as many neutrinos as baryons in the universe. Although the

neutrino masses are now known to be too small to explain the amounts of nonluminous dark matter, they may have affected the evolution of the universe. See BIG BANG THEORY; COSMIC BACKGROUND RADIATION; COSMOLOGY; UNIVERSE.

Oscillations. If at least one flavor of neutrino has mass, and if any of the laws of conservation of lepton number is violated, then neutrino oscillations are possible. This can happen if the experimentally observed neutrino flavors (ν_e , ν_μ , ν_τ) are not pure quantum-mechanical mass eigenstates but are, in fact, superpositions of mass-eigenstate neutrinos (ν_1 , ν_2 , ν_3). As the neutrinos travel, the admixture components ν_i , which have different masses, propagate with different frequencies. Hence, an initially pure beam of ν_e neutrinos could transform over time into a beam with ν_e and admixtures of ν_μ and ν_τ (vacuum neutrino oscillation). For the two-neutrino oscillation case, the oscillation wavelength λ_{osc} is proportional to $E_\nu/\Delta m^2$, where E_ν is the neutrino energy and Δm^2 is the squared mass difference of the two neutrino mass eigenstates. The oscillation effect is maximum when the distance from a neutrino source to the detector is similar to λ_{osc} . See QUANTUM MECHANICS.

Experiments searching for this oscillatory behavior have been performed by using terrestrial neutrino sources at nuclear reactors and at particle accelerators and by observing extraterrestrial neutrinos, atmospheric neutrinos, and solar neutrinos. The accelerator-based experiments look for the disappearance or appearance of neutrinos of a specific flavor, while the reactor experiments are limited to searches for the disappearance of $\bar{\nu}_e$ neutrinos. The atmospheric neutrino experiments look for the disappearance or appearance of ν_μ or ν_e neutrinos, and the solar neutrino experiments look for disappearance of ν_e neutrinos. All these experiments are extremely difficult to perform, requiring careful attention to possible uncertainties of backgrounds and to the presence of unexpected systematic effects. Although there have been hints of neutrino oscillations both in the measurements of solar neutrinos (the solar neutrino problem) and in the study of atmospheric neutrinos (the atmospheric neutrino anomaly), which both observed less neutrinos than predicted, the definitive evidence of neutrino oscillations was found in 1998 in the study of atmospheric neutrinos by the Super-Kamiokande experiment. The experiment obtained strong evidence, free from the uncertainty of backgrounds and systematic effects. The existence of neutrino oscillations indicates that neutrinos have masses and new physical phenomena such as lepton flavor-violation processes like $\mu \rightarrow e\gamma$, neutrinoless double beta decay, and so on may also exist.

Atmospheric neutrinos. High-energy primary cosmic rays, consisting mostly of protons, hit the Earth's atmosphere and produce pions and kaons (mesons). The lower-energy mesons subsequently decay to muons and ν_μ neutrinos. Muons further decay into electrons and ν_e and ν_μ neutrinos. Higher-energy muons hit the Earth's surface before they decay.

Therefore, in the low-energy regions, where muons decay before reaching the surface of the Earth, the atmospheric neutrino ratio of $(\nu_\mu + \bar{\nu}_\mu)/(\nu_e + \bar{\nu}_e)$ is nearly equal to 2 and the ratio goes up as the energy becomes higher.

Measurements by large underground detectors indicate a discrepancy between the measured ratio of $(\nu_\mu + \bar{\nu}_\mu)/(\nu_e + \bar{\nu}_e)$ and the predicted ratio. Although calculation of the absolute neutrino fluxes can be done to only about 20% accuracy, it has been generally thought that the calculation of the ratio should be good to around 5%. However, combined data yield a value for the ratio of observed fluxes that is about 40% less than the predicted value. This observation can be explained by neutrino oscillations, but could also reflect theoretical model assumptions in the flux calculations.

Definitive evidence of neutrino oscillations was obtained by measuring the neutrino flux as a function of the distance from the neutrino production points to the detector. Cosmic rays approach the Earth very uniformly, and therefore the zenith-angle distribution of the incoming direction of atmospheric neutrinos is symmetric. This property does not depend upon how the neutrino flux is calculated. The zenith angle corresponds approximately to the distance from the neutrino source to the detector. Indeed, the measured zenith-angle distribution of electron neutrinos is very symmetric. However, the measured muon-neutrino distribution is very asymmetric (Fig. 2). Muon neutrinos coming up through the Earth, having traveled about 13,000 mi (8000 mi), are strongly suppressed, and those coming from above, with a flight distance around 10 km (6 mi), are not suppressed. This distance-dependent effect is seen only in the muon-neutrino data, and cannot be explained by phenomena other than neutrino oscillations.

The electron data suggest that the electron neutrinos are not oscillating. An experiment at a reactor, which explored a similar neutrino mass range, also did not see electron neutrino oscillations. Therefore, the atmospheric neutrino oscillation is not an oscillation between muon neutrinos and electron neutrinos, but between muon neutrinos and tau neutrinos. The detailed study shows that the wavelength of the oscillation $\nu_\mu \rightarrow \nu_\tau$ observed in the atmospheric neutrinos is about 500 km/GeV, in other words, about 500 km (300 mi) for 1-GeV neutrinos.

The atmospheric neutrino oscillations can be tested by using the artificial neutrinos produced by particle accelerators with an average neutrino energy of order of 1 GeV and with a detector placed at a distance of a few hundred kilometers—a long-baseline neutrino oscillation experiment. An experiment called K2K (KEK to Kamioka), where the neutrino was produced by the high-energy accelerator at KEK (High Energy Accelerator Organization), in Japan and detected by Super-Kamiokande, 250 km (150 mi) from the neutrino source, was performed and has confirmed atmospheric neutrino oscillations. New experiments to further study neutrino masses by using high-energy accelerators are

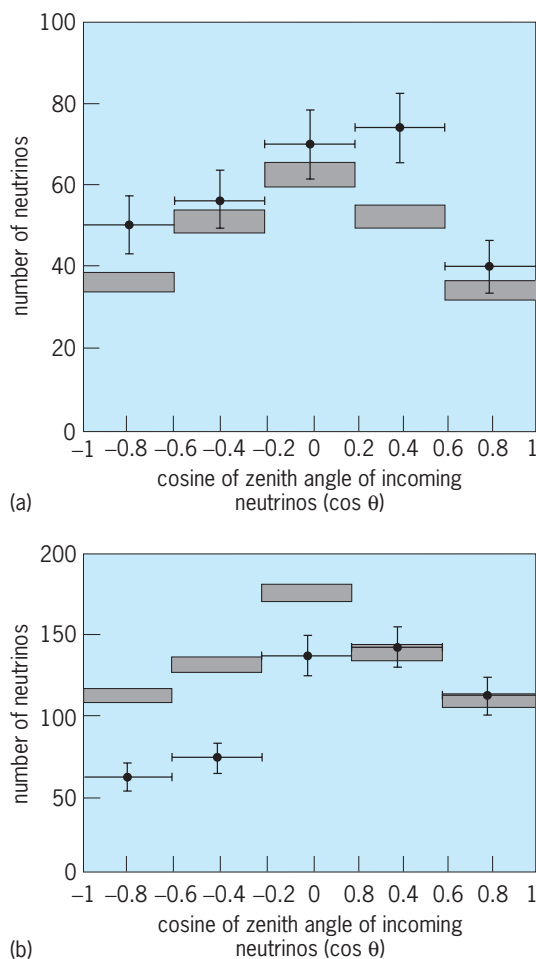
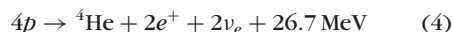


Fig. 2. Zenith-angle (θ) distribution of neutrinos observed at the Super-Kamiokande detector over a period of 535 days. The cross-bars show number of events observed in each range of value of $\cos \theta$, with the horizontal bars indicating the extent of the range and the vertical bars indicating the statistical uncertainty in the frequency of the neutrinos. Rectangles show the expected number of events predicted by a theoretical calculation; the thickness of the rectangles shows the range of uncertainty of the prediction. (a) Electron neutrinos. (b) Muon neutrinos.

underway (2006) and are being constructed in the United States, Europe, and Japan.

Solar neutrinos. Stellar energy production in the Sun is thought to be predominantly powered by proton-proton (pp) chain nuclear reactions, starting with the $p + p \rightarrow d + \nu_e + e^+$ fusion process, that take place at the center of the Sun. This process and additional nuclear reactions following the pp fusion process are predicted to produce abundant quantities of electron neutrinos. The net chain reactions (4)



produce a thermal energy of 26.7 MeV for every two neutrinos. These neutrinos offer an ideal probe to study the physics taking place at the center of the Sun, since, unlike photons (the thermal energy) that take at least 100,000 years to diffuse outward from the core to the surface, neutrinos escape the Sun's core essentially instantaneously with no scattering. See PROTON-PROTON CHAIN.

Before 2001, five experiments, using three different detection methods, had all found significantly fewer neutrinos than are predicted by current theoretical solar models. This discrepancy, called the solar neutrino problem, could arise from an incomplete understanding of solar physics or neutrino properties, or a combination of both. The standard model of the Sun and all proposed nonstandard solar models are unable to account for the observed features of the data.

However, all the experimental results are nicely explained either by vacuum neutrino oscillations or by matter-enhanced neutrino oscillations occurring as the electron neutrinos propagate through the Sun. (The possible occurrence of matter-enhanced oscillations is known as the MSW effect, after S. P. Mikheyev, A. Yu. Smirnov, and L. Wolfenstein.) Unfortunately, there are more than one possible sets of allowed masses and mixing parameters that accommodate all the solar neutrino experiments. This interpretation depends upon the solar neutrino flux calculations, and therefore could not provide convincing evidence.

In June 2001, the Sudbury Neutrino Observatory (SNO) experiment, a 1000-ton heavy-water experiment located in Sudbury, Ontario, Canada, announced the results of the charged-current measurement of solar electron neutrinos, involving exclusively the ν_e content of solar neutrinos arriving at the Earth. The result was compared to the results of the Super-Kamiokande experiment located in Kamioka, Japan, which has a small sensitivity to ν_μ and ν_τ as well as ν_e through neutrino-electron elastic scattering. The measured flux of SNO, $1.75 \pm 0.07 \pm 0.12 \times 10^6/\text{cm}^2/\text{s}$, is significantly lower than the result of Super-Kamiokande, $2.32 \pm 0.03 \pm 0.08 \times 10^6/\text{cm}^2/\text{s}$, by 3 standard deviations, which indicates existence of the ν_μ and ν_τ components in solar neutrinos at the Earth—the result of the neutrino oscillation. The solar neutrino problem has been finally resolved.

The solar neutrino oscillation can be tested by a terrestrial experiment. The confirmation of the solar neutrino oscillation and the precise determination of the oscillation parameters was carried out by the KamLAND experiment located in the Kamioka mine in Japan, which measured the neutrinos produced from nuclear power reactors with an average distance of 180 km (110 mi).

Supernova neutrinos. In February 1987, light from a supernova located about 163,000 light-years (9.6×10^{17} mi or 1.54×10^{18} km) away in the Large Magellanic Cloud (SN1987A) reached the Earth. Preceding the optical signal, a burst of neutrinos was independently observed in two water Cerenkov detectors. This detection of 19 neutrino events was the first recorded signature of the physical phenomena that occur at the core of a star during stellar collapse, and is believed to have signaled the creation of a neutron star. The observation confirmed that about 99% of the binding energy of the star is emitted as neutrinos, as predicted from theories of stellar collapse. Model-independent constraints on ν_e lifetime ($>1.6 \times 10^5$ years) and ν_e mass (<30 eV/ c^2) were also determined from the observation of these few

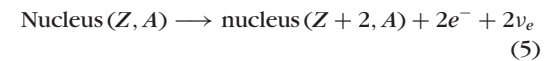
events. See GRAVITATIONAL COLLAPSE; SUPERNOVA.

Direct mass searches. Direct searches for neutrino mass rely on the kinematics of decay reactions (see **table**). The most sensitive mass limit has been set for the electron neutrino (ν_e). This value comes from a measurement of the shape of the beta-decay energy spectrum for tritium (an isotope of hydrogen). Measurements of the electron energy spectrum are made in the energy region where the electron carries away the maximum possible energy E_0 . If the neutrino has a mass, then the number of decays in the

| Lepton masses | | | |
|---------------|--------------------|-----------------|-------------------|
| Neutrinos | | Charged leptons | |
| Particle | Mass | Particle | Mass |
| ν_e | <2.5 eV/ c^2 | e | 0.5110 MeV/ c^2 |
| ν_μ | <0.17 MeV/ c^2 | μ | 105.7 MeV/ c^2 |
| ν_τ | <0.18 MeV/ c^2 | τ | 1777.1 MeV/ c^2 |

region below E_0 will be less than would be expected for the case of a massless neutrino. However, subtle effects such as instrumental resolution, energy loss of the electrons, and so on can cause shape effects that are of similar size to the effect of nonzero neutrino mass. Thus, precise measurement of a value or limit for the neutrino mass requires complete understanding of all systematic effects. A 1999 experiment using a simple gaseous molecular tritium source set an essentially model-independent limit of $m_\nu < 2.5$ eV/ c^2 , and results from three additional experiments are in agreement with this measurement.

Double beta decay. In some radioactive nuclei, ordinary beta decay is energetically forbidden, but the decay given by reaction (5), called two-neutrino (2ν)



double beta decay, is energetically allowed. Here, Z is the atomic number and A the atomic mass. This decay mode, the rarest process observed in nature (first seen in geochemical experiments), is allowed by the standard electroweak theory. Another decay mode, given by reaction (6), called



neutrinoless or 0ν double beta decay, will occur if the neutrino has mass and the neutrino is a Majorana particle (meaning that the particle and antiparticle are indistinguishable, which requires that lepton number is violated). This decay process can be considered a two-step process in which a neutrino emitted in the first step is absorbed in the second.

None of the experimental searches for this process, which employ direct counting experiments that can distinguish two-neutrino and neutrinoless decays, have yet observed the neutrinoless process. Although there is a claim of observing an excess of events in the decay process of ^{76}Ge , this result is very controversial and further studies and confirmation will be needed. The first laboratory observation of the two-neutrino double beta-decay process

in 1986 [$^{82}\text{Se } T_{1/2}(2\nu) = (0.8 - 1.9) \times 10^{20}$ years, where $T_{1/2}(2\nu)$ refers to the half-life for the two-neutrino decay of the parent nuclei] was followed by increasingly sensitive experiments using isotopically enriched detectors and the two-neutrino double beta-decay process is now very well established. Those experiments with higher sensitivity have been searching for the elusive neutrinoless decay. The present limits from these experiments for neutrinoless double beta decay are $T_{1/2}(0\nu) > 10^{23}$ – 10^{25} years. This corresponds, under the assumption that neutrinos are Majorana particles, to deduced Majorana neutrino mass limits of $m_\nu < 0.4 \text{ eV}/c^2$. Yoichiro Suzuki

Bibliography. Q. R. Ahmad et al., Measurement of the rate of $\mu_e + d \rightarrow p + p + e^-$ interactions produced by ^8B solar neutrinos at the Sudbury Neutrino Observatory, *Phys. Rev. Lett.*, 87:071301 (6 pp.), 2001; F. Boehm and P. Vogel, *Physics of Massive Neutrinos*, 2d ed., 1992; E. D. Commins and P. H. Bucksbaum, *Weak Interactions of Leptons and Quarks*, 1983; S. Fukuda et al., Solar ^8B and hep neutrino measurements from 1258 days of Super-Kamiokande data, *Phys. Rev. Lett.*, 86:5651–5655, 2001; Y. Fukuda et al., Evidence for oscillation of atmospheric neutrinos, *Phys. Rev. Lett.*, 81:1562–1567, 1998; R. N. Mohapatra and P. B. Pal, *Massive Neutrinos in Physics and Astrophysics*, 3d ed., 2004; D. H. Perkins, *Introduction to High Energy Physics*, 4th ed., 2000; C. Sutton, *Spaceship Neutrino*, 1992; Y. Suzuki and Y. Totsuka (eds.), *Neutrino 98: Proceedings of the 18th International Conference on Neutrino Physics and Astrophysics*, 1999; K. Zuber, *Neutrino Physics*, 2003.

Neutrino astronomy

The detection and study of neutrinos to learn about astronomical objects and the universe. Almost all current knowledge of the universe derives from the observation of photons. Radio waves, infrared radiation, visible light, ultraviolet waves, x-rays, and gamma rays are all electromagnetic waves composed of photons. Some further knowledge of the cosmos beyond the solar system is gained by observing cosmic rays, which are mostly protons and heavier atomic nuclei. But these positively charged particles do not point to their place of origin because of the magnetic fields of the Milky Way Galaxy, which bend their flight paths. See COSMIC RAYS; ELECTROMAGNETIC RADIATION; GAMMA-RAY ASTRONOMY; INFRARED ASTRONOMY; MAGNETISM; MILKY WAY GALAXY; PHOTON; RADIO ASTRONOMY; ULTRAVIOLET ASTRONOMY; X-RAY ASTRONOMY.

Deep, sharply focused examination of the universe requires a telescope that can see a particle that is not much affected by gas, dust, and magnetic fields. Neutrinos are a candidate. These neutral, weakly interacting particles come almost without any disruption straight from their sources, traveling at very close to the speed of light. A low-energy neutrino in flight would not notice a barrier of lead 50 light-years thick. Neutrino light would provide a wondrous new view of the universe.

Neutrinos in the universe. The fundamental building blocks of the universe, of which all matter is composed, are the fermions: the quarks (up, down, charmed, strange, top, and bottom) and leptons (the electron, muon, and tau-on, plus a neutral particle partner for each, the electron-neutrino, muon-neutrino, and tau-neutrino). Data from the Super-Kamiokande experiment presented in 1998 establish with high probability that some neutrinos have mass, and thus so do all the fermions. See ELEMENTARY PARTICLE; LEPTON; QUARKS.

Neutrinos were made in huge numbers at the time of the big bang. Like the cosmic background radiation, they now possess little kinetic energy as a result of the expansion of the universe. There are expected to be at least 114 neutrinos per cubic centimeter, averaged over all space. There could be many more at Earth because of condensation of neutrinos, moving slowly under the gravitational pull of the Milky Way Galaxy. See BIG BANG THEORY; COSMIC BACKGROUND RADIATION.

The problem with observing these relic neutrinos is the probability of a neutrino interacting within a detector decreases with the square of the neutrino's energy, for low energies. And even when the neutrino does react in the detector, the resulting signal is minuscule. Nobody has been able to detect these lowest-energy neutrinos as yet, and prospects are not good for doing so, at least in the near future.

Stellar neutrinos. Neutrinos also originate in the nuclear fusion in stars. The Sun close by produces a huge flux of neutrinos, which have been detected in five experiments. A long-standing mystery persists in the deficit of about a factor of 2 in the numbers of neutrinos detected compared to expectations—the solar neutrino problem. This deficit is now thought probably to result from neutrino oscillations. The calculation of the neutrino flux from the Sun, in as close agreement with observations as it is, represents a great triumph for the understanding of stellar burning and evolution. However, observations of stellar neutrinos are limited to the Sun. Just as the sky is dark at night despite all the stars, the Sun far outshines all the rest of the cosmos in numbers of detectable neutrinos. See SOLAR NEUTRINOS.

Supernovae. On February 23, 1987, two detectors in deep mines in the United States (the IMB experiment) and Japan (the Kamiokande experiment) recorded a total of 19 neutrino interactions over a span of 13 seconds. Two and a half hours later, astronomers in the Southern Hemisphere saw the first supernova to be visible with the unaided eye since 1604, located in the Large Magellanic Cloud at a distance of 50 kiloparsecs (roughly 150,000 light-years). Many deductions followed about the nature of neutrinos, such as limits on mass, charge, gravitational attraction, and magnetic moment.

Supernovae of the gravitational-collapse type occur when elderly massive stars run out of nuclear fusion energy and can no longer resist the force of gravity. The neutrinos carry off most of the in-fall energy, some 10% of the total mass-energy of the inner part of a star of about 1.4 solar masses. Approximately 3×10^{46} joules of energy is released, along

with about 10^{58} neutrinos, over a few seconds. This is 1000 times the Sun's energy release over its whole lifetime. The awesome visible display consists of a mere one-thousandth of the energy release in neutrinos.

Much can be learned from the final stages of stellar evolution, not only about the process of stellar collapse to a neutron star or black hole (the latter if the progenitor is very massive) but also about properties of neutrinos. For example, the heavier the neutrino, the more slowly it travels, and by studying the structure of the neutrino wave passing by Earth, it may be possible to extract the relative masses of the three types of neutrinos in a direct way that is not dependent on the phenomenon of neutrino oscillations. Four underground detectors have significant capability for supernova detection from the Milky Way Galaxy. From historical records and from observations of distant spiral galaxies, the rate of supernovae in the Milky Way Galaxy is expected to be between one and five per century. Thus experimentalists may have to wait a long time before the next observation, and there is no way of predicting when it will occur. See SUPERNOVA.

High-energy cosmic neutrinos. Higher-energy neutrinos must be made in many of the most luminous and energetic objects in the universe. The most powerful objects seen are active galactic nuclei, which are known to produce particles with energies much higher than the most powerful human-constructed particle accelerators. Enigmatic objects such as gamma-ray bursters, which may be the most energetic explosions observed and are mostly or all at cosmological distances, produce gamma rays up to great energies as well. They may be boun-

tiful neutrino sources or may not be, depending upon the mechanism for the radiation, at present a mystery. Seemingly disallowed cosmic rays, with energies more than 10^8 times greater than terrestrial accelerators (more than 10^{20} eV), have been observed. These particles apparently do not come from the Galaxy, and indeed remain of unknown origin. In fact, the origin of the cosmic rays generally, particularly above about 1 petaelectronvolt (10^{15} eV), remains unknown, though many models have been proposed. Whatever the source, the machinery that accelerates particles to higher energies will inevitably also produce neutrinos. At the highest energies, many speculative models have been proposed as neutrino sources. See ASTROPHYSICS, HIGH-ENERGY; GALAXY, EXTERNAL.

Two things make prospects brighter in the near future for higher-energy neutrino astronomy than for lower energies. First the interaction probability for neutrinos goes up with energy. For the largest present underground detector, Super-Kamiokande, only about one in 10^{12} neutrinos of the typical energy (about 1 GeV) interact when passing through the detector and can be studied. This ratio goes up almost in proportion to the energy of the neutrino, however. Above about 1 PeV, the Earth is opaque to neutrinos and neutrinos could be seen coming only downward. At lower energies, neutrino astronomy must be done backward from optical astronomy, looking downward, using the Earth to filter out anything but neutrinos. The region between 1 TeV (10^{12} eV) and 1 PeV, roughly, is the most promising for attempts to begin regular neutrino astronomy.

The second virtue of seeking higher-energy neutrinos is that the consequences of neutrino interaction with a target (Earth or detector) become more detectable as the energy release is greater. The favored method is to detect muons produced by neutrinos. These muons (unlike electrons or tau-ons) fly a long distance (in closely the same direction as the neutrino) in Earth before stopping, for example about 1 km (0.6 mi) at an energy of a few teraelectronvolts. These charged particles produce Cerenkov radiation, a short flash of light detectable at tens of meters distance by photomultipliers in clear water or ice. Cerenkov radiation occurs when particles exceed the velocity of light in the medium, and is rather like an electromagnetic version of a sonic boom, or of the wake of a ship. Thus a detector can effectively collect the results of neutrino interactions from a target volume much greater than the detector volume itself. See CERENKOV RADIATION; PHOTOMULTIPLIER.

High-energy neutrino telescopes. Neutrino detectors must be placed deep underground or underwater to escape the backgrounds caused by the rain of cosmic rays upon the atmosphere. The cosmic rays produce many muons which penetrate deeply into the Earth, but with decreasing numbers with depth. Hence the first attempts at high-energy neutrino astronomy have been initiated underwater and under ice. The lead project, DUMAND (Deep Underwater Muon and Neutrino Detector), was canceled in 1995, but made great headway in pioneering techniques,

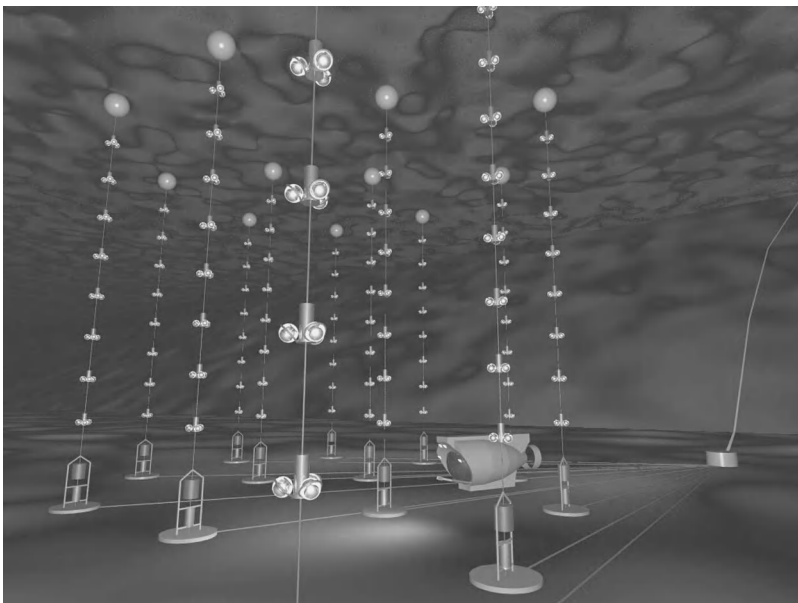


Fig. 1. View of the ANTARES Project from the ocean bottom. The optical detectors consist of modules of a cluster of photomultipliers and electronics, spaced along vertical buoy strings. Spherical floats at the top of each detector string keep the string close to vertical; anchors and releases are at the bottom. Fiber-optic cables go from each string to a junction box, which is serviced by the submarine at the far side of the array. A cable descends the slope from shore in the background. Scales are exaggerated. (ANTARES Collaboration)

studying backgrounds, exploring detector designs, and stimulating interest in astrophysical neutrinos. Another long-running project exists in Lake Baikal, the largest (in volume) and deepest lake in the world, in Siberia. That instrument consists of large light detectors lowered on cables from the winter ice and connected to shore by cable.

Two projects similar to DUMAND are under way in the Mediterranean, the more developed NESTOR (Neutrino Experimental Submarine Telescope with Oceanographic Research) Project located off Pylos in the southwest of Greece, and the ANTARES (Astronomy with a Neutrino Telescope and Abyss Environmental Research) Project located offshore of Marseilles, France. These projects employ basically the same method of bottom-anchored cables, with photomultipliers protected in spherical glass pressure housings, as developed for DUMAND. The deep ocean water is remarkably clear with optical absorption lengths of 40–50 m (130–160 ft), and instruments can be spaced a few tens of meters apart to detect light flashes from most muons passing nearby. An array of vertical strings of such detectors can cover a cubic kilometer (Fig. 1). Both the NESTOR and ANTARES projects aim at prototype neutrino detectors with effective areas for muon collection in the range of 20,000–50,000 m² (200,000–500,000 ft²). This may be compared with the largest present un-

derground instruments, which are about 1000 m² (10,000 ft²) in area, and the desired size for real astronomy of about 10⁶ m² (10⁷ ft² or 1 km²).

A different type of neutrino telescope, the AMANDA (Antarctic Muon and Neutrino Detector Array) Project, is under construction in ice at the South Pole (Fig. 2). Ice below about 1.4 km (0.9 mi) depth is quite clear (100 m or 330 ft attenuation length) and bubble-free, though optical scattering is still somewhat of a problem (25 m or 80 ft effective scattering length). Experimenters have worked out a method to use hot water to drill 2-km-deep (1.2-mi) holes, down which they lower strings of photomultipliers. The instruments become permanently frozen-in after about a day, but the cables can be accessed at the surface, so no complex and expensive electronics need be placed in the inaccessible holes. See NEUTRINO.

John G. Learned

Bibliography. P. Amram et al., Background light in potential sites for the ANTARES undersea neutrino telescope, *Astropart. Phys.*, 13:127–136, 1999; E. Andrs et al., The AMANDA neutrino telescope: Principle of operation and first results, *Astropart. Phys.*, 13:1–20, 2000; V. A. Balkanov et al., The Lake Baikal neutrino experiment, *Nucl. Phys. Proc. Suppl.*, 87:405–407, 2000; J. G. Learned and K. Mannheim, High energy neutrino astrophysics, *Annu. Rev. Nucl. Part. Sci.*, vol. 50, December 2000; L. K. Resvanis, NESTOR status report, in M. Baldo-Ceolin (ed.), *8th International Workshop on Neutrino Telescopes*, Venice, February 23–26, 1999.

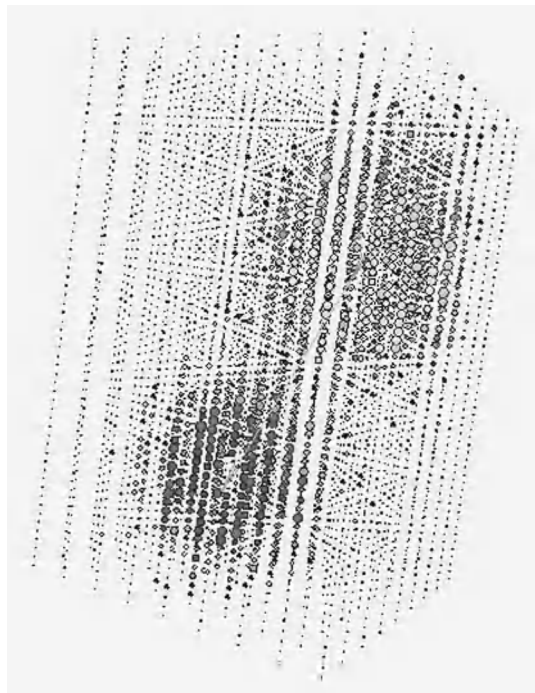


Fig. 2. Simulation of ultra high-energy (10-petaelectronvolt) tau neutrino interaction in IceCube, a proposed follow-up experiment to AMANDA, 1 km³ (0.24 mi³) in size. The circles represent the array of optical modules, and their sizes correspond to the number of photons detected. The event has the “double-bang” signature associated with the production and decay of a tau lepton. The early hits correspond to the initial electromagnetic shower created by the tau neutrino interaction, and the later hits correspond to the decay of the tau lepton.

Neutron

An elementary particle having approximately the same mass as the proton, but lacking a net electric charge. It is indispensable in the structure of the elements, and in the free state it is an important reactant in nuclear research and the propagating agent of fission chain reactions. Neutrons, in the form of highly condensed matter, constitute the substance of neutron stars. See NEUTRON STAR.

Neutrons in nuclei. Neutrons and protons are the constituents of atomic nuclei. The number of protons in the nucleus determines the chemical nature of an atom, but without neutrons it would be impossible for two or more protons to exist stably together within nuclear dimensions, which are of the order of 10⁻¹³ cm. The protons, being positively charged, repel one another by virtue of their electrostatic interactions. The presence of neutrons weakens the electrostatic repulsion, without weakening the nuclear forces of cohesion. In light nuclei the resulting balanced, stable configurations contain protons and neutrons in almost equal numbers, but in heavier elements the neutrons outnumber the protons; in ²³⁸U, for example, 146 neutrons are joined with 92 protons. Only one nucleus, ¹H, contains no neutrons. For a given number of protons, neutrons in several different numbers within a restricted range often yield nuclear stability—and hence the isotopes of an element. See ISOTOPE; NUCLEAR STRUCTURE; PROTON.

Sources of free neutrons. Free neutrons have to be generated from nuclei, and since they are bound therein by cohesive forces, an amount of energy equal to the binding energy must be expended to get them out. Usually the binding energy for each neutron amounts to 1–1.4 picojoules (6–8 MeV). Nuclear machines, such as cyclotrons and electrostatic generators, induce many nuclear reactions when their ion beams strike target material. Some of these reactions release neutrons, and these machines are sources of high neutron flux. If the accelerator is sharply pulsed, the time-of-flight method can be used for accurate energy resolution of the neutrons up to energies of about 0.3 pJ (2 MeV), the flight paths being up to 660 ft (200 m) long. *See* NEUTRON SPECTROMETRY; NUCLEAR BINDING ENERGY; NUCLEAR REACTION.

There are several kinds of portable neutron sources. Some consist of an intimate mixture of an alpha-emitting radionuclide with beryllium powder. Neutrons are released from the nuclear reaction ${}^9\text{Be}(\alpha, n){}^{12}\text{C}$, which is the reaction by which the neutron was discovered in 1932. Intense sources that emit about 5×10^{10} per second are now made, using mixtures of beryllium with, for example, ${}^{238}\text{Pu}$ (half-life 89 years) or ${}^{241}\text{Am}$ (half-life 458 years). Such a source generates several hundred watts of heat. Pure ${}^{252}\text{Cf}$ (half-life 2.65 years) needs no admixture of beryllium, because neutrons are emitted in its spontaneous fission; such a source is especially compact. The neutrons emitted by the foregoing sources have energies that extend up to 0.8–1 pJ (5–6 MeV). A source that gives neutrons of lower energy is the Sb-Be photoneutron source. Here the 1.70-MeV gamma rays of ${}^{124}\text{Sb}$ (half-life 60 days) slightly exceed the binding energy of neutrons in beryllium (1.67 MeV), so the neutrons have an energy of $1.70 - 1.67 = 0.03$ MeV (4.6 femtojoules).

Neutrons are released in the act of fission, and nuclear reactors are unexcelled as intense neutron sources. The absorption of one neutron by a ${}^{235}\text{U}$ nucleus is required to induce fission, but 2.5 neutrons are on the average released; this regeneration makes possible the nuclear chain reaction. A powerful research reactor may generate neutrons in such abundance that 1 cm^2 (0.15 in^2) of a sample placed therein would be traversed by 10^{15} neutrons per second. A hole through the surrounding shield can yield a collimated beam having a unidirectional flux of 10^9 neutrons/ $(\text{cm}^2)(\text{s})$ or 6.5×10^9 neutrons/ $(\text{in}^2)(\text{s})$. The explosion of a 10-kiloton (4×10^{13} J) nuclear bomb releases about 10^{30} neutrons in about 1 microsecond. *See* ATOMIC BOMB; CHAIN REACTION (PHYSICS); NUCLEAR FISSION.

Neutrons occur in cosmic rays, being liberated from atomic nuclei in the atmosphere by collisions of the high-energy primary or secondary charged particles. They do not themselves come from outer space. For information on neutrons of another origin *see* DELAYED NEUTRON; COSMIC RAYS.

Penetrating power. Neutrons resulting from nuclear reactions usually possess kinetic energies of the order of 0.2 pJ (1 MeV). Having no electric charge, they interact so slightly with atomic electrons in mat-

ter that energy loss by ionization and atomic excitation is essentially absent. Consequently they are vastly more penetrating than charged particles of the same energy. The main energy-loss mechanism occurs when they strike nuclei. As with billiard balls, the most efficient slowing-down occurs when the bodies that are struck in an elastic collision have the same mass as the moving bodies; hence the most efficient neutron moderator is hydrogen, followed by other light elements: deuterium, beryllium, and carbon.

The great penetrating power of neutrons imposes severe shielding problems for reactors and other nuclear machines, and it is necessary to provide walls, usually of concrete, several feet in thickness to protect personnel. The currently accepted health tolerance levels for an 8-h day correspond for fast neutrons to a flux of 20 neutrons/ $(\text{cm}^2)(\text{s})$ or 130 neutrons/ $(\text{in}^2)(\text{s})$; for slow neutrons, 700/ $(\text{cm}^2)(\text{s})$ or 4500/ $(\text{in}^2)(\text{s})$. On the other hand, fast neutrons are useful in some kinds of cancer therapy. *See* RADIATION DAMAGE TO MATERIALS; RADIATION INJURY (BIOLOGY); RADIATION SHIELDING; RADIOLOGY.

Detection of neutrons. In pulse counting, neutrons are allowed to produce exothermic (energy-releasing) nuclear reactions, the ionizing products of which are made to generate electrical impulses, in a proportional counter, ionization chamber, or scintillation counter, that can be amplified for individual counting. A proportional counter containing boron, either as a coating on the inner walls or as a filling gas (boron trifluoride), counts neutrons by virtue of the reaction ${}^{10}\text{B}(n, \alpha){}^7\text{Li} + 0.451$ pJ (2.78 MeV). An ionization chamber coated internally with ${}^{235}\text{U}$ gives ionization pulses from the energy of fission fragments as they travel through the gas. A lithium iodide crystal (europium-activated) scintillates because of the energy released by the reaction ${}^6\text{Li}(n, \alpha){}^3\text{H} + 0.765$ pJ (4.78 MeV). The light pulses (scintillations) are reflected onto a photomultiplier, which transforms them to electrical pulses. Capture gamma-rays emitted from strong neutron absorbers, such as cadmium, can similarly be registered by scintillation counting. Large and sensitive neutron detectors have been made by dissolving cadmium or boron salts in tanks containing scintillating liquids. *See* IONIZATION CHAMBER; PARTICLE DETECTOR; SCINTILLATION COUNTER.

In detection by activation, advantage is taken of the fact that many elements become radioactive under neutron irradiation. A sample is exposed, and its radioactive strength is subsequently measured by conventional counting equipment. Gold and indium foils are convenient and sensitive detectors of this kind. Their applications can be further specialized by taking advantage of resonance absorption. If, for example, gold foil is enclosed in cadmium, the cadmium will exclude thermal neutrons, and the gold will be activated mainly by neutrons with an energy of 0.78 attojoule (4.9 eV) because gold has a neutron capture resonance at that energy. Other elements can be similarly used for other selected energies. The converse also occurs; for example, if a thick plug of ${}^{57}\text{Fe}$ is placed in a beam of fast neutrons, it will

preferentially transmit neutrons with an energy of 4 pJ (25 keV) because at that energy the neutrons interact only weakly with the ^{57}Fe nuclei.

Detection by recoil is particularly applicable to the counting of fast neutrons. A counter with hydrogenous walls or filling gas (for example, methane) gives pulses because the protons produce ionization when they recoil after being struck by the fast neutrons.

Intrinsic properties. Free neutrons are themselves radioactive, each transforming spontaneously into a proton, an electron (β^- particle), and an antineutrino. The energy release is 0.125 pJ (0.782 MeV) per event, and the half-life is 10.61 ± 0.16 min. This instability is a reflection of the fact that neutrons are slightly heavier than hydrogen atoms. The neutron's rest mass is 1.0086652 atomic mass units on the unified mass scale (1.67495×10^{-24} g), as compared with 1.0078252 atomic mass units for the hydrogen atom.

