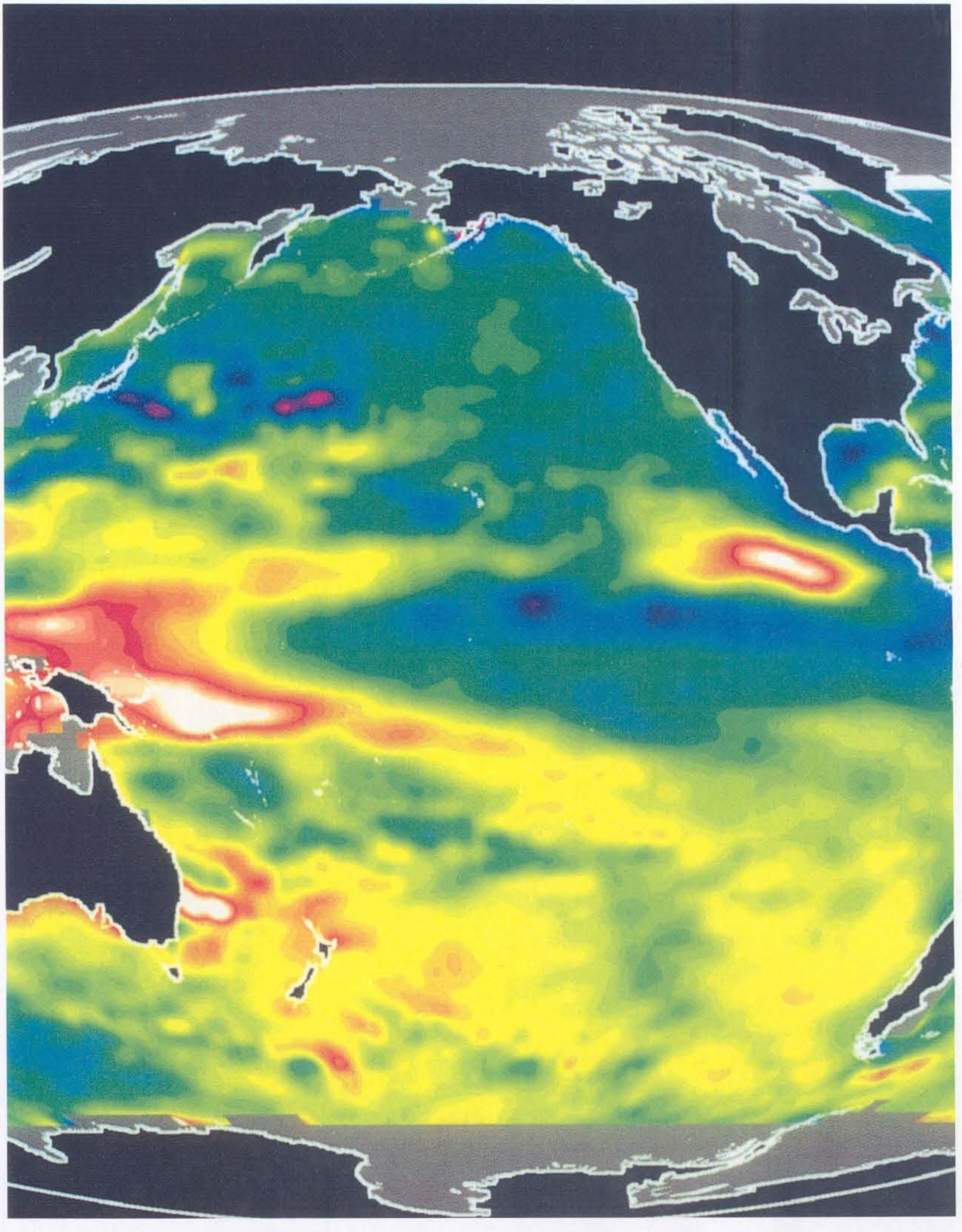
California Institute
of Technology

Engineering & Science

Spring 1995

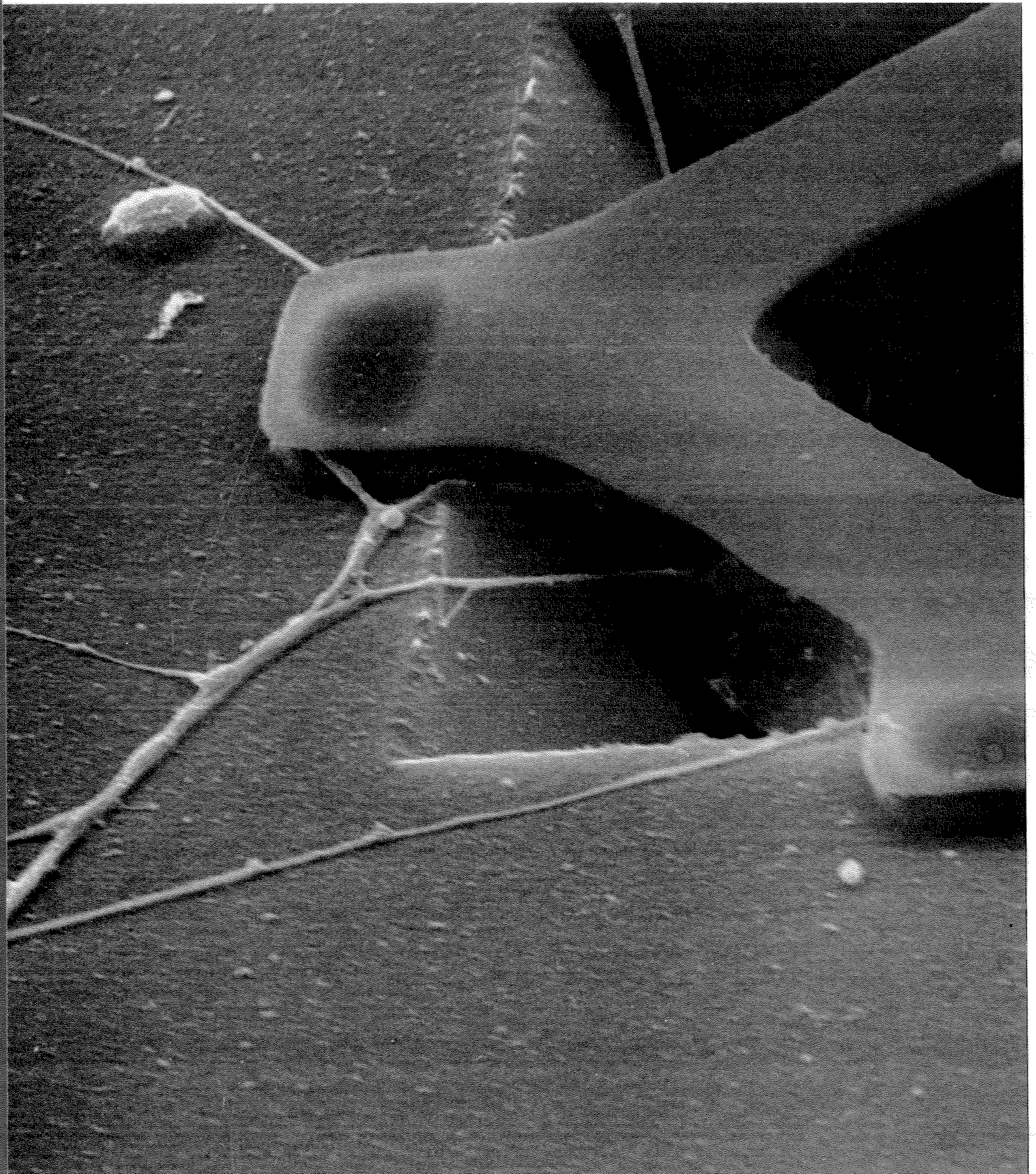


In this issue

Changing Currents

Tiny Baubles

Ubiquitous Ubiquitin

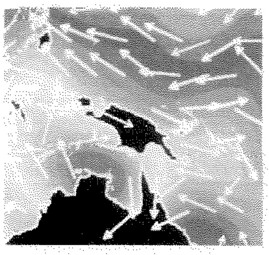


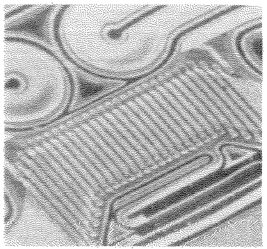
Like seedlings growing up through a steel grating in a Manhattan sidewalk, nerve-cell fibers—called processes—grow out through a silicon grillwork under which their parent nerve cell is imprisoned. (The grillwork's bars are about half a micron thick and four microns wide.) This nerve cell is one of 16 such cells arrayed on a silicon chip for a study on how nerve cells communicate. The hopes are that the processes will connect with the chip's other nerve cells to form a functioning network. The chip, grillwork and all, was built in Caltech's micromachining laboratory, which is making all sorts of tiny gadgets from silicon. For more on what the lab is up to, see the story beginning on page 14.

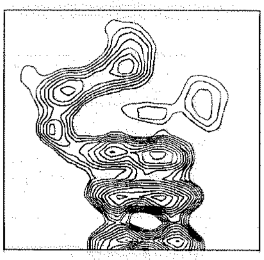
California Institute
of Technology

Engineering & Science

**Spring 1995
Volume LVIII,
Number 3**







On the cover: Events that precipitated last winter's El Niño could already be discerned in April 1994. This map, from TOPEX/POSEIDON satellite data, shows a pileup of warm water (correlated with higher than normal sea surface height, indicated by yellow through red and up to white) in the western Pacific. When the trade winds shifted by autumn, the warm water moved eastward, changing the path of the jet stream and causing heavy rains in California. For more on observing oceans from space, see page 2.

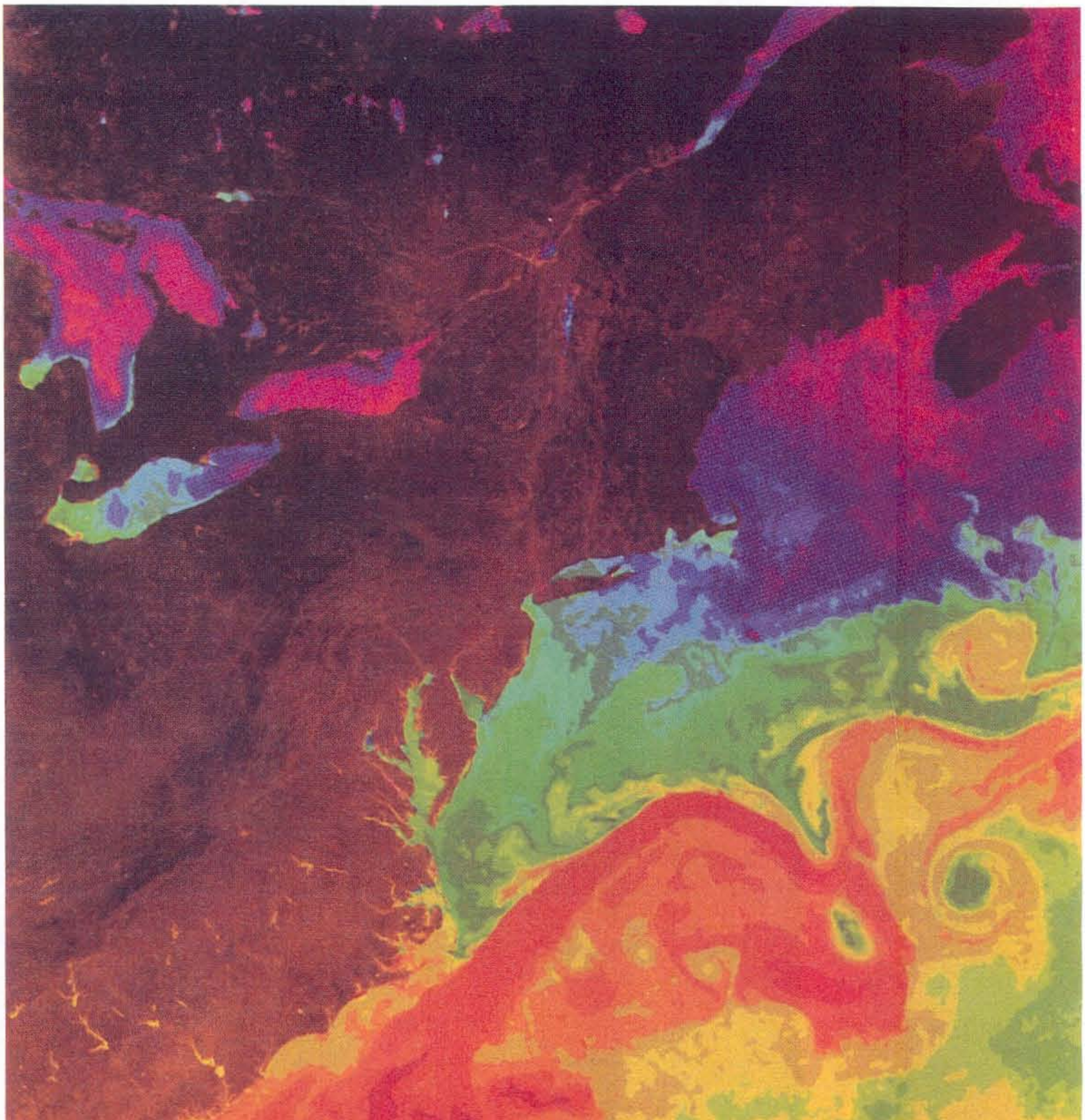
Departments

Engineering & Science (ISSN 0013-7812) is published quarterly, Fall, Winter, Spring, and Summer, at the California Institute of Technology, 1201 East California Boulevard, Pasadena, California 91125. Annual subscription $10.00 domestic; $20.00 foreign air mail; single copies $3.00. Third class postage paid at Pasadena, California. All rights reserved. Reproduction of material contained herein forbidden without authorization. © 1995 Alumni Association, California Institute of Technology. Published by the California Institute of Technology and the Alumni Association. Telephone: 818-395-3630. Postmaster: Send change of address to Caltech 1-71, Pasadena, CA 91125.