Neutrons are, individually, small magnets. This property permits the production of beams of polarized neutrons, that is, beams of neutrons whose magnetic dipoles are aligned predominantly parallel to one direction in space. The magnetic moment is -1.913042 nuclear magnetons. The magnetic structure has a finite size, being roughly exponential in intensity, with a root-mean-square radius of 0.9×10^{-13} cm. Neutrons spin with an angular momentum of $1/2$ in units of $\hbar/2\pi$, where \hbar is Planck's constant. The negative sign attached to the magnetic moment indicates that the magnetic moment vector and the angular momentum vector are oppositely directed. *See* MAGNETON; NUCLEAR MOMENTS; NUCLEAR ORIENTATION; SPIN (QUANTUM MECHANICS).

Despite its overall neutrality, the neutron does have an internal distribution of electric charge, as has been revealed by scattering experiments. On a still finer scale, the neutron can also be presumed to have a quark structure in analogy of that of the proton. *See* QUANTUM CHROMODYNAMICS; QUARKS.

If the centers of the $+$ and $-$ charge distributions in the neutron should be slightly displaced from each other, the neutron would have an electric dipole moment. This possibility has a fundamental importance because it is linked through various interaction theories with the conservation of parity and with the symmetry of time reversal. (If time-reversal invariance holds, the neutron should have no electric dipole moment.) So far it has been found that if the separated charges are equal to $\pm e$ (the electronic charge), the distance between their centers must be less than 10^{-24} cm. This limit is not yet sufficiently small to give a conclusive choice between the various forms of theoretically possible interactions, but it is likely that the sensitivity of the experiments can be increased through the use of ultracold neutrons. *See* PARITY (QUANTUM MECHANICS); SYMMETRY LAWS (PHYSICS).

Ultracold neutrons. When neutrons are completely slowed down in matter, they have a maxwellian distribution in energy that corresponds to the temperature of the moderator with which they are in

equilibrium. At room temperature their mean energy is about 0.004 aJ (0.025 eV), their mean velocity is about 2200 m/s (7300 ft/s), and their de Broglie wavelength is about 0.2 nm. (The approximate coincidence of this wavelength with the interatomic distances in solids is the basis for the science of neutron diffraction.) The maxwellian distribution has a tail extending to very low energies, and a few neutrons (about 10^{-11} of the main neutron flux) at this extreme have energies less than 5×10^{-8} aJ (3×10^{-7} eV), and hence velocities of less than about 7 m/s (23 ft/s). The de Broglie wavelength of these ultracold neutrons is greater than 50 nm, which is so much larger than interatomic distances in solids that they interact with regions of a surface rather than with individual atoms, and as a result they are reflected from polished surfaces at all angles of incidence. *See* NEUTRON DIFFRACTION; THERMAL NEUTRONS.

A typical source of ultracold neutrons consists of a "converter" in the reflector of a neutron reactor, together with an internally smooth, evacuated tube several centimeters in diameter that leads the neutrons out through the shield. The lead-out duct has three or four bends; the ultracold neutrons are preferentially reflected at these bends and are thus selected from the numerous faster neutrons. The neutrons can be further slowed either by sending them upward against gravity [they can rise only 2–3 m (6–9 ft), and the lead-out duct can be vertical if desired], or by means of a "neutron turbine," which is a paddle wheel whose curved blades move in the same direction as the neutrons, but with lower velocity. The neutrons can be polarized by passage through or reflection from a sheet of magnetized material.

The ultracold neutrons can be stored in "neutron bottles," of which there are two kinds. One is simply a vacuum vessel with a door that can be closed after a batch of neutrons has entered. Populations of about 100 neutrons have been retained in such vessels, but the storage times are considerably shorter than the half-life of the neutrons against their natural radioactive decay, and the nature of the extra loss mechanisms is not yet fully understood. The other kind of bottle is again a vacuum vessel, but it uses a multipolar magnetic field that contains the neutrons, because the field gradients act upon the neutrons' magnetic dipole moments so as to keep the neutrons away from the walls. Such a configuration is realized in a torus with hexapole windings around its major circumference, or a sphere with polar and equatorial windings carrying opposing currents.

Ultracold neutrons are important in basic physics and have applications in studies of surfaces and of the structure of inhomogeneities and magnetic domains in solids. *See* ELEMENTARY PARTICLE. Arthur H. Snell Bibliography. Yu. A. Alexandrov, *Fundamental Properties of the Neutron*, 1992; J. Byrne (ed.), *Neutrons, Nuclei and Matter: An Exploration of the Physics of Slow Neutrons*, 1994; R. Golub, *Ultracold Neutrons*, 1991; V. K. Ignatovich, *The Physics of Ultracold Neutrons*, 1990.

Neutron diffraction

The phenomenon associated with the interference processes which occur when neutrons are scattered by the atoms within solids, liquids, and gases. The use of neutron diffraction as an experimental technique is relatively new compared to electron and x-ray diffraction, since successful application requires high thermal-neutron fluxes, which can be obtained only from nuclear reactors. (A thermal neutron is defined as a neutron possessing a kinetic energy of about 0.025 eV.) These diffraction investigations are possible because thermal neutrons have energies with equivalent wavelengths near 0.1 nm and are therefore ideally suited for interatomic interference studies. *See* THERMAL NEUTRONS.

The scattering of low-energy neutrons is generally considered a tool for the study of solid-state phenomena, but many significant investigations have also been performed to obtain information necessary for the understanding of nuclear processes. Diffraction techniques have been employed to measure numerous coherent neutron-scattering amplitudes under special conditions, and these determinations have provided details on the interaction between nuclear forces, potential scattering, and resonance effects. Experiments have also helped to establish upper limits for values of a possible small neutron electric charge and a possible small neutron electric dipole moment. The most important investigations by neutron diffraction have been concentrated on solid-state problems, because these experiments offer unique methods to obtain information on crystallographic properties, magnetic phenomena, and the dynamics of crystal lattices. In its applications to solid-state problems, neutron diffraction is very similar in both theory and experiment to x-ray diffraction, but its importance arises from the significant differences in the scattering of the two types of radiation. *See* CRYSTAL STRUCTURE; CRYSTALLOGRAPHY; LATTICE VIBRATIONS; X-RAY DIFFRACTION.

The scattering of x-rays by atoms results from a scattering interaction with the atomic electrons, and the scattering amplitudes are approximately proportional to the atomic number of the scatterer. Since the electrons are distributed within the atom at distances comparable to the x-ray wavelength, interference effects occur which produce an angular distribution of the scattering, usually referred to as a form factor, that is descriptive of the spatial distribution of the electrons. In the scattering of neutrons by atoms, there are two important interactions. One is the short-range, nuclear interaction of the neutron with the atomic nucleus. This interaction produces isotropic scattering because the atomic nucleus is essentially a point scatterer relative to the wavelengths of thermal neutrons. Strong resonances associated with the scattering process prevent any regular variation of the nuclear scattering amplitudes with atomic number. These resonances can cause the scattering amplitudes to have imaginary components of appreciable size, and they can affect the phase changes between the incident and scattered neutron waves

so that the scattering amplitudes can be either positive or negative. The other important process for the scattering of neutrons by atoms is the interaction of the magnetic moment of the neutron with the spin and orbital magnetic moments of the atom. The amplitude of the interaction varies with the size and orientation of the atomic magnetic moment, and the intensity of scattering has a form-factor angular dependence that is representative of the magnetic electrons within the atom. *See* SCATTERING EXPERIMENTS (ATOMS AND MOLECULES); SCATTERING EXPERIMENTS (NUCLEI).

Techniques. Although thermal neutron beams from nuclear reactors have intensities that are lower than those obtained from efficient x-ray tubes, most of the methods developed for x-ray diffraction can be used with neutrons. Furthermore, since the neutron absorption cross section for many materials is very small, diffraction effects can also be investigated by observations of the neutrons transmitted through a sample. In most experiments the sample is irradiated with monochromatic neutrons, and the scattered radiation is measured with a neutron detector, such as a proportional counter filled with boron trifluoride gas, BF_3 . In structure determinations and other experiments where there is no energy change in the scattering process, only the angular distribution of the scattered neutrons is required. In inelastic scattering experiments, that is, experiments involving an increase or decrease in neutron energy, the energy distribution of the scattered neutrons must also be measured. The neutron energies can be determined with a crystal spectrometer or by analyzing the time of flight of the scattered neutrons. Both polycrystalline and single-crystal specimens can be examined, and the type of specimen is usually determined by the conditions of the experiment. However, since single-crystal techniques provide much better resolution of the diffraction peaks, this method is required for the study of complicated structures.

For those experiments requiring monochromatic neutron beams, the neutrons must be obtained by isolating a narrow slice of the neutron spectrum from the reactor, because these neutrons have a continuous energy distribution with no pronounced peaks. Monochromatization is usually accomplished by diffraction of the reactor neutrons from large single crystals, but filters and mechanical neutron velocity selectors also can be used. In investigations of certain magnetic properties, it is frequently necessary to use neutron beams that are both monochromatic and polarized. Such beams can be obtained in the monochromating process by using single crystals, because the neutrons scattered under specific conditions from particular ferromagnetic crystals are almost completely polarized. Diffraction experiments with polarized neutrons have been particularly important for precise determinations of magnetic form factors, and techniques using polarization analysis provide a unique method for separating neutrons scattered by magnetic and nuclear interactions.

Another technique for neutron diffraction utilizes pulses containing the entire spectrum of reactor

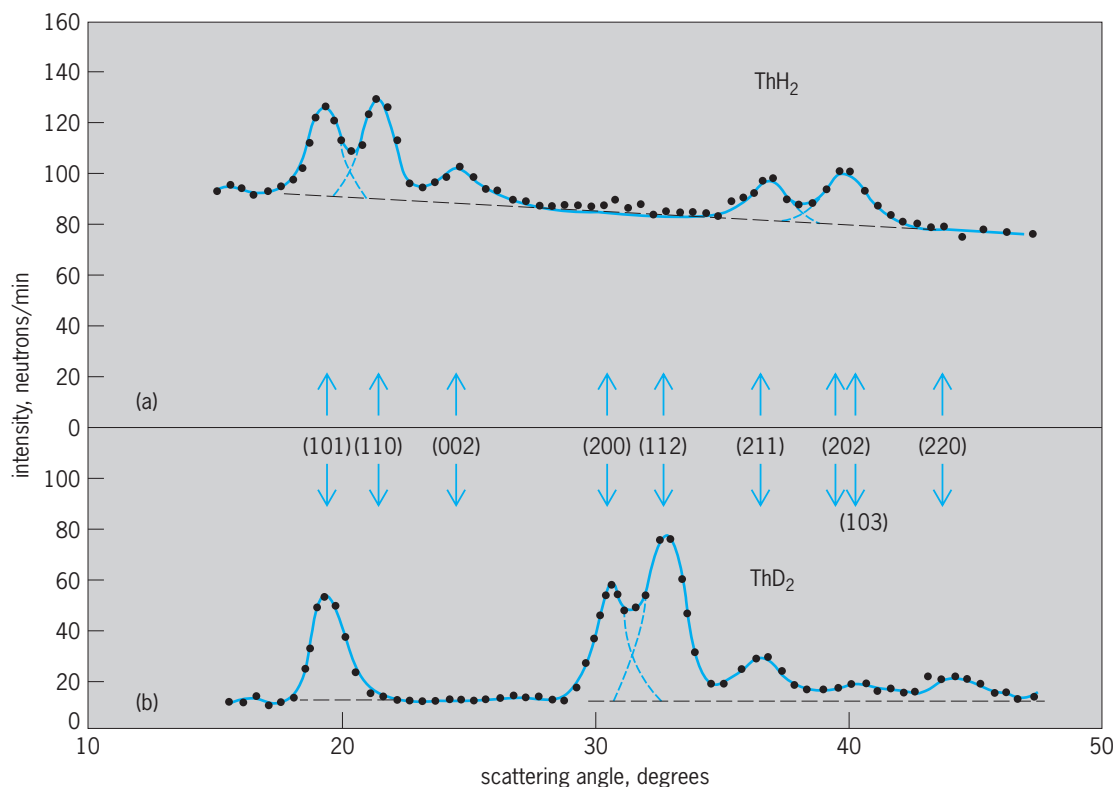


Fig. 1. Neutron diffraction patterns from (a) polycrystalline thorium hydride, ThH_2 , and (b) polycrystalline thorium deuteride, ThD_2 . Differences in the patterns are caused primarily by differences in the nuclear scattering from hydrogen and deuterium atoms.

neutrons and employs time-of-flight energy analysis of the neutrons scattered at a fixed angle. This technique is particularly useful with pulsed neutron sources, and it offers definite advantages in certain types of experiments where the range of scattering angles is limited. However, in most investigations with continuous neutron sources, it is not competitive with the conventional diffraction methods. Photographic techniques can also be used in neutron experiments, but they have been restricted almost completely to quick qualitative examinations.

Auxiliary apparatus for controlling the conditions of the samples can be constructed easily because of the relatively low neutron cross sections of most materials. Consequently, many diffractometers and spectrometers have low-temperature cryostats, furnaces, and electromagnets as integral parts of the instruments. Furthermore, the methods of investigation are readily adaptable to automation, and some of the newer instruments at high-flux research reactors are controlled directly by on-line computers.

Chemical crystallography. Since the nuclear scattering amplitudes for neutrons do not vary uniformly with atomic number, there are certain types of chemical structures which can be investigated more readily by neutron diffraction than by x-ray diffraction. Moreover, since neutron scattering is a nuclear process, when the scattering amplitude of an element is not favorable for a particular investigation, it is frequently possible to substitute an enriched isotope which has scattering charac-

teristics that are markedly different.

The most significant application of neutron diffraction in chemical crystallography is the structure determination of composite crystals which contain both heavy and light atoms, and the most important compounds in this general classification are the hydrogen-containing substances. Since hydrogen and deuterium have neutron scattering amplitudes that are comparable to those of other atoms, their positions in crystals can be determined by this technique, whereas x-ray diffraction usually gives little information about them (Fig. 1). Most of the early investigations concerned relatively simple inorganic compounds, but with the construction of higher-flux reactors and the use of computers in data collection and processing, more complex crystal structures have been examined. The crystal structure of sucrose, $\text{C}_{12}\text{H}_{22}\text{O}_{11}$, which required the measurement and analysis of 2800 independent Bragg reflections, was determined by neutron diffraction, and studies have also been made on some of the simplest biological molecules. In addition to the inherent interest in the materials, all of these investigations have helped to provide a better understanding of hydrogen bonds in crystals. For a discussion of Bragg reflections see X-RAY DIFFRACTION.

Neutron diffraction techniques have also been applied to many other compounds with special chemical or physical properties and with unfavorable x-ray scattering amplitudes. These investigations include the ionic displacements associated with ferroelectric

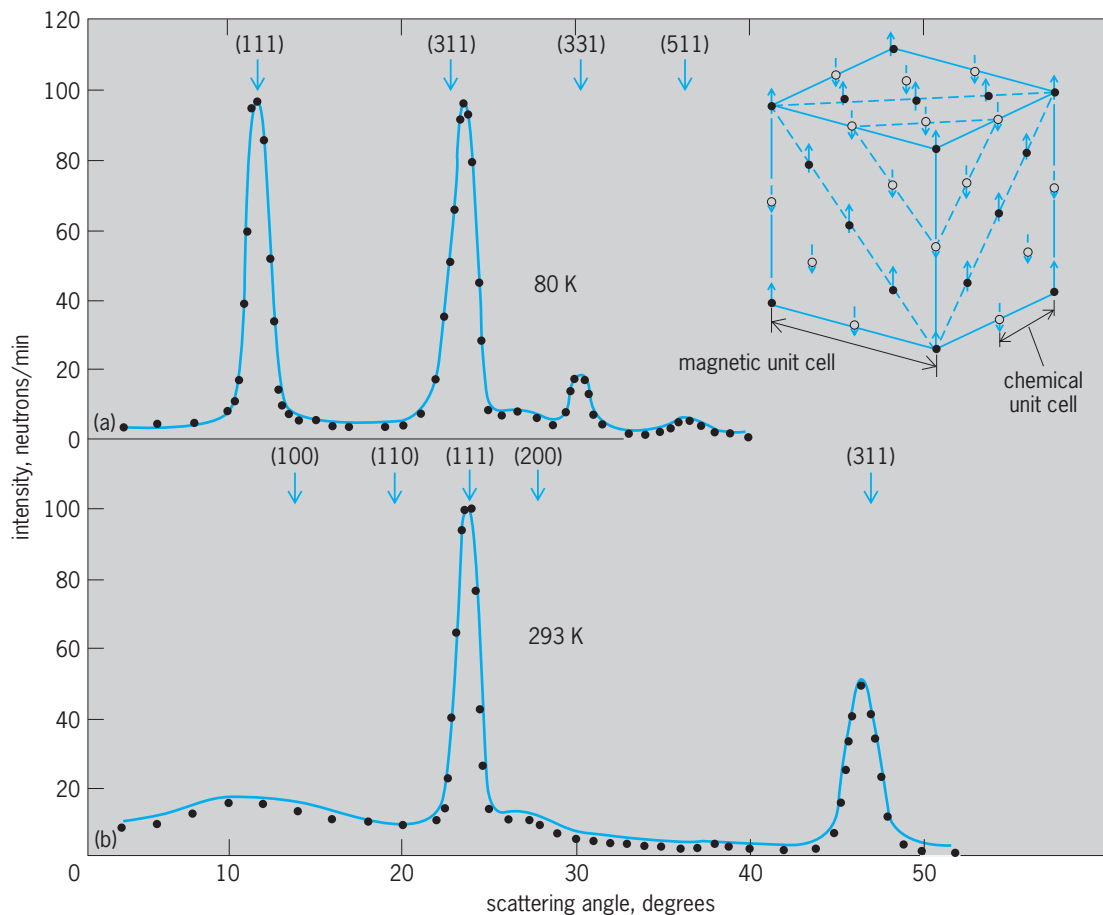


Fig. 2. Neutron diffraction patterns from polycrystalline manganese oxide, MnO , at temperatures (a) below and (b) above the antiferromagnetic transition at 122 K (-240°F). At 293 K (68°F), only nuclear reflections are observed, while at 80 K (-316°F), additional reflections are obtained from the indicated antiferromagnetic structure. The atomic magnetic moments in this structure are directed along a magnetic axis within the (111) planes.

transitions, the rotation of molecular groups in compounds, and order-disorder phenomena in alloy systems composed of atoms with almost the same atomic number. Furthermore, since the scattering of neutrons by the nucleus is isotropic, the neutron technique is advantageous in investigations of liquids, gases, amorphous materials, and other structures where the features of the diffraction pattern at large scattering angles are significant.

Magnetic scattering. The interaction of the magnetic moment of the neutron with the orbital and spin moments in magnetic atoms makes neutron scattering a unique tool for the study of a wide variety of magnetic phenomena, because information is obtained on the magnetic properties of the individual atoms in a material. This interaction depends on the size of the atomic magnetic moment and also on the relative orientation of the neutron spin and of the atomic magnetic moment with respect to the scattering vector and with respect to each other. Consequently, detailed information can be obtained on both the magnitude and orientation of magnetic moments in any substance which displays magnetic properties.

Each type of magnetic lattice has a characteristic neutron diffraction pattern. For paramagnetic ma-

terials, where the atomic moments are uncoupled and randomly oriented in direction, the magnetic scattering is diffuse. For ordered magnetic lattices the magnetic scattering is found in Bragg reflections. Magnetic reflections from ferromagnetic materials occur superimposed on the nuclear reflections, but for antiferromagnetic materials, in which the atomic moments are oriented with no net magnetization per unit volume, superlattice reflections are observed at other scattering angles, as shown in Fig. 2. Since ferrimagnetic materials have atomic moments with antiparallel components but also possess a net ferromagnetic moment, magnetic reflections are observed at both nuclear and other positions. Thus neutron diffraction experiments can determine the magnetic transition temperature, type of magnetic order, temperature variation of the magnetic order, and detailed magnetic configuration in the ordered lattice. This information is basic to understanding the magnetic exchange interactions that are responsible for producing an ordered magnetic structure. See ANTIFERROMAGNETISM; FERRIMAGNETISM; FERROMAGNETISM; PARAMAGNETISM.

The investigation of antiferromagnetic and ferrimagnetic substances is one of the most important applications of the neutron diffraction technique,

because detailed information on the magnetic configuration in these systems cannot be obtained by other methods. Several hundred antiferromagnetic structures have been investigated, and various types of systems that have been determined are shown schematically in Fig. 3. In most antiferromagnetic substances the magnetic moments are found in truly antiparallel arrays, but more complicated systems have been encountered. Structures have been determined in which the magnetic moments are canted with respect to each other, and a number of systems have been found to have a long-range modulation of the moment distribution. The latter configurations require long-range magnetic interactions for stability, and in the heavy rare-earth metals and alloys which have such configurations, a long-range interaction through the conduction electrons can explain many of their unusual magnetic properties. Similar complex structures have also been observed in certain types of ferrimagnetic materials.

The magnetic moments of a simple ferromagnet can usually be obtained from saturation magnetization experiments; but in substances such as ferromagnetic alloys, although the moments are arranged in parallel alignment, different types of atoms have different moment values. Since magnetic measurements can give only the average moment of the alloy, the determination of the individual magnetic moments of the constituent atoms has been another important aspect of neutron diffraction. Alloys with both ordered and disordered arrangements of the atoms can be studied, and experiments on very dilute concentrations provide information on the effect of magnetic and nonmagnetic impurities.

The form factor for the magnetic scattering of neutrons can be interpreted in terms of the spatial distribution and angular momentum characteristics of the magnetic electrons within the atoms. Determinations of these form factors can be made from either measurements of magnetic intensities in coherent reflections or from measurements of paramagnetic scattering. However, the most precise measurements of this type have been made on reflections from ferromagnetic and ferrimagnetic materials, utilizing polarized neutrons. This technique takes advantage of cross terms in the combined nuclear and magnetic scattering to provide an accuracy not readily obtainable with an unpolarized beam. Measurements on the ferromagnetic iron-group metals have provided detailed maps showing the distribution of magnetic electrons throughout the unit cells.

A variety of changes can be produced in neutron diffraction patterns by application of magnetic fields sufficiently strong to change the orientation of atomic magnetic moments within the sample. The use of magnetic fields in these experiments can therefore provide information on ferromagnetic and antiferromagnetic domains, on the magnetic anisotropy within magnetic structures, and on the nature and strength of magnetic exchange interactions.

Inelastic scattering. Scattering investigations of thermal neutrons, in which the neutrons undergo a

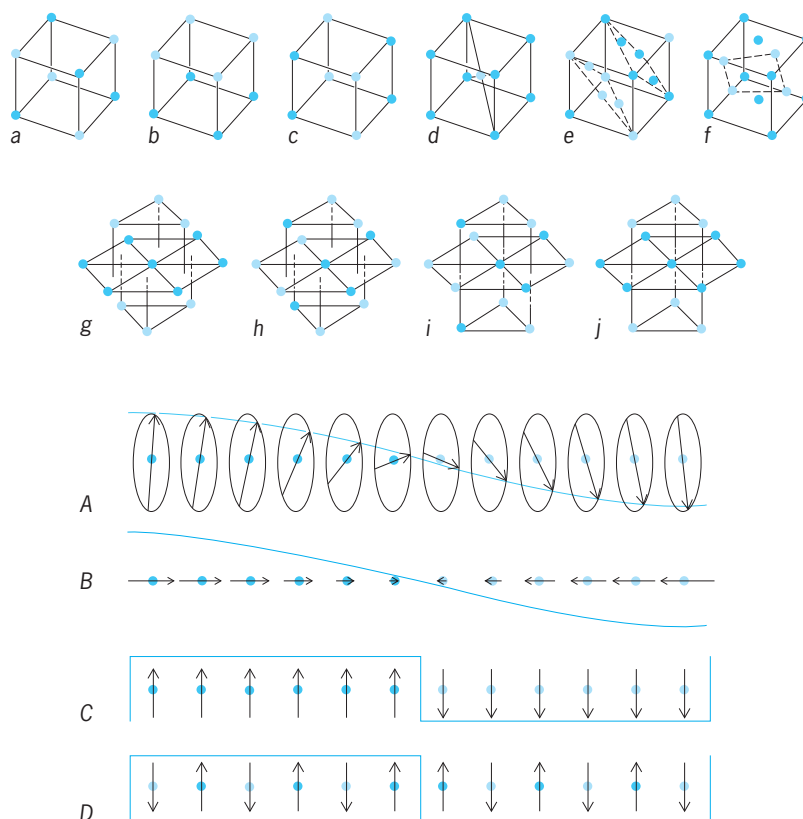


Fig. 3. Schematic representation of various antiferromagnetic systems studied by neutron diffraction. In structures *a* through *j* there is a single magnetic axis and the atomic moments at the darker circles are antiparallel to those at the lighter circles. Types of antiferromagnetism with a long-range modulation of the moment distribution are indicated by *A* through *D*.

energy change, fall into the broad scope of neutron diffraction. However, since both the angular distribution and the energy distribution of the scattered neutrons must be determined, these investigations are different from those usually associated with diffraction experiments. Because of the favorable values of energies and wavelengths associated with thermal neutrons, these measurements provide a method for studying many physical properties of solids and liquids that cannot be studied by any other method. The wavelengths are comparable with atomic separations, and the energies are of the order of the characteristic energies of solids and liquids, so that the energy and momentum changes resulting from many interactions can be measured easily by diffraction techniques. Furthermore, analogous to the case for elastic scattering, these inelastic scattering processes can result from a nuclear interaction or from a magnetic interaction, and the dynamical properties of both atomic systems and magnetic systems can be investigated.

One of the most important uses of inelastic neutron scattering is the study of thermal vibrations of atoms about their equilibrium positions, because lattice vibration quanta, or phonons, can be excited or annihilated in their interactions with low-energy neutrons. The measurements provide a direct determination of the dispersion relations for the normal vibrational modes of the crystal and do not

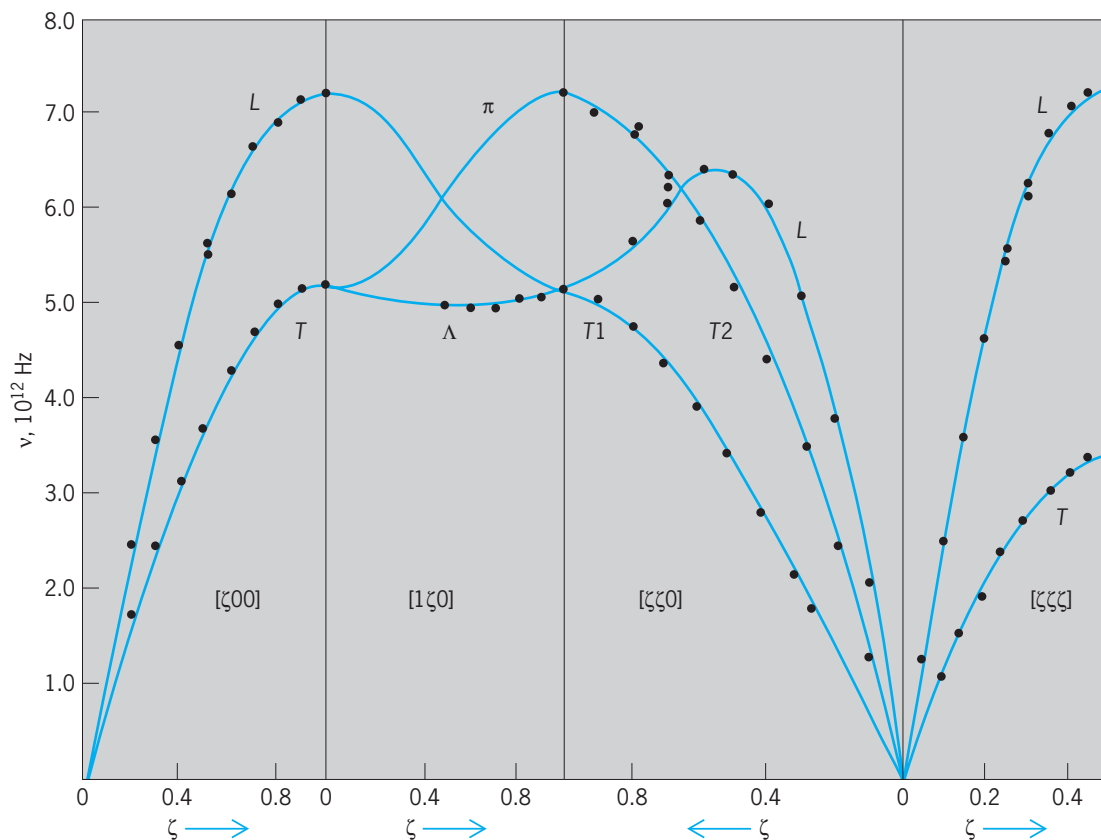


Fig. 4. Phonon dispersion curves for copper at 49 K (-371°F), which relate phonon frequency ν to phonon wave vector ζ along major symmetry directions indicated in brackets. Solid circles are results from inelastic neutron scattering experiments and smooth curves are calculations based on an axially symmetric interatomic force model extended to six nearest neighbors. (*L* and *T* correspond to longitudinal and transverse modes of vibration, respectively, while π and Δ represent modes of vibration with both longitudinal and transverse components.)

require the large corrections necessary in similar x-ray investigations. These measured dispersion relations furnish the best experimental information available on interatomic forces that exist in crystals (**Fig. 4**). Similar measurements can be made on the quantized motion of magnetic moments about the equilibrium direction in an ordered magnetic lattice, and these magnon dispersion curves can be interpreted in terms of the magnetic forces between atoms. Furthermore, neutron scattering techniques are not restricted to solids but can be used to investigate details of atomic motion in liquids.

With the availability of higher neutron fluxes and more sophisticated techniques, it has been possible to extend these investigations to more difficult problems, such as the effect of impurities on interatomic forces. Localized vibrational modes associated with impurities can be observed, and in certain types of experiments information can be obtained on the degree of spatial localization of the modes in the crystal. Dispersion curves have been measured with sufficient precision to permit observation of additional effects, including interaction between phonons and magnons and interaction between phonons and conduction electrons. The latter observations may prove useful for determining the Fermi surface in metals as a function of crystallographic direction. See FERMI SURFACE; MAGNON; PHONON. Michael K. Wilkinson

Bibliography. E. Balcar and S. W. Lovesey, *Theory of Magnetic Neutron and Photon Scattering*, 1989; Y. A. Izyumov, *Neutron Diffraction of Magnetic Materials*, 1991; Y. A. Izyumov and N. A. Chernoplekov, *Neutron Spectroscopy*, 1994; S. W. Lovesey, *Theory of Neutron Scattering from Condensed Matter*, 2 vols., 1986.

Neutron optics

The general class of experiments designed to emphasize the wavelike character of neutrons. Like all elementary particles, neutrons can be made to display wavelike, as well as particlelike, behavior. They can be reflected and refracted, and they can scatter, diffract, and interfere, like light or any other type of wave. Many classical optical effects, such as Fresnel diffraction, have been performed with neutrons, including even those involving the construction of Fresnel zone plates. A perfect double-slit neutron interference pattern is shown in **Fig. 1**. This pattern not only conforms closely to the predictions of scalar diffraction theory, but can be used to put a stringent upper limit on possible nonlinear contributions to the Schrödinger equation, which determines how the neutron propagates. See DIFFRACTION; INTERFERENCE OF WAVES; REFLECTION OF ELECTROMAGNETIC

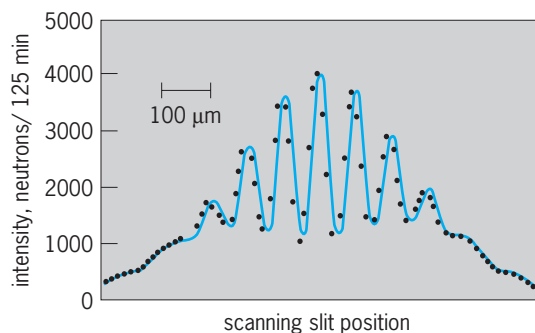


Fig. 1. Diffraction pattern made by passing a collimated neutron beam through a two-slit system. The solid line is the prediction of scalar diffraction theory. (After A. Zeilinger et al., 1981 Symposium on Neutron Scattering, Argonne National Laboratory, American Institute of Physics, 1962)

RADIATION; REFRACTION OF WAVES; SCATTERING OF ELECTROMAGNETIC RADIATION; WAVE (PHYSICS).

The typical energy of a neutron produced by a moderated nuclear reactor is about 0.02 eV, which is approximately equal to the kinetic energy of a particle at about room temperature (80°F or 300 K), and which corresponds to a wavelength of about 10^{-10} m. This is also the typical spacing of atoms in a crystal, so that solids form natural diffraction gratings for the scattering of neutrons, and much information about crystal structure can be obtained in this way. However the wavelike properties of neutrons have been confirmed over a vast energy range from 10^{-7} eV to over 100 MeV. See NEUTRON DIFFRACTION.

Neutron interferometer. Neutrons, being uncharged, can be made to interfere over large spatial distances, since they are relatively unaffected by the stray fields in the laboratory that deflect charged particles. This property has been exploited very beautifully by using the neutron interferometer, invented by H. Rauch, W. Treimer, and U. Bonse in 1974. This device is made possible by the ability to grow essentially perfect crystals of up to 4 in. (10 cm). The typical interferometer is made from a single perfect crystal cut so that three parallel “ears” are presented to the neutron beam (Fig. 2). This allows the incident beam to be split and subsequently recombined coherently. See COHERENCE; INTERFEROMETRY; SINGLE CRYSTAL.

The incident beam is split by Bragg scattering at the first ear (through angle θ off planes parallel to a - b in Fig. 2b), redirected at the second ear, and recombined at the third ear. The beam is then detected at counters A and B . Because the crystal is perfect, the de Broglie wavelength of the neutrons that are Bragg-scattered and reach the counters is very accurately determined, and its magnitude is of the same order as the crystal spacing. Therefore, a neutron wave packet that passes through the interferometer has the remarkable property that it is split into two subpackets, which travel about 4 in. (10 cm) along separate paths, I and II, during which time they are separated by several centimeters, until they are recombined at d . The relative phase at d determines the counting rates at A and B . An experimental apparatus

can be inserted at c to change the relative phase. A change of a half wavelength will vary counter A from a maximum to a minimum reading (with the opposite effect on B). While separated, the beams have traveled a distance of about 10^9 wavelengths, and yet, if the packet in one beam is perturbed by a small fraction of a wavelength, the relative counting rates of counters A and B will significantly change. Thus, this instrument is capable both of detecting very minute forces acting on the neutron, and of demonstrating the wave behavior of the neutron over a macroscopic distance. Using these properties, experiments have been carried out with the interferometer to illustrate that neutrons are ruled by the laws of quantum theory and show very dramatic wavelike and unclassical behavior.

Gravity experiment. One of the most significant experiments performed with the interferometer (sometimes referred to as the COW experiment, after R. Colella, A. W. Overhauser, and S. A. Werner) involved rotating the interferometer about the incident beam so that one neutron path was higher than the other, creating a minute gravitational potential difference (of 10^{-9} eV) between the paths. This was sufficient to cause a path difference of 20 or so wavelengths between the beams, which could easily be seen with the interferometer (Fig. 3). This remains the only type of experiment that has ever seen a quantum-mechanical interference effect due to gravity. It also verifies the extension of the equivalence principle to quantum theory (although in a form more subtle than its classical counterpart). See GRAVITATION; RELATIVITY.

Rotation of wave function by 360°. Another significant experiment involved the first direct verification of a highly unclassical prediction of quantum theory, namely, that a neutron wave function (or that of any spin- $1/2$ particle such as the electron or proton) would change its sign if the neutron was rotated by 360°. Classically, of course, if an object is rotated by 360° it will be totally unaffected. But according to quantum theory, if the neutron in one leg of the

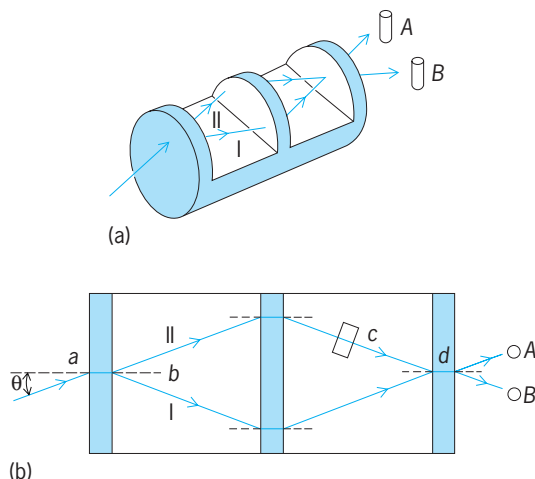


Fig. 2. Neutron interferometer. (a) Oblique view. (b) Top view. (After D. M. Greenberger, *The neutron interferometer*, *Rev. Mod. Phys.*, 55:875-905, 1983)

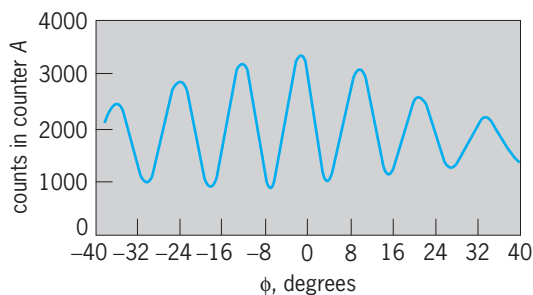


Fig. 3. Variation of counting rate in one counter with rotation angle ϕ in the gravity experiment, due to phase shift resulting from gravitational potential difference. Neutron wavelength was 0.106 nm. (After J.-L. Staudenmann et al., *Gravity and inertia in quantum mechanics*, *Phys. Rev.*, **A21**:1419–1438, 1980)

interferometer is rotated by 360° , then if the two beams would have constructively interfered on recombining they will now destructively interfere and vice versa. If the neutron is rotated by another 360° , it will revert to its original state. This result was known to be true through many indirect measurements, but it was shown directly by placing a magnetic field in one leg of the interferometer which caused the neutron to precess by a known amount.

Superposition principle. One of the most important properties of quantum theory is the superposition principle, according to which the wave functions of coherent states combine additively. One of the most simple, yet most profound examples of this involves the spin for spin- $1/2$ particles. Such particles have two spin states, spin “up” (spin is directed upward along the z axis, $s_z = +1/2$), and spin “down” ($s_z = -1/2$). If the z component of spin is measured, it is found that the particle is either spin up or spin down. Yet a spin along the x axis can be described as a linear combination of these two states. By a judicious use of magnetic fields, the neutron was polarized so that it was spin up in one leg of the interferometer and spin down in the other. Then when these two beams of equal amplitude, each polarized along the z axis, were recombined after passing through the interferometer, it was seen that they were polarized in the x - y plane, exactly according to the laws of quantum theory. This polarization was perpendicular to that of either of the beams separately, which proved both the coherence of the beam and the superposition principle for spin. This is the first time a beam of spin- $1/2$ particles has ever been spatially separated and then coherently recombined in this fashion. See SUPERPOSITION PRINCIPLE.

Ultracold neutron resonances. Many noninterferometer experiments have also been done with neutrons. In one experiment, resonances were produced in transmitting ultracold neutrons (energy about 10^{-7} eV) through several sheets of material. This is theoretically similar to seeing the few lowest states in a square-well potential in the Schrödinger equation. See NEUTRON; NONRELATIVISTIC QUANTUM THEORY; QUANTUM MECHANICS.

Daniel M. Greenberger

Bibliography. U. Bonse and H. Rauch (eds.), *Neutron Interferometry*, 1979; D. M. Greenberger and A. W. Overhauser, The role of gravity in quantum

theory, *Sci. Amer.*, 242(5):66–67, May 1980; V. E. Sears, *Neutron Optics*, 1989; S. A. Werner, Neutron interferometry, *Phys. Today*, 33(12):24–30, December 1980.

Neutron spectrometry

A generic term applied to experiments in which neutrons are used as the probe for measuring excited states of nuclides and for determining the properties of these states. The term neutron spectroscopy is also used. The strength of the interaction between a neutron and a target nuclide can vary rapidly as a function of the energy of the incident neutron, and it is different for every nuclide. At particular neutron energies the interaction strength for a specific nuclide can be very strong; these narrow energy regions of strong interactions are called resonances. The strength of the interaction, expressing the probability that an interaction of a given kind will take place, can be considered as the effective cross-sectional area presented by a nucleus to an incident neutron. This cross-sectional area is expressed in barns ($1 \text{ barn} = 10^{-28} \text{ m}^2$) and is represented by the symbol σ . The neutron total cross section of the nuclide ^{231}Pa from 0.01 to 10 eV is shown in **Fig. 1**. Even though the neutron has zero charge, neutron energies are measured in electronvolts ($1 \text{ eV} = 1.60 \times 10^{-19}$ joule). Neutron spectroscopy covers the vast energy range from 10^{-3} eV to 10^3 MeV.

Unbound and bound states of nuclides. Each resonance corresponds to an unbound excited state of the compound nucleus (**Fig. 2**) at an excitation energy that is the sum of the energy of the neutron and the binding energy of the neutron (4–11 MeV) which has been added to the target nuclide. The compound nucleus has a mass number which is one more than that of the target nuclide. Near the ground state of the compound nuclide the spacing of energy levels may be 10^4 to 10^6 eV. However, for a heavy nuclide, such as the compound nuclide ^{232}Pa , the excited states at an excitation energy just above the binding energy of approximately 5.5 MeV are less than 1 eV apart. To observe the individual states, the neutron energy resolution must be smaller than the level spacing. This can be achieved with low-energy neutrons, because they provide the requisite resolution, and there is no Coulomb repulsion to prevent them from entering the target nucleus. Neutron spectroscopy is presently the only technique which can provide this detailed information. For light nuclei (atomic weights ≤ 40) the spacing between the excited states can be many keV, and resonances can be resolved up to neutron energies of several MeV (**Fig. 3**). Lower-energy bound excited states of the compound nucleus below the neutron binding energy can also be studied by gamma-ray spectroscopy, observing the energy of the gamma rays emitted after the capture of neutrons at resonances or at thermal energy (**Fig. 2**).

Nuclear energy levels of the target nuclide can also be determined by measuring the energy spectrum of

neutrons which are inelastically scattered by the target under bombardment from monoenergetic MeV incident neutrons (Fig. 4). The energy of an excited state is equal to the difference in energies between the incident and scattered neutrons, $E_n - E_n'$. If the incident neutron in Fig. 4 has energy E_{n1} , it has enough energy to excite any of the six lowest levels and emit a neutron of lesser energy than E_{n1} . A neutron of energy E_{n2} could excite only the two lowest levels and emit a neutron of lesser energy than E_{n2} . Information on these same low-energy states can be obtained by measuring the energies and intensities of the gamma rays from the deexcitation of these states excited by inelastically scattered neutrons.

Neutron reactions and resonance parameters. The abbreviated notation for neutron reactions, (n,n) and so on, lists the bombarding particle before the comma, and the emitted particle or particles after the comma. The standard symbols are n (neutron), p (proton), d (deuteron), α (alpha particle), γ (gamma ray), f (fission), and T (total). A more complete description of the reaction lists also the target and product nuclides, for example, ${}^AZ(n,\gamma){}^{A+1}Z$. The reactions most useful for neutron spectroscopy are the total interaction; elastic scattering (n,n) ; radiative capture (n,γ) ; fission (n,f) ; inelastic scattering (n,n') ; charged particle emission (n,p) , (n,α) , and (n,d) ; and three-body breakup or sequential decay $(n,2n)$ and (n,np) . See NUCLEAR REACTION.

The resonances observed in these various reactions can be fitted by a theoretical formula to give parameters of the resonances (E_0 , Γ , Γ_f , Γ_γ , Γ_n , and so forth) which correspond to detailed properties of the excited states in the compound nucleus. For example, E_0 is the resonance energy; the fission width, Γ_f , is obtained from the fission cross section; the radiation width, Γ_γ , from the capture cross section; and so forth. The neutron width, Γ_n , can be obtained from the scattering cross section or the total cross section. The total width, Γ , can be obtained if the energy resolution is less than or equal to Γ . In addition, two other properties, the angular momentum and the spin, J , of the state can often be determined. For narrow resonances where Γ (in eV) $\leq 0.05 \sqrt{E_0}$ (in eV), it is necessary to consider the Doppler broadening of resonances due to the thermal motion of the target nuclides.

Neutron cross sections. The measurement of a cross section for a particular reaction consists of measuring the number of such reactions produced by a known number of neutrons incident on a known number of target nuclides. When the probability of all neutron interactions with the target nucleus is small, the number of reactions of a particular process i per unit area and unit time using a beam of neutrons equals $(nw)(Nx)\sigma_i$. The quantity nw is the number of incident neutrons per unit area normal to their direction per unit time, N is the number of target nuclei per unit volume, x is the thickness of the target in the direction of the incident neutrons, and σ_i is the cross section per target nucleus for a particular reaction expressed in units of area. If the probability of

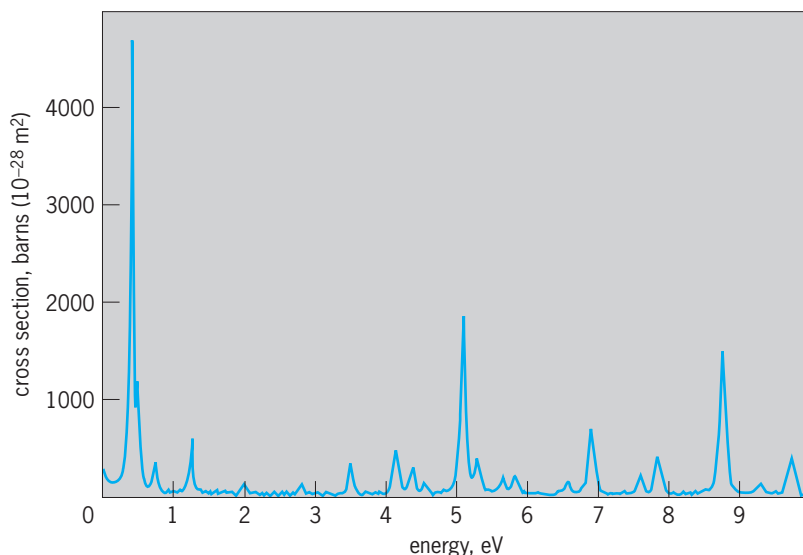


Fig. 1. Neutron total cross section of ${}^{231}\text{Pa}$ + neutron. The variation in sizes of the resonances and the nonuniform spacing of resonances are apparent.

all interactions is not small, the incident beam will be attenuated exponentially, as $\exp(-Nx\sigma_T)$, in passing through the sample, where Nx is the number of target nuclei per unit area normal to the beam and σ_T is the neutron total cross section.

The most common type of neutron cross-section measurement, which can usually be made with the

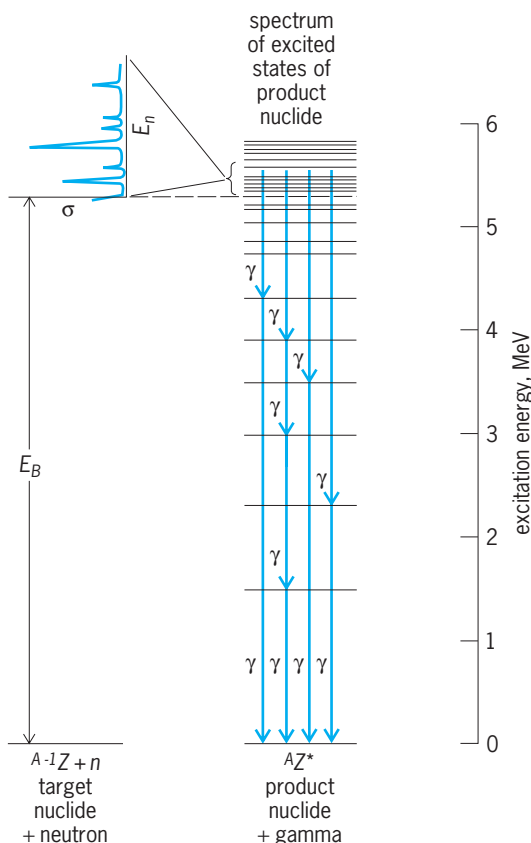


Fig. 2. Energy-level diagram for the product nucleus ${}^AZ^*$. The asterisk emphasizes that the product nucleus is in an excited state.

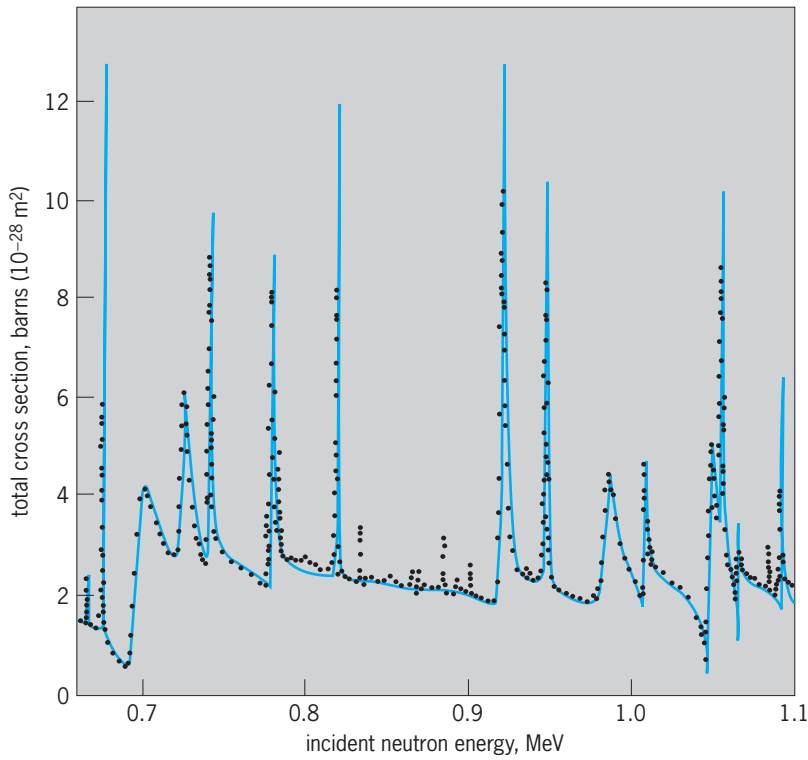


Fig. 3. Experimental neutron total cross section of sulfur compared to a theoretical fit. The fit does not include contributions from small resonances and minor isotopes of sulfur. The asymmetry of some resonances arises from the interference of resonance and potential scattering.

highest neutron energy resolution and usually with the most accuracy, is that of the total cross section. This measurement consists of measuring the transmission of a well-collimated beam of neutrons through a sample of known thickness; the transmission through the sample is simply the ratio of the intensity of the beam passing through the sample to that incident on the sample. The intensity of the incident beam is reduced in passing through the sample because the incident neutrons are absorbed or scattered by the target particles. The total cross section, σ_T , is determined from the equation $\sigma_T = -[\log_e(\text{transmission})]/Nx$.

In order to measure partial cross sections, more elaborate equipment is needed, in general, than for total cross-section measurements, and the measurements are considerably more difficult. For example, to measure the differential elastic scattering cross section, it is necessary to measure the number of elastically scattered neutrons as a function of angle of the scattered neutron relative to that of the incident neutron. MeV-energy neutrons, in addition to being elastically scattered, can also lose energy when scattered from a target nucleus [inelastic scattering, $(n, n' \gamma)$]. Several techniques have been developed for determining inelastic cross sections, both by measuring the energy spectrum of the inelastically scattered neutrons and by measuring the energies and intensities of the gamma rays emitted from the excited nuclei. These cross sections can also be measured as a function of the angle relative to the direction of the incident neutron.

Techniques for neutron spectroscopy. Neutron spectroscopy can be carried out by two different techniques (or a combination): (1) by the use of a time-pulsed neutron source which emits neutrons of many energies simultaneously, combined with the time-of-flight technique to measure the velocities of the neutrons; this time-of-flight technique can be used for neutron measurements from 10^3 eV to about 200 MeV; (2) by the use of a beam of nearly monoenergetic neutrons whose energy can be varied in small steps approximately equal to the energy spread of the neutron beam; however, useful “monoenergetic” neutron sources are not available from about 10 eV to about 10 keV.

Time-of-flight neutron spectrometers. Time-of-flight neutron spectrometers are the most widely used spectrometers for most neutron cross-section measurements. The time-of-flight technique requires an intense pulse of neutrons which contains neutrons of many energies and a flight path to measure the velocities of the neutrons. Various detectors are placed at the end of the flight path depending on the type of cross-section measurement. Burst widths from 10^{-5} to 10^{-9} s and flight paths from 3 to 3000 ft (1 to 1000 m) have been used. The resolution of a time-of-flight spectrometer is often quoted in microseconds or nanoseconds per meter. The energy resolution ($\Delta E/E$) is equal to $2\Delta t/t$, where Δt is the time width of the neutron burst plus the time spread in the detector, and t is the time of flight of the neutron [t (in microseconds) = $72.3 \times \text{path length (in meters)}/\sqrt{E}$ (in eV)]. The time between pulses, the flight path length, and filters in the beam must be selected so that low-energy neutrons from previous pulses do not interfere with the high-energy

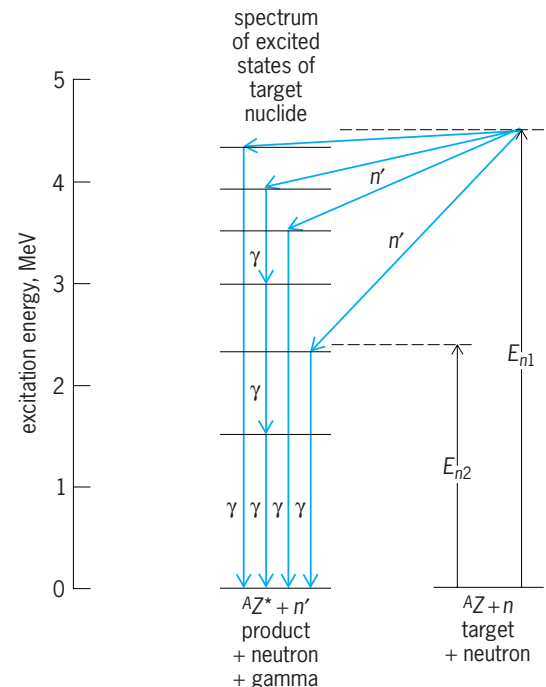


Fig. 4. Energy diagram of the target nuclide showing excitation of levels by the inelastic scattering of MeV neutrons.

neutrons from following pulses. By the use of multichannel storage, usually a computer, the complete neutron spectrum (or cross section) can be obtained in one measurement with good energy resolution over a broad energy range.

The most valuable neutron source for neutron time-of-flight spectroscopy is an electron or other charged-particle accelerator capable of producing intense pulses of neutrons of short duration (on the order of 10^{-9} s). Excellent neutron cross-section measurements can be made with these spectrometers from 0.01 eV to 200 MeV. For example, a beam of 140-MeV electrons incident on a tantalum target produces neutrons with an energy distribution that has a peak at about 1 MeV and extends up to about 80 MeV. The peak of the neutron distribution for protons or deuterons incident on a heavy target occurs at a higher neutron energy than for electrons; for deuterons a broad peak occurs at about half the energy of the deuterons. Lower-energy neutrons from these accelerators are obtained by placing a moderator (about 0.8 in. or 2 cm thick) around or near the target. The duration of these moderated neutron pulses in nanoseconds is approximately $2/\sqrt{E}$ (in MeV). The flux distribution of these moderated neutrons approximately follows the relation $E^{-0.8}$ down to thermal neutron energies. With a moderated neutron source the pulse repetition must be sufficiently low, depending on the flight path length, beam filter, and energy range, to prevent overlap of neutrons from successive pulses. Typical high-resolution results obtained using a moderated source are shown in Figs. 1 and 3. See PARTICLE ACCELERATOR.

Before the development of short-pulse accelerators, a mechanical chopper rotating at high speed in a well-collimated beam from a moderated fission reactor was used to produce bursts of electronvolt-energy neutrons. The neutron pulses were sufficiently short (about 10^{-6} s) and of sufficient intensity for measurements to be made up to a neutron energy of about 10^4 eV using neutron flight paths up to about 300 ft (100 m) in length. In order to produce pulses of only 10^{-6} s duration, the neutron beam had to be collimated to narrow slits (0.02 in. or 0.05 cm) to match the narrow slits through the chopper. Only when the rotation of the chopper was such that the slits in the rotor lined up with the slits in the collimator was a neutron beam with a broad energy spread passed. A fast-chopper time-of-flight spectrometer is particularly useful for transmission measurements on samples which are available in small amounts, since the sample only needs to be large enough to cover the beams passing through the narrow slits in the collimator.

For time-of-flight measurements in the energy region from about 10 keV to 1 MeV, a pulsed electrostatic accelerator using the Li^7 (p,n) reaction is capable of producing neutron pulses of short duration (10^{-9} s) with sufficient intensity for measurements with flight paths of a few meters. By selecting the proton-bombarding energy and a suitable target thickness, neutrons produced in the reaction

at a given angle can have well-defined upper and lower energy limits. With no low-energy neutrons and short flight paths, rather high repetition rates of 10^6 Hz can be used and an energy resolution of about 1% can be realized.

The most intense pulsed neutron source used for neutron time-of-flight spectroscopy is that achieved from an underground nuclear explosion. The burst duration is about 80 nanoseconds, and the neutron distribution extends down to about 20 eV. Fission cross-section measurements have been made on very small samples of many radioactive heavy nuclides using such a source and an approximately 980-ft (300-m) flight path length. The availability of this source is obviously rather restricted, but it is unique for measurements on highly radioactive samples.

Neutron time-of-flight measurements have also been made using a pulsed fission reactor where the duration of burst is about 40 μs . Finally, subcritical boosters have been used to multiply the intensity of the neutron pulses from electron accelerators by factors of 10–200, which results in pulse durations of 0.08–4 μs .

Monoenergetic neutron spectrometers. The best technique for obtaining an intense beam of low-energy neutrons (≤ 10 eV) with an energy spread of only about 1% is to use a crystal monochromator placed in a well-collimated beam of neutrons from a high-flux moderated fission reactor. If a single crystal (such as beryllium, copper, or lead) is properly oriented in a collimated neutron beam, neutrons of a discrete energy, E , will be elastically scattered from a particular set of planes of atoms in the crystal through an angle 2θ given by Bragg's law $n\lambda = 2d \sin \theta$. In this equation, the integer n is the order of the reflection, λ is the neutron wavelength, d is the spacing between the planes of atoms of the particular set in the crystal, and θ is the angle of incidence between the direction of the neutron beam and the set of planes of atoms being considered. The neutron wavelength λ in centimeters equals $0.286 \times 10^{-8} \sqrt{E}$, where E is in eV. The energy of the diffracted beam can be continuously varied by changing the angle of the crystal. Measurements of many rare-earth nuclides and heavy nuclides have been made with crystal spectrometers up to 10 eV neutron energy. Capture gamma-ray spectra have also been studied as a function of neutron energy, specifically from different neutron resonances. See NEUTRON DIFFRACTION.

"Monoenergetic" neutrons in the energy range from a few keV to 20 MeV can be obtained by bombarding various thin targets with protons or deuterons from a variable-energy accelerator such as an electrostatic accelerator. The most useful (p,n) reactions to cover the energy range from a few keV to a few MeV are those on lithium and tritium targets. The (d,n) reaction on deuterium is useful from about 1 to 10 MeV, and the (d,n) reaction on tritium from 10 to 20 MeV. In the energy range up to 1 MeV an energy resolution of about 1 keV is possible, but this resolution is usually not adequate for neutron spectroscopy for neutrons with energies less than 10^4 eV. The measurement of a complete cross-section

spectrum up to 1 MeV may require 1000 sequential measurements at slightly different neutron energies. Monoenergetic neutron sources are also useful for measurements such as activation, which cannot be done with the time-of-flight technique.

Applications. Neutron spectroscopy has yielded a mass of valuable information on nuclear systematics for almost all nuclides. The distribution of the spacings between nuclear levels and the average of these spacings have provided valuable tests for various nuclear theories. The properties of these levels, that is, the probabilities that they decay by neutron or gamma-ray emission, or by fission, and the averages and distribution of these probabilities have stimulated much theoretical effort.

In addition, knowledge of neutron cross sections is fundamental for the optimum design of thermal fission power reactors and fast neutron breeder reactors, as well as fusion power reactors now in the conceptual stage. Cross sections are needed for nuclear fuel materials such as ^{235}U or ^{239}Pu , for fertile materials such as ^{238}U , for structural materials such as iron and chromium, for coolants such as sodium, for moderators such as beryllium, for shielding materials such as concrete. The optimum choice of materials for the energy region under consideration is critical to the success of the project and is of great economic significance. See NUCLEAR STRUCTURE; REACTOR PHYSICS.

John A. Harvey

Bibliography. Y. A. Izyumov and N. A. Chernoplekov, *Neutron Spectroscopy*, 1994; S. F. Mughabghab, M. Diradeenam, and N. E. Holden (eds.), *Neutron Cross Sections*, vol. 1, pt. A 1981, pt. B 1984.

Neutron star

A star containing about $1\frac{1}{2}$ solar masses of material compressed into a volume approximately 6 mi (10 km) in radius. (1 solar mass equals 4.4×10^{30} lbm or 2.0×10^{30} kg.) Neutron stars are one of the end points of stellar evolution and are the final states of stars that begin their lives with considerably more mass than the Sun. The density of neutron star material is 10^{14} to 10^{15} times the density of water and exceeds the density of matter in the nuclei of atoms. Neutron stars are pulsars (pulsating radio sources) if they rotate sufficiently rapidly and have strong enough magnetic fields.

Neutron stars play a role in astrophysics which extends beyond their status as strange, unusual types of stellar bodies. The interior of a neutron star is a cosmic laboratory in which matter is compressed to densities which are found nowhere else in the universe. Precise measurements of the rotation of neutron stars can probe the behavior of matter at such densities. Neutron stars in double-star systems can emit x-rays when matter flows toward the neutron star, swirls around it, and heats up. Neutron stars are almost certainly formed in supernova explosions, events in which a dying star becomes more luminous than an entire galaxy, up to 10^{12} times as powerful as the Sun. A few pulsars are found in double-star systems, and careful timing of the pulses they emit

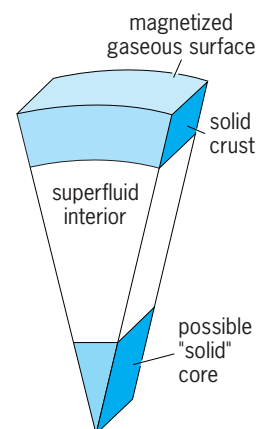
can test Einstein's general theory of relativity. See GRAVITATION; RELATIVITY.

Dimensions. The radius of a neutron star, while not precisely known, is about 6 mi (10 km). A few neutron stars exist in double-star systems, wherein it is possible to measure the strength of the gravitational pull of the neutron star on the other star in the system and hence determine the mass of each. The masses of these neutron stars are slightly more than the mass of the Sun, with the measured values ranging from 1.4 to 1.8 solar masses. See BINARY STAR.

The highest mass that a neutron star can have is about 2 solar masses. More massive neutron stars might possibly exist if current ideas about the behavior of neutron star matter turn out to be wrong. Objects that end their lives with masses higher than this neutron star mass limit will become black holes. See BLACK HOLE.

Internal structure. The state of matter beneath the surface of a neutron star is completely different from matter on the surface of the Earth and cannot be reproduced in terrestrial laboratories. The gaseous surface, only several feet thick, is different from ordinary gases, for the structure of atoms is dominated by the strong magnetic field that certainly exists in all pulsars and probably is present in all neutron stars. The magnetic field distorts the shape of the electron orbits which form the outer part of the atoms. Several feet below the surface, the material solidifies. The solid crust is about 0.6 mi (1 km) thick (see **illus.**) and is approximately 10^{17} times stiffer than steel. At greater depths, the densities increase, and a progressively greater fraction of the electrons are forced into atomic nuclei, where they combine with protons to form neutrons. See ATOMIC STRUCTURE AND SPECTRA.

Most of the interior of a neutron star consists of matter which is almost entirely composed of neutrons. In the bulk of the star, this matter is in a superfluid state, where circulation currents can flow without resistance. In rotating superfluids maintained in physics laboratories, small vortices form; such vortices may also exist in neutron stars. This material is under pressure, since it must be able to support the tremendous weight of the overlying layers at each



Cross section of a neutron star, with various layers.

point in the neutron star. This pressure, called degeneracy pressure, is caused by the close packing of the neutrons rather than by the motion of the particles. As a result, neutron stars can be stable no matter what the internal temperature is, because the pressure that supports the star is independent of temperature. *See* SUPERFLUIDITY.

The central regions of neutron stars are poorly understood. For example, the role played by subnuclear particles such as the *K* meson, pi meson, or other hyperons is unknown. It is remotely possible that quark matter, material composed of the postulated fundamental particles of matter, exists in the cores of neutron stars. Claims of some astronomers to have actually detected strange quark stars have not generally been accepted. *See* DELTA RESONANCE; ELEMENTARY PARTICLE; NUCLEAR PHYSICS; QUARKS.

Observations. Theoretical calculations form the basis for the above description of a neutron star. A variety of astronomical observations may eventually confirm or disprove various aspects of this model. The pulses from a pulsar (a rotating neutron star) can be timed very precisely and can suggest the way that neutron stars change their shape as they spin more slowly. A single, isolated neutron star loses energy in the form of high-speed particles emitted from its surface, and its rotation rate slows down. A newly formed pulsar spins very fast—the pulsar in the Crab Nebula, formed a little less than 1000 years ago as observed from Earth, rotates 30 times every second. Eventually it will rotate more slowly, and its structure will readjust, occasionally producing abrupt changes in the rotation period which are called glitches. *See* CRAB NEBULA.

Telescopes launched above the atmosphere have detected x-rays and extreme ultraviolet radiation from the hot surfaces of a number of neutron stars. In some cases, these neutron stars are extremely young and hot, the product of recent supernova explosions. Unfortunately, it is not yet possible to clearly determine the mass and radius of a neutron star and compare those results to observations, as is the case for white dwarf stars. *See* SATELLITE (ASTRONOMY); ULTRAVIOLET ASTRONOMY; WHITE DWARF STAR; X-RAY ASTRONOMY.

Some neutron stars exist in double-star systems, where matter flows from one star to the neutron star, forms a whirling disk around it, and causes it to rotate faster as it gains more mass. These neutron stars emit pulses of x-rays, which can be detected and timed by satellite observatories orbiting above the Earth's atmosphere. Sudden and irregular changes in the arrival times of these x-ray pulses show that solid matter exists somewhere in the interior of neutron stars, but the data are sufficiently poor that it is not yet possible to test detailed models of neutron stars.

In the late 1980s, the *EXOSAT* satellite, launched by the European Space Agency and containing an x-ray telescope, detected what were described as quasiperiodic oscillations from a number of x-ray-emitting double stars. The x-ray intensity increased and decreased with a reasonably well-defined period, though the frequency of these oscillations changed

from one observation time to another. It is possible that these oscillations can indicate something about the neutron star itself, though it is more likely that they originate from the disk surrounding the neutron star and that analysis of them will provide information on accretion disk physics rather than neutron star physics.

Beginning in 1979, space satellites, some originally intended to monitor nuclear bomb tests, detected sudden blasts of gamma rays coming from space. Their origin was a mystery. Many explanations were proposed, but none gained support. In the 1990s, new data led to the widespread acceptance of the "magnetar" explanation. Magnetars are neutron stars with fields of 10^{10} to 10^{11} tesla (10^{14} to 10^{15} gauss), fields 10 to 100 times stronger than the fields of normal neutron stars. *See* GAMMA-RAY BURSTS.

Some neutron stars rotate so fast that their pulse periods are milliseconds, rather than the pulse periods of typical pulsars which are roughly seconds. Someone listening to the pulses from these millisecond pulsars would hear a musical tone rather than a series of drumbeats (which would be heard from the conventional pulsars). The current interpretation of these rapid rotation rates is that most millisecond pulsars are in double star systems, and have been spun up to high speeds because they accrete matter. One of these ultrafast pulsars was the first object outside the solar system to have a planet detected around it. *See* PLANET.

Origin. In the 1930s, Fritz Zwicky first suggested that isolated neutron stars originate in supernova explosions. Stars more massive than approximately 10 times the mass of the Sun end their lives in a tremendous explosion. It is now clear that in many cases the remnant of this explosion is a tiny neutron star, whose formation provided the energy for the supernova. The discovery of a pulsar in the Crab Nebula, a known supernova remnant, strongly supported the connection between neutron stars and supernovae. The 1987 supernova explosion in the Large Magellanic Cloud both confirmed the genesis of neutron stars in supernovae and provided some important new details in the story. When the light from this supernova explosion arrived at the Earth, so did a burst of subnuclear particles called neutrinos. These neutrinos were observed by underground detectors located in the United States and Japan. The detection of these supernova neutrinos firmly cemented the link between supernovae and neutron stars, and also is helping astronomers understand how the formation of a neutron star in the center of a massive star is connected to these celestial explosions. *See* GRAVITATIONAL COLLAPSE; NEUTRINO; PULSAR; STELLAR EVOLUTION; SUPERNOVA. Harry L. Shipman

Bibliography. R. Kippenhahn, *100 Billion Suns*, trans. by J. Steinberg, 1983, reprint 1993; R. P. Kirshner, *The Extravagant Universe: Exploding Stars, Dark Energy, and the Accelerating Cosmos*, 2002; L. Marschall, *The Supernova Story*, 1988, paper 1994; S. L. Shapiro and S. A. Teukolsky, *Black Holes, White Dwarfs, and Neutron Stars*, 1983; D. J. Yakovlev and C. J. Pethick, Neutron star cooling, *Ann. Rev. Astron. Astrophys.*, 42:159–210, 2004.

New Zealand

A landmass situated in the Southern Hemisphere, bounded by the South Pacific Ocean to the north, east, and south and the Tasman Sea to the west, with a total land area of 103,883 mi² (269,057 km²). The exposed landmass represents about one-quarter of a subcontinent, with three-quarters submerged. This long, narrow, mountainous country, oriented northeast to southwest, consists of two main islands, surrounded by a much greater area of crust submerged to depths reaching 1.2 mi (2 km) [Fig. 1].

Geology. The major modern geological structural features of the subcontinent have developed over the last 25 million years, during which the New Zealand region formed from interactions along the boundary between the Australia and Pacific crustal plates. Off the east coast of North Island, the Pacific plate of oceanic crust is subducting (sinking) beneath that edge of the Australia plate that forms North Island.

The results of subduction include earthquakes at progressively greater depths westward, folding and faulting of rocks beneath the eastern continental shelf and ranges, and volcanic activity along an arc that extends from Mount Ruapehu 9174 ft (2797 m) through the Taupo and Rotorua regions to White Island (Fig. 2). See EARTHQUAKE; FAULT AND FAULT STRUCTURES.

The plate boundary subduction zone terminates in the northeast of South Island, from where the boundary extends to the southwest as the Alpine Fault, along which plate motion is in the form of lateral shearing. The Alpine Fault extends offshore north of Fiordland, where plate convergence results in subduction of the edge of the Australia plate beneath the New Zealand subcontinent, and formation of a southern seismic zone. See SUBDUCTION ZONES.

Plate boundary processes are responsible for the crustal deformations that have resulted in still-active uplift of the Southern Alps of South Island with

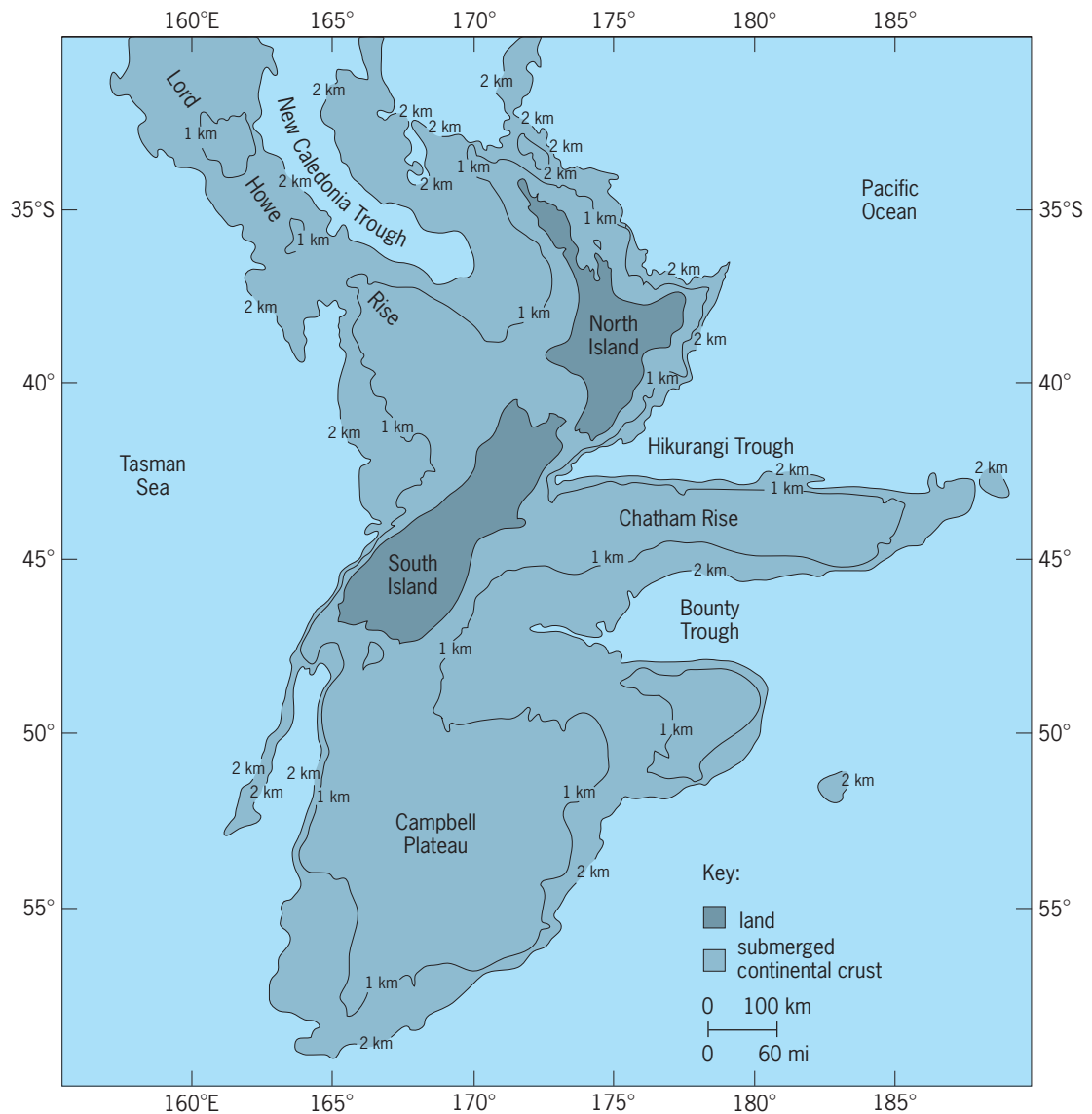


Fig. 1. New Zealand subcontinent and its exposed landmass. Contours represent water depth. (After I. Wards, ed., *New Zealand Atlas*, Government Printer, 1976)

attendant erosion of the young cover rocks and exposure of older basement rocks of graywacke (sandstone and argillite) and schist. In contrast, the subduction of the Pacific plate below eastern North Island has produced uplift and inversion of the overlying sedimentary basins containing mudrocks and limestones. North Island rocks are therefore of three groups: Cretaceous to Cenozoic age sedimentaries, Cenozoic volcanics, and exposed basement graywackes in the zones of greatest uplift and highest relief in the ranges extending northward from Wellington to eastern Bay of Plenty. See GRAYWACKE; LIMESTONE; SANDSTONE; SCHIST.

The relief of New Zealand has been created by tectonic forces. Two-thirds of the country has an elevation between 650 and 3600 ft (200 and 1100 m), about one-sixth lies below 660 ft (200 m), and the remaining sixth is above 3600 ft (1100 m). The highest mountains in South Island reach over 8200 ft (2500 m), and 18 peaks reach over 9840 ft (3000 m); Mount Cook, the highest, is 12,346 ft (3764 m). The alpine relief is rugged, with deep glaciated valleys; and from Milford Sound northeastward to Arthur's Pass, there are active high-basin glaciers, and around Mount Cook valley glaciers that form the basis of a tourism and alpine sports industry. See OROGENY; PLATE TECTONICS.

Soils. South Island lowlands are either alluvial plains as in Otago, Southland, and Nelson, or glacial outwash fans as in Westland and Canterbury. North Island lowlands such as Hawke's Bay, Wairarapa, and Manawatu are alluvial; the Waikato, Hauraki, and Bay of Plenty lowlands occupy structural basins that contain large volumes of reworked volcanic debris from the central volcanic region. The alluvial lowlands of both main islands form the most agriculturally productive areas of the country. See PLAINS.

Lowland soils in South Island may develop in alluvium, or loess derived from dry river beds, and in North Island in alluvium or beds or volcanic ash and loess. Hill-country soils have, generally, low fertility and shallow depths. Soils of the lower hills and plateaus form the basis of much pastoral farming and, increasingly, of plantation forests; these soils are greatly influenced in their characteristics by rainfall regimes and the original natural vegetation. See LOESS; SOIL.