PICTURE CREDITS: Cover, 7, 8, 9, 11, 13 – JPL; inside front cover, 22, 23 – John Wright; 2 – NASA; 5 – World Meteorological Organization; 13 – R. Smith, Los Alamos National Laboratory; 14 – Michael Maher; 16–19 – Raanan Miller; 20, 24, 25 – Chang Liu; 22 – Hannah Dvorak; 30–34, 36 – Alexander Varshavsky; 37 – Heidi Aspaturian; 38 – Mark Davis; 39 – Kam Wan; 40 – Floyd Clark; 43 – Bob Paz; inside back cover – Los Alamos National Laboratory

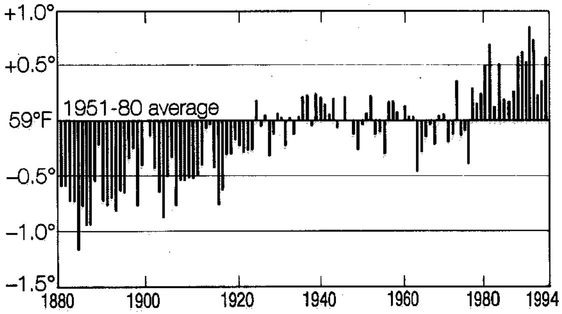
Peter V. Mason
President of the Alumni Association
Thomas W. Anderson
Vice President for Institute Relations
Robert L. O'Rourke
Assistant Vice President for Public Relations

STAFF: *Editor* — Jane Dietrich
Managing Editor — Douglas Smith
Copy Editors — Michael Farquhar,
Danielle Gladding, Julie Hakewill
Production Artist — Barbara Wirick
Business Manager — Debbie Bradbury
Circulation Manager — Susan Lee
Photographer — Robert Paz



The Ocean and Climate: Observations from Space

by Lee-Lueng Fu



Left: This infrared image, taken by a meteorological satellite, provides a colorized view of the sea surface temperature in the North Atlantic. The Gulf Stream (deep red at 25°C) flows up along the southeastern coast of the United States, bringing warm water from the tropics. Veering out to sea at Cape Hatteras, it mixes with the cooler waters offshore (blue is 10–16°C and magenta is 2–9°C).
Above, right: Globally averaged temperatures over the past 115 years show a net increase of about 1°F.

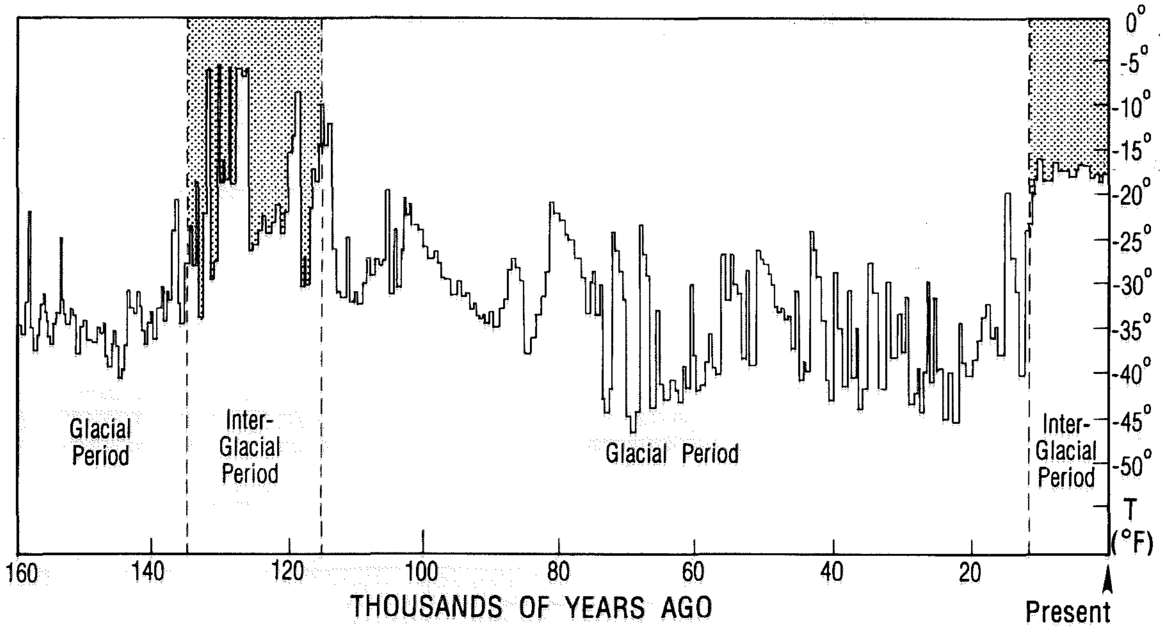
Eight hundred miles above the earth, a satellite named TOPEX/POSEIDON is observing the sea surface with radar, studying the ocean's currents and how they change with time. From this vantage point, all the world's oceans are in the satellite's view within a very short period of time. The radar can see through clouds, day and night, under all weather conditions, detecting even small movements of water to a high degree of accuracy. Now you're probably wondering: Why are ocean currents so important? And why do we bother to fly a satellite to study them?

The short answer is that we want to decipher the ocean's role in global climate change. Climate is long-term weather averaged over a season, a year, a decade, or even longer. It's not about rain or shine tomorrow, or even two weeks from now. It's about whether next winter will be warm or cold, dry or wet; it's about whether we're going to have frequent El Niño conditions in the next five years; and it's about the extent of global warming in the next 50 years. The ocean is the key to our understanding of climate and ultimately to our ability to predict it.

First, let's consider some realities of climate change, both at present and in the past. The Southern California floods this past winter were blamed on the returning El Niño in the Pacific. El Niño (named for the Christ child because its first noticeable effect usually comes around Christmas, when it causes warm currents to appear along the west coast of Central and South America) is an unusual warming of the tropical Pacific Ocean. The warm ocean alters the path of the jet stream in the upper atmosphere, which then changes the weather patterns all around the world. El Niño usually occurs once every three to five years, but lately the Pacific Ocean doesn't seem to be able to shake off a lingering condition that has prevailed for three winters in the past four years.

Many experts believe that this increased frequency of El Niños is caused by global warming, because the tropical ocean–atmosphere system is most sensitive to warming. As you can see from the record of temperatures averaged globally (above), there has been a net increase of about 1 degree Fahrenheit over the past hundred years. Half of this increase occurred in the 1980s, making it the warmest decade in this record. The warmest year was 1990, and after that came a few years of cooling caused by the eruption of Mount Pinatubo, which sent volcanic dust into the upper atmosphere, blocking sunshine. But the heat came back in 1994, making that the fifth warmest year of the century. This warming trend is believed to be a direct consequence of the buildup of carbon dioxide in the atmosphere, mostly due to the burning of fossil fuels, over the past hundred years or so.

An apparent result of this rising temperature is the increased frequency of severe weather—such as the deep freeze experienced in the eastern United States in the winter of 1994—despite the fact that that year as a whole was the fifth warmest year on record. Are we entering an unusual period of time, with three El Niños in four years as well as the warmest summer and the coldest winter in the same year? The answer depends on what we are comparing it to. In the



A 160,000-year temperature record from a Greenland ice core shows that temperatures over the past 10,000 years, during which human civilization developed, have been warm and stable. This was not so in earlier times, however, and frequent, sudden temperature swings were the rule. The temperatures here—from -5° down to -55°F —may seem a bit chilly (this *is* Greenland, after all), but other evidence indicates that the pattern of these fluctuations was typical of the whole planet.

ancient past, frequent abrupt climate change was actually the rule rather than the exception. A record of Greenland's temperature over the past 160,000 years, obtained from an ice core drilled more than 3,000 meters into the Greenland ice sheet, shows an interesting history over geological time. From the chemical properties of the ice, scientists can determine the temperature of the air when the ice was formed; other evidence suggests that these fluctuations were not just a local characteristic, but typical of the entire globe.

As you can see from the graph above, temperatures have been relatively warm and stable for the past 10,000 years, over which human civilizations flourished. The rest of the record, before the last 10,000 years, is characterized by frequent and abrupt change. This tells us that global temperature swings of more than 10 degrees Fahrenheit could happen in a period as short as 20 years, which is quite alarming. This record raises many questions: Why have the temperatures of the past 10,000 years been so stable? How long are we going to enjoy this present stability? What would trigger the instabilities and rapid climate swings that were so common in the past?

The answers to all these questions have a great deal to do with the ocean. The ocean is the flywheel of the climate engine, because it is the biggest repository for key elements of climate change such as water, heat, and carbon dioxide. The giant currents of the ocean transport these elements from one ocean to another, from the equators to the poles. They also control their exchange with the atmosphere, which ultimately affects the earth's climate and therefore our own lives. A few facts about the ocean will help illustrate its power and influence. The upper three meters of the ocean (of its average depth of 4,000 meters) stores the same amount of heat as does the entire atmosphere. The heat transport of the North Atlantic Ocean is a hundred times the man-made energy production of the entire world. And 99 percent of all the carbon dioxide that has ever existed in the atmosphere now resides in the sediments at the bottom of the ocean.

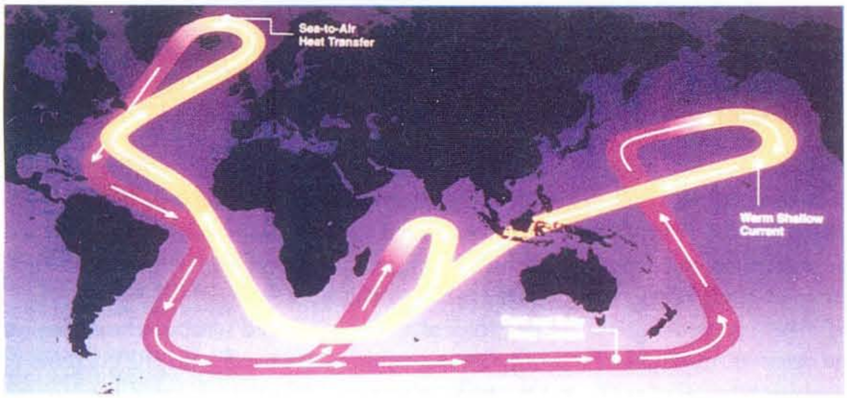
Let's focus first on heat transport, shown as a conveyor belt in the schematic diagram on the opposite page. This is an overly simplistic picture of a highly complex process, including only one of many important components. The warm surface water brings heat from low latitudes all the way to the northern North Atlantic, where it transfers the heat to the atmosphere. Then the water gets cold and heavy, begins to sink to the deep ocean, and returns to the tropics, where the cycle begins again. The efficiency of this conveyor belt controls the climate, especially in the northern hemisphere. The faster the water sinks in the north, the more efficient the belt, and the warmer the climate. Conversely, if the water sinks slowly in the north, the efficiency decreases, and the climate becomes colder. That was indeed the case during the Ice Ages.

The rate at which water sinks in the north is controlled by the ocean's temperature and salt content. The salt content may be the real key to the switch of this conveyor belt. If the water is too fresh, it's not heavy enough to sink. The salinity, in turn, is controlled by many things, including the patterns and the rate of ocean currents, the mixing, the precipitation and evaporation and, perhaps most important, the formation and melting of the ice in the region. These processes are all interrelated, making the conveyor belt potentially prone to instabilities and rapid changes.

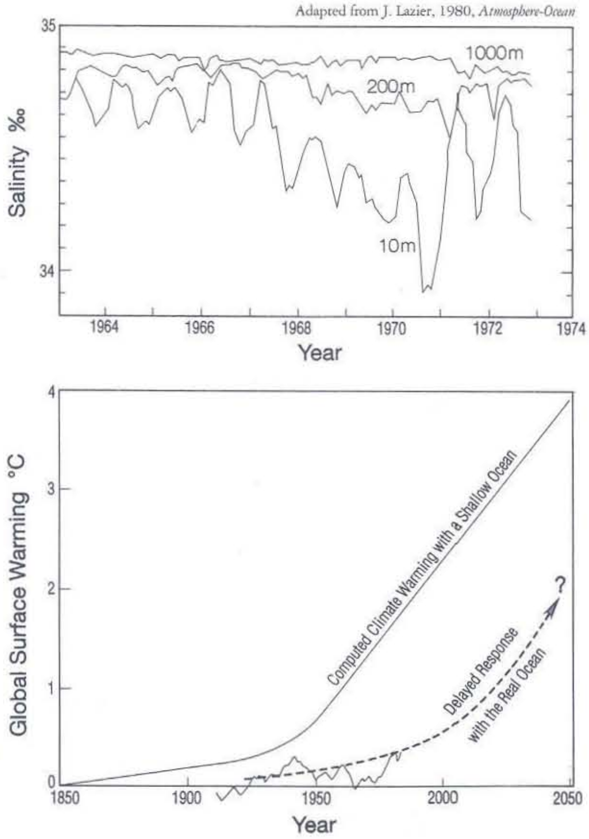
An alarming event in the Labrador Sea (in the western corner of the northern North Atlantic) in the late sixties illustrates this delicate balance. The graph in the middle at left shows the salinity at three levels: 10 meters (that's almost surface water), 200 meters, and 1,000 meters. The salinity increases with depth, so the surface water is fresher. At the beginning of the record in 1964 the temperature is low enough and the salinity high enough during the winter, so that the water sinks to mix the upper water column, making the salinity almost the same at all three levels. This wintertime convection process suddenly disappeared between 1968 and 1971, probably due to a temporary increase in unusually fresh water input to the region from the Greenland Sea to the north. You can see that during wintertime the ocean was still stratified; the salinity varied at the different depths, and the sinking process stopped. The extent of the event was small, and it had no significant effects on climate, but it was alarming nonetheless. We don't know the complete story of this incident. In order to diagnose such a problem you need to know what's going on in the entire North Atlantic and its overlying atmosphere for a long period of time, and at that time we didn't have that knowledge. Even now we don't yet have the observations and understanding required to predict whether a full-blown shutdown of the conveyor belt, possibly bringing the Ice Age back, is likely or not in the near future. This is because the actual process of oceanic heat transport is far more complicated than the schematic diagram indicates. It involves currents of very complex, three-dimensional structures, which are difficult to construct in a computer model. Current climate models usually treat the ocean as a shallow swamp and describe an oversimplified coupling with the atmosphere. When they try to make predictions, more often than not these models fail.