Climate. The climate of New Zealand is influenced by three main factors: a location in latitudes where the prevailing airflow is westerly; an oceanic environment; and the mountain chains, which modify the weather systems as they pass eastward, causing high rainfalls on windward slopes and sheltering effects to leeward.

Weather is determined mostly by series of anticyclones and troughs of low pressure that produce alternating periods of settled and variable conditions. Westerly air masses are occasionally replaced by southerly airstreams, which bring cold conditions with snow in winter and spring to areas south of 39° S, and northerly tropical maritime air, which brings warm humid weather to the north and east coasts. See MARITIME METEOROLOGY.

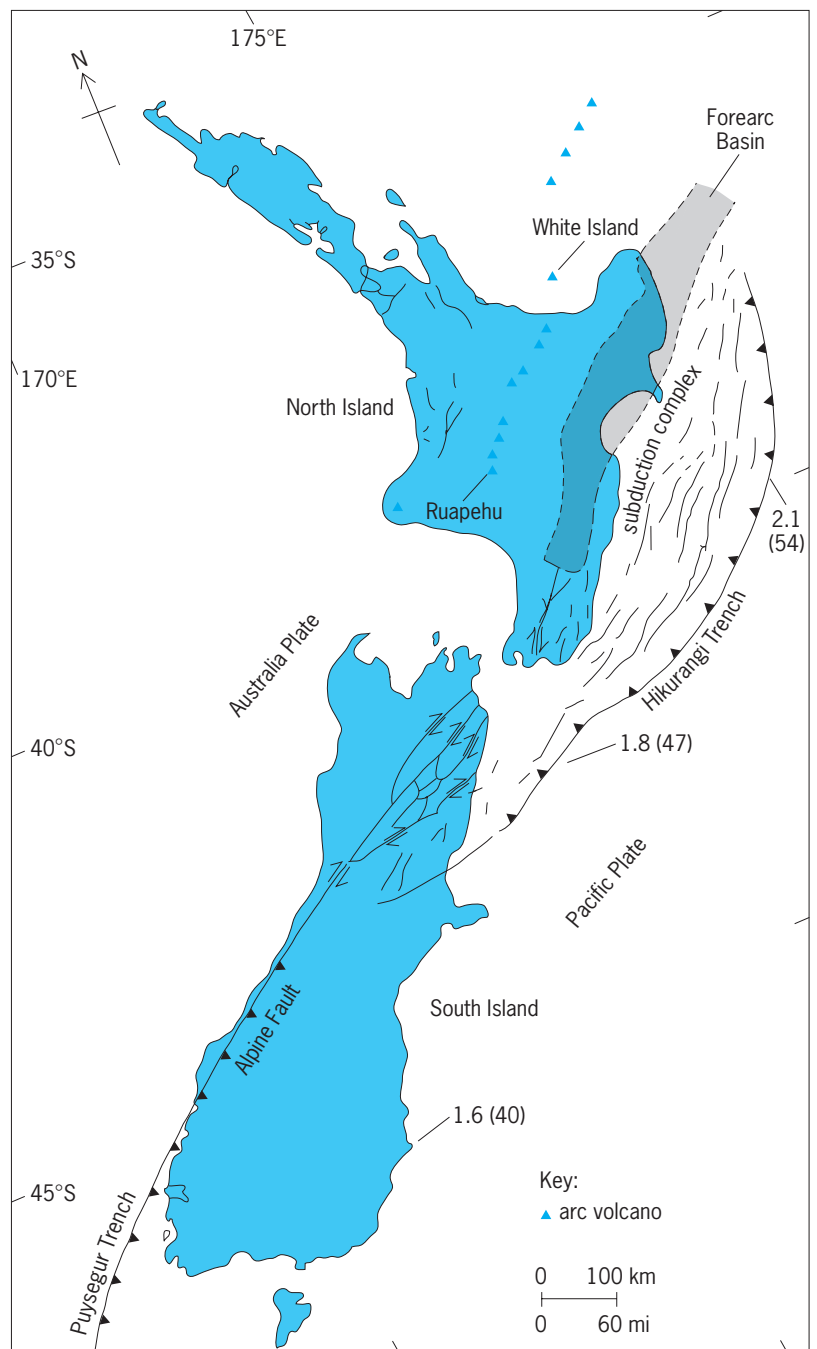


Fig. 2. Some structural features of New Zealand, including the boundary between the Australia and Pacific plates, the Alpine Fault, the line of arc volcanoes, and the zone of inverted sedimentary basins (Forearc Basin) of eastern North Island. The numerical values are rates of convergence of the crustal plates in inches (millimeters) per year. (After P. J. J. Kamp, *Tectonic architecture of New Zealand*, in J. M. Soons and M. J. Selby, *Landforms of New Zealand*, 2d ed., Longman Paul, 1992)

Rainfall over the open ocean is commonly in the range of 25–30 in. (600–800 mm) per year, but on land is 16–470 in. (400–12,000 mm) per year (Fig. 3), with the highest rainfalls being on the western windward slopes of the mountains, and the lowest on the eastern basins in the lee of the Southern Alps in Central Otago and south Canterbury. Annual rain days are at least 130 for most of North Island, but the South Island the totals are far more variable, with over 200

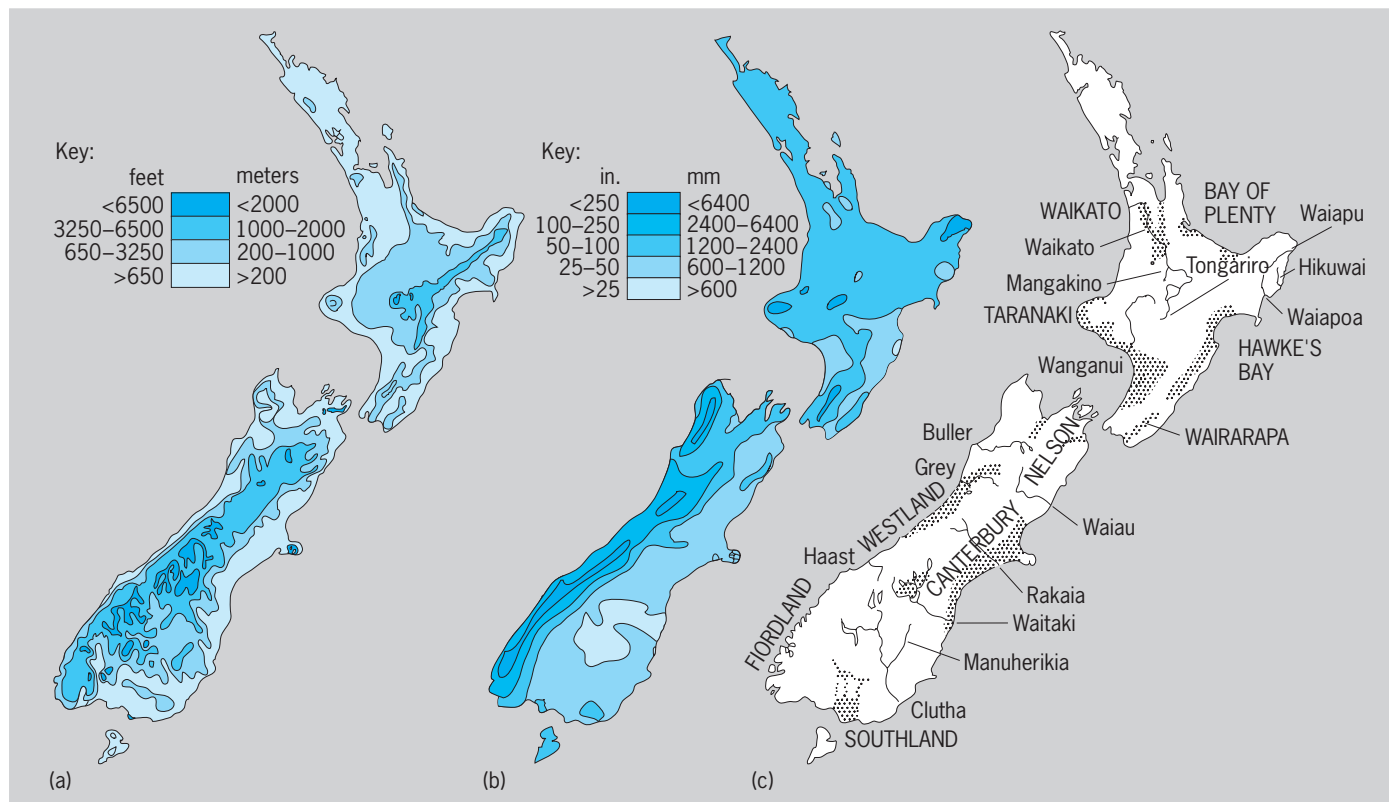


Fig. 3. Maps showing (a) relief, (b) annual rainfall, and (c) main rivers, and sedimentary lowlands of New Zealand. Note the close relationship between relief and rainfall. (After I. Wards, ed., *New Zealand Atlas*, Government Printer, 1976)

occurring in Fiordland, 180 on the west coast, and fewer than 80 in Central Otago. Summer droughts are relatively common in Northland, and in eastern regions of both islands. See DROUGHT; PRECIPITATION (METEOROLOGY).

Mean annual temperatures decrease progressively from north to south: 59°F (15°C) in Northland, 55°F (13°C) in Wellington, 50°F (10°C) in Southland; some inland parts of South Island have averages of 46°F (8°C). January and February are the warmest months and July the coldest. Highest summer temperature exceed 86°F (30°C) in Central Otago, lowest temperatures are rarely below 14°F (−10°C) except in the mountains. Average annual duration of bright sunshine is greatest on the northern lowlands of South Island, where it exceeds 2400 h around Nelson and Blenheim, and on the eastern lowlands of North Island and Bay of Plenty.

Droughts, springtime air frosts, and hailstorms are the major common climatic hazards for the farming industry, but floods associated with prolonged intense rainstorms are the major general hazard.

Rivers and lakes. The rivers of New Zealand have a total discharge of about 70 mi³ (300 km³) per year. Some 12 mi³ (50 km³) of water is stored in perennial snow and glacier ice. Flow variability and specific yield (flow per unit area of catchment) is very high by comparison with most rivers in the world. The Clutha River, with a mean flow of 20,000 ft³ (570 m³)/s is the largest river in New Zealand, and

the Waikato is the largest in North Island. The highest specific discharges are those of rivers draining westward from the Southern Alps, where steep mountainous catchments, lack of lake storage, exposed rock, thin soils, and very high rainfall produce very large flood flows (Table 1); the Grey River has a specific discharge that is three times higher than those of the two longer North Island rivers, Waikato and Wanganui, which have larger catchments but receive lower rainfall.

In an average year, New Zealand's rivers carry 4.4×10^6 tons of sediment from the land to the ocean. Two major factors account for the extraordinarily high suspended sediment yields of many rivers (Table 2): amount and intensity of rainfall, and catchment geology with land use. These factors are illustrated by the yields for the Cropp River with its mountainous catchment and very high rainfall, and the Waiapu River with erodible sandstone and mudstone, compared with the low yields from the Manuherikia with low rainfall but erodible rock, and the Mangakino with moderate rainfall and very permeable rock. See RIVER.

Seven hundred seventy-six lakes, with lengths exceeding 0.3 mi (0.5 km), have been cataloged in New Zealand; of these 291 were formed by glacial action and occur in South Island, while 30 are of volcanic origin and occur in the central North Island. The other major categories are dune, riverine, and artificial impoundments. Because of the importance of

TABLE 1. Major rivers ranked by water discharge*

| River | Mean annual discharge | | Catchment area | | Specific discharge | | Length | |
|----------|-----------------------|-------------------|-----------------|-----------------|--|---------------------------------------|--------|-----|
| | ft ³ /s | m ³ /s | mi ² | km ² | ft ³ /(s)(mi ²) | m ³ /(s)(km ²) | mi | km |
| Clutha | 20,128 | 570 | 7,947 | 20,582 | 3 | 0.03 | 200 | 322 |
| Waiau | 15,432 | 437 | 3,141 | 8,134 | 5 | 0.05 | 105 | 169 |
| Buller | 15,114 | 428 | 2,452 | 6,350 | 7 | 0.07 | 110 | 177 |
| Waitaki | 12,960 | 367 | 3,768 | 9,760 | 4 | 0.04 | 130 | 209 |
| Grey | 11,900 | 337 | 1,479 | 3,830 | 9 | 0.09 | 75 | 120 |
| Waikato | 11,547 | 327 | 4,400 | 11,395 | 3 | 0.03 | 264 | 425 |
| Wanganui | 7,910 | 224 | 2,565 | 6,643 | 3 | 0.03 | 181 | 290 |

*After M. Mosley (ed.), *Waters of New Zealand*, 1992.

tourism and hydroelectricity, these lakes not only are significant features of the landscape but also are of major economic importance. See LAKE.

Flora and fauna. New Zealand gained its earliest flora and fauna from Gondwanaland before the separation that occurred in the Cretaceous Period. Consequently its plants and animals are highly endemic—over 80% for the flowering plants and well over 90% for the arthropods. Separation, as well as the opening of the Tasman Sea, was early enough to exclude snakes and marsupial mammals but late enough to include podocarps, tree ferns and other ferns, southern beech (*Nothofagus* sp.), and conifers related to the modern kauri (*Agathis australis*). The only terrestrial reptiles in the New Zealand fossil record are the tuatara (*Sphenodon*), which still survives on small offshore islands, and many species of gecko and skink; the only indigenous mammals are two species of bat. See CRETACEOUS.

The modern bird population is extremely varied, including many migratory species, some Australian recent settlers, and many European species introduced since the mid-nineteenth century. Native species had no predators and consequently there were many flightless, or virtually flightless, as well as flying species. But since the arrival of the first Polynesian settlers about 1000 years ago, and the subsequent early nineteenth century Europeans, many species of flightless birds have become extinct as a result of hunting, destruction of habitat, predation

by the Polynesian dog and rat (*Rattus exulans* or *Kiore*) and then the European dogs and ships' rat (*R. norvegicus*), and by mustelids. Both Polynesian forest burning and the later European destruction of forest have had catastrophic consequences for native birds, including all 24 species of moa (Dinornithiformes) and even the kiwi—the national symbol—is under threat because of loss of habitat and predation by dogs. See RATITIS.

The natural vegetation of New Zealand below the tree line is mostly evergreen forest, with kauri-podocarp (rimu, miro, totara)—hardwood (rewarewa, tawari) north of latitude 38°S; podocarp-hardwood in central North Island; podocarp-hardwood-beech in southern North Island and the wetter parts of South Island (with variations of species with soils and climate); and beech forests on poorer soils, and cooler and steeper sites of higher altitudes on both main islands.

Forest clearing since about 1840 has removed much of the 54,040 mi² (14 million hectares) of forest that was present at that date. Forests cover about 28%, or 29,680 mi² (8 million hectares), of New Zealand's land area; of this, about 23,900 mi² (6 million hectares) are in natural forest, and 5780 mi² (1.5 million hectares) in plantations of introduced species. Of the plantations, 89% is *Pinus radiata*; only 1% is in hardwood species. The *P. radiata* is the basis of construction timber and of paper and board manufacturing.

TABLE 2. Selected rivers ranked by suspended sediment yield from their catchments*

| River | Suspended sediment yield | | Catchment area | | Precipitation | | Catchment rock type |
|-------------|----------------------------|-----------------------------------|-----------------|-----------------|---------------|--------|-------------------------|
| | tons/(mi ²)(y) | metric tons/(km ²)(y) | mi ² | km ² | in./y | mm/y | |
| Cropp | 83,768 | 29,600 | 11 | 29 | 393 | 10,070 | Schist |
| Waipua | 56,515 | 19,970 | 532 | 1,378 | 94 | 2,400 | Sandstone and mudstone |
| Hikuwai | 39,308 | 13,890 | 119 | 307 | 74 | 1,900 | Sandstone and mudstone |
| Haast | 36,042 | 12,736 | 394 | 1,020 | 253 | 6,500 | Schist |
| Waipaoa | 16,515 | 5,836 | 611 | 1,582 | 62 | 1,600 | Sandstone and siltstone |
| Rakaia | 4,644 | 1,641 | 1,019 | 2,640 | 117 | 3,000 | Graywacke and argillite |
| Tongariro | 1,570 | 555 | 298 | 772 | 78 | 2,000 | Graywacke and andesite |
| Grey | 1,562 | 552 | 248 | 642 | 117 | 3,000 | Granite |
| Wanganui | 922 | 326 | 2,565 | 6,643 | 70 | 1,800 | Sandstone and siltstone |
| Manuherikia | 99 | 35 | 786 | 2,036 | 32 | 830 | Schist |
| Mangakino | 99 | 35 | 144 | 373 | 55 | 1,400 | Igimbrite |

*After M. P. Mosley (ed.), *Waters of New Zealand*, 1992.

Natural resources and the economy. The New Zealand economy is heavily dependent on the natural resources soil, water, and plants. It has few exploitable minerals, but it does possess a climate generally favorable for agriculture, pastoral farming, renewable forestry, and tourism. With a relatively small population (3.4 million people), much of its manufacturing is concerned with processing produce from the land and surrounding seas, and supplying the needs of those industries.

Natural hazards. Because of its high relief and its location on an active crustal plate boundary in the zone of convergence between Antarctic air masses and tropical air masses, New Zealand is prone to high-intensity and high-frequency natural hazards—earthquakes, volcanic eruptions, large and small landslides, and floods.

Earthquakes are felt throughout the country, but they are most common and most severe in a central seismic zone that includes the northern half of South Island and the southern half of North Island. Volcanic eruptions are of limited magnitude and are confined to the central North Island volcanic zone, where Ngauruhoe, Ruapehu, and White Island are mildly active; in 1886, Tarawera erupted more violently. Of far greater potential power and hazard are ignimbrite-producing eruptions from calderas of the central North Island, the last of which occurred 1900 years ago from the Taupo caldera. *See* CALDERA; VOLCANO.

Very large landslides in the Southern Alps and in the eastern areas of Marlborough (South Island) and North Island have tectonic origins due to land tilting and earthquakes in highly sheared or weak rocks. Such landslides occur infrequently, three to four per thousand years, but each may involve $3\text{--}500 \times 10^6 \text{ m}^3$ ($1\text{--}1.8 \times 10^{10} \text{ ft}^3$) of rock. Climatically induced landslides are far more frequent, but involve only a few tens of cubic meters of regolith material in each failure. They may, however devastate some tens or hundreds of hectares of land in a single storm. Such failures are most common on the North Island and northeastern South Island hills. *See* LANDSLIDE.

Since the end of the nineteenth century, natural hazards have been limited in their impact on New Zealand by the wide dispersal of the small human population and, with the exception of the Napier earthquake of 1931, the relative low intensity, and location, of the largest landslides and volcanic eruptions. Fewer than 500 people have been killed in all extreme events of the last century. M. J. Selby

Bibliography. M. Gage, *Legends in the Rocks: An Outline of New Zealand Geology*, 1980; G. Kuschel (ed.), *Biogeography and Ecology in New Zealand*, 1975; L. Molloy, *Soils in the New Zealand Landscape: The Living Mantle*, 1988; M. P. Mosley (ed.), *Waters of New Zealand*, 1992; New Zealand Department of Statistics, *New Zealand Official Yearbook*, annually; J. M. Soons and M. J. Selby (eds.), *Landforms of New Zealand*, 2d ed., 1992; I. Wards (ed.), *New Zealand Atlas*, 1976.

Newcastle disease

A viral infection that affects the digestive, intestinal, and respiratory tracts and the neurological system of birds. The causative agent is an enveloped ribonucleic acid (RNA) virus that is classified as a paramyxovirus. *See* PARAMYXOVIRUS.

Newcastle disease occurs in five forms based on a virulence in chickens ranging from inapparent infection to severe disease and death. Viscerotropic-velogenic Newcastle disease causes a very severe infection, producing hemorrhagic lesions in the intestinal tract and high mortality. The neurotropic-velogenic type is also highly lethal and produces neurologic and respiratory signs in infected birds. The mesogenic form causes an acute respiratory or neurologic infection that may be lethal only in young birds. The lentogenic type is a mild or inapparent respiratory infection of chickens. The last group includes the viruses causing inapparent or asymptomatic infections of the digestive tract.

The wide susceptibility of avian species to infection with Newcastle disease has complicated control. Newcastle disease is spread worldwide by the international transportation of live birds disseminating the virus. In many countries, the velogenic form has been eradicated in poultry by quarantine and slaughter. More difficult to control are reservoirs of natural infection in wild free-flying birds. Waterfowl (ducks and geese) are naturally infected with the avirulent intestinal strain of the virus. Newcastle disease has been identified in racing pigeons and has spread worldwide through racing activity and the extensive trade in these birds. Another important reservoir is exotic pet birds, particularly those from tropical regions where the virulent virus is endemic. These birds are more resistant and can be asymptotically infected with the virus virulent for poultry. Countries free of velogenic Newcastle disease impose quarantine and testing of imported birds to prevent release of infected birds.

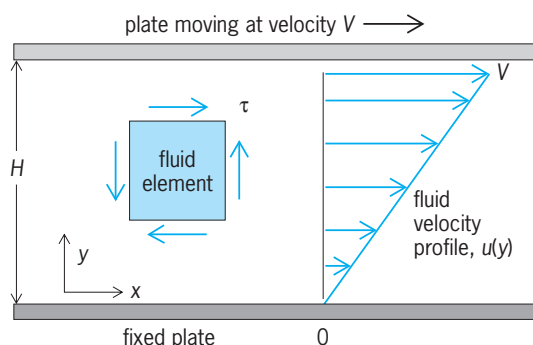
Control of and protection from Newcastle disease can be achieved by the correct use of vaccines. Lentogenic and some mesogenic strains are used to produce vaccines that can be administered by aerosol, intranasal drops, or intramuscular injection, or as an additive to the drinking water. *See* ANIMAL VIRUS.

Mary Lynne Vickers

Bibliography. D. J. Alexander (ed.), *Newcastle Disease: Developments in Veterinary Virology*, 1988; B. W. Calnek (ed.), *Diseases of Poultry*, 10th ed., 1997.

Newtonian fluid

A fluid whose stress at each point is linearly proportional to its strain rate at that point. The concept was first deduced by Isaac Newton and is directly analogous to Hooke's law for a solid. All gases are newtonian, as are most common liquids such as water, hydrocarbons, and oils. *See* HOOKE'S LAW; STRESS AND STRAIN.



A fluid sheared between two plates. The resulting strain rate equals V/H .

A simple example, often used for measuring fluid deformation properties, is the steady one-dimensional flow $u(y)$ between a fixed and a moving wall (see **illus.**). The no-slip condition at each wall forces the fluid into a uniform shear strain rate ϵ , given by Eq. (1), which is induced by a uniform shear

$$\epsilon = \frac{\partial u}{\partial y} = \frac{V}{H} \quad (1)$$

stress τ . Here V is the speed of the moving wall, H is the perpendicular distance between the walls, and u is the fluid velocity at distance y from the fixed wall.

If the fluid is newtonian, the experimental plot of τ versus ϵ will be a straight line. The constant of proportionality is called the viscosity μ of the fluid, as stated in Eq. (2). For consistency, the dimensions of

$$\tau = \mu \epsilon \quad (2)$$

μ must be (mass)/(length)(time), or $\text{kg}/(\text{m} \cdot \text{s})$. The viscosity coefficients of common fluids vary by several orders of magnitude (see **table**).

When flowing near a solid boundary, all viscous fluids, except for extremely rarefied gases, stick to that surface and move at the velocity of the surface (see **illus.**). This is the no-slip condition, used in formulating mathematical models of fluid flow.

If the plot of τ versus ϵ forms a curved line, the fluid is termed non-newtonian and can exhibit more complex effects. See NON-NEWTONIAN FLUID.

The simple one-dimensional relation, Eq. (2), can easily be extended to fully three-dimensional unsteady viscous flows. When the result is combined with the momentum principle, the fundamental equation for newtonian fluid flow emerges. See BOUNDARY-LAYER FLOW; FLUID-FLOW PRINCIPLES; NAVIER-STOKES EQUATION.

Viscosities of common fluids

| Fluid | Viscosity (μ), $\text{kg}/(\text{m} \cdot \text{s})^*$ |
|------------|--|
| Air | 1.80×10^{-5} |
| Methanol | 5.98×10^{-4} |
| Water | 0.001 |
| Mercury | 0.00156 |
| SAE 10 oil | 0.104 |

* At 20°C (68°F) and 1 atmosphere (101.3 kilopascals).

The viscosity μ of a fluid is a true thermodynamic property and can be measured or tabulated for each substance. Generally, μ for liquids decreases sharply with temperature and increases moderately with pressure. The viscosity of gases increases moderately with temperature and increases slightly with pressure. See FLUIDS; VISCOSITY.

Frank M. White
Bibliography: F. S. Sherman, *Viscous Flow*, 1990;
E. M. White, *Viscous Fluid Flow*, 2d ed., 1991.

Newton's laws of motion

Three fundamental principles which form the basis of classical, or newtonian, mechanics. They are stated as follows:

First law: A particle not subjected to external forces remains at rest or moves with constant speed in a straight line.

Second law: The acceleration of a particle is directly proportional to the resultant external force acting on the particle and is inversely proportional to the mass of the particle.

Third law: If two particles interact, the force that is exerted by the first particle on the second particle (called the action force) is equal in magnitude and opposite in direction to the force that is exerted by the second particle on the first particle (called the reaction force).

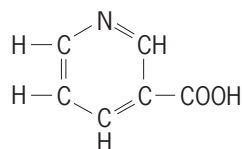
The first law, sometimes called Galileo's law of inertia, can now be regarded as contained in the second. At the time of its enunciation, however, it was important as a negation of the aristotelian doctrines of natural placement and continuing force.

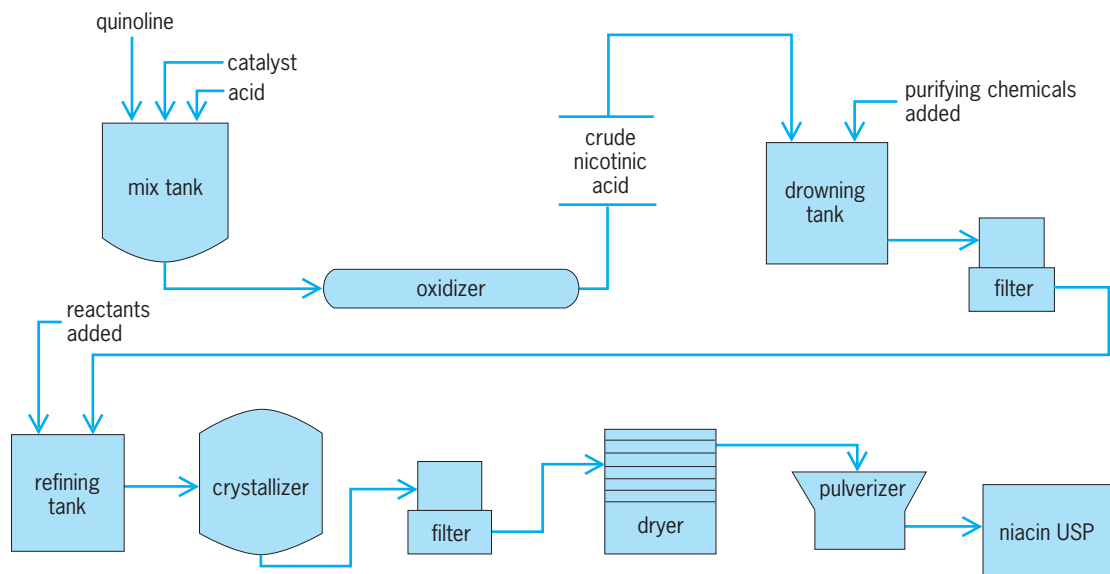
The third law, sometimes called the law of action and reaction, was also to some extent established prior to Newton's statement of it. However, Newton's formulation of the three laws as a mutually consistent set, with the nature of force clearly defined in the second law, provided the basis for classical dynamics.

The newtonian laws have proved valid for all mechanical problems not involving speeds comparable with the speed of light (approximately 186,000 mi/s or 300,000 km/s) and not involving atomic or subatomic atomic particles. The more general classical methods of Lagrange and of Hamilton are elaborations of the newtonian principles. See DYNAMICS; FORCE; HAMILTON'S EQUATIONS OF MOTION; LAGRANGE'S EQUATIONS; MOTION. Dudley Williams

Niacin

A vitamin also known as nicotinic acid. It is a white water-soluble powder stable to heat, acid, and alkali, with the structure shown below. It is found





Flow sheet for the manufacture of niacin from quinoline.

in biochemically active combinations as the amide, niacinamide. Analyses for niacin are usually done microbiologically using *Lactobacillus arabinosus* as the test organism. Chemical methods of analysis are not usually satisfactory. All living cells studied have enzymic systems involving niacin. Many animals, including humans, are capable of synthesizing niacin in varying degrees from the amino acid tryptophan. Niacin is widely distributed in foods. Yeasts, wheat germ, and meats, particularly organ meats, are rich sources of the vitamin. Some foods such as milk are relatively poor sources of niacin but contain generous quantities of tryptophan.

Niacin-deficiency disease is known as pellagra and is particularly prevalent among people whose diet is largely corn. Pellagra is characterized by dermatitis, dementia, diarrhea, and death. Skin lesions are usually observed in areas exposed to the sun. The disease is accompanied by gastrointestinal lesions.

Niacin is present in enzymes in the form of two coenzymes, diphosphopyridine nucleotide (DPN), also known as nicotinamide adenine dinucleotide (NAD) or coenzyme I, and triphosphopyridine nucleotide (TPN), also known as nicotinamide adenine dinucleotide phosphate (NADP) or coenzyme II. Enzymes containing DPN or TPN function in oxidation-reduction systems by virtue of their ability to accept hydrogen ions (protons) and electrons from substrates and transfer them to other hydrogen acceptors, such as the flavo-proteins. Niacin-containing enzymes catalyze about 40 reversible biochemical reactions, many of different types, as illustrated by the following: acetaldehyde \rightleftharpoons ethanol; 1,3-diphosphoglyceric acid \rightleftharpoons 3-phosphoglyceric acid; pyruvic acid \rightleftharpoons lactic acid; imino glutamic acid \rightleftharpoons L-glutamic acid; and acetic acid \rightleftharpoons acetaldehyde. See BIOLOGICAL OXIDATION; COENZYME; RIBOFLAVIN.

Unlike the other B vitamins, little niacin is excreted in the urine. Humans excrete niacin mostly

as *N*¹-methylnicotinamide and the 6-pyridone of *N*¹-methylnicotinamide. The pellagra-preventive potency of a diet is related not only to its niacin and tryptophan content but also to the availability of its niacin. There is evidence that some of the niacin of foods cannot be released by digestive enzymes. The existence of an antiniacin material in corn has been suggested. The effect of the carbohydrate content of the diet on the synthesis of niacin by intestinal bacterial may also be important. The use of urinary excretion data to determine nutritional status with regard to niacin has been disappointing. In humans an average of 60 mg of dietary tryptophan is equivalent to 1 mg of niacin. The recommended dietary allowance of the National Research Council is 6.6 niacin equivalents per 1000 kcal. See NICOTINAMIDE ADENINE DINUCLEOTIDE (NAD); NICOTINAMIDE ADENINE DINUCLEOTIDE PHOSPHATE (NADP); VITAMIN.

Stanley N. Gershoff

Niacin is produced to United States Pharmacopoeia (USP) requirements. The principal route for manufacture of niacin is oxidation of quinoline, obtained from coal tar (see *illus.*). Oxidation of 2-methyl-5-ethylpyridine is nearly as important; other processes involve oxidation of 3-picoline, hydrolysis of 3-cyanopyridine, or oxidation and hydrolysis of nicotine. Many patents cover the field. Nicotinamide is produced by amidation of niacin or its esters and by hydrolysis of 3-cyanopyridine.

Niacin, the least expensive vitamin, is used in foods, feed, and pharmaceutical preparations to supplement limited amounts available naturally. A major outlet for niacin is swine and poultry feeds, especially those based on corn. In human nutrition, enrichment of flour and bread with niacin is required in most states. Rice and corn products, alimentary pastes, and milk are fortified with niacin in some areas. Both niacin and nicotinamide are used in single and multivitamin capsules and tablets, and are prescribed for various clinical applications.

Adrian H. Cubberley

Nickel

A chemical element, Ni, atomic number 28, a silver-white, ductile, malleable, tough metal. The atomic mass of naturally occurring nickel is 58.71. See PERIODIC TABLE.

| | | | | | | | | | | | | | | | | | | | | | |
|-------------------|----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|----|----|----|----|----|----|
| 1 | | | | | | | | | | | | | | | | | 18 | | | | |
| 2 | | | | | | | | | | | | | | | | | 2 | | | | |
| 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 | | | | | | |

Nickel consists of five natural isotopes having atomic masses of 58, 60, 61, 62, 64. Seven radioactive isotopes have also been identified, having mass numbers of 56, 57, 59, 63, 65, 66, and 67.

Nickel metal. Most commercial nickel goes into stainless steel and other corrosion-resistant alloys. Nickel is also important in coins as a replacement for silver. Finely divided nickel is used as a hydrogenation catalyst. See NICKEL ALLOYS.

Nickel is a fairly plentiful element, making up about 0.008% of the Earth's crust and 0.01% of the igneous rocks. Appreciable quantities of nickel are present in some kinds of meteorite, and large quantities are thought to exist in the Earth's core. Two important ores are the iron-nickel sulfides, pentlandite and pyrrhotite ($(\text{Ni,Fe})_x\text{S}_y$); the ore garnierite, $(\text{Ni,Mg})\text{SiO}_3 \cdot n\text{H}_2\text{O}$, is also commercially important. Nickel occurs in small quantities in plants and animals. It is present in trace amounts in sea water, petroleum, and most coal.

Nickel metal is of moderate strength and hardness (3.8 on Mohs scale). When viewed as very small particles, nickel appears black. The density of nickel is 8.90 times that of water at 20°C (68°F). Nickel melts at 1455°C (2651°F) and boils at 2840°C (5144°F). Nickel is only moderately reactive. It resists alkaline corrosion and does not burn in the massive state, although fine nickel wires can be ignited. Nickel is above hydrogen in the electrochemical series, and it dissolves slowly in dilute acids, releasing hydrogen. In metallic form nickel is a moderately strong reducing agent.

Nickel is usually divopositive in its compounds, but it can also exist in the oxidation states 0, 1+, 3+, and 4+. Besides the simple nickel compounds, or salts, nickel forms a variety of coordination compounds or complexes. Most compounds of nickel are green or blue because of hydration or other ligand bonding to the metal. The nickel ion present in water solutions of simple nickel compounds is itself a complex, $[\text{Ni}(\text{H}_2\text{O})_6]^{2+}$.

William E. Cooley

Biology. Depending on the chemical species and concentration, nickel compounds exhibit a wide variety of biological effects. Selected nickel-containing compounds are harmful to humans and other animals. Nickel, however, also appears to be a beneficial trace element for many animals. Similarly, high concentrations of nickel are toxic to both plants and microorganisms, while trace levels of nickel are essential to many species. A small group of nickel-containing enzymes have been identified and characterized to varying extents, and nickel has been substituted for the normal metal in several metalloproteins.

Robert L. Blakeley; Robert P. Hausinger

Bibliography. F. A. Cotton et al., *Advanced Inorganic Chemistry*, 6th ed., Wiley-Interscience, 1999; J. R. Davis (ed.), *Metals Handbook: Desk Edition*, 2d ed., ASM International, 1998; R. P. Hausinger, *Biochemistry of Nickel*, 1993; D. R. Lide, *CRC Handbook Chemistry and Physics*, 85th ed., CRC Press, 2004; M.-H. Yu, *Environmental Toxicology: Biological and Health Effects of Pollutants*, 2d ed., 2004.

Nickel alloys

Combinations of nickel with other metals. Nickel has been used in electroplating since 1843 and as an alloying addition to steels since about 1889. It was first used as a base for alloys with the introduction of Monel nickel-copper alloy in about 1905. The nominal compositions of some of the alloys containing more than 50% nickel are given in **Table 1**.

Nickel-base alloys may be melted in open-hearth, electric-arc, or induction furnaces in air, under inert gas, or in vacuum. Casting may also be done under these same ambient conditions. Cast shapes are made in sand or investment molds or by shell molding. Ingots for wrought products are cast in metal molds and are hot-worked by forging, rolling, or extruding. In some instances further work may be done cold by rolling or drawing. Nickel-base alloys, made available in this way in bar, rod, wire, plate, strip, sheet, and tubular forms, may be fabricated into finished products by using conventional metalworking and metal-joining techniques.

Alloyed nickels. Nickel 211 and Duranickel alloy 301 are essentially binary alloys with 4.75% manganese and 4.5% aluminum, respectively. Manganese, in the first of these, extends the range of applicability in the presence of sulfur by about 300°F (150°C) to a limiting temperature in the neighborhood of 1000°F (540°C). A characteristic use of this material is as wire for spark-plug electrodes.

Aluminum and titanium confer age-hardening characteristics on Duranickel alloy 301 and a tensile strength in excess of 200,000 lb/in.² (1400 megapascals) is attainable in this alloy by appropriate cold work and heat treatment. In this condition it is well suited to the manufacture of springs and diaphragms.

Monel alloy 400. This alloy contains about two-thirds nickel and one-third copper and is the oldest of the commercial nickel-base alloys, dating from about 1905 when it was directly smelted from the

TABLE 1. Nominal composition of some nickel-base alloys, wt %

| Trademark | Ni | Cu | Cr | Co | Mo | Ti | Al | Cb | Fe | Mn | Si | C | Other |
|----------------------|------|------|------|------|------|-----|------|-----|------|-------|-------|-------|------------------|
| Nickel 211 | 95 | — | — | — | — | — | — | — | — | 4.75 | — | 0.08 | — |
| Duranickel alloy 301 | 93.7 | 0.05 | — | — | — | 0.4 | 4.4 | — | 0.35 | 0.3 | 0.5 | 0.17 | — |
| Monel alloy 400 | 66 | 31.5 | — | — | — | — | — | — | 1.35 | 0.9 | 0.15 | 0.18 | — |
| Monel alloy K-500 | 66 | 29 | — | — | — | 0.5 | 2.75 | — | 0.9 | 0.75 | 0.5 | 0.15 | — |
| Chromel P | 90 | — | 10 | — | — | — | — | — | — | — | — | — | — |
| Nichrome V | 80 | — | 19.5 | — | — | — | — | — | — | 2.5* | 1 | 0.25 | — |
| Alumel | 94 | — | — | — | — | — | 2 | — | — | 3 | 1 | — | — |
| Nimonic 75 | Bal | 0.5* | 19.5 | — | — | 0.4 | — | — | 5* | 1* | 1* | 0.12 | — |
| Nimonic 80A | Bal | — | 19.5 | 2* | — | 2.2 | 1.1 | — | 5* | 1* | 1* | 0.1* | — |
| Inconel alloy 600 | Bal | 0.5* | 15.5 | — | — | — | — | — | 8 | 1* | 0.5* | 0.15* | — |
| Inconel alloy X-750 | Bal | 0.5* | 15 | — | — | 2.5 | 0.9 | 0.9 | 7 | 0.7 | 0.4 | 0.04 | — |
| Inconel alloy 718 | 53 | 0.3* | 19 | 1.0* | 3 | 0.9 | 0.5 | 5 | Bal | 0.35 | 0.35* | 0.08* | — |
| Alloy 713C | Bal | — | 12 | — | 4 | 0.5 | 6 | 2 | 5* | 1* | 1* | 0.2* | 0.012 B, 0.10 Zr |
| Udimet 500 | Bal | — | 17.5 | 16.5 | 4 | 2.9 | 2.9 | — | 4* | 0.75* | 0.75* | 0.15* | — |
| Waspaloy | Bal | — | 19 | 14 | 3 | 2.5 | 1.2 | — | 2 | 0.7 | 0.4 | 0.05 | — |
| M252 | 55 | — | 19 | 10 | 10 | 2.5 | 0.75 | — | 2 | 1 | 0.7 | 0.1 | — |
| GMR 235 | Bal | — | 15.5 | — | 5 | 2.5 | 3 | — | 10 | 0.25* | 0.6* | 0.15 | 0.06 B |
| Hastelloy B | 61 | — | 1* | 2.5* | 27.5 | 2 | — | — | 5.5 | 1* | 1 | 0.05* | 0.4V |
| Hastelloy C | 54 | — | 15.5 | 2.5* | 15.5 | — | — | — | 5.5 | 1* | 1* | 0.08* | 0.35 V*, 4 W |
| Hastelloy D | 82 | 3 | 1* | 1.5* | — | — | — | — | 2* | 1 | 9 | 0.12* | — |

*Maximum.

copper-nickel matte obtained from Sudbury, Ontario, sulfide ore. The good fabricating characteristics and corrosion resistance of this alloy have made it widely used in marine applications and in the chemical-processing and petroleum industries. It also has applicability in the field of nuclear propulsion. As with nickel, this alloy can be made age-hardenable by the addition of aluminum and titanium. Monel alloy K-500 has corrosion-resisting characteristics similar to those of the non-age-hardenable composition, and is widely specified for such applications as marine propellers, shafting, valves, pump parts, and springs. A usable tensile strength of about 175,000 lb/in.² (1200 MPa) is obtainable in cold-drawn and age-hardened wire.

Nickel-chromium alloys. Nickel-chromium binary alloys are used primarily in specialty high-temperature service. An 80 nickel-20 chromium is a common high-quality resistance-heating-element material possessing good resistance to oxidation up to about 2100°F (1150°C), superior to either of its two component elements. The alloy is used both in industrial-furnace and household-appliance heating elements. Chromel P is used with the chromium-free nickel-base Alumel in temperature-sensing devices known as thermocouples. This alloy couple has favorable thermoelectric characteristics for applicability in the measurement of temperatures up to 2000°F (1090°C).

Nickel-chromium and related complex alloys are widely used for structural and general-purpose applications at high temperatures and in certain corrosive environments, particularly where freedom from stress-corrosion cracking is essential. In this latter instance Inconel alloy 600 has applicability in nuclear propulsion units.

This class of alloys encompasses a broad range of high-temperature properties. Nimonic 75 nickel-chromium alloy is used as a scale-resistant sheet

material. Neither this material nor Inconel alloy 600 responds to age hardening, and hence they are on the low side of the elevated-temperature mechanical property range. By contrast, Inconel alloy X-750, which contains added aluminum, titanium, and columbium, develops greatly improved high-temperature strength after proper heat treatment. The more highly alloyed Alloy 713C, a nickel-chromium cast alloy, exhibits further strength improvement at the high side of the temperature range of applicability for the complex nickel-chromium alloys. For comparison, 100-h rupture strengths of these three alloys are listed in **Table 2**.

The aforementioned materials and a number of similar proprietary alloys which combine high strength and oxidation resistance, have found application in jet engine and gas turbines for such parts as combustion liners, blades, vanes, and disks.

Inconel alloy 718 is an age-hardenable alloy with a slow aging response, a characteristic which permits welding and annealing without spontaneous hardening. It is suitable for service from cryogenic temperatures (−423°F or −253°C) to moderately elevated temperatures (1300°F or 704°C).

In the interest of optimizing properties, these complex nickel-chromium alloys are being produced in increasing quantities by vacuum-melting and vacuum-pouring techniques.

TABLE 2. Rupture strengths (100 h) of nickel-chromium alloys, lb/in.²*

| Alloy | 1500°F (815°C) | 1700°F (927°C) |
|---------------------|----------------|----------------|
| Inconel alloy 600 | 8000 | 3800 |
| Inconel alloy X-750 | 25,000–30,000 | 8000–10,000 |
| Alloy 713C | 60,000 | 30,000 |

*1 lb/in.² = 6.9 kPa.

A cast heat-resisting alloy carrying the Alloy Castings Institute designation HW (nominally 60% nickel, 12% chromium, 23% iron) is used principally for furnace parts and heat-treating fixtures. It has good resistance to oxidation and carburization, only modest hot strength, but good thermal shock resistance.

Hastelloy alloys. Hastelloy alloys B, C, and D are used primarily in corrosive environments. Hastelloy B is resistant to hydrochloric and sulfuric acids within certain limits of concentration, temperature, and degree of aeration. It is not recommended for service involving strong oxidizing acids or oxidizing salts. Hastelloy C is unusually resistant to oxidizing solutions and to moist chlorine. Hastelloy D has exceptional resistance to hot concentrated sulfuric acid.

Nickel-iron alloys. Alloys containing more than 50% nickel are used in various applications involving controlled thermal expansivity or certain magnetic requirements. In the range 50–52% nickel, the balance iron, the alloys have thermal expansion characteristics useful in making some types of glass-to-metal seals.

In the range 77–80% nickel, with or without about 4% molybdenum, the balance iron, the alloys have very high initial and maximum magnetic permeabilities when properly processed. *See* ALLOY; IRON ALLOYS; NICKEL; NICKEL METALLURGY; STAINLESS STEEL.

E. N. Skinner; Gaylord Smith

Nickel metallurgy

The extraction and refining of nickel from its ores. Nickel's properties of strength, toughness, and resistance to corrosion have been used to advantage in alloys since ancient times. Paktong, similar to modern nickel silver, was used in the sixteenth century in China, and early weapons were often fashioned from tough, nickel-bearing meteoric iron. A. F. Cronstedt first isolated nickel as an element in 1751, and in 1804 H. T. Richter prepared it in relatively pure form.

Occurrence. Although nickel ranks twenty-fourth in order of abundance of the elements, and igneous rocks average 0.01% nickel content, there are relatively few nickel deposits of commercial importance. Nickel ores are of two generic types, sulfides and laterites. Explorations of the ocean bottoms have revealed vast deposits of manganese oxide nodules which contain significant values of nickel, copper, and cobalt.

Sulfides. In ores of this type, nickel is present chiefly as the mineral pentlandite, a nickel-iron sulfide, usually in association with pyrrhotite and chalcopyrite. The most important known deposits, at Sudbury, Canada, have provided the major portion of the world's supply since 1905. Other substantial deposits have been developed in the Thompson-Moak Lake area of northern Manitoba, Canada, and in Western Australia. Russia exploits deposits on the Kola Peninsula and near Norilsk in Siberia. Smaller deposits are

worked in South Africa and in Finland. *See* PENTLANDITE.

Laterites. Lateritic nickel ores occur in two main forms, oxides and silicates. In the oxide form the nickel is dispersed through limonite, a hydrated form of iron oxide, while in the silicate form nickel partially replaces magnesium in the lattice of a hydrated magnesium silicate. Lateritic ores are widely distributed throughout the tropics and constitute the world's largest known reserves of nickel. Deposits in New Caledonia, Cuba, and Oregon have been worked commercially for many years. Commercial installations are also operating in Greece, the Dominican Republic, the Philippines, and Australia. The Japanese nickel industry is almost exclusively based on imported ores from New Caledonia and Indonesia and concentrates from Australia and Canada. *See* LATERITE.

Ocean nodules. The existence of manganese oxide nodules on the bottoms of the oceans, notably the deep valleys of the Pacific and the Atlantic, has been known for many years. Exploration revealed that the nickel and cobalt values contained in these nodules, averaging about 1% nickel and 0.2% cobalt, rival the metals content of the land-based lateritic ores. *See* MANGANESE NODULES.

Production and uses. Nickel gained commercial prominence late in the nineteenth century, when substantial reserves were found in New Caledonia and at Sudbury, and the world's naval powers adopted nickel-bearing armor. Until about 1920 nickel markets depended upon military requirements. Following World War I, research into industrial applications of nickel was greatly increased, and the success of the continuing program is evident from nickel's numerous diversified and expanding uses.

Nickel is marketed in various forms. In the United States, nickel consumption is estimated at more than 45% for stainless and other steels; over 25% for nickel alloys; approximately 15% in electroplating, 7% in high-temperature and electrical-resistance alloys, 2% in cast irons, and the remainder in miscellaneous applications such as catalysts, magnets, and ceramics.

Extractive metallurgy. Selection of processes for nickel extraction is largely determined by the type of ore to be treated. Sulfide ores are amenable to concentration by such methods as flotation or magnetic separation. The state of combination of nickel in the lateritic ores usually precludes such enrichment, thus requiring treatment of the total ore.

Sulfide ores. **Figure 1** illustrates the major processes for the extraction and recovery of nickel from sulfide ores. These ores, usually containing from 1 to 3% nickel and varying amounts of copper, are first crushed and ground to liberate the mineral values and then subjected to froth flotation to concentrate the valuable constituents and reject the gangue or rock fraction. Depending on the ore's copper and pyrrhotite contents, it is sometimes appropriate to produce a separate copper concentrate and a separate pyrrhotite concentrate. The three examples chosen are typical in the industry and, with some minor

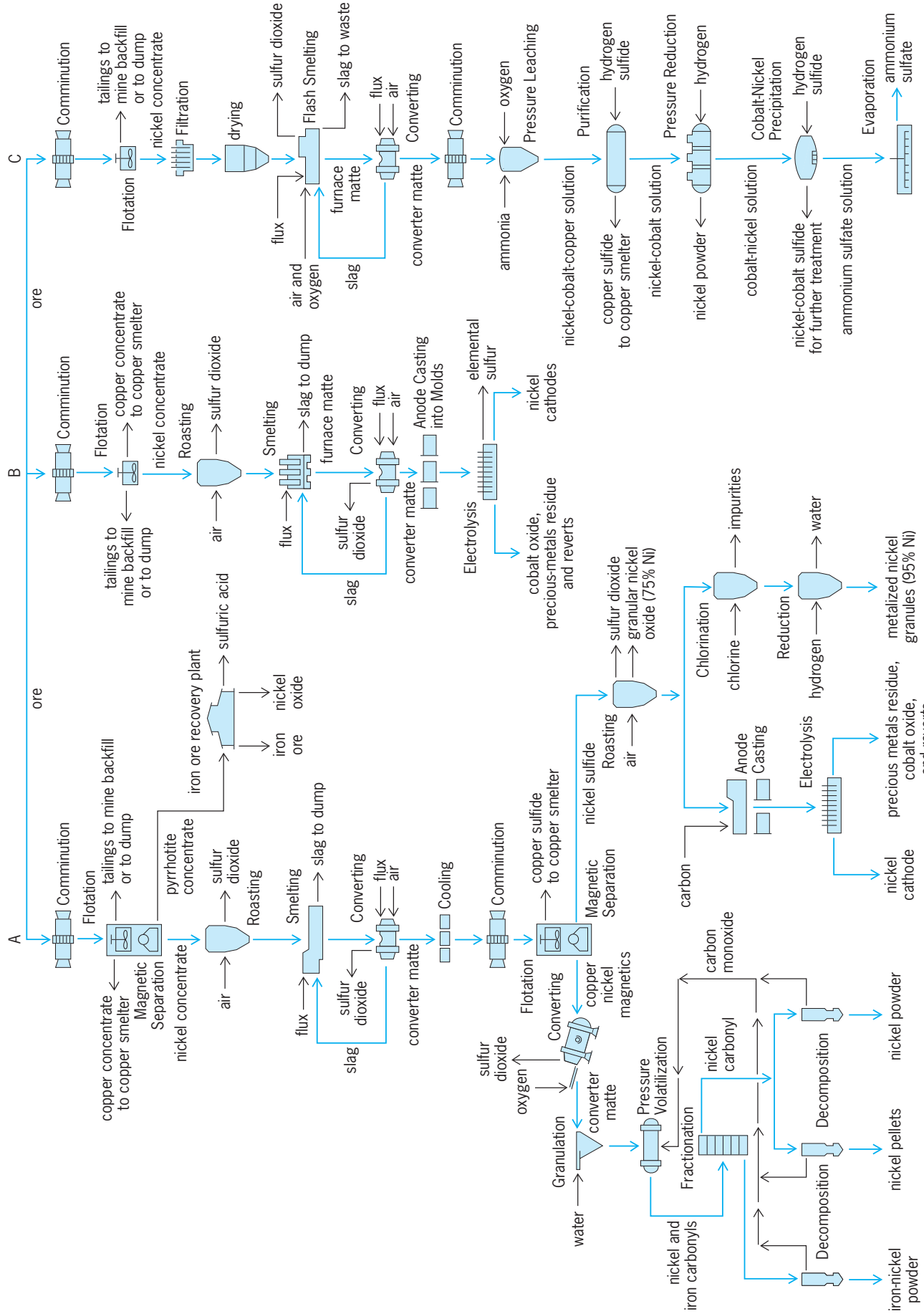


Fig. 1. Flowsheets for three methods of sulfide ore treatment.

variations in the processing steps, account for about 90% of the world's nickel production from sulfide ores. *See* ORE DRESSING.

Selective flotation and magnetic separation may be employed to divide the bulk concentrate into nickel, copper, and iron-rich fractions for separate treatment (Fig. 1, flowsheet A). A high-grade iron ore, nickel oxide, and sulfuric acid are recovered from the iron concentrate. The nickel concentrate is treated by pyrometallurgical processes. The major portion undergoes partial roasting in multihearth or fluidized-bed furnaces to eliminate about half of the sulfur and to oxidize the associated iron. The hot calcine, plus flux, is smelted in natural gas-coal-fired reverberatory furnaces operating at about 2200°F (1200°C) to produce a furnace matte, enriched in nickel, and a slag for discard. The furnace matte is transferred to Pierce-Smith converters and blown with air in the presence of more flux to oxidize the remaining iron and associated sulfur, yielding Bessemer matte containing nickel, copper, cobalt, small amounts of precious metals, and about 22% sulfur. The slag which is generated in the converting operation is returned to the smelting furnace to recover its metal values.

The molten Bessemer matte is cast into 25-ton (22.5-metric-ton) molds in which it undergoes controlled slow cooling to promote formation of relatively large, discrete crystals of copper sulfide, nickel sulfide, and a small quantity of a metallic phase, a nickel-copper alloy which collects most of the precious metals. After crushing and grinding, the metallics are removed magnetically and treated in a refining complex for recovery of metal values, and the main sulfide fraction is separated into copper sulfide and nickel sulfide concentrates by froth flotation. *See* FLOTATION; MAGNETIC SEPARATION METHODS.

The nickel sulfide is converted to granular nickel oxide sinter in fluidized-bed reactors. A portion of this product is marketed directly for alloy steel production. Another part of the granular oxide is treated by chlorination at high temperature (2200°F or 1200°C) to lower its copper content to about 0.5%, and then reduced by hydrogen at about 930°F (500°C) to yield a highly metallized product (95% Ni) for market. Two processes are employed to convert the remaining oxide to pure metal for market. One involves reduction smelting to metal anodes which is followed by electrolytic refining, by using a sulfate-chloride electrolyte with divided cells and continuous electrolyte purification. The product of this operation is electrolytic nickel cathodes, and the by-products are cobalt and precious metals.

The other process is the atmospheric-pressure carbonyl process which is used in Great Britain. The nickel oxide sinter is reduced with hydrogen and treated with carbon monoxide at about 122°F (50°C) to volatilize nickel as gaseous nickel carbonyl. This compound is decomposed at about 390°F (200°C) to yield high-purity nickel in pellet form. Copper and cobalt salts and precious metals are recovered from the residue. Nickel powder is also produced at this plant in a pressure carbonyl system.

The nickel-copper alloy from the matte separation step, containing significant platinum group metal values, is melted in a top-blown rotary converter, its sulfur content is adjusted by blowing with oxygen at temperatures up to 2900°F (1600°C), and the metal product is granulated with water. The dried metal granules are treated in high-pressure (6.7 megapascals) reactors with carbon monoxide at 300°F (150°C) to form nickel carbonyl and some iron carbonyl. The mixed carbonyls are separated by fractionation, and the pure nickel carbonyl is decomposed at about 390°F (200°C) to yield high-purity nickel pellets and nickel powder for market.

There is a much simpler procedure that can be used to process nickel sulfide ores because the associated copper values are too low to warrant local processing (Fig. 1, flowsheet B). Selective flotation is employed to produce a nickel concentrate low in copper and a small amount of copper concentrate for treatment elsewhere. The dewatered nickel concentrate is fluid-bed-roasted for partial elimination of sulfur, and the calcine plus flux is smelted in arc-type electric furnaces. Waste furnace slag is granulated for disposal while the molten matte is transferred to Pierce-Smith converters for further upgrading to Bessemer matte. The conventional procedure of flux addition and blowing with air removes all but traces of iron from the matte, and the slag produced is returned to the electric furnace for recovery of metal values. The converter matte, essentially nickel sulfide (Ni₃S₂), is cast into anodes and electrolyzed to yield elemental sulfur at the anode and pure nickel at the cathode. The refining operation also produces cobalt oxide and precious-metal residues.

In the unique all-hydrometallurgical process (Fig. 1, flowsheet C), the nickel ore is concentrated by conventional froth flotation, and the dried nickel concentrate is flash-smelted with oxygen-enriched air and flux to produce furnace matte and waste slag. The furnace matte is cooled, crushed, and finely ground as feed for the hydrometallurgical plant, where it is leached under pressure with a strong ammonia solution and air to solubilize the base metal values, with the simultaneous production of ammonium sulfate. The pregnant leach liquor is treated to remove impurities and then reduced with hydrogen at elevated pressure (435 lb/in.² or 3 MPa) and temperature (374°F or 190°C) to yield a granular nickel powder product. The tail liquor from this operation is treated further to recover ammonium sulfate crystals and a mixed nickel-cobalt sulfide.

There are many other processes for the treatment of sulfide nickel ores. However, the differences from the illustrated examples are believed to be minor.

Lateritic ores. The bulk of the nickel originating from lateritic ores is marketed as ferronickel. The process employed is basically simple and involves drying and preheating the ore usually under reduction conditions. The hot charge is then further reduced and melted in an electric-arc furnace, and the crude metal is refined and cast into ferronickel pigs. A typical operation is shown in Fig. 2, flowsheet A.

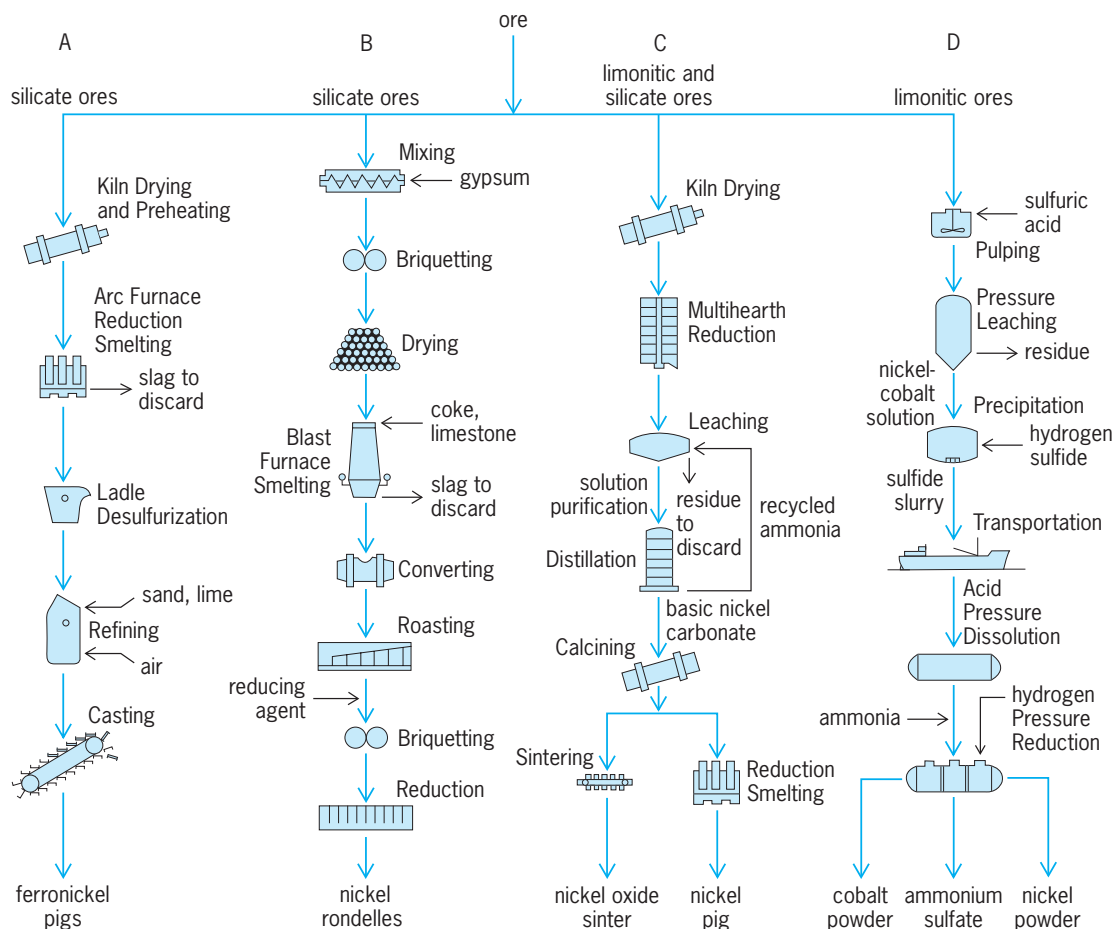


Fig. 2. Flowsheets for four methods of lateritic ore treatment.

A substantial amount of nickel is produced from lateritic ores by the nickel sulfide matte technique. In this process the ore is mixed with gypsum or other sulfur-containing material such as high-sulfur fuel oil, followed by a reduction and smelting operation to form matte. As in the treatment of sulfide concentrates, the molten furnace matte is upgraded in either conventional or top-blown rotary converters to a high-grade matte, which can be further refined by roasting and reduction to a metallized product. An example of this procedure is shown in Fig. 2, flowsheet B.

Other large-scale operations are based on selective reduction of the ore followed by ammoniacal leaching at atmospheric pressure to dissolve the nickel values. The pregnant liquor is treated to remove impurities and then heated in suitable vessels to drive off ammonia and carbon dioxide and to precipitate a basic nickel carbonate. This material may be sintered under reducing conditions to yield a metallized nickel oxide sinter, or the carbonate may be redissolved in ammoniacal solution and then treated with hydrogen under pressure to yield nickel powder for briquetting. This process is depicted in Fig. 2, flowsheet C. Flowsheet D in Fig. 2 depicts a process wherein limonitic-type ores are leached with sulfuric acid at elevated temperature and pressure to solubilize nickel and cobalt. The pregnant solution is then

treated with hydrogen sulfide to precipitate mixed nickel-cobalt sulfides. This precipitate may be treated by the Sherritt-Gordon pressure ammonia leach process to yield separate nickel and cobalt powders. See NICKEL; NICKEL ALLOYS; PYROMETALLURGY, NONFERROUS.

Alexander Illis

Bibliography. W. Betteridge, *Nickel and Its Alloys*, 1977; A. R. Burkin, *Extractive Metallurgy of Nickel*, 1987; F. Habashi, *Pyrometallurgy*, 1986; J. J. Jacobs, M. Allard, and S. Behmo, *Nickel and Cobalt Extraction Using Organic Compounds*, 1985; G. Tyroler and C. Landolt (eds.), *Extractive Metallurgy of Ni and Co*, 1988.

Nicotinamide adenine dinucleotide (NAD)

An organic coenzyme and one of the most important components of the enzymatic systems concerned with biological oxidation-reduction reactions. It is also known as NAD, diphosphopyridine nucleotide (DPN), coenzyme I, and codehydrogenase I. NAD (Fig. 1) is found in the tissues of all living organisms. See COENZYME.

The nicotinamide, or pyridine, portion of NAD can be reduced chemically or enzymatically with the formation of reduced or hydrogenated NAD, or NADH (Fig. 2). NAD functions as the immediate oxidizing

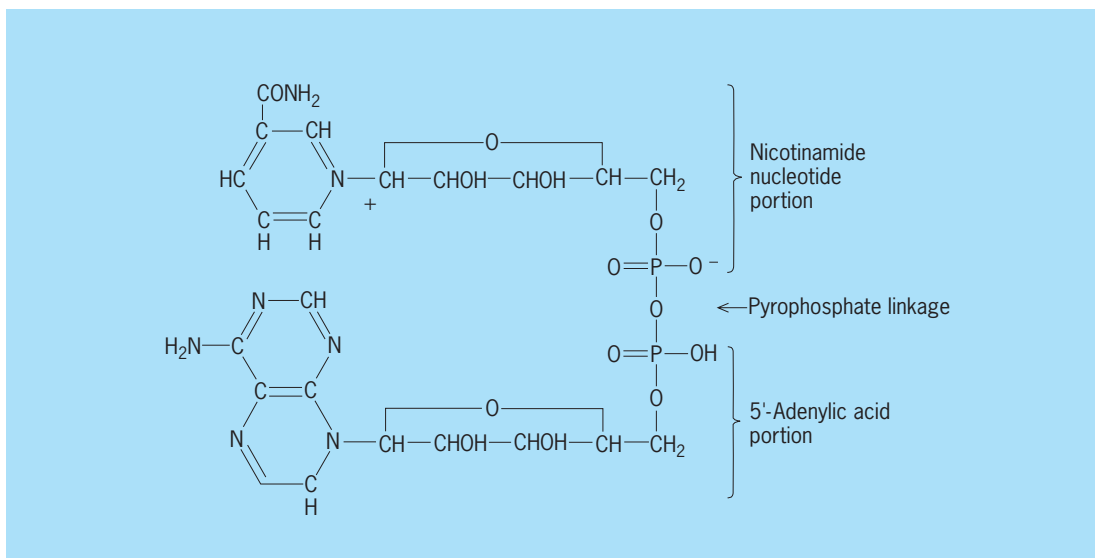
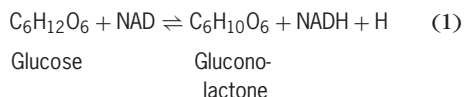


Fig. 1. Oxidized form of nicotinamide adenine dinucleotide (NAD).

agent for the oxidation, or dehydrogenation, of various organic compounds in the presence of appropriate dehydrogenases, which are specific apoenzymes, or protein portions of the enzyme. In the dehydrogenase reactions one hydrogen atom is transferred from the substrate to NAD, while another is liberated as hydrogen ion. For instance, in a reversible reaction, catalyzed by glucose dehydrogenase, glucose is oxidized to gluconolactone, as shown in reaction (1).



The NADH, formed in biological oxidations, is reoxidized to NAD (Fig. 2) in coupled reductions, also catalyzed by specific enzymes. In respiration the NADH is reoxidized through a sequence of reactions in which a flavoprotein enzyme, diaphorase, and the cytochrome system of iron-porphyrin catalysts transfer the electrons from NADH to molecular oxygen, the overall reaction being expressed as reaction (2).



In fermentations NADH is reoxidized with the concomitant reduction of organic molecules, which are usually produced in intermediary metabolism. For example, in the metabolism of muscle tissue, lactic acid is produced by the reduction of pyruvic acid,

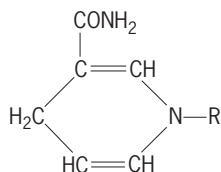
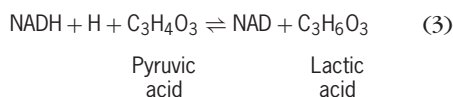


Fig. 2. Reduced form of nicotinamide portion in hydrogenerated NAD (NADH).

as shown in reaction (3). The enzyme catalyzing

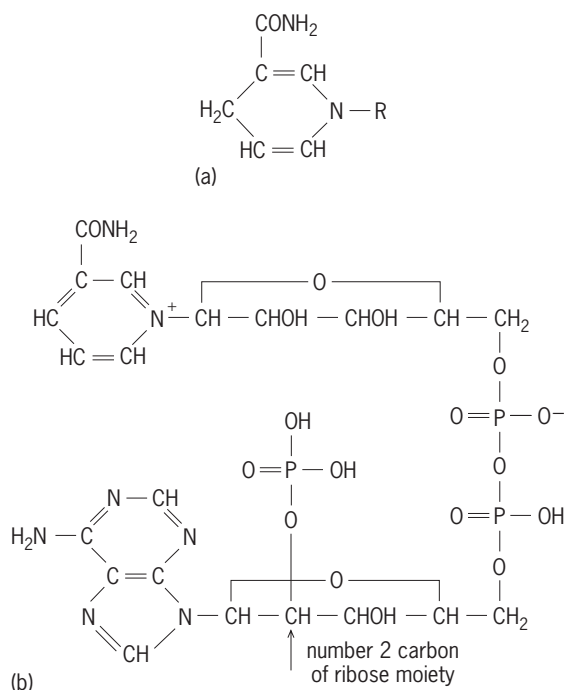


this reaction is called lactic dehydrogenase, since the process is reversible, and lactic acid can be oxidized with NAD. NAD and its reduced form, NADH, serve to couple oxidative and reductive processes and are constantly regenerated during metabolism. Hence, they serve as catalysts and NAD is referred to as a coenzyme. In some enzymatic reactions a different coenzyme, nicotinamide adenine dinucleotide phosphate (NADP), also called triphosphopyridine nucleotide, or coenzyme II, is required. Dehydrogenases are generally quite specific with respect to the coenzyme which they can utilize. See BIOLOGICAL OXIDATION; CYTOCHROME; ENZYME; NIACIN; NICOTINAMIDE ADENINE DINUCLEOTIDE PHOSPHATE (NADP).

Michael Doudoroff

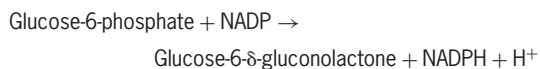
Nicotinamide adenine dinucleotide phosphate (NADP)

A coenzyme and an important component of the enzymatic systems concerned with biological oxidation-reduction systems. It is also known as NAD, triphosphopyridine dinucleotide, triphosphopyridine nucleotide (TPN), coenzyme II, and codehydrogenase II. The compound is similar in structure and function to nicotinamide adenine dinucleotide (NAD). It differs structurally from NAD in having an additional phosphoric acid group esterified at the 2' position of the ribose moiety of the adenylic acid portion. In biological oxidation-reduction reactions the NADP molecule becomes alternately reduced to its hydrogenated form (NADPH) and reoxidized to its initial state (see *illus.*). See NICOTINAMIDE ADENINE DINUCLEOTIDE (NAD).



Nicotinamide adenine dinucleotide phosphate. (a) Reduced form of the nicotinamide portion of NADPH. (b) Oxidized form of the molecule (NADP).

NADP is specifically required in some enzymatic reactions, just as NAD is in others. There are, however, a few reactions in which either compound can serve as a coenzyme. In the carbohydrate metabolism of yeast and mammalian tissue, NADP acts as the oxidizing agent for glucose-6-phosphate, with the enzyme glucose-6-phosphate dehydrogenase, as in the reaction below. NADP is formed, enzymatically, from



NAD by phosphorylation with adenosine triphosphate. See BIOLOGICAL OXIDATION; CARBOHYDRATE METABOLISM; COENZYME; ENZYME. Michael Doudoroff

Nicotine alkaloids

Alkaloids found in various species of the genus *Nicotiana*. The species most often used for the production of tobacco because of its high level of nicotine is *N. tabacum*, which is cultivated in many parts of the world for the preparation of cigarettes, cigars, and pipe tobacco. Nicotine is the most abundant alkaloid in *N. tabacum*, occurring to the extent of 2–8% based on the dry weight of the cured leaf. Other alkaloids that are found in this species are nornicotine, anabasine, and anatabine. The chemical structures of these alkaloids are shown in the **illustration**. See ALKALOID; TOBACCO.

The relative amounts of the alkaloids depend on the variety of *N. tabacum* and the various species of *Nicotiana*. In tree tobacco, *N. glauca*, (RS)-anabasine is the main alkaloid. (S)-Nicotine

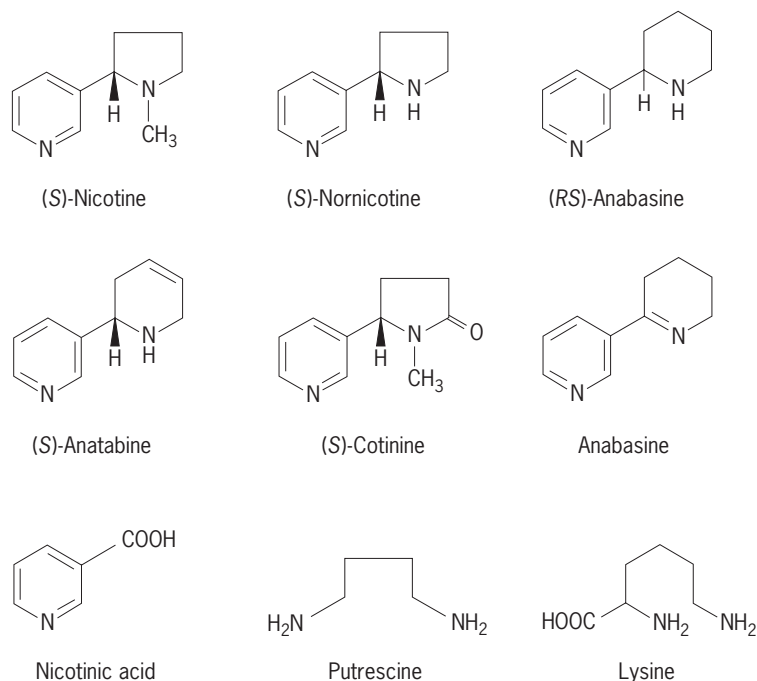
(C₁₀H₁₄N₂) is readily extracted from tobacco roots and stalks that remain after the leaves have been picked for tobacco production. (S)-Nicotine is a water-soluble colorless liquid when freshly distilled at reduced pressure (123–125°C or 253–257°F; 17 mmHg or 2.3 kPa). However, it rapidly turns yellow and then brown on exposure to air. It is optically pure when obtained from the tobacco plant, specific rotation $[\alpha]_D^{25} = -169^\circ$ (pure liquid). It is racemized on heating with a small amount of potassium tertiary butoxide at its boiling point (250°C or 482°F) for 15 min. Nicotine is dibasic [$pK_a = 7.9$ (pyrrolidine); 3.1 (pyridine)], and it forms crystalline salts with many acids. See PK.

Nicotine forms a crystalline salt with two molecules of tartaric acid and a salt, sparingly soluble in ethanol, with two molecules of picric acid.

Nicotine has a bitter taste and a sharp odor. It is quite poisonous, with 250–350 mg being fatal for an adult when taken orally, and less than fatal amounts causing fever, trembling, nausea, and convulsions. It is toxic to animals, since it binds to the receptor site for acetylcholine, a nerve transmitter, and so it is used as an insecticide for killing cockroaches, houseflies, mosquitoes, and aphids. A typical commercial nicotine insecticide is a 40% solution of nicotine sulfate in water. See ACETYLCHOLINE; INSECTICIDE.

The sensation that is obtained from smoking tobacco is due to the presence of nicotine in the tobacco leaf. Some is destroyed during burning, but some is converted to nitroso compounds, such as *N*-nitrosornicotine, by reaction with nitrous acid, which is formed in the burning tobacco by pyrolysis of the nitrates in the unburnt leaf. While these nitroso compounds are carcinogenic, the main source of carcinogens in tobacco smoke is probably in the tars, which contain polynuclear aromatic hydrocarbons. Smokers metabolize nicotine to cotinine, which is much less toxic and is excreted in the urine. The special chewing gum that is prescribed to alleviate the withdrawal symptoms of those who stop smoking contains about 2 mg nicotine per tablet of gum. See MUTAGENS AND CARCINOGENS; NITRO AND NITROSO COMPOUNDS; PYROLYSIS.

Nicotine is formed initially in the roots of the tobacco plant and is then translocated into the leaves, where it is relatively inert and is present as salts of organic acids such as citric acid and malic acid. The precursors of nicotine are nicotinic acid (the vitamin niacin), which is the source of the pyridine ring, and putrescine, which is the source of the pyrrolidine ring. The putrescine is derived from the α -amino acids ornithine and arginine. The *N*-methyl group is derived from the *S*-methyl group of the amino acid methionine. This same biosynthetic route operates in the *Duboisia* species. Nornicotine is a metabolite of nicotine resulting from an *N*-demethylation that occurs in both the growing plant and the dried leaf. The pyridine ring of anabasine is also formed from nicotinic acid, and the saturated ring is derived from lysine. Both rings of anatabine are derived from nicotinic acid. See PYRIDINE; PYRROLE.



Structures of nicotine and some other alkaloids found in tobacco.

Nicotine is readily isolated from tobacco leaves by blending with a mixture of concentrated ammonia and an organic solvent such as chloroform or methylene chloride. Nicotine and the minor alkaloids are extracted from the organic solvent with dilute hydrochloric acid. This aqueous solution is made basic with ammonia and reextracted with chloroform. Evaporation of this extract affords the crude alkaloids, which can be further purified and separated into the individual components by various kinds of chromatography. The simplest method for analyzing this mixture is by capillary gas chromatography. *See* CHROMATOGRAPHY.