One model, using a very shallow ocean, predicted global temperature from 1850 to 2050, based on the recorded and projected levels of carbon dioxide. This model (left) predicted the temperature over the past 100 years to increase by 2 degrees C, but the actual temperature increase was quite small—0.6 degrees C, or about 1 degree F. We know from this that the current climate models don't work, because they cannot reproduce what we observe has happened. It may

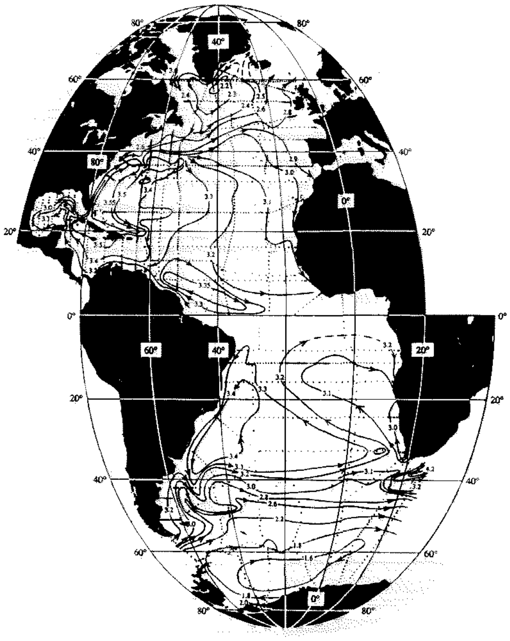
Below: The heat transport system of the ocean is like a conveyor belt. Currents of warm surface water bring tropical heat to the North Atlantic, where it's exchanged with the atmosphere. The now-cooler water sinks, returns to the tropics, and begins the cycle anew.

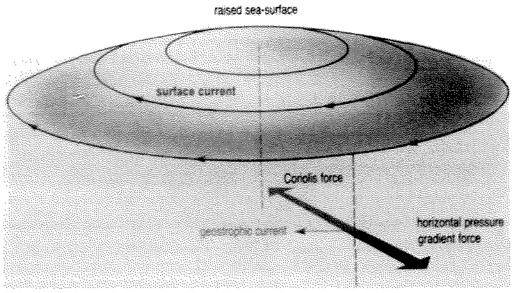


Adapted from J. Lazier, 1980, *Atmosphere-Ocean*



Right (top): The drop in salinity (shown here in parts per thousand) of an area in the North Atlantic threatened to disrupt the heat-transport conveyor belt in the late sixties. Usually, wintertime sinking of surface water leaves the salinity well mixed and fairly equal at 10 m, 200 m, and 1,000 m below the surface. But between 1968 and 1971 surface water was quite fresh all year round, indicating that the sinking process had mysteriously stopped. Right (bottom): Based on the record of carbon dioxide since 1850, this model predicted a 2°C rise in temperature by the year 1990. It has actually risen only 0.6°C. This might be a delayed response due to the ocean's high heat capacity.


From J. Reid, 1994, *Progress in Oceanography*



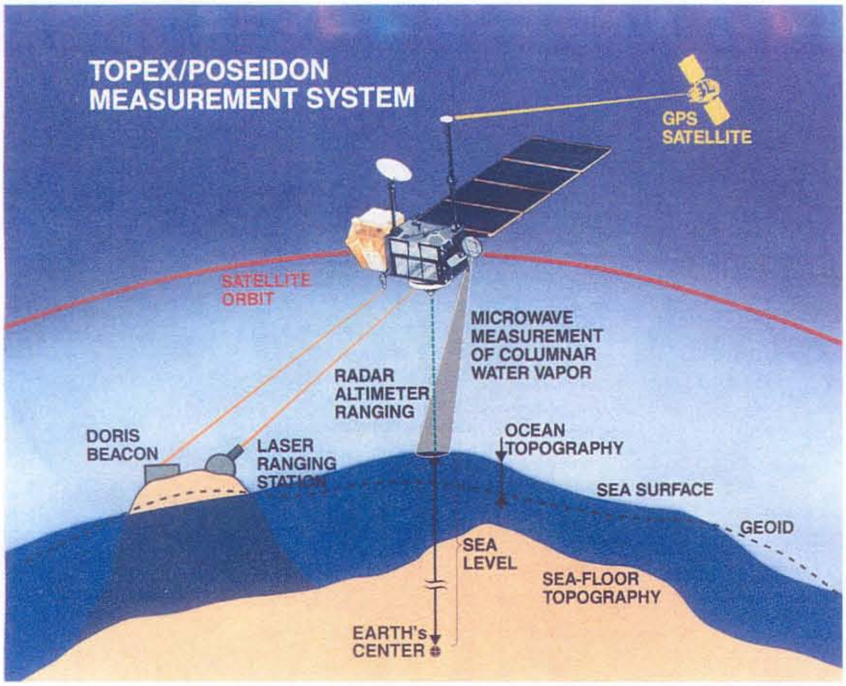
Right: Before satellites, oceanographers could draw only very simple diagrams of the circulation of the ocean's currents from instruments placed in the ocean itself.
Far right: The balance (called the geostrophic current) between the Coriolis force (from the earth's rotation) and the horizontal pressure created by an ocean current pushes the water up into a hump. A current's speed can be calculated from the slope of the hump, or, in other words, the shape of the sea surface elevation.

be that the response to global warming is delayed because of the high heat capacity of the real ocean, but the kind of model that could incorporate this complexity doesn't yet exist. A major reason for the slow development of ocean models is the lack of adequate global observations. In the past we learned about the ocean from piecemeal data taken from ships—ships that take months to cross the ocean at the speed of a bicycle. And during this time the ocean is constantly changing.

Using sparse data taken in different seasons, in different years, and with the radical assumption that the ocean doesn't change, oceanographers have been trying to draw ocean circulation diagrams for the past hundred years. An example is shown above. The resulting picture is inevitably distorted or much too smoothed out, but it has proven useful for a *qualitative* climatological description. In fact, most of our knowledge of the ocean circulation has been obtained this way. But for a *quantitative* analysis of a very complex system like climate, it is totally inadequate. It can't come anywhere near the spatial scale of the infrared image of the sea surface temperature in the North Atlantic taken from space, shown on page 2. The temperature reflects the pattern of ocean currents to some extent; you can see the Gulf Stream and the eddies that surround it. Every week this current system changes. To resolve the ocean currents in both space and time, you would have to sample the ocean every 50 kilometers. A rough calculation suggests that you would need 200,000 permanent stations in the ocean in order to do adequate sampling for quantitative analysis. This is out of the question. Satellites turn out to be the only way to study the global ocean at the required resolution.

But what can we observe about ocean currents from space? Although sea surface temperature is easy to detect using infrared sensors, the relationship between this temperature and the currents is not straightforward. We need three-dimensional ocean currents to solve climate problems, but sea surface temperature doesn't reveal much information about what's going on below the surface. So we use a radar altimeter to measure the shape of the sea surface. This approach is based on a simple principle that can be explained with an analogy to a cup of coffee. If you stir coffee in a cup, circularly, it will create a depression in the surface. Smart undergraduate physics students can calculate the velocity of the coffee everywhere in the cup just by looking at the shape of the surface, because there's a balance of forces between the pressure (caused by the depression in the surface) and the centrifugal force (caused by the circular velocity of the coffee). The only difference in the ocean is that the balance is between the pressure force (caused by the currents we want to measure) and the Coriolis force, a force that is exerted on every moving object in a rotating frame. If you roll a marble on the floor of a merry-go-round, for example, the marble cannot roll straight; it has to deflect either to the left or to the right, depending on which direction the merry-go-round is rotating. Similarly, in the rotating frame of the earth currents are deflected to the right in the northern hemisphere and to the left in the southern hemisphere until the

Rough seas sometimes have waves several meters high; how are you going to determine mean sea level of rough seas to within a few centimeters?



TOPEX/POSEIDON's radar altimeter bounces pulses off the sea surface, measuring the distance between the satellite and the sea surface. By subtracting that distance from the radial orbit height (the distance from the satellite to the earth's center) you can calculate the sea level. Then the geoid (the influence of gravity on sea level) has to be subtracted from the sea level to obtain ocean topography. To pick up a signal of a couple of inches, the satellite also has to compensate for water vapor, using a microwave radiometer, and to establish its own position in space within a couple of inches, using lasers, the DORIS microwave system, and the global positioning system.

Coriolis force is balanced by the pressure force.

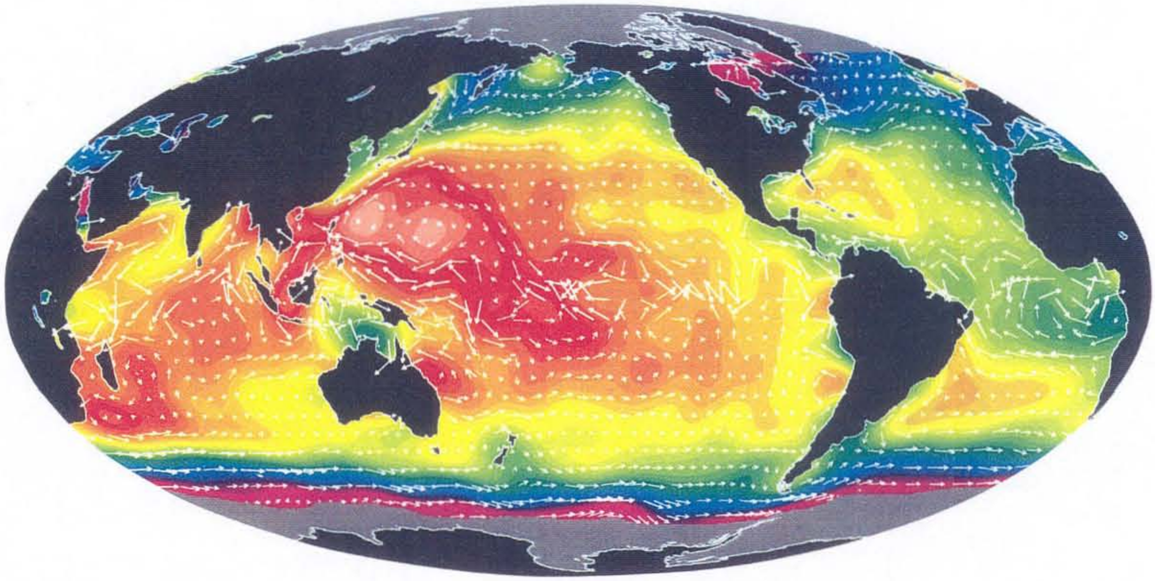
So the combination of the current's pressure and the Coriolis force pushes the sea up into a mound or a dip, and scientists can calculate the current's speed from the mound's, or dip's, slope. This sea surface elevation is to oceanographers what air pressure is to meteorologists; a map of the sea surface elevation is the equivalent of a map of surface pressure. Just as from the lows and highs on the surface pressure chart, meteorologists can tell you the wind speed and direction, with a chart of sea surface elevation oceanographers can tell you the speed and direction of ocean currents, not only at the surface, but at depths with the aid of a model, which I'll discuss later.

Measuring the shape of the sea surface elevation from space is also based on a simple principle. A radar altimeter on the satellite sends radar pulses to the surface of the sea, which bounces the pulses back. We can measure the round-trip travel time of the pulse and calculate the distance between the radar and the sea surface. But what we really want to know is the elevation of the sea surface relative to the center of the earth, so we have to know the precise height of the satellite, called the radial orbit height. Then we subtract the distance we measured with the altimeter from the radial orbit height to get the sea level relative to the center of the earth. Ocean currents do not actually control the shape of the sea level. The most important force is the earth's gravity field: variations in gravity caused by uneven density distributions in the earth's crust create sea level changes of hundreds of meters in different parts of the ocean. The ocean currents deflect the sea surface from the gravity surface (which we call the geoid) by only two meters, or 1 percent of the total variation of the sea level. It's this 1 percent that we're looking at. Temporal changes in ocean currents, which is what we're *really* interested in, create a change of only 10 to 20 centimeters—10 percent of the total 1 percent signal. So, to measure global changes in ocean currents, we have to be able to measure the sea level to within a few centimeters, or a couple of inches. That's a challenge. Rough seas sometimes have waves several meters high; how are you going to determine the mean sea level of rough seas to within a few centimeters?

In the early eighties two groups of scientists and engineers (one from France and the other from the United States) believed this could be done. They eventually joined forces and proposed a mission called TOPEX/POSEIDON. TOPEX, for "ocean topography experiment," was the original name of the U.S. mission; the French scientists named their mission after the Greek god of the sea. The two governments approved the mission in 1987, and the satellite was launched in 1992 by a French Ariane rocket. The satellite contains several instrument systems: one of them, the radar altimeter, sends pulses to measure the range to the sea surface. Because of the rough seas, it sends thousands of pulses every second to average out the wave effects. Tides, on the other hand, which move the sea surface up and down by about one meter, are quite easy to deal with, because the frequencies of the tides are well known. The satellite's orbit, which determines how the ocean is sampled in time, was planned so that the tides could be determined precisely by the satellite and removed from the signal.

Many things interfere with this signal; for example, the free electrons in the upper atmosphere and the water vapor in the lower atmosphere slow it down. To correct for the first we send the pulses in two radio frequencies. Because the delay is a function of frequency, if we combine these two frequency measurements, we can retrieve the signal's delay and make corrections for the electron effects. To correct for the second, a radiometer measures the total water vapor content of the atmosphere. Actually, only a tiny portion of the atmosphere has water vapor, but it's enough to slow down the signal and we have to correct for it.

We also need to know where the satellite is in space to within a few centimeters. We have three systems to do that job. One is traditional laser range finding, which uses the round-trip travel



This map from TOPEX/POSEIDON data represents the average relief of the ocean topography from September 1992 to September 1993. With the geoid (the large variation caused by gravity) removed, the variation covers a range of two meters, from the lowest (magenta and blue) near Antarctica, to the highest (red and pink) in the western Pacific, which stands about half a meter higher than the Atlantic. The Pacific's larger size allows the winds room to raise the western Pacific and create the highest sea surface elevation. Calculated currents are shown by the white arrows (each arrow is about 10 cm/sec). Gyres, the large recirculating cells in the western ocean basins, are part of the permanent system of circulation—the climatology of the ocean.

time of light to determine the distance between the satellite and the laser station. A second system, called DORIS, consists of an antenna that receives microwave signals from a ground network of beacons. From the change of the frequency due to the motion of the satellite (the Doppler effect) you can determine its velocity. The third system is the global positioning system, which has many applications, including determining the position of tanks to within a few meters during the Persian Gulf war. That was good enough for the military, but we have to determine the center of mass of the satellite, which is about the size of a Greyhound bus, within about an inch.