Nicotine has also been isolated from species other than *Nicotiana* such as *Duboisia* (native to Australia), *Equisetum* (horsetails), *Lycopersicum* (tomatoes), *Lycopodium* (club mosses), *Sedum* (succulent plants), and *Solanum* (potato family). Remarkably, anabasine, a minor alkaloid in tobacco, has been isolated from a species of marine worms (*Amphiporus*) and some ants (*Apbaenogaster*). *See* HETEROCYCLIC COMPOUNDS. Edward Leete

Bibliography. L. E. Ember, The nicotine connection, *Chem. Eng. News*, 72(48):8-18, 1994; J. W. Gorrod and J. Wahren (eds.), *Nicotine and Related Alkaloids: Distribution, Metabolism, and Excretion*, 1993; S. W. Pelletier (ed.), *Alkaloids: Chemical and Biological Perspectives*, vol. 3, 1990, vol. 8, 1992.

Niobium

A chemical element, Nb, atomic number 41 and atomic weight 92.906. In the United States this element was originally called columbium. The metal-

lurgists and metals industry still use this older name. *See* PERIODIC TABLE.

Most niobium is used in special stainless steels, high-temperature alloys, and superconducting alloys such as Nb₃Sn. Niobium is also used in nuclear piles.

Niobium metal has a density of 8.6 g/cm³ (5.0 oz/in.³) at 20°C (68°F), a melting point of 2468°C (4474°F), and a boiling point of 4927°C (8900°F). Metallic niobium is quite inert to all acids except hydrofluoric, presumably owing to an oxide film on the surface. Niobium metal is slowly oxidized in alkaline solution. It reacts with oxygen and the halogens upon heating to form the oxidation state V oxide and halides, with nitrogen to form NbN, and with carbon to form NbC, as well as other elements such as arsenic, antimony, tellurium, and selenium.

The oxide Nb₂O₅, melting point 1520°C (2768°F), dissolves in fused alkali to yield a soluble complex niobate, Nb₆O₁₉⁸⁻. Normal niobates such as NbO₄³⁻ are insoluble. The oxide dissolves in hydrofluoric acid to give ionic species such as NbOF₅²⁻ and

| | | | | | | | | | | | | | | | | | |
|----|----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|----|----|----|----|----|
| 1 | | | | | | | | | | | | | | | | | 18 |
| 2 | | | | | | | | | | | | | | | | | 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 | | | | | |
| 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 |

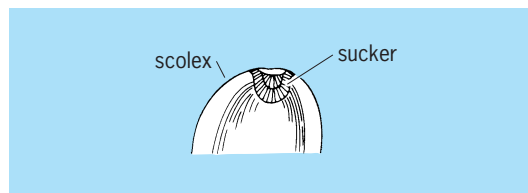
| | | | | | | | | | | | | | | |
|-----------------|----|----|----|----|----|----|----|----|----|----|----|-----|-----|-----|
| actinide series | 89 | 90 | 91 | 92 | 93 | 94 | 95 | 96 | 97 | 98 | 99 | 100 | 101 | 102 |
| | Ac | Th | Pa | U | Np | Pu | Am | Cm | Bk | Cf | Es | Fm | Md | No |

NbOF_6^{3-} , depending on the fluoride and hydrogen concentration. The highest fluoro complex which can exist in solution is NbF_6^- . Edwin M. Larsen

Bibliography. F. A. Cotton et al., *Advanced Inorganic Chemistry*, 6th ed., Wiley-Interscience, 1999; J. Hala (ed.), *Halides, Oxyhalides and Salts of Halogen Complexes of Titanium, Zirconium, Hafnium, Vanadium, Niobium and Tantalum*, 1989; J. Stephens and I. Ahmad (eds.), *High Temperature Niobium Alloys*, 1991.

Nippotaeniidea

An order of tapeworms of the subclass Cestoda. The few known species are intestinal parasites of Eurasian fresh-water fishes. The head (scolex) bears a single terminal sucker (see *illus.*). The segmental



Scolex of *Nippotaenia*.

anatomy shows relationships to the Pseudophyllidea and Cyclophyllidea. The life history is unknown. It is probable that this order is related to the proteocephalids. See EUCESTODA; PSEUDOPHYLLIDEA. Clark P. Read

Niter

A potassium nitrate mineral with chemical composition KNO_3 . Niter crystallizes in the orthorhombic system, generally in thin crusts and delicate acicular crystals; it also occurs in massive, granular, or earthy forms. It has good cleavage in three directions; fracture is subconchoidal to uneven; it is brittle; hardness is 2 on Mohs scale; specific gravity is 2.109. The luster is vitreous, and the color and streak are colorless to white.

Niter is commonly found, usually in small amounts, as a surface efflorescence in arid regions and in caves and other sheltered places. It is usually associated with soda niter, epsomite, nitrocalcite, and gypsum. The mineral may occur as an efflorescence on soils rich in organic matter from the action of certain bacteria on nitrogenous or animal matter. Niter occurs associated with soda niter in the desert regions of northern Chile, and in similar occurrences in Italy, Egypt, Russia, the western United States, and elsewhere. It was formerly found in some abundance in limestone caves in Tennessee, Kentucky, Alabama, and Ohio, and was used for the manufacture of gunpowder during the War of 1812 and the Civil War. See NITRATE MINERALS; SODA NITER. George Switzer

Nitrate minerals

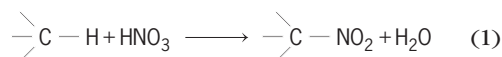
These minerals are few in number and with the exception of soda niter are of rare occurrence. Normal anhydrous and hydrated nitrates occurring as minerals are soda niter, NaNO_3 ; niter, KNO_3 ; ammonia niter, NH_4NO_3 ; nitrobarite, $\text{Ba}(\text{NO}_3)_2$; nitrocalcite, $\text{Ca}(\text{NO}_3)_2 \cdot 4\text{H}_2\text{O}$; and nitromagnesite, $\text{Mg}(\text{NO}_3)_2 \cdot 6\text{H}_2\text{O}$. In addition there are three known naturally occurring nitrates containing hydroxyl or halogen, or compound nitrates. They are gerhardtite, $\text{Cu}_2(\text{NO}_3)(\text{OH})_3$; buttgenschite, $\text{Cu}_{19}(\text{NO}_3)_2\text{Cl}_4(\text{OH})_{32} \cdot 3\text{H}_2\text{O}$; and darapskite, $\text{Na}_3(\text{NO}_3)(\text{SO}_4) \cdot \text{H}_2\text{O}$. See NITER; SODA NITER.

The natural nitrates are for the most part readily soluble in water. For this reason they occur most abundantly in arid regions, particularly in South America along the Chilean coast. See FERTILIZER; NITROGEN. George Switzer

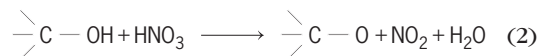
Nitration

A process in which a nitro group ($-\text{NO}_2$) becomes chemically attached to a carbon, oxygen, or nitrogen atom in an organic compound. A hydrogen or halogen atom is often replaced by the nitro group. Three general reactions summarize nitration chemistry:

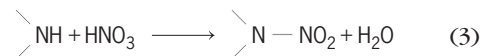
1. C-nitration, in which the nitro group attaches itself to a carbon atom [reaction (1)].



2. O-nitration (an esterification reaction), in which an O-N bond is formed to produce a nitrate [reaction (2)].

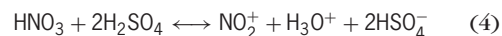


3. N-nitration, in which a N-N bond is formed [reaction (3)].



The chemical steps of nitration differ based on the type of organic compound nitrated, the nitration agent, and the operating conditions. Additionally, there are three important types of nitration processes: ionic, free-radical, and the Victor Meyer process.

Ionic-type nitrations. Aromatics, alcohols, glycols, and amines are generally nitrated with mixed acids via ionic reactions. The mixture includes nitric acid, a strong acid such as sulfuric acid which acts as a catalyst, and a small amount of water. With sulfuric acid, nitric acid is ionized to the nitronium ion, NO_2^+ , which is the nitrating agent [reaction (4)]. The nitro-



anium ion concentration increases in the acid mixture as the water content of the acids decreases and as the

molar ratio of sulfuric acid to nitric acid approaches 2.0. A mathematical model based on theoretical considerations has been developed by Albright, Good, and Eckert to predict the nitronium ion concentration in the mixed acids. Nitronium ions react with aromatics, alcohols, and amines to form intermediate ions, which release protons to complete the nitration and regenerate sulfuric acid. Nitrations with mixed acids are often conducted at temperatures ranging from 50 to 120°C.

Other strong acids sometimes used to produce nitronium ions include hydrogen fluoride, boron trifluoride (plus a trace of water), perchloric acid, acetic acid, and acidic-type ion-exchange resins. A mixture of nitric acid, sulfuric acid, and liquid sulfur trioxide, is also effective at relatively low temperatures. Nitrations to form dinitro or trinitro aromatics are harder to achieve, and require stronger acids or higher temperatures. The ease of nitrating an aromatic compound decreases as the number of nitro groups attached to the ring increase. *See* NITRO AND NITROSO COMPOUNDS; NITROAROMATIC COMPOUND.

Dinitrogen pentoxide, N_2O_5 , is also a highly effective nitrating agent, particularly for compounds that cannot be nitrated with mixed acids. N_2O_5 ionizes, forming NO^+_2 plus NO^-_3 . Nitric acid is often a by-product when N_2O_5 is used. *See* NITROGEN OXIDES.

For easily nitrated aromatics such as phenols or phenolic ethers, a mixture of nitrous acid (or a nitrite salt) and nitric acid is often used. Nitrosonium ions (NO^+), generated by ionization, react with aromatics to produce nitrosoaromatics, which are then oxidized by nitric acid to form both the nitroaromatic compound and nitrous acid. Nitrous acid permits further production of nitrosonium ions. Once these aromatics have been mononitrated, mixtures of nitric and sulfuric acid are generally required to di- and trinitrate them.

When mixed acids are used, two immiscible liquid phases—acid and hydrocarbon—are present in the reactor. Since nitration reactions occur mainly at the interface between the two phases, a large interfacial area between them is required for rapid nitration. To obtain a large interfacial area, high levels of agitation or special injection devices are often provided. No known experimental data are available concerning interfacial areas in commercial or laboratory reactors, but mixing models, such as VisiMix 2000, can often be used to predict these areas.

During nitration reactions, some undesired oxidation always occurs. Oxidation results when small amounts of nitric acid are extracted from the acid phase to the hydrocarbon phase. The amount of extracted nitric acid becomes significant as the amount of nitrated hydrocarbons in the hydrocarbon phase increases. In the absence of sulfuric acid and water, nitric acid is an effective oxidizing agent, capable of producing phenolic compounds from aromatics. Such phenolic compounds are easily nitrated and may even form di- and trinitrophenols, which are explosives. *See* NITRIC ACID; OXIDIZING AGENT.

Several nitrated hydrocarbons, such as trinitrotoluene, dinitrotoluenes, picric acid, and nitroglycerine, are explosives; major accidents, including explosions, have been reported in chemical plants that manufacture them. Several serious explosions have also occurred in plants producing mononitroaromatics and nitroalcohols. Other hazards in nitration plants include handling and recovery of acids, flammability of the hydrocarbon feeds and products, side reactions including undesired oxidations, and the toxicity of some hydrocarbons.

Batch reactors were used until perhaps 50 years ago to produce most, if not all, nitroaromatics, nitroalcohols, nitroglycerine, and nitroamines. Improper design and/or operation of these batch reactors and auxiliary equipment has on several occasions contributed to serious accidents. Temperature control of a highly exothermic reaction mixture is always a concern, especially in larger reactors that often have less heat transfer area per volume of reactants. To minimize accidents, the rates of nitrations are generally maintained at low levels: reaction times in the batch reactor are often in the 2–5-hour range.

Continuous-flow reactors are currently widely used to produce nitrobenzene, dinitrotoluenes, nitroglycerine, plus several other products. Continuous-flow reactors often provide large interfacial areas between the two liquid phases, resulting in much higher rates of nitration. Smaller reactors can be used due to their improved heat-transfer capabilities. In addition, by properly controlling the compositions and feed rates of the feed streams the reactors can often be operated adiabatically with only a small temperature increase resulting from the reaction, because only small amounts of nitrated products are in the reactor at any given time.

The Noram process is used for worldwide production of most nitrobenzene. In this process, the mixed acids contain only several percent of nitric acid, which is completely ionized to form nitronium ions. In one example, the feed acid contained 2.8% nitric acid, 27.7% water, and 69.5% sulfuric acid. Injectors are used to obtain large interfacial areas between the two phases. The process offers the following features: (1) The nitration reaction is completed in about 70 seconds; much shorter times have been reported in other flow processes. (2) Over 99% of the nitric acid reacts to produce nitrobenzene, using inlet temperatures of about 97–120°C. (3) The exit acid is a diluted sulfuric acid, which can be concentrated and reused. (4) The unreacted benzene is recovered and is recycled into the feed benzene. (5) Undesired side reactions, including oxidations, are of minor importance. (6) Operating costs are relatively low in part because the heat of reaction is used in the preheating and separation steps.

Free-radical nitrations. Propane is commercially nitrated in relatively large amounts using nitric acid in gas-phase free-radical reactions at temperatures of about 380–420°C. Nitric acid decomposes at these temperatures to produce nitrogen dioxide radicals

(actually a mixture of $\cdot\text{NO}_2$ and $\cdot\text{ONO}$) and a hydroxy radical ($\cdot\text{OH}$). In the free-radical reaction, about 35–40% of the nitric acid reacts to form four $\text{C}_1\text{--C}_3$ nitroparaffins; C–C bonds are broken during the nitration. The remaining nitric acid acts mainly as an oxidizing agent to form aldehydes, alcohols, carbon monoxide, carbon dioxide, water, and small amounts of other oxidized materials. Commercially, an adiabatic reactor is used, and the heat of reaction is employed to preheat and vaporize the nitric acid feed (containing water).

Separation steps used with the gaseous product steam include almost a complete condensation of mixed nitroparaffins. This liquid mixture is washed to remove the aldehydes, and then is distilled to recover each of the four nitroparaffins—nitromethane, nitroethane, 1-nitropropane, and 2-nitropropane. The unreacted propane is recovered, combined with the feed propane, and returned to the reactor. The oxides of nitrogen are converted back to nitric acid. Carbon monoxide, carbon dioxide, and water are discarded.

Vapor-phase nitrations of methane and ethane produce relatively low yields of nitroparaffins. These two paraffins contain only primary C–H bonds, which are less reactive than the secondary C–H bonds in propane, heavier normal paraffins, and cycloparaffins.

Gaseous nitrogen dioxide was previously used to commercially nitrate propane. This nitration could be run at lower temperatures, such as 220–250°C. Only 20–27% of the nitrogen dioxide, however, was converted to the four nitroparaffins. Most of the remaining nitrogen dioxide acted as an oxidizing agent.

Propane and other paraffins can also be nitrated in a liquid-phase process using a dispersion of liquid paraffins and liquid nitric acid. Relatively low temperatures (140–200°C) are used, and sufficiently high pressures are required to maintain the reactants as liquids. Dinitroparaffins, such as 2,2-dinitropropane, can be produced; they have never been produced in a vapor-phase process probably because of the higher temperature. See NITROPARAFFIN.

Victor Meyer and other processes. In the classical Victor Meyer process, an organic halide (often a bromide) is reacted with silver nitrite to produce a nitrohydrocarbon and silver halide. In a modified process, sodium nitrite, dissolved in a suitable solvent, is substituted for the more expensive silver nitrite. The desired nitroalkanes are produced in high yields by these processes, whereas they are produced in rather low yields in free-radical nitrations.

Nitrations can also often be performed by addition reactions using unsaturated hydrocarbons with nitric acid or nitrogen dioxide.

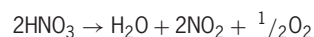
Lyle F. Albright

Bibliography. L. F. Albright, *Chem. Eng.*, pp. 149–156, June 6, 1966; L. F. Albright, R. V. C. Carr, and R. J. Schmitt, Nitration: Recent Laboratory and Industrial Developments, *ACS Symp. Ser.*, no. 623, American Chemical Society, Washington, DC, 1996; A. A. Guenkel, J. M. Rae, and E. G. Hauptmann, U.S. Pat. 5,313,009, May 14, 1994.

Nitric acid

A strong mineral acid having the formula HNO_3 . Pure nitric acid is a colorless liquid with a density of 1.52 at 25°C (77°F); it freezes at -47°C (-53°F). Nitric acid is used in the manufacture of ammonium nitrate and phosphate fertilizers, nitro explosives, plastics, dyes, and lacquers. The principal commercial process for the manufacture of nitric acid is the Ostwald process, in which ammonia, NH_3 , is catalytically oxidized with air to form nitrogen dioxide, NO_2 . When the dioxide is dissolved in water, 60% nitric acid is formed. Production of 90–100% nitric acid is based on processes such as the reaction of sulfuric acid with sodium nitrate (an older method of nitric acid manufacture), dehydration of 60% acid, and oxidation of nitrogen dioxide in a solution of dilute nitric acid.

Nitric acid decomposes readily as shown in the reaction below. It is a strong oxidizing agent, oxidizing



carbon to carbon dioxide, sulfur to sulfuric acid, and phosphorus to phosphoric acid. It reacts with most metals; products depend on the metal's electromotive series position and nitric acid concentration. See AMMONIA; NITROGEN; NITROGEN OXIDES; OXIDIZING AGENT.

Francis J. Johnston

Bibliography. G. Agam, *Industrial Chemicals, Their Characteristics and Development: Chemicals in the Real World*, 1994; F. A. Cotton et al., *Advanced Inorganic Chemistry*, 6th ed., 1999.

Nitric oxide

An important messenger molecule in mammals and other animals. It can be toxic or beneficial, depending upon the amount and where in the body it is released. Initial research into the chemistry of nitric oxide (NO) was motivated by its production in automobile emissions and other combustion processes, which results in photochemical smog and acid rain. In the late 1980s, researchers in immunology, cardiovascular pharmacology, neurobiology, and toxicology discovered that nitric oxide is a crucial physiological messenger molecule. Nitric oxide is now thought to play a role in blood pressure regulation, control of blood clotting, immune defense, digestion, neuronal signaling, the senses of sight and smell, and possibly learning and memory. Underproduction or unregulated overproduction of nitric oxide may also contribute to disease processes such as diabetes, atherosclerosis, stroke, hypertension, carcinogenesis, multiple sclerosis, transplant rejection, damage associated with reperfusion in ischemic (oxygen-deprived) tissues, impotence, septic shock, and long-term depression.

Most cellular messengers are large, unreactive biomolecules that make specific contacts with their targets. In contrast, nitric oxide is a small molecule that can diffuse freely throughout biological tissues,

reacting with target molecules at remote sites that are as far away as ~ 300 micrometers from its site of generation, a distance corresponding to 10–20 cell lengths. Nitric oxide is a free radical—that is, it contains an unpaired electron. However, unlike most free radicals, it is remarkably unreactive toward other compounds, and its chemistry is limited primarily to bond-forming combination with other radicals and coordination to metal centers. This contrasts markedly with the other common nitrogen oxide free radical, nitrogen dioxide (NO_2), which is reactive toward a wide variety of organic compounds and is frequently used in large-scale commercial syntheses. This dramatic difference in reactivity has been attributed to the capacity for NO_2 to dimerize to dinitrogen tetroxide (N_2O_4) and undergo autoionization to nitrate (NO_3^-) and the very reactive nitrosonium (NO^+) ion. In contrast, NO does not dimerize and its autoionization to NO^+ and the nitroxyl anion (NO^-) is energetically highly unfavorable, precluding formation of these reactive species. The chemical inertness of NO extends to the biological components of cells, where its direct reactions are limited primarily to coordination of transition metals such as iron centers in heme (iron porphyrin-containing) proteins and reaction with the diradical O_2 , O_2 -derived free radicals, and organic free radicals that are formed upon oxidation of various biomolecules. See FREE RADICAL; NITRO AND NITROSO COMPOUNDS; NITROGEN; NITROGEN OXIDES; PORPHYRIN.

Production. Nitric oxide is produced in the body by enzymes called nitric oxide synthases, which convert the amino acid L-arginine to nitric oxide and L-citrulline (Fig. 1). There are three recognized types of nitric oxide synthase, often designated nNOS, eNOS, and iNOS for the neuronal, endothelial, and inducible forms, respectively. Both nNOS and eNOS are constitutive—that is, they are always present in cells—while iNOS is normally not present but is synthesized in response to biochemical signals generated within the cellular environment. Nitric oxide synthases are structurally organized into two domains called the reductase and oxygenase domains. The reductase domain contains two redox-active flavin cofactors (FAD and FMN) that deliver electrons one at a time from the physiological electron donor (NADPH) to the oxygenase domain that

contains the active site for arginine oxidation. The oxygenase domain contains a biologically unique combination of heme and tetrahydrobiopterin (H_4B) cofactors. Oxidation of arginine to citrulline plus NO requires two cycles of oxidation with O_2 (Fig. 1). Recent structural and kinetic evidence strongly suggests that H_4B plays essential roles in both cycles to deliver electrons to reactive oxo-heme intermediates, thereby directing substrate oxidation and minimizing formation of unwanted reactive oxygen species, such as superoxide ion (O_2^-) and nitroxyl anion (NO^-), as side products. A calcium-calmodulin binding site is located between the two domains. The constitutive enzymes, nNOS and eNOS, are activated by Ca^{2+} -calmodulin binding at this site, which in turn is regulated by the amount of Ca^{2+} in the medium. In contrast, Ca^{2+} -calmodulin binds so tightly to iNOS that this enzyme is effectively “turned on” under all physiological conditions. The practical consequence of this difference in binding affinities is that the constitutive enzymes generate only low levels of NO, whereas, once its formation has been induced, iNOS generates relatively massive amounts of NO. These differences in enzymatic activity are consistent with the presumed functions of the enzymes, which are secondary messenger generation and cytotoxin production, respectively. See AMINO ACIDS; ENZYME.

Blood pressure regulation. In the cardiovascular system, NO is produced by eNOS located in the endothelial layer, the cells that line the blood vessels. The activation of eNOS is triggered by the binding of hormones to receptors on the surface of these cells, opening up calcium channels that release Ca^{2+} into the endothelial cells. The Ca^{2+} then attaches to the cofactor calmodulin, which binds to and activates eNOS. The NO that is produced diffuses to the smooth muscle and binds to the iron atom of the heme group in an enzyme known as soluble guanylyl cyclase (sGC). This enzyme converts guanosine triphosphate to cyclic guanosine monophosphate and pyrophosphate within cells. The binding of NO to the heme of sGC leads to a structural change in the enzyme which causes a dramatic increase in the production of cyclic guanosine monophosphate (cGMP). It is thought that the cGMP activates a protein kinase (an enzyme that adds a phosphate

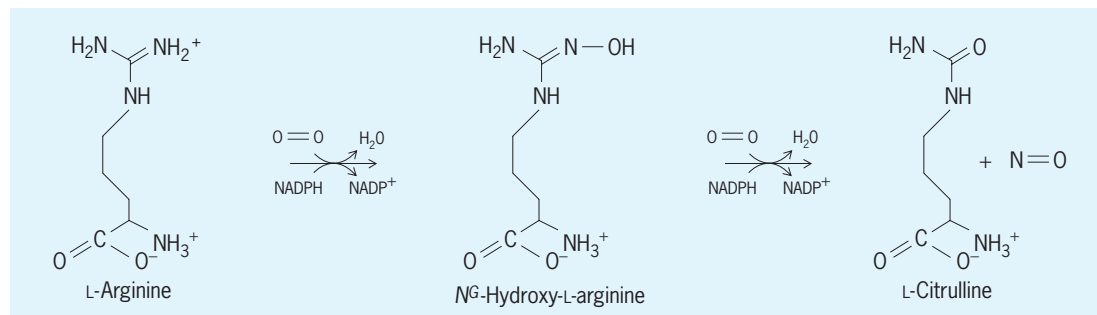


Fig. 1. Conversion of L-arginine to L-citrulline and nitric oxide (NO) catalyzed by the enzyme nitric oxide synthase. (After P. L. Feldman et al., *The surprising life of nitric oxide*, C&EN News, 71(51):26–38, 1993)

group to another protein) that phosphorylates calcium transporters, causing Ca^{2+} to be returned to its cellular storehouse within the smooth muscle cells. The decrease in calcium levels causes relaxation of blood vessels since the muscle cells require Ca^{2+} for contraction. The treatment of hypertension, heart attacks, and other blood pressure abnormalities often involves using chemical compounds that can release NO in the body. These compounds are thought to bypass the nitric oxide synthase pathway and directly release NO into smooth muscles, causing relaxation and thus decreasing the blood pressure. Direct inhalation of gas mixtures containing low levels of NO has also been used effectively to combat pulmonary hypertension in neonatal and adult respiratory distress syndrome patients. *See* CARDIOVASCULAR SYSTEM.

The protein hemoglobin, the O_2 carrier in the blood, has also been implicated in NO transport and regulation of blood pressure. Hemoglobin contains four heme groups that function as O_2 -binding sites. Nitric oxide reacts with hemoglobin in several distinct ways. First, it reacts rapidly with O_2 bound to the iron atom of the heme to form NO_3^- and methemoglobin, an oxidized form of hemoglobin which cannot bind O_2 . Second, the iron atom of the heme in hemoglobin can bind NO when O_2 is not bound. Since hemoglobin is present in high concentrations in the vascular system and hemoglobin can effectively scavenge NO, it has been a mystery how enough NO can reach the smooth muscle cells to cause blood pressure relaxation. There is now evidence that hemoglobin may protect NO by forming S-nitrosothiols (RS-NO compounds, where R represents a hydrocarbon group). Research suggests that two conserved cysteine residues in hemoglobin react with nitric oxide to form nitrosothiols, preventing the oxidation of NO to NO_3^- . The NO could then be released to the smooth muscle cells, causing blood vessel relaxation. Nonetheless, the physiological relevance of protein S-nitrosothiols is controversial, and the issue of its importance as a delivery pathway for NO is unresolved. *See* HEMOGLOBIN.

Neurotransmitter. Within the central nervous system, nitric oxide synthase (nNOS) is found only in discrete populations of neuronal cells in brain tissue, and is not found in the glial cells that make up the major portion (85%) of the cellular mass. Although specific physiological roles for nNOS-containing neurons have not been established, nNOS is activated in some cells by Ca^{2+} -calmodulin binding during neurotransmission. In this case, glutamate released from a stimulated (presynaptic) neuron diffuses to binding sites on an unstimulated adjacent (postsynaptic) neuron resulting in influx of Ca^{2+} into the neuron, increasing as well the Ca^{2+} -calmodulin concentration and activating nNOS. The NO formed appears to promote additional release of neurotransmitters and activates sGC to form the secondary messenger cGMP. It has been suggested that these reactions lead to changes in the neuronal pathways that contribute to development of memory. The overproduction of nitric oxide in brain tissues has

been implicated in stroke and other neurological problems initiated by oxidative damage to deoxyribonucleic acid (DNA). *See* DEOXYRIBONUCLEIC ACID (DNA).

The peripheral nervous system also uses the production of NO in neurons to regulate blood flow. Neurons that are sensitive to NO but unresponsive to conventional neurotransmitters have been found in many peripheral tissues, including those of the cardiovascular, urogenital, respiratory, and digestive systems. Nitric oxide has also been implicated in vision and the sense of smell. Penile erection is mediated by NO production in peripheral neurons, causing relaxation of smooth muscles in that organ. Thus, drugs that stimulate release of NO have been developed as a noninvasive way of treating impotence. *See* NEUROBIOLOGY.

Cytotoxicity and cytoprotection. Macrophages are phagocytic (devouring) cells within the immune system whose primary functions are to ingest and destroy invading microbes, kill targeted tissue cells, and digest cellular debris. They also contain iNOS, whose biosynthesis is induced by small molecules derived from the microbes and/or other host cells. Nitric oxide released by macrophages can inhibit cellular processes in nearby cells, including DNA synthesis and respiration, as well as selective inhibition of intracellular enzymes involved in other essential metabolic pathways. Primary target sites in these biomolecules include nonheme iron-sulfur clusters, sulfhydryl substituents of cysteine amino acids at metal-binding and enzyme active sites, and the phenolic substituent of tyrosine, as well as various sites within DNA. The resulting damage is sufficient to induce death in many types of cells by either of the two recognized pathways—programmed cell death (apoptosis) or necrosis. The extent to which these reactions are caused by direct reaction of the biomolecules by NO or by secondary oxidants derived from NO is an area of active ongoing research. It appears that most of the oxidative damage mediated by NO is actually caused by secondary oxidants, which is consistent with the limited chemical reactivity displayed by NO. Compounds thought to be major contributors are indicated in **Fig. 2**. These include (1) dinitrogen trioxide (N_2O_3), a potent nitrosating (that is, NO^+ -transferring) agent that is formed during reaction of NO with O_2 (a diradical); (2) the peroxynitrite ion (ONOO^-), which is formed by radical coupling of NO and O_2^- , and when protonated or in the presence of carbon dioxide undergoes homolytic cleavage of its peroxy O-O bond to give the strongly oxidizing radicals NO_2 and OH or CO_3^- , respectively;

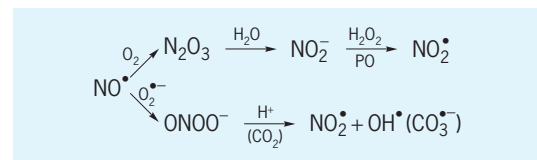


Fig. 2. Secondary oxidants formed upon reaction of NO with O_2 and superoxide ion. PO = peroxidase enzyme.

(3) the nitrite anion (NO_2^-), which can be oxidized to NO_2 radical by H_2O_2 in reactions catalyzed by peroxidase enzymes. Additional potential oxidants are metal nitrosyl complexes, which could act as nitrosating agents or could release nitroxyl (NO^-) to the medium, where it could react rapidly with O_2 to form peroxynitrite. Confounding attempts to identify the origins of cellular oxidative damage caused by NO is the recognition that under some circumstances NO can also protect cells from damage by other oxidants. For example, NO has been shown to inhibit lipid peroxidation, most likely by coupling with reactive lipid peroxy radical intermediates, and thereby terminating membrane-destroying radical chain reactions with O_2 . See APOPTOSIS; IMMUNOLOGIC CYTOTOXICITY; PEROXYNITRITE.

Although nitric oxide production in the immune system serves a crucial biological function, there can be adverse effects when too much NO is produced. During a massive bacterial infection, excess NO can go into the vascular system, causing a dramatic decrease in blood pressure, which may lead to fatal septic shock. Thus, scientists are working on drugs that can selectively inhibit iNOS in order to avoid the harmful effects produced by excess NO without interfering with nitric oxide signaling pathways. See IMMUNOLOGY.

James K. Hurst; Judith N. Burstyn; Mark F. Reynolds

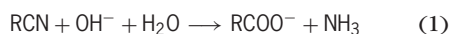
Bibliography. P. L. Feldman, O. W. Griffith, and D. J. Stuehr, The surprising life of nitric oxide, *Chem. Eng. News*, 71(51):26-38, 1993; R. Rawls, Bioinorganic reactions of nitric oxide underlie diverse roles in living systems, *Chem. Eng. News*, 74(19):38-42, 1996; S. H. Snyder and D. S. Bredt, Biological roles of nitric oxide, *Sci. Amer.*, 266(5):68-77, 1992.

Nitrile

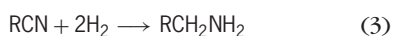
One of a group of organic chemical compounds of general formula $\text{RC}\equiv\text{N}$. A nitrile is named from the acid to which it can be hydrolyzed by adding the suffix -nitrile to the acid stem, for example, acetonitrile from acetic acid. An alternative system names the group attached to CN , thus CH_3CN is also named methyl cyanide. In more complex structures the CN group is named as a substituent, cyano.

Nitriles may be identified by a distinctive weak-to-medium infrared absorption band due to triple-bond stretching at $2260\text{--}2222\text{ cm}^{-1}$.

Nitriles are hydrolyzed to acids in either basic or acidic solution, as shown in reactions (1) and (2).

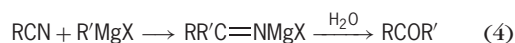


By catalytic hydrogenation in the presence of nickel or cobalt, nitriles are reduced to primary amines, reaction (3). Nitriles are also reduced to pri-

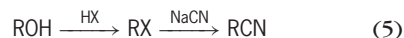


mary amines by lithium aluminum hydride.

Grignard reagents added to nitriles give ketones (after hydrolysis), as shown in reaction (4).

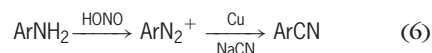


The formation of cyanides from alkyl halides is an important chain-lengthening reaction in organic synthesis. The starting compound is most often an alcohol, and the sequence of reactions is shown in (5).

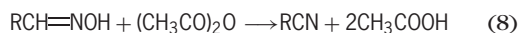
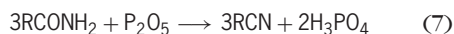


The reaction is practical only with primary aliphatic halides, since the alkali cyanides are fairly strong bases and eliminate HX from secondary or tertiary alkyl halides.

Aromatic nitriles are made by displacement of a diazotized primary amino group with the cyanide group in the presence of copper cyanide or copper powder, reaction (6).



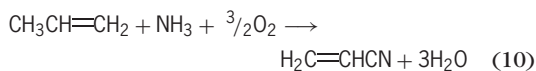
The dehydration of acid amides or oximes with phosphorus pentoxide or acetic anhydride in either the aliphatic or aromatic series serves as another preparative method, reactions (7) and (8).



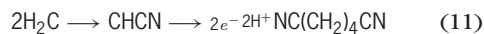
Industrially, nitriles are formed by heating carboxylic acids with ammonia and a dehydration catalyst under pressure. The amide is an intermediate but need not be isolated, reaction (9).



For the preparation of acrylonitrile, which is used on a large scale in the plastics industry, a vapor-phase catalytic ammoxidation of propylene has been developed, reaction (10).



Acrylonitrile is electrochemically coupled "tail to tail" to give the hydrodimer adiponitrile, which is a key intermediate in the commercial production of nylon-6,6, reaction (11).



See ACRYLONITRILE; AMINE; CARBOXYLIC ACID; OXIME; POLYAMIDE RESINS.

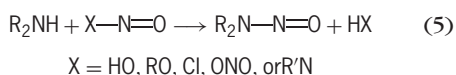
Paul E. Fanta
Bibliography. S. N. Ege, *Organic Chemistry*, 4th ed., 1998; R. T. Morrison and N. Boyd, *Organic Chemistry*, 6th ed., 1992.

Nitro and nitroso compounds

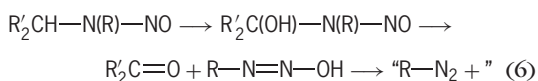
Derivatives of organic hydrocarbons having one or more $-\text{NO}_2$ groups bonded via nitrogen to the carbon framework (nitro compounds) or an $-\text{NO}$ group attached to carbon or nitrogen (nitroso compounds).

Consequently, aryl nitroso compounds (for example, nitrosobenzene) are largely monomeric in solution. Almost all other C-nitroso compounds are dimeric (and colorless) in the solid state. See CONJUGATION AND HYPERCONJUGATION; ENTHALPY; ENTROPY; FREE ENERGY; MOLECULAR ISOMERISM.

N-nitroso compounds (*N*-nitrosamines) have the general structure R^aR^bN-NO . Secondary *N*-nitrosamines are moderately stable, monomeric, yellow substances. The structure is intermediate between R^aR^bN-NO and $R^aR^bN^+=N-O^-$. *N*-nitroso compounds are easily formed by reaction of a secondary amine with various nitrosating agents [reaction (5). Many *N*-nitroso compounds are



strong carcinogens, and some occur widely in the environment. The carcinogenicity is thought to be associated with biological conversion (alpha-hydroxylation) of an *N*-nitroso compound into an alkylating agent as in reaction scheme (6), followed



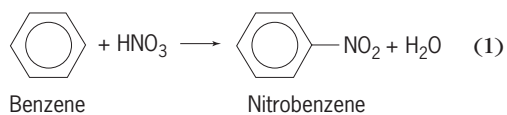
by alkylation of nucleophilic sites of cellular constituents such as the N-7 and O-6 positions of guanine in deoxyribonucleic acid (DNA). See DEOXYRIBONUCLEIC ACID (DNA); NITROGEN.

Frederick D. Greene

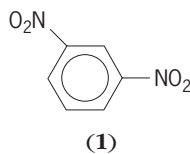
Bibliography. A. G. M. Barrett and G. G. Graboski, Conjugated nitroalkenes: Versatile intermediates in organic synthesis, *Chem. Rev.*, 86:751, 1986; D. H. Barton, *New Principles in Organic Chemistry: The Discovery and Invention of Chemical Reactions*, 1995; D. H. R. Barton and W. D. Ollis (eds.), *Comprehensive Organic Chemistry*, vol. 2, 1979; J. March, *Advanced Organic Chemistry: Reactions, Mechanisms, and Structures*, 4th ed., 1992; S. Patai (ed.), *The Chemistry of Functional Groups*, suppl. F, *The Chemistry of Amino, Nitroso, and Nitro Groups and Their Derivatives*, pt. 1 and 2, 1982; S. Patai (ed.), *The Chemistry of Nitro and Nitroso Groups*, pt. 1, 1969, pt. 2, 1970.

Nitroaromatic compound

A member of the class of organic compounds in which the nitro group ($-NO_2$) is attached directly to the cyclic, aromatic nucleus. The prototypal compound is nitrobenzene, which was first synthesized in 1834 and produced commercially in England in 1856. It is prepared by the reaction of benzene with nitric acid in the presence of sulfuric acid, as shown in reaction (1). The reaction is very efficient, with



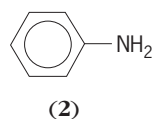
yields approaching 100%. It is an exothermic reaction, and careful control of the temperature is necessary to avoid the formation of *m*-dinitrobenzene (1).



See BENZENE.

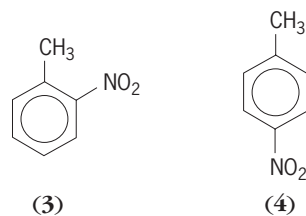
Nitrobenzene is a pale-yellow liquid with a characteristic sweet but unpleasant odor. It boils at 211°C (412°F), and therefore it has a fairly high vapor pressure at room temperature. It is a very toxic substance that is absorbed by contact of the liquid with the skin or by inhalation of the vapor. In the body it reacts with the red blood cells to form methemoglobin, resulting in cyanosis. Chronic exposure can lead to spleen and liver damage, jaundice, and anemia.

The most significant use of nitrobenzene is in the manufacture of aniline (2). About 97% of the ni-

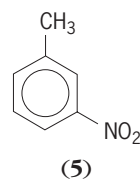


trobenzene produced in the United States is converted to aniline, which is used in the manufacture of plastics, rubber additives, dyes, drugs, and other products.