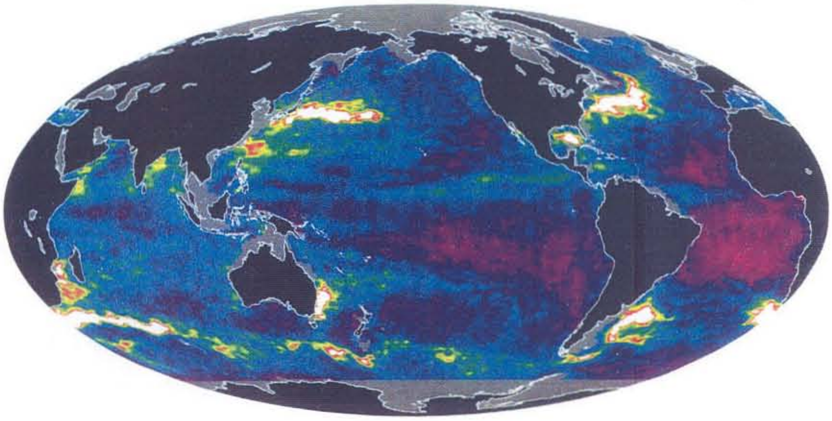
Satellite radar altimetry began with SEASAT, launched by JPL in 1978. The uncertainty of SEASAT's radial orbit height was one meter, so it couldn't resolve (nor could the satellites that followed it) the changing part of the ocean's large-scale signal, which is about 10–20 centimeters. With TOPEX/POSEIDON we achieved a measurement accuracy of better than five centimeters for the first time, and were able to resolve the changing sea surface elevation at even the largest scales.

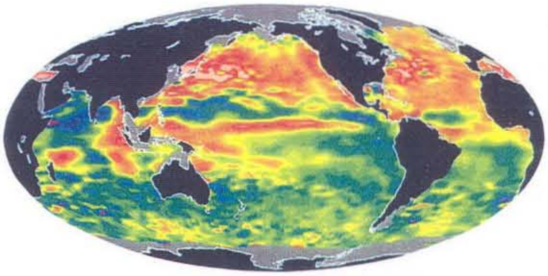
Every 10 days the satellite makes measurements along exactly the same ground track, so that we can compare one cycle's measurement with the next and then determine precisely how the ocean changes with time. The moment we got our first map from the satellite was very exciting; it was the first snapshot of the ocean's currents from space. No more waiting for months for a ship to cross the ocean just to collect one single section of the ocean. Now, in 10 days we could have it all. The amount of data contained in one 10-day record is equivalent to all the data collected over the past 100 years. The map above shows, in false color, the relief of the ocean topography, which covers a range of two meters. And every 10 days we get a map like this. They all show basically similar features—the semipermanent systems analogous to such features as the Aleutian low and the Siberian high in the atmosphere. The gyres, the large circulating systems of water on the western sides of the ocean basins, are permanent ocean systems, although their details change.

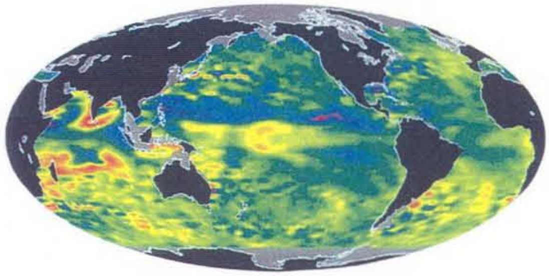
When we remove the average elevation, as calculated from the first year's data, what is left is the temporal change. Then we average that for each season to get the deviation, or the change of the sea level from its mean, during the four different seasons. The scale here, in the maps at right, is no longer two meters, but ranges from minus 15 to plus 15 centimeters. It is these small changes in the ocean that carry the signal for climate consequences.

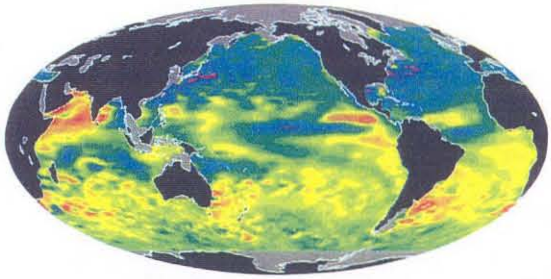
Sea level changes inherently with the seasons. The highest sea level occurs in the fall because it takes time to heat the ocean. After a whole summer's heating of the sea surface, the heat content reaches a maximum in the fall, and thermal expansion raises the sea level to its highest point. And, conversely, after a whole winter's cooling, the lowest sea level occurs in spring. Again, the maximum seasonal change occurs in the western part of the ocean, because of the rotation of the earth. If the earth rotated the other way, you

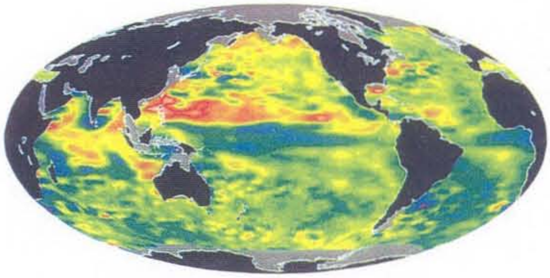
The map at right illustrates a year's summary of random fluctuations of ocean currents—the ocean's storms. Magenta (0 to 5 cm) and blue (10 cm) represent the most stable regions of the ocean, while the red (20 cm) and white (30 cm) show areas of turbulence and instability, most notably the warm Kuroshio current off Japan and the Gulf Stream in the North Atlantic.







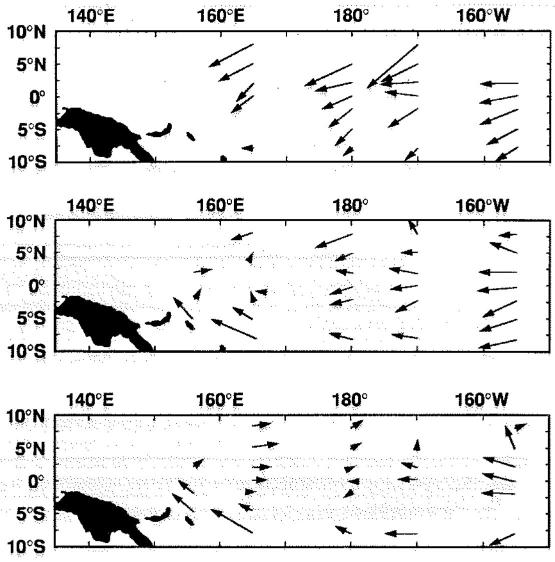




Averaged for season, the TOPEX/POSEIDON data show a deviation from the mean sea surface height of from –15 cm (magenta) to +15 cm (pink); yellow-green means zero change, and red is +10 cm. After a summer's heating, the highest sea surface elevation occurs in the fall (top); then the surface cools off in winter and reaches its lowest point in spring before starting to warm up again in summer (bottom). The highest seasonal change occurs in the western part of the oceans because of the earth's rotation. Greater land mass in the northern hemisphere makes for greater variation.

would see the gyres and the highest seasonal variation on the eastern side of the ocean. Note, too, that the southern hemisphere has a similar seasonal change, but its intensity is much lower. This is because there is less land in the southern hemisphere to provide the severe cold air that blows out from the continental interiors during the winter and cools the oceans in the northern hemisphere. The southern hemisphere contains mostly ocean, creating a steadier climate with less seasonal change.

In addition to seasonal change, the ocean has its own weather. In the atmosphere weather consists of random fluctuations of air flow; the ocean weather is random fluctuations of ocean currents. These are the ocean's storms. A summary of a year's observations shows, in the map above, the typical magnitudes of sea surface change resulting from ocean storms. The range from red to white represents about 20 centimeters. Off Japan you can see the famous Japan Current (the Japanese call it the Kuroshio, which means "black current"), which is the ocean's version of the atmospheric jet stream. It has a lot of the same turbulence and instability as the jet stream. Typically this causes a 20–30 centimeter change in this region's sea level, but the maximum can be as high as one or even two meters from very severe storms. This is also the case in the Gulf Stream region and in the Antarctic Circumpolar Current. Ocean storms are much smaller than atmospheric storms, with a diameter of 50–100 kilometers, as opposed to the 1,000-plus kilometers of atmospheric storms. So, to resolve all these ocean storms in a giant



Wind patterns in the El Niño of 1994 show the trade winds blowing strongly westward in April (top), pushing the warm surface water to the western Pacific (that's New Guinea at lower left). In July (middle) the winds started to grow disorganized in the west, and by October (bottom) had reversed direction.

computer model, we have to have a much higher spatial resolution than the atmospheric models have.

In addition to seasons and weather, the oceans also have unusual events on larger scales of space and time. One is the famous El Niño phenomenon, which we in Southern California have become very familiar with in recent years. In a normal December, the strong trade winds, blowing westward, push the warm surface water against the western boundary of the Pacific Ocean. The air rises in the warm western Pacific, and the rainfall comes down in Indonesia and Australia. If most of the warm water is pushed westward, the cold water has to come up to compensate for it, welling up along the west coast of South America and bringing the nutrients that make for good fishing here in a normal season.

During an El Niño year, the trade winds weaken and even reverse direction. (The trades are controlled by an inherent oscillation mechanism between the atmosphere and the ocean, which is caused by the sea surface temperature.) So this huge mass of warm water in the western Pacific is no longer pressed against the ocean boundary and begins moving eastward. As it does so, it sends a large number of wave pulses called Kelvin waves after Lord Kelvin, the British scientist who first studied them. These waves send a signal back east, changing the internal density structure of the ocean, and allowing the warm water to continue on its path. As the warm water moves eastward, it occupies the entire tropical ocean, and as the tropical Pacific Ocean warms up, convection occurs in the middle of the Pacific. Then torrential rain falls in places like the Christmas Islands and the Marshall Islands; Indonesia and Australia experience severe drought. Australia is experiencing its fifth year now of drought due to a lingering El Niño.

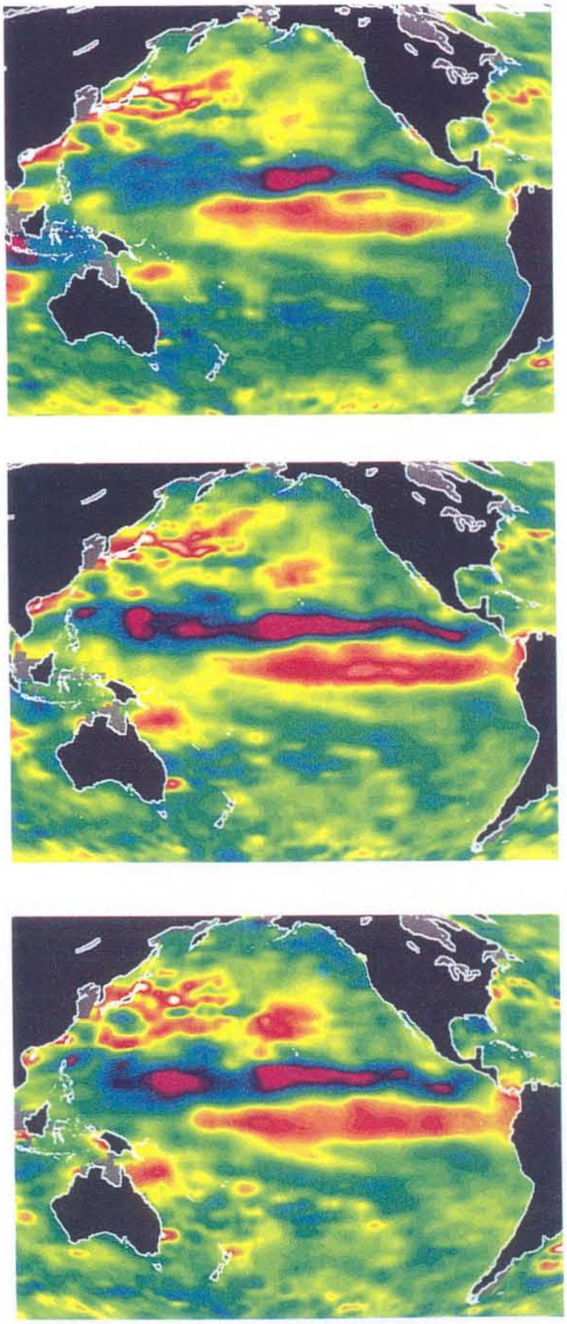
At left you can see the progression of events that led to the heavy rains in California this past winter. The wind, reported from an array of buoys on the equator in the Central Pacific, was normal in April 1994. The trade winds were blowing strongly westward. In July the trade winds in the western part of the ocean became disorganized, and in October they changed direction. This is the classic sequence leading to El Niño. Last April, when the trade winds were blowing strongly, the highest sea level (15 to 20 cm above normal) occurred in the western Pacific because the wind was piling up the warm water there (see cover). Cold water welled up along the South American coast. After the disorganized wind in July, the warm water in the western Pacific moved to the east (opposite page) in the form of Kelvin waves in the late fall—four pulses of them, the largest in November—setting the stage for the heavy rains we experienced in January. As late as January, these conditions were still lingering, but by March they were beginning to disperse.

As the currents bring the warm water to the colder part in the east, they feed the heat to the atmosphere, changing the path of the atmosphere's jet stream. During normal times, the jet stream's path goes across America's northern states and brings the winter storms along with it. When El Niño occurs, the warm sea surface temperature diverts the jet stream to the south, bringing heavy rainfalls to California and the Gulf States, as well as relatively warm winters to the northeastern states.