Second to nitrobenzene in commercial importance are the mononitrotoluenes, particularly the ortho and para isomers, (3) and (4), respectively. Re-

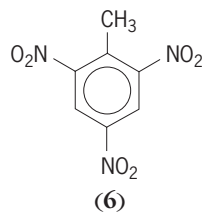


action of toluene with a mixture of nitric and sulfuric acid at about 40°C (104°F) gives a high yield of a mixture of the three isomers containing about 58% ortho, 4% meta (5), and 38% para. The isomers



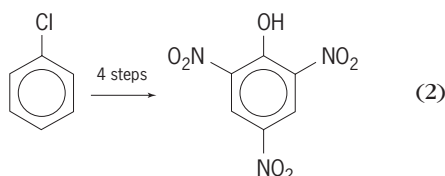
are separated by a combination of fractional distillation and crystallization. The nitrotoluenes are important intermediates in the preparation of dyes, rubber chemicals, and agricultural chemicals.

2,4,6-Trinitrotoluene (TNT; **6**) is a military explo-



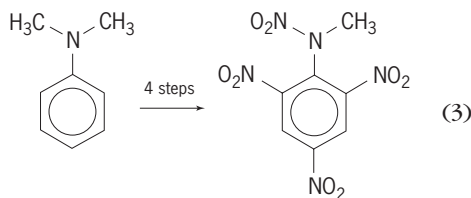
sive that is stable, nonhygroscopic, and relatively insensitive to impact, friction, shock, and electric spark. It is produced by nitration of toluene in successive stages at progressively higher temperatures and concentrations of acid. It can be melted safely at 80°C (176°F) for cast loading of military projectiles. See EXPLOSIVE.

Picric acid (2,4,6-trinitrophenol) is an example of a nitroaromatic compound that cannot be prepared by direct nitration, since phenol is sensitive to oxidation. The commercial method consists of a series of reactions starting with chlorobenzene, including the hydrolysis of the chlorine to give the desired hydroxyl group, reaction (2). Picric acid has been used

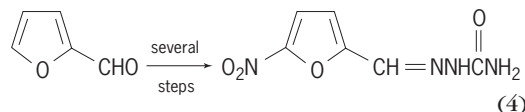


as a dye, an explosive, and a bactericide.

Similarly, the useful explosive tetryl is formed by a sequence of nitration steps starting with *N,N*-dimethylaniline and including the replacement of one of the methyl groups by the nitro group, reaction (3).



Furacin (5-nitro-2-furaldehyde semicarbazone) is an example of a heterocyclic nitroaromatic compound that has attained commercial success. It is prepared in several steps from readily available furfural, reaction (4), and is an antibacterial agent that



is particularly effective against poultry coccidiosis. See HETEROCYCLIC COMPOUNDS.

Although literally thousands of other aromatic ring compounds, including the heterocyclics, have been converted to their nitro derivatives, few such compounds have achieved any significant industrial importance. See AROMATIC HYDROCARBON; NITRATION.

Paul E. Fanta

Bibliography. F. A. Carey and R. J. Sundberg, *Advanced Organic Chemistry*, pt. A: *Structure and Mechanisms*, 3d ed., 1990; L. F. Fieser and M. Fieser, *Advanced Organic Chemistry*, 1961; *Kirk-Othmer Encyclopedia of Chemical Technology*, 4th ed., 1995; *Merck Index*, 11th ed., 1989.

Nitrogen

A chemical element, N, atomic number 7, atomic weight 14.0067. Nitrogen, a gas under normal conditions, is the lightest element of periodic group 5 (nitrogen family). See PERIODIC TABLE.

| | | | | | | | | | | | | | | | | | |
|----|----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|----|----|----|----|----|
| 1 | 2 | | | | | | | | | | | 13 | 14 | 15 | 16 | 17 | 18 |
| 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 | | | | | |
| 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 |

At standard temperature and pressure, elemental nitrogen exists as a gas with a density of 1.25046 g/liter. This value indicates that the molecular formula is N₂. Some physical properties of elemental nitrogen are listed in **Table 1**.

Elemental nitrogen has a low reactivity toward most common substances at ordinary temperatures. At high temperatures, molecular nitrogen, N₂, reacts with chromium, silicon, titanium, aluminum, boron, beryllium, magnesium, barium, strontium, calcium, and lithium (but not the other alkali metals) to form nitrides; with O₂ to form NO; and at moderately high temperatures and pressures in the presence of a catalyst, with hydrogen to form ammonia. Above 1800°C (3300°F), nitrogen, carbon, and hydrogen combine to form hydrogen cyanide.

Table 2 lists the principal classes of inorganic nitrogen compounds. Thus, in addition to the typical oxidation states of the family (−3, +3, and +5), nitrogen forms compounds with a variety of additional

TABLE 1. Properties of nitrogen

| Property | Value |
|---|--|
| Heat of transformation (α - β) | 54.71 cal/mole |
| Heat of fusion | 172.3 cal/mole |
| Heat of vaporization | 1332.9 cal/mole |
| Critical temperature | 126.26 \pm 0.04 K |
| Critical pressure | 33.54 \pm 0.02 atm |
| Density: α form | 1.0265 g/ml at -252.6°C |
| β form | 0.8792 g/ml at -210.0°C |
| Liquid | 1.1607–0.00457 ($T = \text{abs temp}$) |

TABLE 2. Compounds of nitrogen

| Oxidation state | Examples |
|-----------------|---|
| +5 | N ₂ O ₅ , HNO ₃ , nitrates, NO ₂ X |
| +4 | N ₂ O ₄ ⇌ 2NO ₂ |
| +3 | N ₂ O ₃ , HNO ₂ , nitrites, NOX, NX ₃ |
| +2 | NO, Na ₂ NO ₂ , nitrohydroxylamates |
| +1 | N ₂ O, H ₂ N ₂ O ₂ , hyponitrites |
| 0 | N ₂ |
| -1/3 | HN ₃ , acids |
| -1 | NH ₂ OH, hydroxylammonium salts |
| -2 | NH ₂ NH ₂ , hydrazinium salts, hydrazides |
| -3 | NH ₃ , ammonium salts, amides, imides, nitrides |

oxidation states. See AMINE; AMMONIA; HYDRAZINE; NITRIC ACID; NITROGEN COMPLEXES; NITROGEN OXIDES.

Molecular nitrogen is the principal constituent of the atmosphere (78% by volume of dry air), in which its concentration is a result of the balance between the fixation of atmospheric nitrogen by bacterial, electrical (lightning), and chemical (industrial) action, and its liberation through the decomposition of organic materials by bacteria or combustion. In the combined state, nitrogen occurs in a variety of forms. It is a constituent of all proteins (both plant and animal) as well as of many other organic materials. Its chief mineral source is sodium nitrate.

The methods for the preparation of elementary nitrogen may be grouped into two classes, separation from the atmosphere and decomposition of nitrogen compounds. The industrial method for the production of nitrogen is the fractional distillation of liquid air. Nitrogen containing about 1% argon and traces of other inert gases may be obtained by the chemical removal of oxygen, carbon dioxide, and water vapor from the atmosphere by appropriate chemical reagents.

Because the importance of nitrogen compounds in agriculture and chemical industry, much of the industrial interest in elementary nitrogen has been in processes for converting elemental nitrogen into nitrogen compounds. The principal methods for doing this are the Haber process for the direct synthesis of ammonia from nitrogen and hydrogen, the electric arc process, which involves the direct combination of N₂ and O₂ to nitric oxide, and the cyanamide process. Nitrogen is also used for filling bulbs of incandescent lamps and, in general, wherever a relatively inert atmosphere is required. Harry H. Sisler

Bibliography. T. Chivers, *A Guide To Chalcogen-nitrogen Chemistry*, 2004; J. S. Clark (ed.), *Nitrogen, Oxygen and Sulfur Ylide Chemistry*, 2002; F. A. Cotton et al., *Advanced Inorganic Chemistry*, 6th ed., 1999.

Nitrogen complexes

Compounds containing the dinitrogen molecule, N₂, bound to a metal (also called dinitrogen complexes). While it was long known that molecular nitrogen avidly binds to the surface atoms of many metallic

phases, no reaction of molecular nitrogen has been observed under the conditions of ordinary solution chemistry. Indeed, chemists had been so accustomed to regard nitrogen as a completely inert gas at ordinary temperatures that the formation of coordination compounds with molecular nitrogen aroused general surprise when first reported in the mid-1960s. Since then the interaction of molecular nitrogen with coordination compounds has been the subject of intensive research. Today the capability of molecular nitrogen to enter into coordination compounds as a ligand is established beyond any doubt, and it is at least partly understood under which conditions the dinitrogen molecule can lose its customary inertness. See COORDINATION CHEMISTRY.

Formation. Outstanding in their ability to form coordination compounds with nitrogen are a number of metals which belong to the group 18 transition metal family (Fig. 1). For each metal of this group, several nitrogen complexes have been identified. Nitrogen complexes of these metals occur in low oxidation states, such as Co(I) or Ni(0). The other ligands present in these complexes besides N₂ are usually of a type known to stabilize low oxidation states; phosphines appear to be particularly prominent bonding partners in this respect. Figure 2 shows the structure of a typical N₂ complex, elucidated by a crystal structure determination. The N-N bond axis in this complex is aimed, within the limits of experimental error, directly toward the position of the metal atom. The Co-N₂ bond length, 0.18 nanometer, is within the normal range of comparable metal-ligand bonds.

| | | | | | | | | | |
|----|----|----|----|----|----|----|----|----|----|
| Sc | Ti | V | Cr | Mn | Fe | Co | Ni | Cu | Zn |
| Y | Zr | Nb | Mo | Tc | Ru | Rh | Pd | Ag | Cd |
| La | Hf | Ta | W | Re | Os | Ir | Pt | Au | Hg |

Fig. 1. Transition metals which form coordination compounds with N₂ as ligand are mostly members of groups shown in colored squares. Coordination compounds which reduce N₂ to NH₃ in approximately stoichiometric yields are derived from the metals in the left half of the transition series (circles).

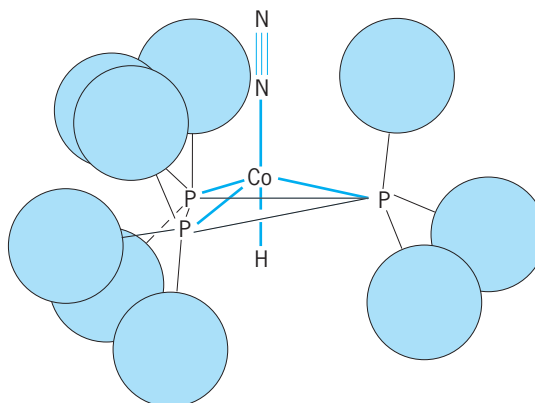


Fig. 2. Structure of a coordination compound with N₂ (circles represent phenyl groups).

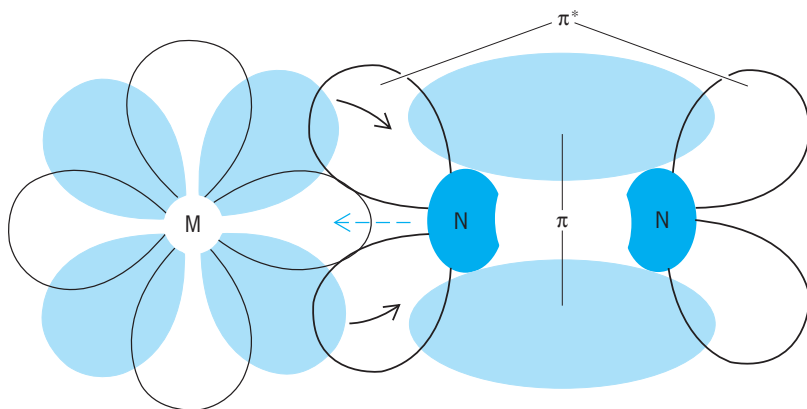


Fig. 3. Factors contributing to the stability of N_2 coordination compounds: dative bond from an electron pair on one of the N atoms to a vacant metal orbital (broken arrow) and the opposite process of back-donation of loosely held metal electrons to the π^* acceptor orbital of N_2 (solid arrows).

The question arises as to the kind of bonding force that is able to attract the neutral, dipole-free, and not easily polarizable N_2 molecule to its position in the coordination sphere of the metal. The usual kind of dative bond between the electron pair on one of the N atoms and an acceptor orbital of the metal (Fig. 3) is likely to be very weak, just as the basicity of N_2 is unmeasurably low. A second and probably more important contribution to the stability of nitrogen complexes is thought to arise from the opposite process—from back-donation of loosely held metal electrons into an acceptor orbital of the N_2 molecule. Such a shift of electron density into the normally unoccupied, antibonding π^* orbitals of the ligand is considered to be an important stabilizing factor in many coordination compounds with unsaturated ligands. While strengthening the metal-ligand bond, the flow of electrons from the metal into these antibonding ligand orbitals would necessarily cause a weakening of the bond between the two N atoms, and this is exactly what is observed. Without exception, the stretching vibration of coordinated nitrogen occurs at a substantially lower frequency than that of free, gaseous nitrogen. Moreover, the known prerequisites for an efficient back-donation explain at least some of the peculiarities of nitrogen coordination. A large number of d electrons, usually six or more, has to be available on the metal; in addition, the oxidation state of the metal has to be low so that these electrons are easily released toward the ligand. Accordingly, it is understandable that group VIII metals with their large number of d electrons dominate among the known N_2 complexes.

In all the respects discussed above, the N_2 ligand exhibits a strong resemblance to the well-known ligand molecule carbon monoxide, CO, which accommodates the same number of electrons in orbitals of the same type as N_2 . Although N_2 by virtue of its symmetry is a much weaker coordinating agent than the lopsided CO molecule, there is no doubt that these two molecules can under certain conditions substitute for each other as ligands. The weaker ligand N_2 appears to be more selective, however. Other ligand partners in the complex, while having to accommo-

date enough of the metal's excess electrons to stabilize a low oxidation state, should not compete too strongly with N_2 for back-donation of the metal electrons. Phosphines, which are known to draw less heavily on metal d electrons than other π acceptors, seem to strike this delicate balance right. That this is not the sole possibility to obtain an appropriate electron distribution, however, is demonstrated by the stability of the complex $[Ru(N_2)(NH_3)_5]^{2+}$, where five NH_3 ligands "condition" the Ru(II) center.

A number of dinitrogen complexes with entirely different structures have been isolated or characterized in solutions and solid matrices. These complexes appear to have their N_2 ligand molecule bound to the metal in an "edge-on" fashion (similar to that in Fig. 4, with both of its N atoms at the same bonding distance from the central metal atom). The general conditions for the occurrence of this type of N_2 complex are not clear at present.

Chemical reactions. Even in most favorable cases the binding of the dinitrogen molecule to the metal is fairly labile; all the compounds lose their nitrogen on mild heating. Some of the nitrogen complexes are only metastable to loss of dinitrogen even at room temperature; accordingly, they cannot be obtained by direct uptake of gaseous nitrogen. In the synthesis of these metastable complexes, hydrazine or azide compounds serve as a source of nitrogen molecules within the coordination sphere of the metal. Addition of other coordinating agents to the nitrogen complexes usually results in a displacement of N_2 from the metal. The cobalt compound in Fig. 2 exchanges its N_2 ligand quite reversibly for other ligand molecules, such as NH_3 and $H_2C=CH_2$. Whereas these ligands are easily displaced again by an excess of N_2 , an irreversible exchange occurs with carbon monoxide. The bulky organic groups on the phosphine ligands are likely to interfere with the approach to the metal of all but the slimmest ligands and thereby help the "thin" dinitrogen molecule to maintain or regain its position on the metal in competition with most other ligands.

An interesting question concerns the reactivity of the coordinated nitrogen molecule. Is the weakening of the bond in $N\equiv N$, which accompanies coordination and back-donation, sufficient to render the otherwise inert nitrogen molecule susceptible to attack and cleavage? Much systematic research has yet to be done if scientists are to understand the perplexing chemical reaction paths involved in nitrogen fixation.

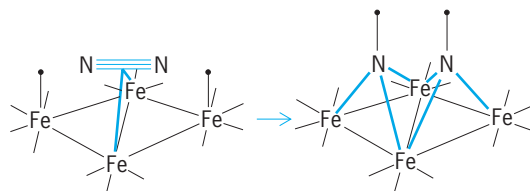


Fig. 4. Conceivable reaction path for the conversion of a surface-bound N_2 molecule to two surface nitride particles. Surface hydrogen may be involved in this reaction (indicated by black dots).

The conversion of inert, molecular nitrogen to ammonium salts, which can then be utilized by plants in their production of proteins and other indispensable nutrients, is one of the limiting factors of life on Earth. Nature has put the burden of this life-sustaining process on the few strains of microorganisms which possess an enzyme capable of reacting with nitrogen. The enzyme, nitrogenase, actually consists of two protein particles, both of which contain iron in a peculiar form associated with sulfide; one of the enzyme particles also contains molybdenum. There is some support for the hypothesis that this latter particle forms a complex with molecular nitrogen, while the other molybdenum-free enzyme particle serves as a sort of energizer, accepting reducing electron equivalents from ferredoxin and making them available for the N_2 -reduction process at a substantially increased negative reduction potential at the expense of adenosine triphosphate hydrolysis. As for the chemical pathways by which the enzyme entrains its N_2 ligand into its smooth reduction to two NH_3 molecules, there is a still-open debate about the possible occurrence of enzyme-bound intermediates such as N_3H_2 (diimide) and N_2H_4 (hydrazine). See NITROGEN FIXATION.

Hardly any better understood is the process by which humans supplement natural nitrogen fixation—the production of fertilizer ammonia from N_2 and H_2 by a high-temperature reaction on metal catalysts. Bonding of nitrogen to the surface of some metals, such as nickel, is spectrally reminiscent of the complexes discussed above and probably involves the same type of end-on coordination of N_2 to surface atoms of the metal. In many other cases, however, there seem to be different types of surface coordination, perhaps with the $N\equiv N$ molecule parallel to the surface, which then can probably rearrange directly to two surface-bound nitride particles (Fig. 4). While many metals form this nitride type of surface compounds with nitrogen, the distinction of catalytically active metals is their tendency to detach the nitride particles from the surface again, together with surface hydrogen particles, to form gaseous NH_3 . The metal surface is thereby restored in its original form and is ready to renew the catalytic cycle by coordination of another N_2 molecule. Because of the complexity of these catalyst systems, some important questions have not been completely solved. It is not clear, for instance, whether the breakdown reaction of the coordinated N_2 molecule is just a rearrangement of bonds and electrons between the metal and nitrogen atoms, or if the participation of surface hydrides assists in this breakdown step, which might then involve adsorbed N_2H_2 or N_2H_4 intermediates.

A related problem is encountered with a series of reactions in which an N_2 ligand bound to some metal compound in solution is reduced to NH_3 or N_2H_4 . These reactions involve mostly transition metals that are located in the left half of the transition series (Fig. 1). N_2 reductions of this kind have been known since about 1965 to occur in the presence of strong reducing agents. Subsequently some reactions have been found where a mere transfer of protons to the

metal-bound (and hence presumably negatively polarized) N_2 molecule of an otherwise stable dinitrogen complex brings about reduction of this molecule to N_2H_4 or NH_3 , while leaving the metal behind in a correspondingly increased oxidation state. In view of their obvious relation to enzyme and surface catalysis, it would be highly interesting to establish the relevant intermediates of these reactions—for example, the bonding geometry of the N_2 ligand molecules involved—and the general conditions for the occurrence of N_2 reduction reactions of this kind. The possibility that reactive intermediates such as N_2H_2 or N_2H_4 derived from catalytic reactions of this type might be trapped into useful synthetic reactions is presently under exploration in many laboratories; its realization would be a step forward in the direction of making the vast supply of atmospheric nitrogen more efficiently available for utilization. See COORDINATION COMPLEXES; MOLECULAR ORBITAL THEORY; NITROGEN.

Hans Brintzinger

Bibliography. G. Henrici-Olive, *Coordination and Catalysis*, 1977; M. M. Khan and A. E. Martell, *Homogeneous Catalysis by Metal Complexes*, 2 vols., 1974.

Nitrogen cycle

The collective term given to the natural biological and chemical processes through which inorganic and organic nitrogen are interconverted. It includes the processes of ammonification, ammonia assimilation, nitrification, nitrate assimilation, nitrogen fixation, and denitrification.

Nitrogen exists in nature in several inorganic compounds, namely N_2 , N_2O , NH_3 , NO_2^- , and NO_3^- , and in several organic compounds such as amino acids, nucleotides, amino sugars, and vitamins. In the biosphere, biological and chemical reactions continually occur in which these nitrogenous compounds are converted from one form to another. These interconversions are of great importance in maintaining soil fertility and in preventing pollution of soil and water.

Nitrogen reserves exist in five major sinks: the primary rocks, the sedimentary rocks, the deep-sea sediment, the atmosphere, and the soil-water pool. Although primary rocks contain as much as 97.8% of the Earth's total (2.1×10^{17} tons or 1.9×10^{17} metric tons), their N_2 contributes little to the nitrogen cycle. Of the remaining nitrogen, 2% is in the atmosphere as N_2 , and about 0.2% is in sedimentary rocks. The biosphere, consisting of the soil-water pool, contains only a small portion of the Earth's total nitrogen (2.6×10^{13} tons or 2.4×10^{13} metric tons), and even here the predominant species (2.4×10^{13} tons or 2.2×10^{13} metric tons) is N_2 dissolved in the sea. In spite of this, it is in this soil-water pool that the major reactions of the nitrogen cycle occur.

An outline showing the general interconversions of nitrogenous compounds in the soil-water pool is presented in Fig. 1. The reactions are much more complex than in the outline, and biological agents

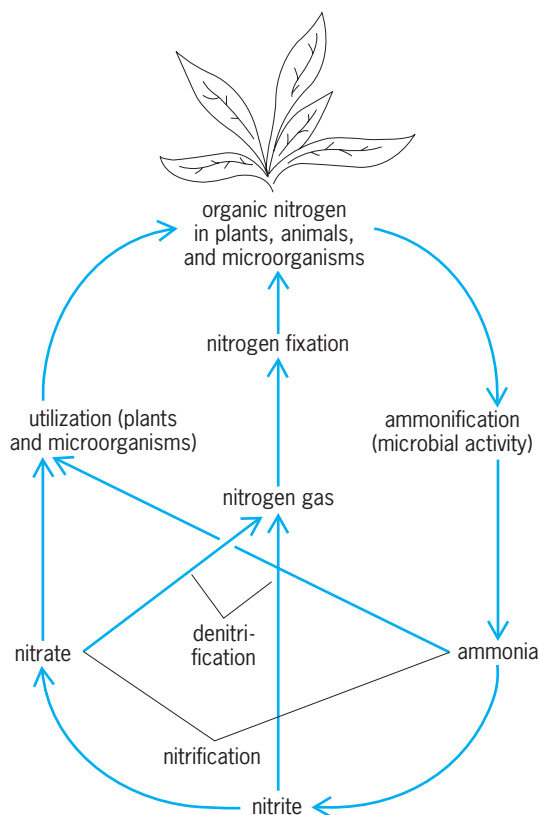


Fig. 1. Diagram of the nitrogen cycle.

have evolved intricate ways to manipulate these nitrogenous compounds for their own use. There are three primary reasons why organisms metabolize nitrogen compounds: (1) to use them as a nitrogen source, which means first converting them to NH_3 , (2) to use certain nitrogen compounds as an energy source such as in the oxidation of NH_3 to NO_2^- and of NO_2^- to NO_3^- , and (3) to use certain nitrogen compounds (NO_3^-) as terminal electron acceptors under conditions where oxygen is either absent or in limited supply. The reactions and products involved in these three metabolically different pathways collectively make up the nitrogen cycle and are discussed below.

Nitrogen compounds as nutrients. The synthesis of organic nitrogen compounds from inorganic nitrogen and carbon compounds begins with NH_3 incorporation. One major reaction, catalyzed by glutamic acid dehydrogenase, involves 2-ketoglutarate, NADH (reduced nicotinamide adenine dinucleotide) and ammonia, and the product is glutamic acid. There are two ways in which organisms obtain ammonia. One is to use nitrogen already in a form easily metabolized to ammonia. Thus, nonviable plant, animal, and microbial residues in soil are enzymatically decomposed by a series of hydrolytic and other reactions to yield biosynthetic monomers such as amino acids and other small-molecular-weight nitrogenous compounds. These amino acids, purines, and pyrimidines are decomposed further to produce NH_3 which is then used by plants and bacteria for

biosynthesis, or these biosynthetic monomers can be used directly by some microorganisms. The decomposition process is called ammonification. Not all organic nitrogen is ammonified easily, and resistant nitrogenous residues constitute humus, a complex component of great importance to soil structure and water-holding capacity.

The second way in which inorganic nitrogen is made available to biological agents is by nitrogen fixation (this term is maintained even though N_2 is now called dinitrogen), a process in which N_2 is reduced to NH_3 . Since the vast majority of nitrogen is in the form of N_2 , nitrogen fixation obviously is essential to life. The N_2 -fixing process is confined to prokaryotes (certain photosynthetic and nonphotosynthetic bacteria). The major nitrogen fixers (called diazotrophs) are members of the genus *Rhizobium*, bacteria that are found in root nodules of leguminous plants (Fig. 2), and of the cyanobacteria (originally called blue-green algae). Even though rhizobia have recently been cultured so that they can fix nitrogen in the absence of the plant, the conditions needed for this fixation, such as a very low O_2 concentration supplied by oxyleg-hemoglobin and the necessary carbon and energy sources supplied by the plant, are ideal in the nodule and are not easily met in the laboratory or elsewhere in nature. There are many "free-living" diazotrophs, and even though they appear to contribute little to soil nitrogen, most knowledge of the biochemistry of nitrogen fixation comes from studies of three of them, *Clostridium pasteurianum*, *Azotobacter vinelandii*, and *Klebsiella pneumoniae*.

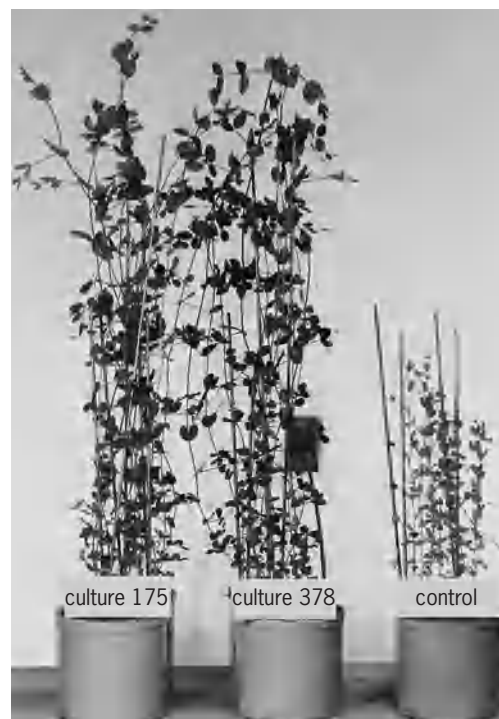


Fig. 2. Effect of inoculation with two strains (cultures 175 and 378) of the nitrogen-fixing *Rhizobium leguminosarum* on the development of peas.

The nitrogen-fixing system of all these organisms is made up of two protein components. One, called the Fe protein, is a dimer of molecular weight about 60,000 daltons, and it contains an Fe_4S_4^* center (similar to that of ferredoxin) that is involved in accepting and transferring the electrons needed for N_2 reduction. It is the Fe protein that initially binds the magnesium adenosine triphosphate (MgATP) needed for N_2 reduction. The second component, called the MoFe protein, is a tetramer of about 220,000 daltons and contains four Fe_4S_4^* centers and two $\text{Fe}_8\text{Mo}_1\text{S}_6$ centers, with the latter centers probably being the sites where N_2 is reduced to NH_3 . For N_2 reduction the MoFe protein accepts electrons from the Fe protein, and to facilitate the transfer, ATP is hydrolyzed to adenosine diphosphate (ADP) and inorganic phosphate. The N_2 fixation process consumes as much as a third of the cell's energy supply, so for obvious reasons the N_2 -fixing system is not synthesized if a usable nitrogen compound other than N_2 is available, that is, nitrogenase synthesis is repressed in the presence of NH_3 .

Nitrogen compounds as energy source. Many microorganisms can use organic nitrogen compounds as energy sources, but in most cases the nitrogen of the compound is first removed and excreted as NH_3 , and then the reduced carbon compound remaining is catabolized to yield both energy and organic carbon intermediates. A relatively few microorganisms (including some fungi) are able to aerobically convert the excreted NH_3 to NO_2^- , and NO_2^- to NO_3^- , and couple these oxidations to the production of ATP and the creation of a membrane potential needed for biosynthetic reactions. Thus, NH_3 added to aerobic soils is rapidly converted by such organisms to NO_3^- . Two chemoautotrophs (organisms that grow using an inorganic compound as an energy source and CO_2 as a carbon source) are primarily responsible for NO_3^- production. First *Nitrosomonas* converts the NH_3 to NO_2^- , and then the relatively toxic NO_2^- is rapidly oxidized to NO_3^- by *Nitrobacter*. Plants and microorganisms readily use this NO_3^- as a nitrogen source by first reducing it to NH_3 . The overall process of oxidation of NH_3 to NO_3^- is called nitrification. The process whereby NO_3^- is reduced to NH_3 is called nitrate assimilation.

Nitrogen compounds as electron acceptors. When NO_3^- accumulates in soils in which metabolizable carbon compounds are available and when such soils become anaerobic because growth of aerobic organisms exhausts the O_2 , certain organisms such as *Pseudomonas*, *Micrococcus*, *Achromobacter*, and *Bacillus* use the NO_3^- either as a normal electron acceptor or as an electron acceptor in place of O_2 . The electron acceptor is needed by such cells to allow electrons from cellular oxidations to flow through the array of electron carries in the cell membrane. This electron flow is needed to facilitate proton (H^+) transfer across the membrane, which in turn creates a membrane potential and a pH gradient. The potential energy of the pH gradient is used in conjunction with the cells adenosine triphosphatase (ATPase) to

allow ATP to be synthesized. The NO_3^- is the terminal electron acceptor in the electron flow and becomes reduced to NO_2^- , and the NO_2^- in turn is further reduced to N_2 (some N_2O may also be produced). The N_2 (and some N_2O) is released into the atmosphere. This process, called denitrification or nitrate respiration, is responsible for ridding many bodies of water and soil of excess fixed nitrogen that could lead to pollution by overproduction and decay of algae and other bacteria.

N_2 cycle in the oceans. Most information on the N_2 cycle comes from studies of the soil-water system. Much less is known about the nitrogen cycle in seas, even though it is estimated that as much as 20% of the N_2 fixed on Earth occurs in the ocean. This N_2 seems to be primarily fixed by cyanobacteria, although in localized areas some contribution by other photosynthetic and nonphotosynthetic bacteria also occurs. One calculation estimates that a single bloom of the cyanobacterium *Trichodesmium* could fix as much as 1100 tons (100 metric tons) of N_2 per day. The other processes associated with the nitrogen cycle, ammonification, nitrification, and denitrification, although obviously present in oceans, have been studied even less. Heterotrophic organisms were not shown to be responsible for nitrification in seas, and therefore chemoautotrophic marine nitrifiers, such as *Nitrospira* and *Nitrococcus*, appear to be the major nitrifiers in oceans. The denitrification step of the ocean's nitrogen cycle may account for a third of the Earth's total denitrification, but there is relatively little information on this. One suspects that most of this denitrification takes place in the anaerobic conditions of marine mud.

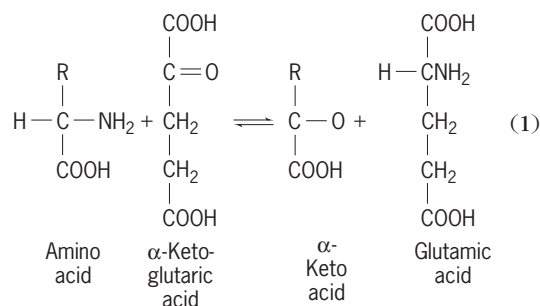
Summary. The major microbial reactions discussed above are the three major contributors to the N_2 cycle. First N_2 is fixed (reduced) to ammonia, and the ammonia is used by N_2 -fixing microorganisms and by plants harboring N_2 -fixing microorganisms. Then these plants and bacteria, and animals and plants living off these plants and bacteria, die and lyse, and the nitrogen of their nitrogenous compounds is converted to ammonia by ammonification. The released ammonia is converted rapidly by nitrifying bacteria to NO_3^- , or it is used directly by microorganisms or plants. The NO_3^- produced by nitrification is used by plants and by bacteria as a nitrogen source; or if anaerobic conditions are created, the NO_3^- is denitrified (reduced) to N_2 and N_2O . This completes the nitrogen cycle. Leonard E. Mortenson

Bibliography. N. S. Rao, *Biological Nitrogen Fixation: Recent Developments*, 1988; J. I. Sprent, *The Ecology of the Nitrogen Cycle*, 1987; G. Stacey, R. H. Burris, and H. J. Evans (eds.), *Biological Nitrogen Fixation*, 1991.

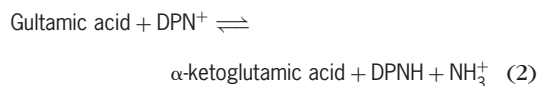
Nitrogen excretion

In the quest for sufficient food energy to meet caloric requirements, animals ingest more nitrogen, largely as amino acids, than they require. Accordingly, the

excess nitrogen ingested must be excreted in some form. Through the action of a series of related enzymes called transaminases, virtually all metabolic nitrogen can be transferred to α -ketoglutaric acid to form glutamic acid, as shown in reaction (1). Under



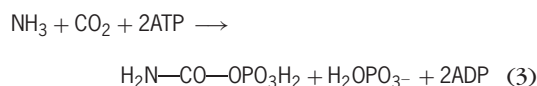
the influence of glutamic dehydrogenase, glutamic acid may be oxidized by the coenzyme diphosphopyridine nucleotide (DPN) with the reformation of α -ketoglutaric acid plus ammonia, as shown in reaction (2).



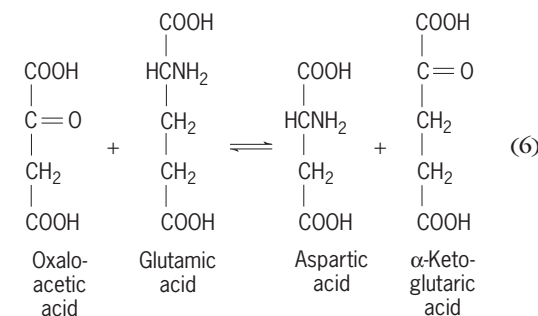
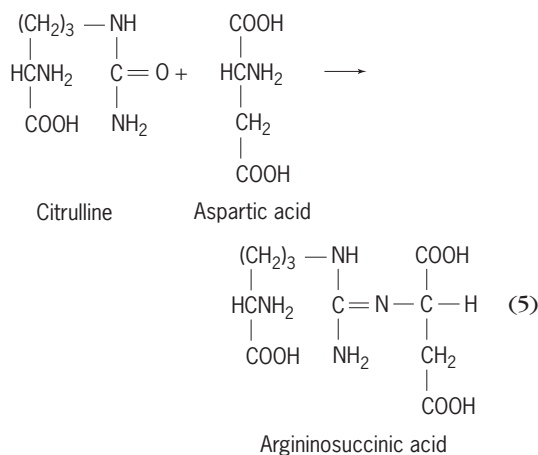
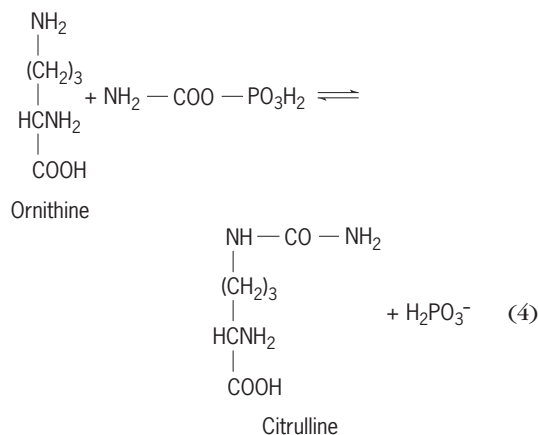
Aquatic animals in general simply excrete the ammonia as such, typically by diffusion across the gills or body surface. Other animals, notably amphibians and mammals, first convert ammonia into urea.

To accomplish the transformation, advantage is taken of the enzymically catalyzed metabolic sequence by which the amino acid arginine is synthesized from ornithine, a sequence common to almost all living forms. It is adapted for urea synthesis by mammals, reptiles, and other forms by the additional presence in liver of the enzyme arginase, which catalyzes the hydrolysis of arginine to urea plus ornithine, which is then available for recycling.

The initial step is catalysis of the formation of carbamyl phosphate from ammonia, CO_2 , and adenosine triphosphate (ATP) by carbamyl phosphate synthetase. The mechanism of this reaction is not understood, but it is known to require the presence of an *N*-acetylglutamate and an additional molecule of ATP. The reaction is shown in (3). The carbamyl phosphate



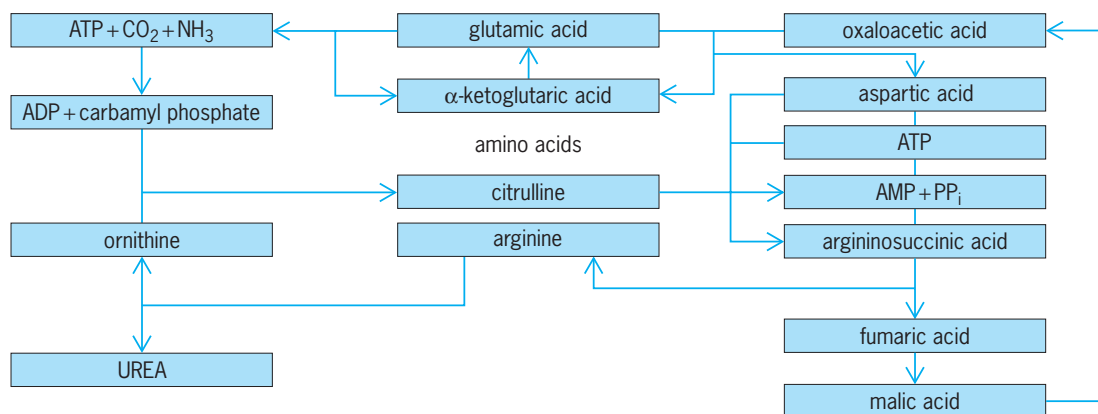
is then caused to react with ornithine under the influence of ornithine transcarbamylase with the formation of citrulline plus orthophosphate, a reaction shown in (4). The second atom of nitrogen necessary for urea synthesis is then introduced in a complex reaction whereby aspartic acid and citrulline condense to form argininosuccinic acid, the energy being derived from another molecule of adenosine triphosphate with the formation of adenosine monophosphate and pyrophosphate, as is shown in reaction (5). It is noteworthy that, like ammonia, the nitrogen of aspartic acid is derived from glutamic acid



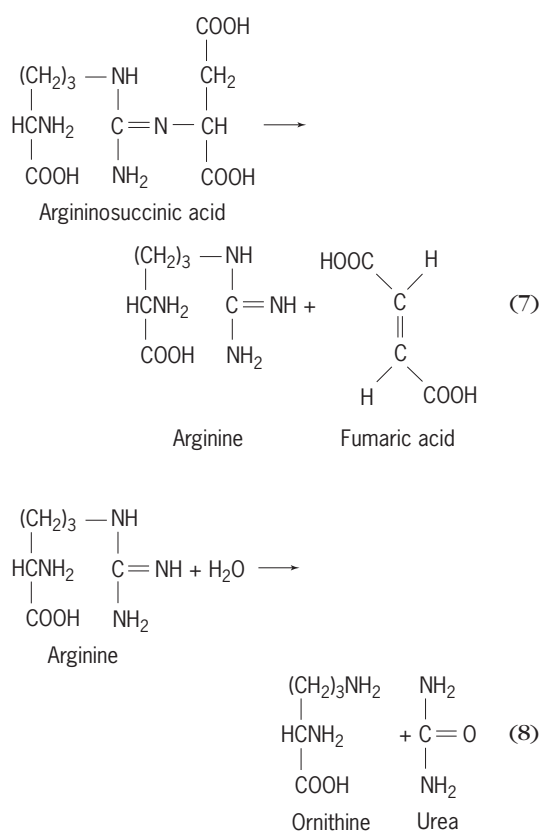
through the operation of a specific aspartic-glutamic transaminase, as shown in reaction (6).