On a larger scale, there's another phenomenon with far vaster potential effects than El Niño, and that is the mean sea level variations in response to global warming. There are two causes of sea level increase. One is thermal expansion: when temperature rises, the ocean occupies a greater volume. Over the past hundred years sea level has risen 15 cm for about a half degree C of temperature rise. Most computer models predict about three degrees (ranging from 1.5 to 4.5°C) of warming under the scenario of doubling of carbon dioxide in the atmosphere by the end of the next century. If we extrapolate this linearly, we get about one meter of sea level rise.

In the past we had to rely on tide gauges sparsely distributed around the ocean. Since many oceanic phenomena such as El Niño can

On a larger scale, there's another phenomenon with far vaster potential effects than El Niño, and that is the mean sea level variations in response to global warming.

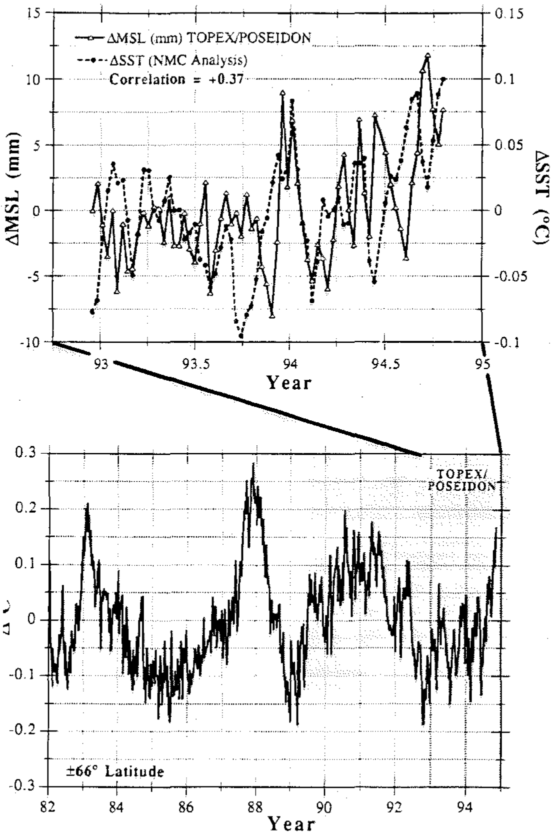


In TOPEX/POSEIDON's measurements of sea surface height, the development of the El Niño during the fall of 1994—October (top), November (middle), and December (bottom)—is clearly visible. Yellow-green represents normal height, shading below normal through blue to magenta (–15 cm), and above normal through yellow, red (+10 cm), to white (+15 cm). When the trade winds reversed in October, warm water pulses moved eastward in the succeeding months, hitting Central and South America, diverting the jet stream, and ultimately bringing heavy rains to California.

create a large local sea level change, the average measurement from such gauges can be distorted. But when we have a satellite giving us half a million observations in just one 10-day cycle, we have a much more accurate measurement of mean sea level rise. If the predicted one-meter sea level rise is correct, it will create enormous problems worldwide. About 3 percent of the earth's land, which is home to about 20 percent of the world's population, will be affected. Dams to hold back the sea would cost hundreds of billions of dollars. But because there's a large element of uncertainty about these predictions for sea level rise, it's an urgent task to obtain a reliable measurement of the sea level trend to determine whether it will be disastrous or relatively benign. Actually, most of the models predict that we will have about a half meter increase in the sea level, even with 3 degrees of temperature increase.

The effects of thermal expansion pale, however, in comparison to the second phenomenon, and that is the melting of ice, in particular the potentially unstable West Antarctic ice sheet, creating a sea level rise of up to five to six meters. Most climatologists assure us that this won't happen in the near future, because the upwelling of cold deep water surrounding Antarctica shields the ice sheet to some extent from the heat of the low latitudes. But there's still a big uncertainty there, which underscores the importance of having a reliable way to monitor sea level rise.

On the following page is the record of mean sea level based on two years of TOPEX/POSEIDON data. You can see a linear trend, with quite a lot of fluctuation, showing about a six-millime-



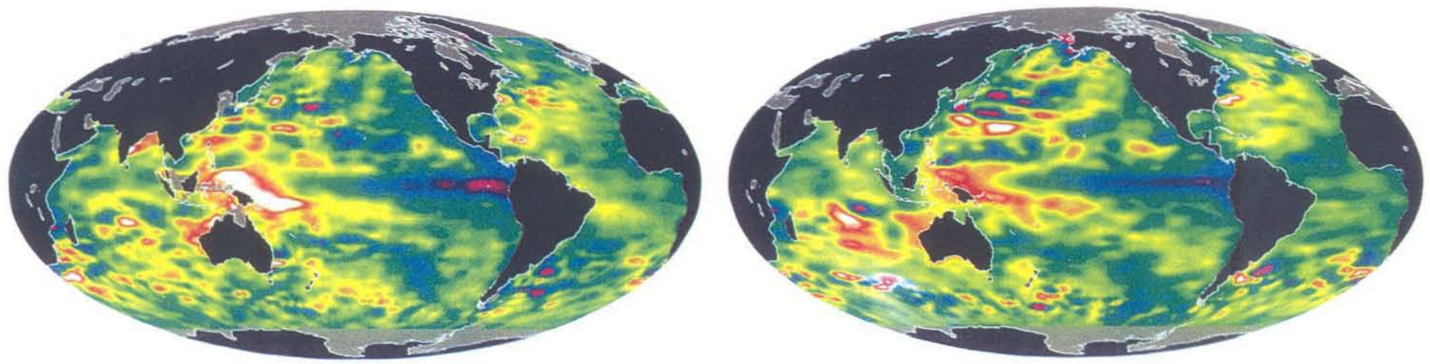
The top graph shows how the change in mean sea level (left axis) follows the change in mean sea surface temperature. The solid line comes from TOPEX/POSEIDON observations over the past two years, and the dotted line is the temperature (right axis in degrees Celsius). The upward trend may not indicate global warming but may be only a transitory expression of El Niño. Over a longer term (lower graph) temperature peaks have corresponded to the El Niños in 1982–83 and 1986–87. (Courtesy of S. Nerem of NASA Goddard Space Flight Center.)

ter sea level rise over this period. The other curve is the mean sea surface temperature in the ocean for the same period of time. You can see that the sea level follows the temperature. In those two years temperature rose about 0.15 of a degree C, but we need to compare this with a longer record to get some perspective on what it means. If we look at the past 10 years, we see that most of the fluctuation in sea surface temperature corresponds to El Niños: 1982–83, the biggest El Niño ever recorded, and 1986–87. So with a short record like this we have to be very cautious; what we see here may not be a long-term trend, but simply a temporary fluctuation caused by El Niño. On the other hand, it's reassuring to have proof that the mean sea level does correspond to the temperature. This lends a lot of credence to the measurements from space. But we will need a long-term record to give us an indicator of how fast sea level is really rising as a result of temperature change.

Now that we have our first global ocean observing system, how are oceanographers going to put this wonderful data stream to work to help improve climate prediction? Meteorologists' methods of weather forecasting make a good comparison here. A successful weather forecast needs three elements: weather satellites, a sophisticated computer model, and a ground network of weather stations. For climate prediction we now have satellite observations of the ocean. Fortunately, in the past five years, parallel to the development of satellite technology, computer technology has also taken off. Massively parallel computing allows oceanographers to resolve all the ocean storms in the system for the first time. We can now produce a credible picture of global ocean circulation just through number crunching. (Compare the computer-model map on the inside back cover with the similar infrared image shown on page 2.)

We can also compare TOPEX/POSEIDON maps and computer models of the intensity of ocean storms, of seasonal change, and of yearly change after subtraction of seasonal fluctuations. The most interesting comparison is that of the yearly, or interannual, change—the change in a particular month from one year to the next. At left on the opposite page is TOPEX/POSEIDON's observations in the difference of the sea level in April 1993 and 1994 (1994 minus 1993). In 1994 you can see the buildup of El Niño and a much higher sea level in the western Pacific than in the same month a year earlier. The map next to it on the right was produced by a state-of-the-art model, and shows very high correlation with the actual observations. So we know now that this interannual change, this climate change in the ocean, can be simulated by a model very well. There are differences between the observations and the models, but the similarities are encouraging, and the differences also tell us that we need to combine these two technologies to achieve an optimal description of the ocean.

But can we just let these models run, to make predictions? The answer is no, because the ocean model, like the atmospheric model, is highly nonlinear; it has a chaotic character. A chaotic system is characterized by the fact that it takes only an infinitesimal change in the initial conditions of a prediction to arrive at entirely different results. That's the famous butterfly effect: a butterfly flapping its wings in the jungles of Brazil sets off an unexplainable chain



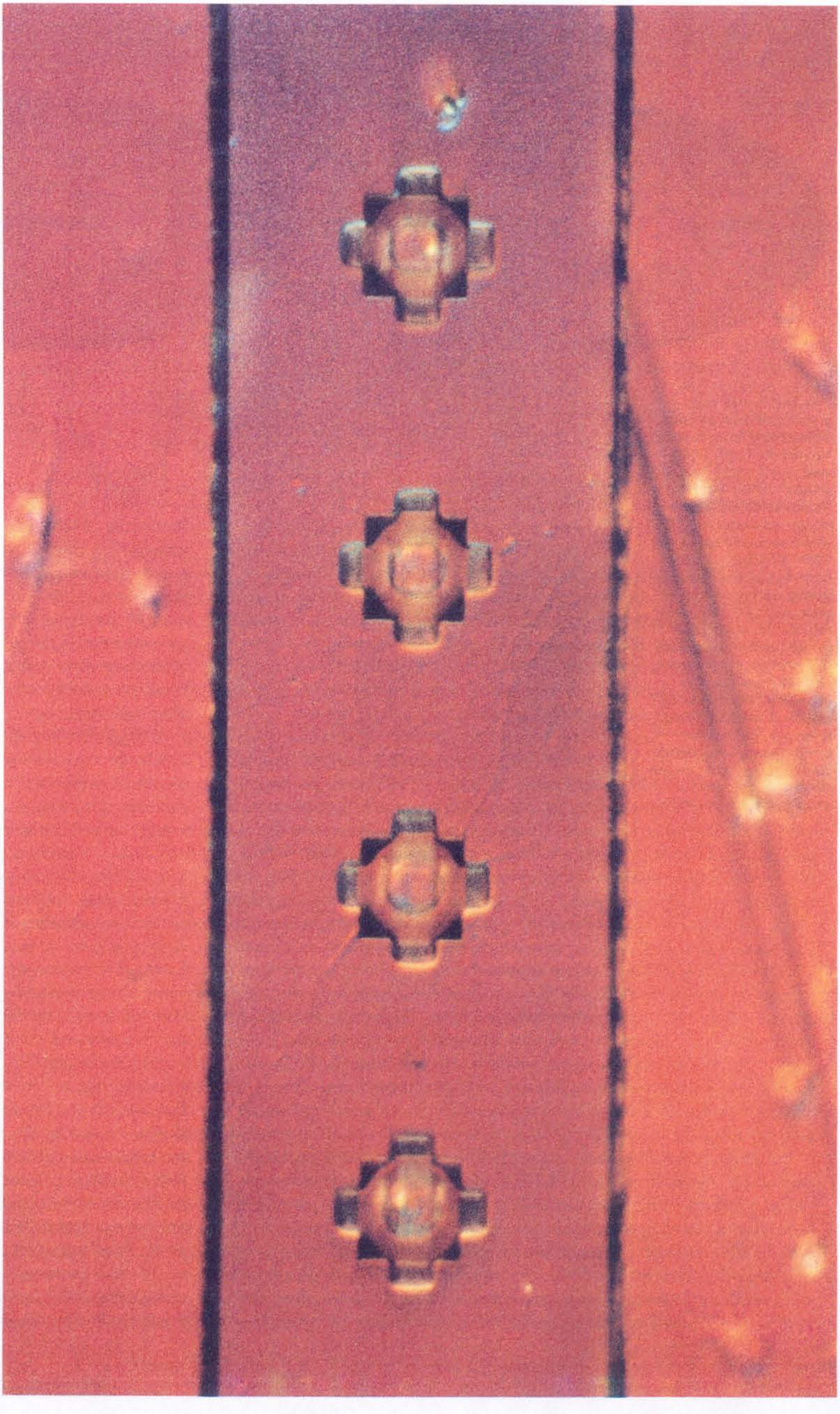
The left-hand map—a comparison of the difference in sea level height in April 1994 and April 1993 (1994 minus 1993) from TOPEX/POSEIDON data—shows the obvious buildup of the 1994 El Niño. Again, zero is yellow-green, going up to yellow (5 cm), red (10 cm), and white (15 cm). The corresponding trough of lower-level magenta and blue can be seen off the coast of South America. The map on the right, which was constructed from a state-of-the-art model, shows a very good correlation.

of events in the atmosphere that produces a storm in China a week later. So, no matter how accurate your model is, you can't just let it run by itself. You will always need observations to adjust the model via a technique called data assimilation, originally used by meteorologists. Like meteorologists, oceanographers depend on fresh data to keep their forecasts on track as well, so that they don't drift away over time.

Now we have global observations and a credible model, so we can assimilate satellite data and make predictions. But we still need the third element—a ground network of in situ observations, to produce reliable three-dimensional pictures of the circulation structure (rather than the shallow swamp that was the basis for earlier models). This is crucial in order to calculate the heat transport and make a correct prediction about the conveyor belt. So we also need to have deep-ocean observations to validate our computer calculations constrained by satellite observations. If it's consistent—great. If there are discrepancies, then we know where to concentrate our ocean observations. We don't have to populate the ocean with a hundred thousand stations, but only need to place them in a few strategic locations—those where the satellite and the model can't reproduce the real features. So in parallel with TOPEX/POSEIDON, we have a field campaign involving 40 nations around the world, called the World Ocean Circulation Experiment. A large number of different types of instruments have been deployed in the ocean over the past three years, an activity that will continue in the years to come. This experiment will provide a framework in which we can combine these observations with those from the satellite and the computer models to define a global climate prediction system. It will rely heavily on models and satellite data, with a minimum requirement of measurement in the sea, but it's the combination of all three of these things that should make a breakthrough in better climate prediction in the years to come.

TOPEX/POSEIDON will probably fly for another three or four years or possibly longer. Ocean climate study, however, is a long-term commitment. The phenomena we need to observe exceed the life cycle of a single mission, and we're not going anywhere unless we obtain at least a 15- or 20-year record. Realizing this, the United States and France are planning to continue precision altimetry measurement into the next century as part of NASA's Mission to Planet Earth. We're entering an era, a very exciting one, in which our investment in space will pay off with the knowledge for predicting the future of our own planet and helping us to prepare for inevitable change. □

Lee-Lueng Fu is a senior research scientist and head of the Ocean Science Group at the Jet Propulsion Laboratory where, since 1980, he has helped develop the new field of the study of oceanography from space. He is also project scientist on the TOPEX/POSEIDON mission, which is managed by JPL. Fu received his BS in physics from National Taiwan University in 1972 and his PhD in oceanography from MIT and Woods Hole Oceanographic Institution in 1980. This article is adapted from his Watson Lecture, given last March.



Report from a Small World

The trick to micromachining—and a big reason why the field is still in its infancy—is to figure out how to make things with moving parts, but using tools designed to manufacture immobile electronic circuits.

by Douglas L. Smith

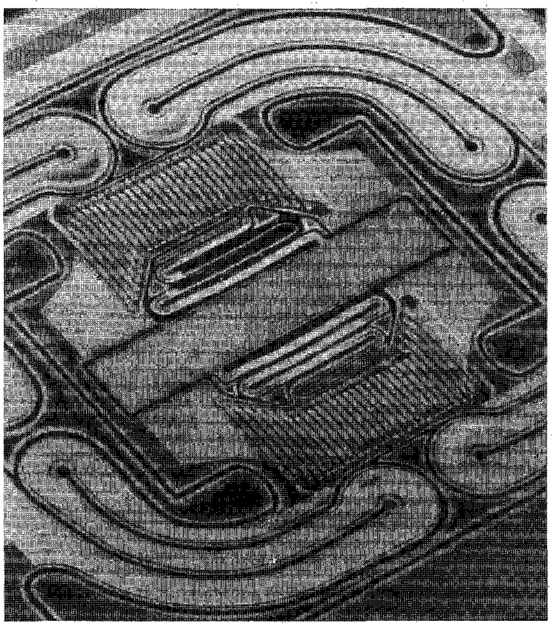
What looks like a Navajo blanket here is actually a probe to study brain function. The reddish-brown strip down the middle of this photomicrograph is part of a silicon needle 0.15 millimeters wide. The four crosses down the center are wells the size of single nerve cells. Immature nerve cells have been implanted in the wells, and the probe will soon be inserted into a living brain, where the researchers hope that the probe cells will wire themselves into the brain's circuitry. (The second nerve cell from the bottom has already begun to send out "feelers" in search of other nerve cells.) The orange-red background is a nerve-cell culture medium; several nerve-cell bodies can be seen in it as light-colored blurs.

You may remember a photo of three intermeshed gears that *Time* magazine ran back in 1989. These gears, made at Bell Labs, were noteworthy in several respects: each tooth was the size of a blood cell; the gears, their axles, and their enclosure had been carved from a silicon chip with standard integrated-circuit-making technology; and they actually worked! Blow a puff of gas across the end one, and all three spun. The accompanying article described how several labs were making tiny springs, itty-bitty motors, and other microcomponents that might some day be assembled into microrobots that would cruise through your bloodstream like roving Public Works Department crews. "Dr. Iwao Fujimasa, a cardiac surgeon at Tokyo University, is building a robot less than one millimeter (0.045 inches) in diameter that could travel through veins and inside organs, locating and treating diseased tissue." The good doctor hoped to have a prototype to test on horses in three years, subject to the availability of parts—robotic, not equine.

Five years have come and gone, and if there *is* a microrobot jackhammering arterial plaque deposits somewhere out there, it's a safe bet that your HMO won't cover the procedure. Although microelectronic circuits are now as cheap as dirt and as pervasive as paper—you can even buy cards that sing "Happy Birthday"—the microfabrication techniques that sparked the electronics revolution have yet to ignite a mechanical one. Nevertheless, micromechanical devices—sensors, primarily—*are* making it out in the real world. The definitive sign that they've "arrived" is that they're now worth stealing—the theft of car stereos is taking a backseat to air-bag extraction as the hottest trend in auto burglaries; and the gadget that makes the air bag possible—the sensor that tells it to inflate when you slam into a tree, but not when you slam on the brakes—is a micromachined accelerometer.

You'd need an accelerometer to keep up with the growth of this field. It was all but nonexistent when Assistant Professor of Electrical Engineering Yu-Chong Tai was a graduate student a few years back. "I'd go to a conference and I'd basically know everybody. Nowadays, you go to a conference, and always more than 50 percent of the faces are newcomers. This society is expanding worldwide. It's like a disease, now—all the high-tech companies have it." If that's the case, then Caltech's biohazard lab is Tai's micromachining laboratory. The lab, currently located in Steele, will nearly double in size with the addition of space in the Moore Laboratory of Engineering, which is currently under construction.

The lab has several micromachine "viruses" in culture, as it were, but the one closest to being released is a micromotor for hard-disk drives. Lyndon Johnson was fond of saying, when told of a scientific advance, "How will this help Grandma?" Well, if Grandma has a computer, it will help her a lot. (Even if she doesn't, her gerontologist and pharmacist assuredly do.) As PCs give way to laptops, and laptops to notebooks, and presumably notebooks to wristwatches, more and more memory gets crammed into less and less space. The hard drive in your average PC is about the size and shape of a Kaiser roll, and stores 400 to 700 million bits per square



The complete microactuator. The large hairpin springs that connect the frame to the actuator curl around its four corners. The two sets of roughly vertical parallel lines just inside where the springs meet are the tops of the two motors' stator coils. Nested between the coils, and perpendicular to them, are the two thin hairpin springs that support the beam to which the read/write head is attached. The beam itself is the horizontal rectangle in the center of the picture. The entire microactuator is about three millimeters square.

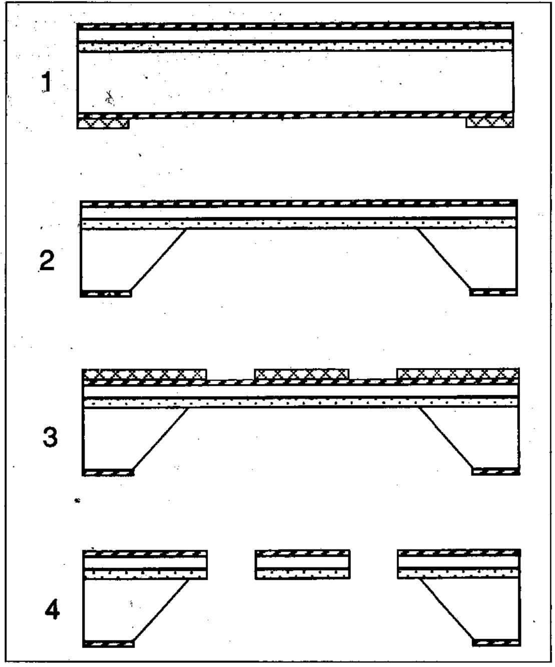
inch of disk. (For comparison, the 44-million-word *Encyclopædia Britannica* runs 2.4 billion bits, not counting the index or the illustrations.) Since 1992, Tai and Denny Miu, an assistant professor of mechanical engineering at UCLA and this year a visiting associate in electrical engineering at Caltech, have been developing the technology needed to make credit-card-sized drives one centimeter thick that will hold one to two billion bits per square inch. The long-term goal is to keep scaling these drives down until they can be arrayed on circuit boards, the way memory chips are mounted, to make the massive storage space of disk drives as instantly accessible as chip memory.

A hard-disk drive works much like a phonograph. In both cases, the information is written on the surface of a disk that spins underneath a stationary arm—the tone arm in your stereo, or a stainless-steel suspension arm in your computer. The arm pivots to reach any part of the disk, from the rim on in. (The disk drive is actually a stack of up to a dozen platters, often with less than an eighth of an inch of space between them, spinning on a common shaft. Each side of each platter has its own suspension arm, so that the drive plays the A and B sides concurrently, without having to flip the record over.)

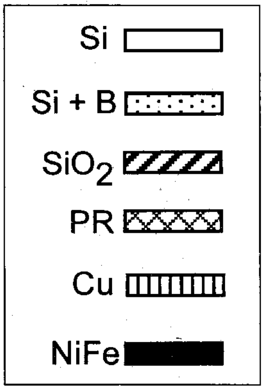
But whereas the sounds of *Sergeant Pepper's Lonely Hearts Club Band* are transcribed onto an LP as a wavy groove, which re-creates the music by vibrating a needle that's inserted in it, the data on a disk are encoded in puddles of magnetic polarity that an electromagnetic transducer—called a read/write head—interprets as a string of zeros and ones. The read/write head doesn't touch the spinning disk, but floats on an air cushion a couple of millionths of an inch thick. And whereas the cuts on a record are segments of one long spiral that takes up the entire album side, allowing any song to be played in its entirety once the needle touches down, the tracks on a hard disk are concentric circles one data bit wide. In order to retrieve a file, the read/write head skitters like a hockey puck from track to track, picking up file segments on the fly.

Current technology squeezes 5,000 tracks into an inch—in other words, 30 tracks would fit within the thickness of this page. Tai's group is initially aiming to cram 10,000 tracks into an inch in the credit-card-sized version, eventually upping that to 25,000 in the chip-sized one. So hitting the correct track is a lot harder than cueing up "Lucy in the Sky with Diamonds"—you can't just squint at the record and drop the tone arm into the dark space between songs. And if you miss your aim on an LP and drop into the song halfway through the third note, the skipped data merely jars your ears. A similar disk error would render the file unreadable. Moreover, the suspension arm is enormous, compared to the tracks—it's like trying to rotate the tower crane at the Moore Lab construction site to within two hundredths of a degree. Imagine trying to do this every 12 milliseconds—the amount of time the suspension arm has to find its next track.

But if the read/write head were mounted on a micromachined actuator, which in turn was attached to the suspension arm, it wouldn't need such exact control—you could just move the arm close, then jockey the actuator to the right track. (Compact-disk players, which pack 18,000 tracks per inch, use such a two-stage gadget, but it's much too big to wedge between the hard drive's platters.) Tai and Miu's actuator is carved from a silicon slab, yet has an almost lacy quality. The read/write head hangs from a beam supported by two impossibly delicate springs—flat, hairpin-turning squiggles that zigzag back and forth. Flanking the beam are two micromotors that pull the read/write head from side to side. The micromotors are what's called variable-reluctance motors. They work in the same way that an electromagnet made by wrapping copper wire around a nail picks up another nail. "In our case," says Miu, "the nails are permalloy, which is 80 percent nickel and 20 percent iron. One nail is the stator, which is fixed to the actuator and has the copper coils; the other nail is the rotor, which is fixed to the beam and moves the read/write head." The whole business is slung by four more hairpin springs—relatively big ones, this time—



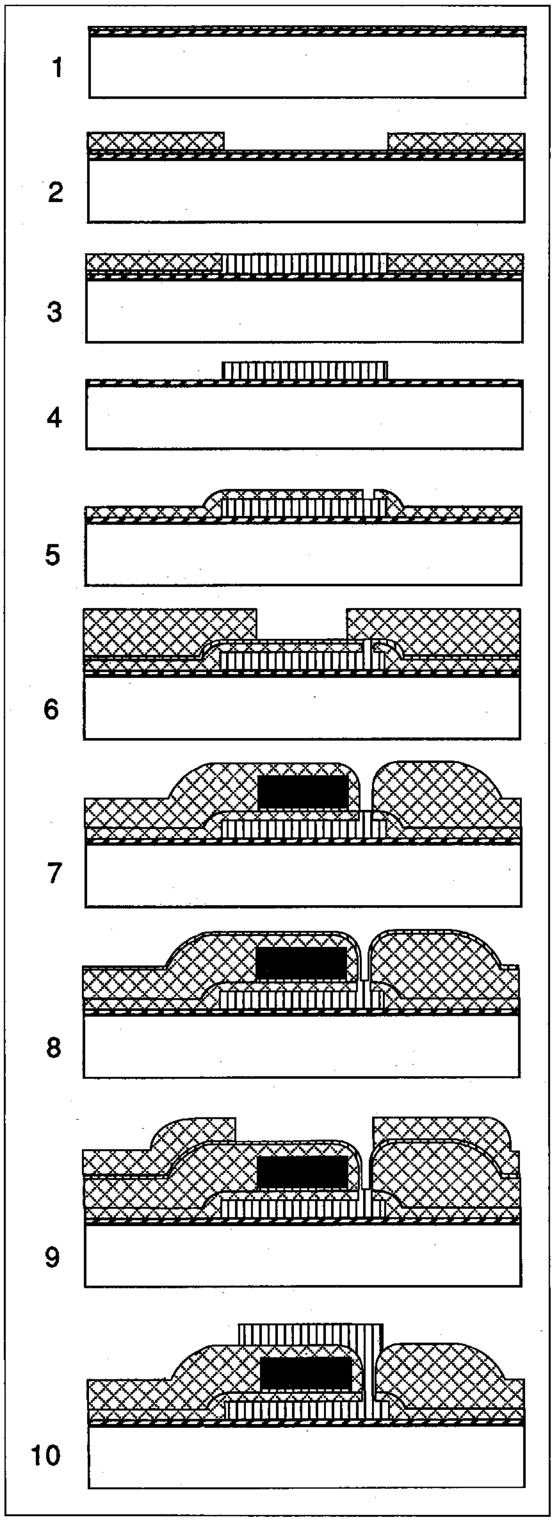
Right: A simplified cross-sectional schematic of how the springs are made. The materials are crosshatched according to the key below. (Si is silicon, B is boron, SiO_2 is silicon dioxide, PR is photoresist, Cu is copper, and NiFe is permalloy.)
1.) The composite wafer's underside is patterned with photoresist in the shape of the diaphragm to be etched.
2.) The etchant eats up through the wafer to the silicon-boron zone.
3.) The wafer's top surface is patterned with photoresist in the shape of the springs.
4.) The springs are cut from above with reactive ion etching.



within a frame that's still part of the same piece of silicon, and the frame is glued under the suspension arm's tip.