The product, argininosuccinic acid, is cleaved by an appropriate enzyme to form one mole each of arginine and fumaric acid, as shown in reaction (7). With the hydrolysis of arginine to ornithine plus urea the cycle is completed and the ornithine is available for the next round, as shown in reaction (8).

The free energy charge ΔG° that is required for the overall reaction $\text{CO}_2 + 2 \text{ ammonia} \rightarrow \text{urea}$ is $-14,000 \text{ cal/mole}$. As in many other biological systems, this energy is provided by utilization of adenosine triphosphate. The cyclic nature of the system is shown in the **illustration**. A second cycle is also operative wherein the 4-carbon dicarboxylic acid, fumaric acid, released in reaction (7) is reconverted to oxaloacetic so that it can be realized for aspartic acid formation and thus for another turn of the urea cycle.



Regenerative cycles in urea biosynthesis beginning with formation of carbamyl phosphate.



The relatively small amounts of urea excreted by aquatic animals are largely derived from the breakdown of purines, derived from the nucleic acids. Terrestrial animals, notably insects, arachnids, reptiles, and birds, convert the ammonia, in a synthesis which requires a greater input of energy per molecule of ammonia than does the synthesis of urea, into the purine xanthine, which is then oxidized to form uric acid, the form in which the ammonia is ultimately eliminated. This mechanism is an adaptation to embryonic development in an egg with closed shell, in which the accumulation of ammonia would be toxic and the formation of urea would cause osmotic difficulties. In addition, for adults in a dry environment uric acid as a nitrogenous end product has the advantage that it is relatively insoluble in water, and may be

eliminated in solid form, without the water which is typically the main component of urine of mammals. See EXCRETION; PROTEIN METABOLISM; URINARY SYSTEM; URINE. Philip Handler; Bradley T. Scheer

Nitrogen fixation

The chemical or biological conversion of atmospheric nitrogen (N₂) into compounds which can be used by plants, and thus become available to animals and humans. In the 1990s, chemical and biological processes together contributed about 260 million tons (230 million metric tons) of fixed nitrogen per year globally. Industrial production of nitrogen fertilizer accounted for about 85 million tons (80 million metric tons) of nitrogen per year, while spontaneous chemical processes, such as lightning, ultraviolet irradiation, and combustion, leading to the synthesis of nitrogen oxides from O₂ and N₂, may have accounted for 44 million tons (40 million metric tons) per year. The remainder, roughly half of the global input of newly fixed nitrogen, arose from biological processes. World agriculture, which is very dependent on nitrogen fixation, is increasingly reliant on chemical nitrogen sources. See NITROGEN.

Chemical fixation. Three chemical processes for fixing atmospheric nitrogen have been developed. All require considerable thermal or electrical energy and yield different products.

In arc processes, such as the Birkland-Eyde process, air is passed through an electric arc and about 1% nitric oxide is formed, which can be chemically converted to nitrates. Inexpensive electricity is required for the industrial use of arc processes, which are now rarely used. In the cyanamide process, which is now obsolete, heating calcium carbide in nitrogen generates calcium cyanamide, which when moistened hydrolyzes to urea and ammonia. In the widely used Haber process, hydrogen (generated by heating natural gas) is mixed with nitrogen (from air), and burned to yield a nitrogen-hydrogen mixture. The nitrogen-hydrogen mixture is compressed (10–80 megapascals) and heated (200–700°C or 390–1300°F) in the presence of a metal oxide catalyst to give ammonia. The Haber process is the major source

of ammonia used for fertilizer. See HIGH-PRESSURE PROCESSES.

Ionization and chemonuclear processes exist for obtaining nitrogen oxides from air but are not used industrially. Laboratory processes also have been developed for fixing nitrogen as metal nitrides or transition-metal complexes, and a procedure has been developed for reducing the nitrogen bound in certain transition-metal complexes to ammonia. Such processes may have industrial applications. See AMMONIA; CYANAMIDE; ELECTROCHEMICAL PROCESS; FERTILIZER; NITROGEN COMPLEXES.

Biological fixation. Only prokaryotes—bacteria, archaea, and cyanobacteria (earlier called blue-green algae)—fix nitrogen. Nitrogen-fixing microbes, called diazotrophs, fall into two main groups, free-living and symbiotic. See ARCHAEA; BACTERIA; CYANOBACTERIA; PROKARYOTAE.

The free-living diazotrophs are subclassified. Aerobic diazotrophs, of which there are over 50 genera, including *Azotobacter*, methane-oxidizing bacteria, and cyanobacteria, require oxygen for growth and fix nitrogen when oxygen is present. *Azotobacter*, some related bacteria, and some cyanobacteria fix nitrogen in ordinary air, but most members of this group fix nitrogen only when the oxygen concentration is low. Free-living diazotrophs, which fix nitrogen only when oxygen is absent or vanishingly low, are widespread. The genera *Bacillus* and *Klebsiella* include many strains of this type, and representatives of symbiotic diazotrophs behave in this way as well. They grow perfectly well in air if prefixed nitrogen, such as an ammonium salt, is present, but they will not fix nitrogen under these conditions. Some *Clostridium* strains, some sulfate-reducing bacteria, and some methane-forming archaea are free-living diazotrophic anaerobes, and do not grow in the presence of air, whether fixing nitrogen or not. There are also free-living diazotrophic phototrophs, including many cyanobacteria, which fix nitrogen in air, but other photosynthetic bacteria are generally unable to fix nitrogen, except when growing anaerobically. See ALGAE; BACTERIAL PHYSIOLOGY AND METABOLISM.

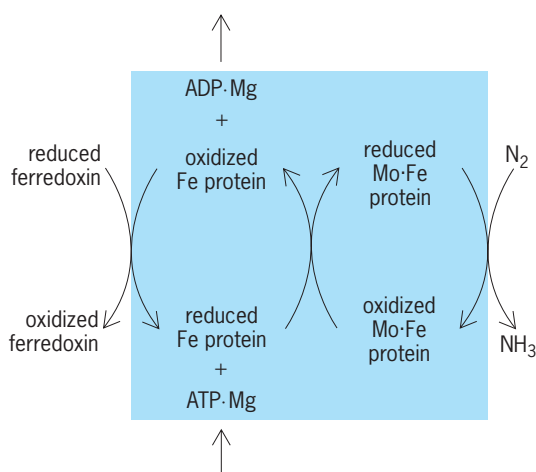
The best-known symbiotic bacteria belong to the genus *Rhizobium*. Species of *Rhizobium*, or related genera, such as *Bradyrhizobium* and *Sinorhizobium*, colonize the roots of leguminous plants and stimulate the formation of nodules within which they fix nitrogen microaerobically. Both plants and bacteria show specificity; for example, certain types of plants require special strains of rhizobia. Some types of rhizobium, such as *Bradyrhizobium*, can fix nitrogen in the absence of plant tissue, but require low oxygen, though most rhizobia fix nitrogen only within the nodules. Effective root nodules are pink due to a protein, called leghemoglobin, related to the hemoglobin of blood, which transports a steady supply of oxygen to the rhizobia. Rhizobial symbioses are of great agricultural importance in legume crops, such as soybean, pulse, and alfalfa. Of comparable ecological importance are *Frankia* species, the diazotrophic symbionts of *Alnus* (alder),

Casuarina (Australian pine), or *Ceanothus*, which are also found in root nodules. Symbiotic cyanobacteria form diazotrophic associations with plants ranging from fungi (as lichens) to ferns (for example, *Azolla*), cycads, and angiosperms (for example, *Gunnera*). Symbiotic associations between several genera of aerobic diazotrophs (such as *Azospirillum*, *Azoarcus*, and a species of *Azotobacter*) and both grasses and cereals have been reported. More casual associations between free-living diazotrophic bacteria and the roots of maize, rice, temperate weeds, and water plants also exist. All of these have limited agricultural potential at present. Diazotrophic associations involving leaf nodules or the mycorrhizae of roots are probably illusory. Associations with ruminant animals, termites, and wood-boring mollusks are real but casual and nutritionally trivial. See SOIL MICROBIOLOGY.

Nitrogenase. The enzymes responsible for nitrogen fixation are called nitrogenases. The most common nitrogenase consists of two proteins, one large (about 220,000 mol wt) containing molybdenum, iron, and inorganic sulfur (the MoFe-protein or dinitrogenase), the other smaller (about 60,000 mol wt) containing iron and inorganic sulfur (the Fe-protein or dinitrogenase reductase). The chemical compositions of both proteins and the natures of their metal-containing sites are known, and the three-dimensional structure of the MoFe-protein has been established by x-ray analysis. The molybdenum atom resides in a unique iron-molybdenum-sulfur cluster, called FeMoco, which can be chemically extracted.

Both nitrogenase proteins are irreversibly destroyed by exposure to oxygen. Therefore, an oxygen-deficient environment is essential for activity. In addition, biological energy in the form of adenosine triphosphate (ATP) is consumed during activity, for which magnesium ions and a reducing agent are also required. The biological reductant is a ferredoxin or a flavodoxin. Nitrogenase reduces one molecule of N_2 to two of ammonia (NH_3), a reaction which is accompanied by the conversion of 16 molecules of ATP to adenosine diphosphate (ADP) and the release of one molecule of H_2 as a by-product. If N_2 is absent (for example, in argon), the enzyme forms hydrogen exclusively. Several small molecules that resemble N_2 in chemical structure will interact with the enzyme. Carbon monoxide blocks nitrogen fixation completely but leaves H_2 formation unimpaired. Other structural analogs of N_2 , such as cyanide, azide, acetylene, or nitrous oxide, can be reduced in its place. The reduction of acetylene to ethylene provides a rapid and highly efficient test for biological nitrogen-fixing systems.

In active nitrogenase, the two proteins form a short-lived complex and their modes of action are largely understood. In the **illustration**, reduced ferredoxin is depicted as reducing the Fe-protein which then, after reacting with a magnesium derivative of ATP, reduces the MoFe-protein that had bound N_2 , displacing a molecule of H_2O . The N_2 is reduced to NH_3 ; and H_2 , magnesium, and ADP are released. Finally, the MoFe-protein is reoxidized and the



A mechanism for the functioning of the two components of nitrogenase. The box signifies that a complex of Fe and Mo-Fe proteins forms the active enzyme.

complex falls apart. The rate of this last step determines the turnover time for N_2 , which is slow at about 0.9 s^{-1} . Because of its slow action, up to 20% of the protein in an active diazotroph can be nitrogenase.

Two less common types of nitrogenase are also known. They occur in certain azotobacters and photosynthetic diazotrophs. Like the Mo-nitrogenase, they consist of two proteins, consume ATP, evolve H_2 , and are destroyed by oxygen. One type contains vanadium (V) and the other type, iron (Fe), in place of Mo, and both metal atoms are in clusters comparable to FeMoco. Known as V-nitrogenase and Fe-nitrogenase, they appear to be less efficient than Mo-nitrogenase and their biological significance is obscure. *Azotobacter vinelandii* possesses all three nitrogenases. See VANADIUM.

Physiology. The irreversible destruction of nitrogenase by air means that all aerobic diazotrophs have developed means of restricting access of oxygen to the active enzyme. The legume root nodule can be viewed as a compartment in which leghemoglobin supplies rhizobial cells with oxygen sufficient for respiration but at concentrations that are harmless to nitrogenase. Some cyanobacteria have comparable compartments or specialized cells, called heterocysts, in which nitrogenase is protected from the oxygen evolved during photosynthesis by an oxygen-impermeable wall. In free-living aerobes, respiration excludes oxygen from the enzyme; and *Azotobacter*, one of the most efficient aerobic diazotrophs, possesses a special iron-sulfur protein which protects nitrogenase in cells subject to oxygen stress, though the enzyme is then nonfunctional until the oxygen stress ceases.

The release of H_2 as a by-product of nitrogen fixation represents a waste of biological energy (as ATP). Efficient aerobic diazotrophs, such as *Azotobacter* and some strains of *Rhizobium*, possess the enzyme hydrogenase, which enables them to use the H_2 as a reductant and source of ATP. There is evidence that hydrogenase-containing rhizobia produce the best

yields of certain legume crops. See NITROGEN CYCLE.

Regulation and genetics. Nitrogenase is an expensive enzyme in terms of its ATP consumption, and most diazotrophs possess physiological mechanisms that enable them to switch off the enzyme should a supply of fixed nitrogen appear. It is also expensive in the sense that the cell is obliged to synthesize large amounts of enzyme protein, and often the apparatus of an oxygen-exclusion mechanism as well. Synthesis of the enzyme is tightly regulated at the genetic level, and it is not made when fixed nitrogen (as NH_3) is available or if O_2 concentrations are unsuitable.

The genes specifying nitrogenase, together with the various ancillary proteins involved in its synthesis and function, called *nif* genes, are regulated at the transcriptional level by ammonia, operating through a genetic system (the *ntr* genes), which also regulates other bacterial nitrogen metabolism functions, such as the use of certain amino acids. Oxygen also regulates *nif* expression. In 1972, *nif* genes were taken from *Klebsiella* and inserted into *Escherichia coli*, which then gained the ability to fix nitrogen. Since then, plasmids (extrachromosomal genetic elements) have been constructed to carry *nif* genes, permitting their transfer to a variety of microorganisms. As a result, entirely new diazotrophs have been created such as strains of *Salmonella*, *Proteus*, and *Serratia*. But not all bacterial recipients of the *nif* genes became diazotrophic. The *nif* genes of *Klebsiella* have been studied exhaustively, and their complete DNA sequences are known. They are present as a cluster of 20 contiguous genes. Expression of the whole cluster is regulated by *ntr* products acting upon genes within it, called *nifA* and *nifL*, and subclusters (operons) within *nif* are regulated by the products of *nifA* and *nifL*. Oxygen bypasses *ntr* and regulates *nif* expression by an indirect action upon the *nifL* gene.

Recombinant DNA clones have been prepared that carry the whole *nif* cluster from *Klebsiella*, the several operons, the separate genes, and their promoter regions. And gene fusions have been prepared to study physiological questions. Clones of *Klebsiella nif* genes have also been used to open up the molecular genetics of *Rhizobium*, *Azotobacter*, *Anabena* (a cyanobacterium), and many other diazotrophs. In several diazotrophs, the *nif* genes occur naturally on plasmids; however, in most diazotrophs they are not in a contiguous cluster but are dispersed about the genome in separate operons. V-nitrogenases and Fe-nitrogenases are coded by distinct but comparable groups of genes (called *vnf* and *anf* respectively), and have been cloned and sequenced. See BACTERIAL GENETICS; GENETIC ENGINEERING.

Genetics of diazotrophic symbioses. In legume symbiosis, and probably in others, both the host plant and the bacteria contribute genetic information. The rhizobia carry the *nif* genes, plus genes called *fix*, which play a part in making functional nitrogenase. In the genus *Rhizobium* these genes are on a plasmid, but in the related *Bradyrhizobium* they are chromosomal. The bacteria also have *sym* genes, which specify the appropriate species of host

plant, and *nod* genes, which induce nodulation in the plant. The plants possess other *nod* genes and contribute products called nodulins. One nodulin is leghemoglobin, and another is a specific attractant that enables the free rhizobia to recognize which plant to colonize. In all, some 50 genes are involved in establishing legume nodule symbiosis. It is likely that analogous genes function in *Frankia* and other diazotrophic symbioses.

Economic considerations. Except in a few areas where extreme climatic conditions are a limiting factor, input of fixed nitrogen determines the productivity of world agriculture and forestry. That input was substantially biological until the early 1900s, during which the use of nitrogen fertilizer produced chemically by the Haber process escalated. By 2000, between 30 and 40% of the world's population depended on chemical nitrogen fertilizer for its food supplies. The global population is expected to reach 8–9 billion by 2100. Substantial increases in nitrogen input, of both chemical and biological origin, into global food production are therefore inescapable. Increased fertilizer use may generate environmental, economic, and sociopolitical problems, ranging from local eutrophication and pollution to increased dependence of world agriculture on sophisticated technology, with consequent economic and social disruption in developing countries. Proposals to ameliorate these problems include (1) developing low-tech “cottage” processes for generating nitrogen fertilizer locally, based perhaps on solar-powered chemistry or immobilized enzyme technology; (2) extending the use of established symbiotic plant associations in food production and forestry, particularly their use as “green manure”; (3) generating, by conventional breeding or genetic manipulation, new or improved diazotrophic symbioses involving wheat and rice, which are the world's major food crops; and (4) transferring, by genetic manipulation, bacterial *nif* genes into the plant genome and rendering them functional therein. See AGRICULTURE.

John Postgate

Bibliography. I. R. Kennedy and E. C. Cocking (eds.), *Global Nitrogen Fixation: The Global Challenge and Future Needs*, University of Sydney SUNfix Press, 1997; J. R. Postgate, *Nitrogen Fixation*, Cambridge University Press, London, 2000; G. Stacey, R. H. Burris, and H. J. Evans (eds.), *Biological Nitrogen Fixation*, Routledge, New York, 1992.

Nitrogen-fixing trees

The ability to fix atmospheric nitrogen into a form that can be used for plant growth is confined to bacteria and cyanobacteria. Plants fix nitrogen only by virtue of associations with these simple organisms. The best-known associations are the symbioses of *Rhizobium* bacteria with agricultural legumes, such as clovers, peas, and beans. Rhizobia stimulate the formation of root nodules, which provide a specialized environment within which high rates of nitrogen fixation can occur. Nitrogen fixation not only

supports plant growth independent of mineral nitrogen in the soil but also can improve soil nitrogen status as plant residues, notably leaves and fine roots, decay and are mineralized. Accurate estimates of amounts of nitrogen fixed are difficult to obtain, but a well-nodulated, young, densely planted stand of nitrogen-fixing trees may fix 80–200 kg of nitrogen per hectare per year, similar to rates reported for some leguminous crops. See BACTERIA; CYANOBACTERIA.

Nitrogen-fixing tree species occur in all three subfamilies of the Leguminosae (Fabaceae). Most of the mimosoid group (such as *Acacia*, *Albizia*, *Leucaena*) and virtually all of the papilionoid group (including *Dalbergia*, *Gliricidia*, *Robinia*) fix nitrogen, but nodulation and nitrogen fixation in the caesalpinoid group (such as *Cbamaecrista*) is very restricted. However, in some genera, for example the acacias, not all the species form nitrogen-fixing symbioses. Most legume trees are found in the tropics and subtropics, but some grow and are of importance in temperate climates, for example the horticultural laburnums or black locust (*Robinia pseudoacacia*), which is used extensively in rehabilitation of derelict land. See LEGUME.

Many species of trees and shrubs other than legumes form nitrogen-fixing root nodules not with rhizobia but with a filamentous bacterium, *Frankia*, or with nitrogen-fixing cyanobacteria, such as *Nostoc*. Symbiosis with cyanobacteria produces a large, coralloid mass of nodules on surface roots of some cycads. Evolutionarily, these plants are ancient members of the gymnosperms; they resemble palm trees when mature, with a stout trunk and a crown of fronds. They are restricted to Central and South America, South Africa, and Australasia, where they are a source of fixed nitrogen for natural ecosystems. Following fires, cycads grow and fix nitrogen in eucalyptus forests. After a few years, they are shaded out by new growth and provide a valuable input of fixed nitrogen to the forest ecosystem as they decay.

Frankia belongs to the bacterial family Actinomycetales, which includes the antibiotic-producing soil actinomycetes *Streptomyces*. *Frankia* forms nitrogen-fixing nodules on the roots of trees and shrubs in eight families other than legumes, most being native to temperate regions (see *illus.*). Species nodulated by *Frankia* are known as actinorhizal plants. They are very important ecologically for their ability to colonize and to improve the nitrogen status of denuded and degraded soils. Some genera, for example *Ceanothus* and *Elaeagnus*, are of horticultural importance, while *Hippophae rhamnoides* (sea buckthorn) berries are a valuable crop in Russia and China. In temperate regions, the most widely utilized actinorhizal tree species belong to the genus *Alnus* (the alders) of the Betulaceae, the family to which the non-nitrogen-fixing birches belong. Different alder species occupy a wide range of natural habitats, from riversides to relatively dry mountain valleys. They were used widely in earlier times to produce high-quality charcoal, dyes, and decay-resistant timber. Their main use now is in land reclamation.



A perennial nodule, several years old, formed on the roots of common or black alder (*Alnus glutinosa*) by symbiosis with the nitrogen-fixing bacterium *Frankia*.

They may be grown also in single-species stands or in mixed stands with non-nitrogen-fixing tree species to improve productivity on nitrogen-deficient sites. There is also considerable interest in utilizing alders as a component of fast-growing, sustainable energy plantations to provide biomass for fuel or paper pulp production.

Only one family of actinorhizal plants, which contains the casuarinas, the Australian sheoaks, is native to the tropics. These trees are a common feature of hot countries, and their needlelike branchlets and conelike fruits frequently lead to confusion with conifers. The family contains species tolerant of a wide range of soil conditions in environments that may be wet, semiarid, or dry. Species such as *Casuarina equisetifolia* or *C. glauca*, which are tolerant of saline conditions and poor soils, are used extensively as wind breaks and in stabilizing coastal and desert sand dunes. *Casuarina equisetifolia* is fast-growing, and the dense wood has a calorific value higher than most other nitrogen-fixing trees. It is grown widely in 5–7-year rotations by small farmers in India and China for fuel wood or charcoal, and it is also valuable for producing small construction timbers such as piles and posts, paper pulp, and fiber boards.

Most species of nitrogen-fixing trees belong to the Leguminosae and have diverse uses. Timber and pulpwood species include *Dalbergia* (rosewood) and *Albizia* (ipil ipil), which is an important component of sustainable forestry in the tropics. Many species are used for fuel wood, and trees able to grow in deforested nitrogen-deficient soils of arid regions are very important in this respect socially, for example some *Acacia* and *Prosopis* (mesquite) species. Other important uses are as shade, nurse, or ornamental trees and as foliage for animal fodder. Some trees also produce edible fruits or gums and tannins and are useful as medicines. Like the actinorhizal trees, nitrogen-fixing legume trees are used widely for land reclamation and stabilization.

Many useful features may be found in a single species. For example, *Leucaena leucocephala* is grown widely in the tropics and can achieve a height of 18 m (60 ft) in 5 years. The wood is excellent for fuel and charcoal and can be used for posts, lumber, and furniture. Dense plantations produce leaf litter

containing up to 300 kg of nitrogen per hectare and valuable as fodder or green manure. Tender vegetative shoots can be eaten, as can seeds if treated carefully to reduce toxicity. The multifaceted properties of nitrogen-fixing trees ensure them a favored role in agroforestry systems, particularly on impoverished lands, where their minimal requirement for fertilizer application and the high nitrogen content of residues are of great value for intercropping practices. See NITROGEN FIXATION; PLANT METABOLISM.

C. T. Wheeler

Bibliography. M. G. R. Cannell, D. C. Malcolm, and P. A. Robertson (eds.), *The Ecology of Mixed Species Stands of Trees*, Blackwell, Oxford, 1992; K. E. Giller and K. J. Wilson, *Nitrogen Fixation in Tropical Cropping Systems*, C. A. B. International, Wallingford, 1991; J. C. Gordon and C. T. Wheeler (eds.), *Biological Nitrogen Fixation in Forest Ecosystems: Foundations and Applications*, Martinus Nijhoff/W. Junk, The Hague, 1983; C. R. Schwintzer and J. D. Tjepkema (eds.), *The Biology of Frankia and Actinorhizal Plants*, Academic Press, San Diego, 1990.

Nitrogen oxides

Chemical compounds of nitrogen and oxygen. Nitrogen and oxygen do not combine when mixed directly (as in air), but they do combine during chemical reactions of compounds containing them. A number of nitrogen oxides can be isolated which differ from one another in the numbers of nitrogen and oxygen atoms present in each molecule.

Table 1 gives data for the five nitrogen oxides which are well established. The structures of these molecules and one laboratory method for the preparation of each oxide are given in **Table 2**. These structures show only the geometry of the molecules. In most cases the N and O atoms are united by complex (double or triple) bonds.

The existence of three higher oxides has been postulated. They are nitrogen trioxide, NO₃, from reaction of ozone with dinitrogen tetroxide or pentoxide; dinitrogen hexoxide, N₂O₆, from reaction of fluorine with nitric acid; and an oxide, NO₄, as an intermediate in the ¹⁸O isotope exchange between dinitrogen pentoxide and oxygen gases. The identity and properties of these three oxides are not fully established.

Nitrous oxide and nitric oxide. When inhaled, nitrous oxide has anesthetic effects; in small amounts it produces mild hysteria and hence is sometimes called laughing gas. It is colorless, is the least reactive of the oxides, and dissolves in water without chemical reaction. Decomposition into nitrogen and oxygen occurs at an appreciable rate above 560°C (1040°F)

The equilibrium in reaction (1) lies entirely to the



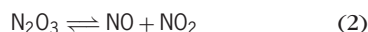
left at low temperatures. Some nitric oxide is formed in an electric arc, as in the technical production of nitric acid.

TABLE 1. Oxides of nitrogen and their properties

| Name | Stoichiometric formula | Melting point, °C (°F) | Boiling point, °C (°F) |
|---|--|------------------------|------------------------|
| Nitrous oxide (dinitrogen monoxide) | N ₂ O | -90.8 (-131) | -88.5 (-127.3) |
| Nitric oxide (nitrogen monoxide) | NO | -163.6 (-262.5) | -151.7 (241.0) |
| Dinitrogen trioxide | N ₂ O ₃ | -103 (-155) | 3.5 (38.3) |
| Dinitrogen tetroxide (⇌ nitrogen dioxide) | N ₂ O ₄ (⇌ NO ₂) | -11.2 (11.8) | 21.2 (70.2) |
| Dinitrogen pentoxide | N ₂ O ₅ | 41 (106) | |

With oxygen or air, nitric oxide is rapidly converted to nitrogen dioxide. Nitric oxide is colorless and is soluble in water without reaction. It is one of the few "odd" molecules which contain an odd number of electrons. Other such molecules (for example, nitrogen dioxide) readily form double molecules, but nitric oxide is exceptional. The gas is monomeric, although dimerization occurs in the liquid, and solid nitric oxide (which is blue) is almost entirely in the form of N₂O₂ molecules. As an odd molecule, it has the ability to lose or gain one electron, thus giving the electrically charged ions NO⁺ and NO⁻. The important nitrosyl compounds contain these ions.

Trioxide. Dinitrogen trioxide exists pure only in the solid state. The dissociation in reaction (2) occurs



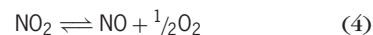
partially in the blue liquid and almost entirely in the vapor state at room temperature. It is the anhydride of nitrous acid; when the oxide (or an equimolecular mixture of NO and NO₂ gases) is dissolved in an alkaline solution, nitrite ion is produced.

Dioxide and tetroxide. The position of the equilibrium between nitrogen dioxide and dinitrogen

tetroxide, reaction (3), depends upon temperature



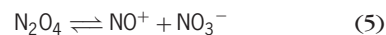
and physical state. The dioxide is a red-brown poisonous gas; the tetroxide is colorless. The colorless solid is entirely in the tetroxide state. In the liquid and gaseous states the tetroxide always contains some dioxide. Thus the liquid tetroxide is brown, although it contains less than 0.1% nitrogen dioxide. The color of the gas becomes more intense with rising temperature; at 100°C (212°F) the tetroxide is 90% dissociated into dioxide. At temperatures above 600°C (1100°F) further decomposition of nitrogen dioxide into nitric oxide occurs, as shown in reaction (4). Dinitrogen tetroxide reacts readily with water to



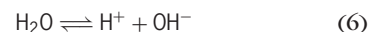
give an equimolecular mixture of nitrous and nitric acids. As temperature is raised, the nitrous acid decomposes to nitric acid and nitric oxide. These reactions are important in the technical production of nitric acid by catalytic oxidation of ammonia. Dinitrogen tetroxide is an oxidizing agent comparable in strength to bromine, and is employed as such in the lead-chamber process for sulfuric acid. In organic chemistry the tetroxide finds use as a special oxidizing agent (for example, in the production of sulfoxides and phosphine oxides) and as a nitrating agent.

The tetroxide forms molecular addition compounds with many simple organic solvents, for example, esters, ethers, ketones, and nitriles.

Liquid dinitrogen tetroxide, alone or mixed with organic solvents, undergoes self-ionization, as in reaction (5). This is to be compared with the aqueous



system shown in reaction (6). For example, liquid



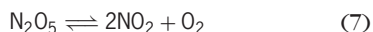
dinitrogen tetroxide attacks some metals (alkali and alkaline-earth metals, zinc, cadmium, and mercury) to produce metal nitrate and evolve nitric oxide. A scheme of reactions has been developed using the liquid tetroxide as a reaction medium, with nitrosyl compounds as acids and nitrates as bases. This medium is therefore valuable for the preparation

TABLE 2. Oxide of nitrogen

| Formula | Structure | Preparation |
|-------------------------------|---|--|
| N ₂ O | N—N—O | Heat ammonium nitrate |
| NO | N—O | Reduce nitric acid with copper |
| N ₂ O ₃ | | Condense gaseous mixture of NO and NO ₂ |
| NO ₂ | | Heat lead nitrate |
| N ₂ O ₄ | | Heat lead nitrate |
| N ₂ O ₅ | | Treat N ₂ O ₄ with ozone |
| Gas | | |
| Solid | NO ₂ ⁺ · NO ₃ ⁻ | |

of anhydrous metal nitrates and nitrate-coordination complexes.

Pentoxide. The ionic nitronium nitrate structure $\text{NO}_2^+\cdot\text{NO}_3^-$ found for N_2O_5 in the solid state accounts for its anomalously high melting point. In solution in sulfuric, nitric, or phosphoric acids the oxide has the same ionic structure. Solid dinitrogen pentoxide readily volatilizes, and the molecular type of structure found in the gaseous state is observed also in solutions of the oxide in low dielectric solvents such as carbon tetrachloride and chloroform. Sodium metal reacts with the liquid oxide, liberating nitrogen dioxide and forming sodium nitrate. Gaseous dinitrogen pentoxide decomposes readily, as in reaction (7), and is a strong oxidizing agent.



With water it is converted to nitric acid. See NITROGEN; OXYGEN. Cyril C. Addison

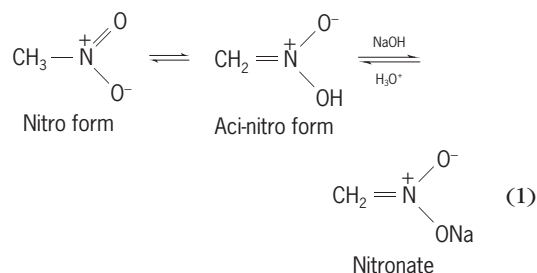
Nitroparaffin

Any derivative of an aliphatic hydrocarbon that contains one or more $-\text{NO}_2$ groups bonded via nitrogen to the carbon framework. Nitroparaffins are also known as nitroalkanes.

Preparation. Low-molecular-weight nitroparaffins are prepared via the vapor-phase nitration of alkanes at $>400^\circ\text{C}$ (750°F). However, the process is not generally satisfactory for higher-molecular-weight nitroparaffins because of polynitration and chain cleavage. The direct nitration of propane is used commercially to prepare nitromethane (boiling point 101°C or 214°F), nitroethane (bp 114°C or 237°F), 1-nitropropane (bp 131°C or 268°F), and 2-nitropropane (bp 120°C or 248°F). Nitroparaffins are prepared in the laboratory by the reaction of nitrite salts with alkyl bromides or iodides, from the oxidation of amines or oximes by using peroxy-carboxylic acids, and by the chain homologation of simple nitroparaffins. See NITRATION.

Nitromethane, nitroethane, and the nitropropanes are useful solvents with high dielectric constants that readily dissolve many polymers. In addition, these simple nitroparaffins are versatile intermediates for the synthesis of specialty chemicals.

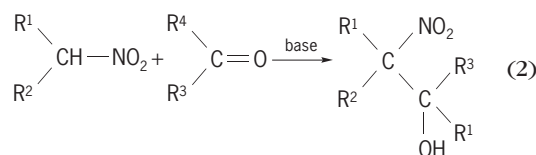
Reactions. Primary and secondary nitroparaffins are in tautomeric equilibrium with the aci-nitro form, as shown in reaction (1). Since tertiary nitroparaffins



do not contain a proton attached to the carbon bear-

ing the nitro group, they are unable to tautomerize in this way. Both primary and secondary nitroparaffins are readily deprotonated with alkali to produce the corresponding nitronate salts [reaction (1)]. The $\text{p}K_a$ of nitromethane is 10.2; and this high acidity, for a carbon acid, is the result of the powerful electron-withdrawing capacity of the nitro substituent. On reaction with alkyllithium bases, primary nitroparaffins (RCH_2NO_2) are converted into α,α -dianions $[\text{RCH}(\text{Li})=\text{NO}_2\text{Li}]$; whereas secondary nitroparaffins, for example, 2-nitropropane, are converted into α,β -dianions $[\text{MeC}(\text{O}^-\text{Li})=\text{CH}_2\text{N}(\text{O}^-\text{Li})_2]$, where Me represents the $-\text{CH}_3$ group. All of these nitro-substituted carbanions are versatile synthetic intermediates. See PK; REACTIVE INTERMEDIATES; TAUTOMERISM.

Primary and secondary nitroparaffins react with aldehydes and ketones under catalysis by a base (such as sodium hydroxide) to produce β -hydroxy-nitroparaffins. This process, known as the Henry or nitro-aldol reaction, is most useful in synthesis [reaction (2)]. Both α -protons of primary nitroparaf-



ins and all three α -protons of nitromethane may be replaced sequentially by hydroxyalkyl groups in this way. For example, nitromethane and formaldehyde may be reacted to produce the alcohol ($\text{HOCH}_2\text{CH}_2\text{NO}_2$), diol $[(\text{HOCH}_2)_2\text{CHNO}_2]$, or triol $[(\text{HOCH}_2)_3\text{CNO}_2]$ adducts. β -Nitroalcohols are easily dehydrated to provide nitroalkenes or reduced to produce important amino alcohols. In consequence of the electronegative nitro group, nitroalkenes are reactive dienophiles toward Diels-Alder reactions with 1,3-dienes. Additionally, they readily undergo addition of nucleophiles to the carbon β to the nitro substituent (Michael addition). See CATALYSIS; DIELS-ALDER REACTION.

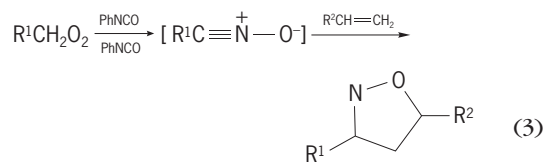
Nitronate salts are hydrolyzed by cold aqueous mineral acids to produce aldehydes or ketones and nitrous oxide. This process, known as the Nef reaction, is also carried out by the oxidation of nitronate salts using ozone or potassium permanganate. Alternatively, primary nitroparaffins are converted into carboxylic acids and hydroxylamine on heating with concentrated mineral acids. Both primary and secondary nitroparaffins ($\text{R}^1\text{R}^2\text{CHNO}_2$) readily add to alkenes ($\text{CH}_2=\text{CHA}$) substituted by electron-withdrawing substituents (A = CN, COMe, CO_2Me , NO_2 , and so forth) to provide the adducts $\text{R}^1\text{R}^2\text{C}(\text{NO}_2)\text{CH}_2\text{CH}_2\text{A}$. The process, which employs catalysis by a base, involves the addition of the nitronate anion.

Nitroparaffins may be reduced to produce amines by heterogeneous catalytic hydrogenation over

transition metals. Transfer hydrogenation from hydrogen donors such as ammonium formate is also effective for the synthesis of amines. The hydrogenation may be stopped at the intermediate hydroxylamine stage. In general, reductions of nitroparaffins by metals such as sodium or potassium (dissolving metals) in acid solution are less efficient for amine preparation. Both secondary and primary nitroalkanes are reduced to produce ketone or aldehyde oximes on reaction with tin(II), chromium(II), or titanium(III) salts. *See* HYDROGENATION.

Primary nitroalkanes (RCH_2NO_2) are dehydrated by reaction with phenyl isocyanate and triethylamine to produce nitrile oxides ($\text{RC}\equiv\text{N}^+\text{—O}^-$). These reactive intermediates dimerize readily, or they are trapped by reaction with unsaturated molecules to

produce heterocyclic adducts [reaction (3)]. For



example, alkenes react with nitrile oxides to produce Δ^2 -isoxazolines. Such heterocycles are useful intermediates in further synthetic transformation.

A. G. M. Barrett

Bibliography. D. H. R. Barton and W. D. Ollis (eds.), *Comprehensive Organic Chemistry*, vol. 2, 1979; R. J. Fessenden and J. S. Fessenden, *Organic Chemistry*, 6th ed., 1998; H. Feuer, *The Chemistry of the Nitro and Nitroso Group*, Interscience, pt. 1, 1969.