Now the trick to micromachining—and a big reason why the field is still in its infancy—is to figure out how to make things with moving parts, but using tools designed to manufacture immobile electronic circuits. The technology is the same—you cover the chip with a mask, then add a layer of something to (or strip a layer of something from) the parts of the chip exposed through the mask. The trick within the trick is planning ahead so that succeeding steps don't mess up what you've already done. Conceptually, there are two basic processes for making the actuator: one for carving the springs, and the other for building the motor. In reality, the two processes are interleaved. The entire procedure requires 20 masks—the equivalent of a memory chip. It's the most complex structure Tai's lab has built.

Cutting the springs is the simpler process. It starts with a silicon wafer 500 microns thick. A thin layer of a silicon-boron mixture is applied to the top surface by chemical vapor deposition silicon epitaxy, meaning that the silicon and boron atoms form a single crystal that blends seamlessly with the pure silicon below. Then comes another 20 microns of pure silicon—which again continues the single crystal—followed by a thin film of silicon dioxide, which acts as the plastic wrap on the sandwich and is applied to both the top and bottom surfaces. Next, the frame is masked off on the wafer's underside and etched from below. Applying the mask is a darkroom process exactly like printing a photograph: you shine a strong light through the negative to project the image onto light-sensitive paper. In this case, a photoresist—a light-sensitive chemical—is spin-coated onto the wafer, and the negative carries the frame pattern. (Spin coating is a neat way to get a very uniform layer of something without much fussing around—you hold the wafer horizontally and put a puddle of the coating in the center, then spin the wafer at several thousand revolutions per minute; centrifugal force does the rest.) When the photoresist is developed, the illuminated stuff doesn't stick to the chip any more and washes off, exposing the areas to be etched. The chip is bathed in hydrofluoric acid, which removes the silicon dioxide in the exposed areas, transferring the mask to the silicon dioxide layer. (The photoresist itself can't stand up to the etchant that follows, but silicon dioxide can.) The photoresist is rinsed off with a solvent, and the chip is then dunked in the etchant (ethylene diamine/pyrocatechol), which eats up through the wafer to the silicon-boron zone. The etchant can't digest the silicon-boron mix, leaving the 500-micron-thick wafer framing a 20-micron-thick diaphragm—the silicon-boron layer and the stuff above it—into which the springs will be carved. Their pattern is masked off on the diaphragm's top surface, using another layer of photoresist, and is cut by reactive ion etching. In this technique, the wafer is bombarded with a sulfur hexafluoride plasma, which consists mostly of fluorine ions that just tear into the unprotected silicon. Once the diaphragm is cut all the way through to make the hairpin springs (which takes about half an hour), another solvent rinse removes the photoresist.



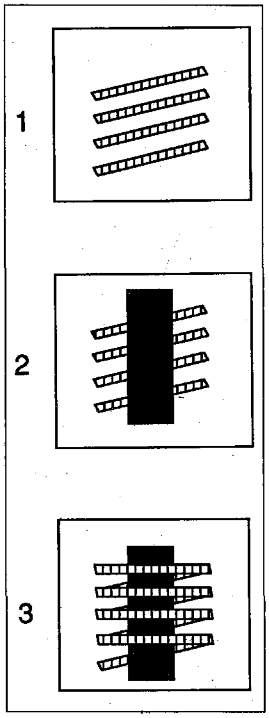
The construction sequence for the motor, using the same crosshatching code.
1.) A "seeder" layer of copper is applied to the wafer's top surface.
2.) A photoresist mold is patterned in the shape of the coil's bottom half.
3.) Electroplated copper fills the mold.
4.) The photoresist is rinsed off, and the exposed (unplated) seeder etched away.
5.) A fresh layer of photoresist is applied, which heat transforms into a permanent insulator.
6.) Another copper seeder layer is deposited, followed by the photoresist mold for the permalloy core.
7.) The core has been plated on, the mold and seeder removed, and a fresh layer of baked-on photoresist insulation added.
8.) A new seeder layer goes on over the insulation.
9.) Another photoresist mold for the top and sides of the coil follows.
10.) The remainder of the coil is plated on, and the photoresist mold and the remaining seeder layer removed.

The springs are flat, but the motor is three-dimensional; consequently, making it is considerably more complicated. You have to wrap a copper coil around a permalloy core, and since the motor is embedded in the actuator, you can't just pick up the core with tweezers and wind wire around it. So the construction proceeds in three stages: first the bottom part of the coil, then the core, and finally the coil's top and sides. The metals are deposited through a process called mold electroplating. Electroplating is commonly used to coat one metal with another—you clip an electrode to a hubcap, for example, dunk it in a bath containing chromium ions and the other electrode, run a current through the circuit, and—zap!—a chrome-plated hubcap. But silicon doesn't conduct electricity very well, so the first step is to apply a "seeder" layer of metal to the whole surface. (This requires yet another technique, called vacuum thermal evaporation, in which you place in a high-vacuum chamber the wafer and a small crucible of the metal to be deposited. The crucible is heated electrically until the metal evaporates, and the vapor then deposits itself on the relatively cool room-temperature wafer like shower steam on your bathroom mirror. Of course, the vapor also deposits itself all over the rest of the vacuum chamber's interior, but oh, well...)

"Electroplating different metals takes different seeders," Tai explains. "For example, for copper we put down 100 Ångstroms of chrome and 1000 Ångstroms of copper." Then comes the photoresist, etc., leaving the seeder exposed where the copper is to go. After copper fills the photoresist mold, the solvent strips the mold away and an acid etch gets rid of the unplated seeder layer. The acid takes a wee bit of the copper, too, but since the copper layer is some 10 times thicker than the seeder, it doesn't matter. What's left is a set of parallel copper lines, slightly slanted, which will be the bottom part of the coil, as shown on the opposite page. (It generally takes about 10 working days from the time you started on the springs to get this far.)

Now you need insulation—if the copper windings touch each other or the core, the motor will short out. It turns out that photoresist is a really good insulator, so a fresh layer of photoresist takes care of that. You have to plan ahead at this point, and remember to pattern this insulating layer to create the holes through which the two halves of the coil will eventually connect. Now you're ready to spend another day mold-electroplating the permalloy core. But how do you remove the *mold* photoresist (and etch off the seeder beneath it) without stripping off the *insu-*

Below: The evolution of a micromotor, as seen from above.
1.) These slanted parallel lines of copper will become the bottom part of the coil.
2.) The permalloy core runs up the middle of the coil.
3.) Other parallel lines of copper arch over the core and connect the slanted copper lines into a continuous coil wrapped around the core.



lating photoresist too? The answer is that you very cleverly baked the insulating photoresist before starting work on the core. The heat turns the photoresist into a long-chain polymer that can withstand the solvents and etchants. "There's a lot of materials science going on here," says Tai. "A lot of these processes are intimately related to the mechanical and electrical properties of the materials. These are the details that actually decide whether the process works or not." Once the core has been laid down, there's another layer of photoresist insulation (again patterned with the holes for the coil's electrical connections). Then another round of mold electroplating for the top half of the coil, and you're done. "That's often a trick I pull on my students when we start a project—'See, it's so easy! That's the way you draw it—now, go make it!' But we know how hard it is. There are a lot of tricky steps."

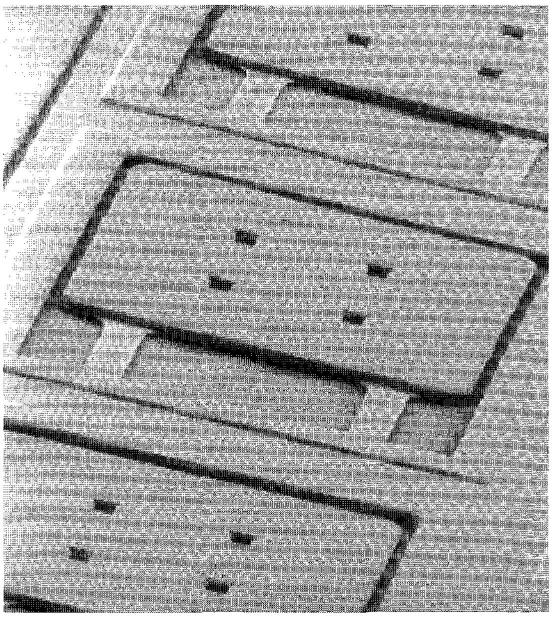
Making integrated circuits is actually easier, because they're only skin deep. The wafer is still 500 microns thick, but the lower 495 just sit there. The working parts don't penetrate any deeper than five microns into the chip, nor do they stick up any higher than five microns above it. "But when we do micromachining, we dig in. We often cut all the way through the wafer. So although the number of masks are the same, the technical issues are very different." For example, you have to treat a 20-micron-thick diaphragm with utmost care to avoid breaking it. "Also, a little bit of force can distort these structures, so we have to develop special expertise to handle them."

Tai's lab has built prototype actuators (a consortium of hardware companies are working on the drives), but there are still some kinks to be worked out. Says Miu, "right now, the arm is still supported on mechanical bearings, which gives you a certain amount of slop. Also, the wire leading to the micromotor behaves as a spring at very small deflections, so we have to account for that error." And there are other subtleties, too—the motors can't be too powerful, for example, or their magnetic fluxes can confuse the sensor that tells the read/write head what track it's on.

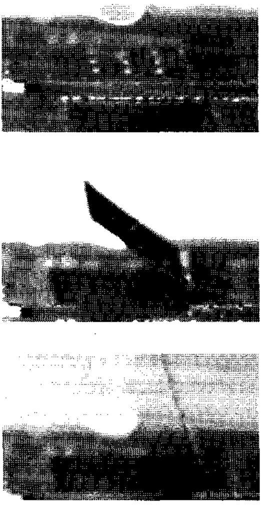
Another project has attracted more attention of late, even though it's much further away from practical application. In collaboration with a group of engineers from UCLA, the micromachine lab has demonstrated a "smart" skin designed to reduce turbulent drag on airplane wings. This isn't the kind of turbulence that makes the pilot turn on the Fasten Your Seat Belt sign and the flight attendants wheel their beverage carts back to the galley just before they get to your row. Instead, it's caused by the airplane itself—the wing's passage spins off swirling pockets of air, called vortices. These vortices start out lying flat against the wing, but they rapidly stand up to become miniature tornadoes, and that's when they cause trouble. Once upright, they pump high-speed air that's trying to rush past the wing down to the wing's surface. Thus the shear stress on that piece of the wing—stress caused by the air moving in one direction while the wing is moving in another—increases. (There's a certain amount of shear stress on the entire wing anyway. The wing essentially peels off a thin layer of the adjoining air and drags it along, but that's just part of the cost of doing business.)

Until now, the fuel that was burned fighting turbulent drag was unavoidable overhead, too. (And it's not trivial—one aerospace industry analyst estimates that a one-percent drag reduction for all commercial aircraft would save the airlines a billion bucks a year worldwide.) The vortices stick to the wing for less than a second before detaching themselves to go sailing harmlessly away, so there's not much time to react. You could shed them sooner by putting a ramp, such as a lifted flap, in their path—they'd hit it like a ski jumper and go roiling off into the wild blue yonder. But at typical wind-tunnel speeds, these vortices begin life larva-sized—about two millimeters wide and one centimeter long—and at jet airplane speeds, they're even smaller. So the sensors that will detect them and the flaps that will punt them need to be small, too. ("In order to test this idea in the wind tunnel, we had to make relatively large devices, and that's not very pleasant," says Tai. "This would actually have been easier on a real airplane. Micromachining technology simply isn't designed to make things that big—it's the inverse of trying to use an enormous power saw to cut a very small part.") And the vortices are all over the wing—the wind-tunnel model looks like it's crawling with maggots. Thus, the entire surface needs to be able to detect them, but only the affected regions should react to them, because if there aren't any vortices, raising the flaps will create them.

The grand design is to tile the wing with four-inch-diameter silicon chips, each of which would incorporate sensors, control circuitry, and flaps. The sensors measure shear—the proximate cause of the drag—by running a steady current through a silicon "wire" whose resistance rises rapidly with increasing temperature. The wire heats up, but the onrushing air carries the heat away. The high shear within a vortex cools the wire faster than usual, causing its resistance to drop below that of its neighbors. The controllers



Right: What all the flap's about. An aerial view of a portion of a flap array, seen from the hinged side. The four holes in the flap allow the etchant to undercut the flap outward from the center as well as inward from the edges, minimizing the time it takes to free the flap. Both this array and the flap shown below are from the steering project.



Below: Three frames from a video of a flap flapping. In the top image, the magnetic field is turned off, and the flap lies flat. (The hinge is to the right.) In the middle picture, the field is at about half strength, and the flap sticks up at a 45-degree angle. The field is at full strength in the bottom frame, and the flap is almost standing straight up.

compare the sensors' outputs to decide where the vortices are, and lift the flaps in that general area. The electronics are still being designed, in collaboration with Professor of Electrical Engineering Rod Goodman's research group, but the lab has built prototype models of the sensors and flaps.

And it's the flaps that are in the limelight. They're thin, flat, multilayer sandwiches that cantilever out over pits etched in the silicon beneath them. One of the layers is a permalloy coil, which, when electrified, raises the flap magnetically—up to a good 65 degrees from the horizontal—by pushing against the field created by another magnet on the floor of the pit below. (The magnets, which operate at 80 gauss, or about the strength of a refrigerator magnet, exert a force some 20 times stronger than gravity on a typical one-millimeter by one-millimeter flap.) Each flap can be raised or lowered individually. Most remarkable of all are the hinges—there aren't any. Instead, two tiny silicon beams connect one side of the flap to the pit's brink. In our world, silicon structures—glass windows and ceramic pots, for instance—are stiff and brittle. They resist stress until they shatter. But in the microworld, silicon behaves differently. If you make thin enough beams of it, they're quite amazingly flexible. This is actually true of most materials, because as you make smaller and smaller crystals of something, the number of lattice defects—places where the atoms don't quite line up, and where fractures can start easily—gets smaller, too. Other people had verified this with millimeter-sized hunks of silicon, says Tai, "but we've gone down to microns, and even nanometers, and we've definitely confirmed the trend. And, of course, we're enjoying it. It's a happy result."

These hinges are not only flexible, they're beefy. In a parallel project, Tai's lab and the UCLA group are building flap arrays that exert ten times more force per flap than the antiturbulence ones—enough muscle to actually steer the airplane. A wing is normally steered by large flaps, called ailerons, along its trailing edge. The microflaps go along the leading edge instead. Both kinds of flaps work by deflecting the boundary layer—the airflow along the wing's surface that causes lift (and drag). The boundary layer is wedge-shaped—very thin along the wing's leading edge, and thickening toward the rear. As the wedge thickens, it contains more air and gets harder to move, so manipulating it from the leading edge makes a lot of sense, says Tai. "You use much less energy to achieve the same degree of control. It's like a transistor—you put a little signal into the leading edge, and it will be amplified automatically as the boundary layer goes back over the wing." Tai's group has demonstrated this approach in UCLA's wind tunnels, using a generic delta-wing model. Aerospace engineers are very interested, but the prospect of aileronless planes is probably too much for the flying public. Don't look for jumbo jets of this design any time soon.

At the moment, each of the three components in the turbulent-drag project—sensor, controller, and flap—are still separate units connected by old-fashioned copper wires. Tai expects to have the three on one chip within six months, but figuring out what sequence to make the magnets and control circuitry in is a chicken-or-egg problem: making the sensors requires heating the wafer to 800°C, which melts the aluminum connections between circuits; making the circuits entails depositing layers of silicon atoms, which clog up the flaps. "The more things you put on, the more headaches you have," Tai says ruefully. "Whenever you try to put a lot of different kinds of devices together, that means you are combining all these processes into a big, long, complicated one. We're constantly thinking about how to solve problems like this."

Tai sees this project as pushing the envelope, not of aircraft design, but of micromachine design. "This may never be used on a real airplane—who knows? The point is that it demonstrates a new technology that combines microsensors with microactuators and microelectronics—what I call M-cubed." Once you've integrated those three components—the eyes, hands, and brain, as it were—there's no mechanical system you can't build, at least in principle. "If we demonstrate that the technology can be developed to include all three things on one chip, we have defined the boundary of microfabrication. That's the ultimate challenge." The *real* ultimate challenge will be to figure out what undreamed-of things you can create with M^3.

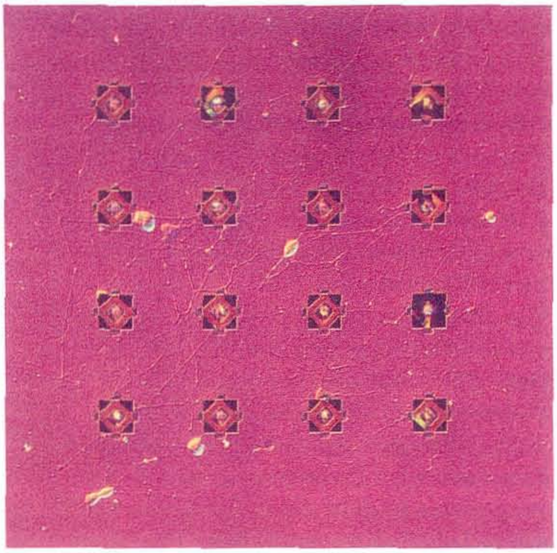
Any hobbyist who has ever been reduced to howling fury while trying to tweezer a balky antiaircraft gun into its mounting on a 17-inch replica of the battleship Missouri *will appreciate the frustrations of trying to do the same sort of thing on something a hundred times smaller.*

For starters, here are a few things that people *have* dreamed of. Like the flap projects, these are distributed systems in which little neighborhoods of components operate independently within large arrays. First, you could use a flap array to create turbulence where you want it—in the combustion zone of a turbine engine, for example, where fuel and air have to mix fast and thoroughly. Or consider active soundproofing, in which a wall detects the sound waves hitting it and adjusts itself to damp them out. Or an array of micromirrors that, properly illuminated, would form a flat-screen TV of unlimited size. Or dish antennas that continuously adjust their surface curvature to focus a signal.

And speaking of communications equipment, the micromachine lab has joined forces with Caltech's Jet Propulsion Laboratory to demonstrate the manufacture of waveguides for millimeter- and submillimeter-wave antennas. These waves fall between microwaves and infrared light, and JPL wants to use them for deep-space communication, radar, and spectroscopy. Waveguides are essentially speaking tubes for electromagnetic waves—tunnels with reflective metal walls down which the waves travel. The waveguide's cross-sectional dimensions need to be within 10 percent of the length of the wave in order to guide it. Accurately machining a metal channel the width of a gnat's eyelash is an art that computers haven't mastered yet, and it takes months for a skilled human to make a submillimeter waveguide that works. Then, to make it into an antenna, you have to glue a transducer on it, which is also done by hand. Any hobbyist who has ever been reduced to howling fury while trying to tweezer a balky antiaircraft gun into its mounting on a 17-inch replica of the battleship *Missouri* will appreciate the frustrations of trying to do the same sort of thing on something a hundred times smaller. Micromachined waveguides avoid these problems. The channel's width is precisely set by the mask, and the depth by the etchable layer's thickness. And the transducer can be micromachined directly into the channel. Silicon doesn't reflect microwaves, but coating the channel with a reflective layer of metal atoms is standard technology, as we've seen.

Robotic spacecraft with silicon hardware are worlds less complex than live mice with protoplasmic circuitry, but a brain's a brain. The micromachining lab is using the construction techniques of the former to help study the workings of the latter. Since 1980, Professor of Physics Jerome Pine has been studying how nerve cells, or neurons, interact in networks. The idea is to grow a small array of neurons connected to one another in their normal fashion, so that you can stimulate one cell and listen in on what it says to its fellows. Growing the arrays in culture is relatively easy, but wiring them for sound is a lot harder. First of all, you can't just jab electrodes in them if you want them to live very long. Pine's first plan was to lay an array of electrodes in the bottom of a Petri dish, and then grow the neurons on it. This was fine in principle, but it was difficult to communicate with a single desired cell after the network grew a cobweb of processes—the filaments that connect nerve cells—all over the array. The next refinement was to make tiny diving-board-shaped silicon electrodes that could be wheeled up to the cell bodies. This proved awkward, but it got Pine thinking about micromachining.

In 1988, Pine's lab began making arrays of shallow wells, each of which was just the size of a mature nerve-cell body and whose floor was an electrode. An immature neuron was injected via micropipette into this dungeon, the ceiling of which was a grating that admitted nutrients and allowed the neuron's processes to grow out. As the cell matured, its body filled the entire volume of its prison and pressed tightly against the electrode in the floor, making a solid contact. The unfettered processes, meanwhile, slipped through



Right: An individual neurodungeon, 30 by 30 microns square and 16 microns deep. The neuron will be injected through the central hole in the diamond-shaped lattice. This cell isn't quite ready for occupancy yet, as there's no electrode in the floor.
Far right: A four-by-four cellblock. The tendrils wandering all over the surface are processes that have formed a neural network. The blobs are nerve-cell bodies, some of which probably escaped from the pen. In fact, if you look carefully, you can see a couple of them in the act of going over the wall.

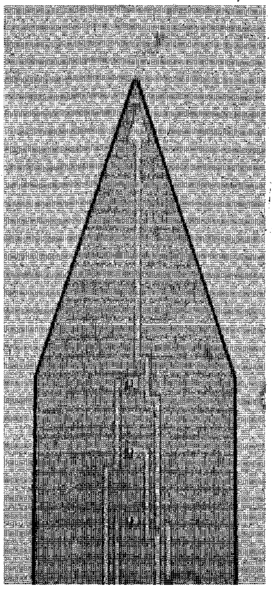
the grillwork and connected with the ward's other inmates. But the fabrication problems were too challenging, says Pine, so he helped recruit Tai to Caltech to collaborate on building a better neurotrap.

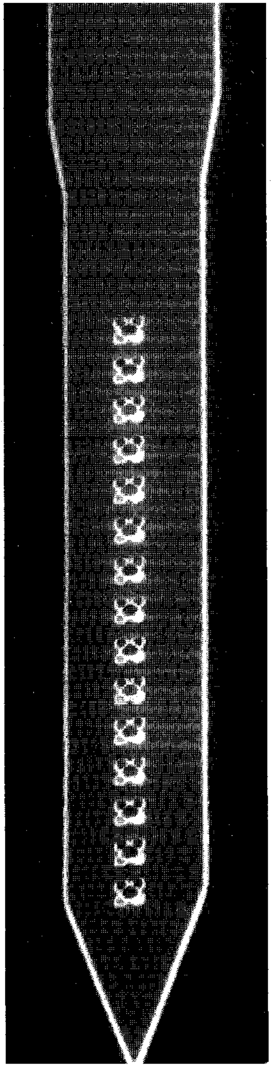
The collaboration is now making 16-neuron cellblocks—arrays of four cells by four cells—in which embryonic nerve cells from rats are incarcerated. The group's record for keeping neurons alive in captivity is about a month, long enough to form a network and start recording its behavior. But Pine would like to keep them alive for about three months, in order to study each network thoroughly—like snowflakes or fingerprints, no two networks are completely alike. The trouble is, the neurons climb through the bars and escape. "They squish like water balloons," says Pine. "It's astonishing how small a hole they'll get through. A 20-micron-diameter neuron can crawl through a one-by-three micron slot. They'll stay alive for three months, easy, just not where we want them." The neuron's growing processes cling to the silicon for support, and one process in particular, called the axon, is known to exert a lot of traction on the cell body—enough, apparently, to pull it through the lattice. The next design will replace the grillwork with narrow channels up to 30 microns long, down which the processes will have to grow. The hopes are that the cell bodies won't be able to stay squeezed long enough to worm through.

Tai and Pine are building similar probes to study neural activity in real, live brains. "We'd love to plant spies in brain tissue to tell us what's going on," says Pine. Multichannel electrodes are a basic tool of such studies, but driving spikes—even wire-thin ones—into the brain tends to kill or maim the cells in the immediate vicinity. This in itself is not bad, as the brain has cells to spare and there are no pain receptors in it, but the signals from the healthy cells on which the researchers wish to eavesdrop are muffled by the dead zone surrounding the probe. And the probe picks up the chatter from everything in its vicinity, while we may only be interested in the conversations over a specific phone line, as it were. Tai and Pine hope that a micromachined probe with a line of neurodungeons will minimize these difficulties. The probe neurons should send their processes out in search of healthy cells to connect to. And, by stocking the probe with a cell type peculiar to the circuit the researchers wish to wiretap, the probe might be encouraged to wire itself into that circuit as the captive neurons instinctively seek out their compatriots. (How nerve cells "know" which connections to make remains one of the great puzzles of neurobiology, but we can still take advantage of the fact that they do.) Of course, all this depends on the assumption that the imprisoned neurosnitches can survive for months or even years in the probe without special attention and "mainstream" themselves into the brain cell population.

The probes are shipped to Rutgers, where Professor of Biology Gyorgy Buzsaki's research group implants them in rat hippocampi. The architecture (although not the function) of the hippocampus is well understood, and collecting embryonic hippocampal neurons and integrating

Below: A couple of months after implantation, probe neurons (stained dark, and marked with arrows) have grown several hundred microns into the surrounding brain tissue. The probe (partly visible as the straight-edged black shape) runs down the photo's right side.





Top: The Art Deco zigzags on the probe's back side are the leads connecting the electrodes (the small squares) to the outside world. The arrowhead is a reference electrode. Bottom: The probe has 15 neurodungeons spaced 50 microns apart at the tip of a 2-millimeter-long shank that's 20 microns thick.

them with host neurons is Buzsaki's specialty. The Rutgers contingent has proven that the probe neurons do, in fact, grow connections to the host cells. Buzsaki's next step is to figure out exactly where those connections go, by stimulating an individual probe cell and monitoring the neighborhood's reaction, or waking the neighbors and seeing which probe cell responds.

Although these probes are strictly for basic research at the moment, Tai sees them eventually getting out into the larger world as controllers for prosthetic limbs. The probe could tap into the brain circuits that would normally move the limb, and send the electrical outputs to servomotors that could flex an artificial knee or clench synthetic fingers into a fist. Such experiments have been going on for 20 years with metal probes, but the neuroprobe offers the chance to make permanent, one-on-one connections. And you wouldn't have to intercept exactly the right circuits—probably an impossible feat in any case—since the patient's brain would automatically rewire the connections as the patient learned to use the prosthesis.

In fact, Tai sees a growth industry in biomedical microdevices of all kinds—not Dr. Fujimasa's *Fantastic Voyage* robot, but less grand schemes. For instance, one company has been making micro blood-pressure sensors for a decade, says Tai, and another is making microvalves "that could revolutionize biomedical instruments. Micromachining can make small systems that function as well as the big ones, or even better. That's terrific for biomedicine, because people want smaller and smaller devices."

We're not talking about teeny-weeny heart valves for preemies here, but something much bigger: a laboratory on a chip. When you visit your doctor for blood work in a few years, you may get away with depositing a few drops, instead of leaving what seems like a gallon's worth. Several organizations are working on scaling down the equipment needed for an arsenal of standard analytical procedures. A technique called capillary electrophoresis, for example, which is used to identify proteins or DNA sequences, separates the constituents in a sample by dissolving them and drawing them through a narrow tube via an electrical gradient. The components pass through the tube at rates depending on their size and charge, allowing each one to be identified when it emerges. Right now, such systems take a lot of fancy plumbing squeezed into a unit about the size of a home bread-making machine. Add the laser sample-detection system that goes with it, and you have another unit the size of a toaster oven. And the workhorse of biotechnology, a technique called polymerase chain reaction (PCR) that takes a snippet of DNA and copies it many times over—a critical step in screening for assorted genetic diseases—requires heating and cooling the sample over and over again, while adding different reagents at specific steps in the cycle. This also means lots of plumbing, *plus* a programmable oven. The current ones are about the size of microwaves, but instead of getting popcorn in five minutes, you get PCR in an hour. Doing the procedure on a chip, with just a smidgen of sample to heat and cool, might cut the processing time to 15 minutes. Eventually, one could design special-purpose chips to do specific blood tests while you wait—can drive-through service be far behind?

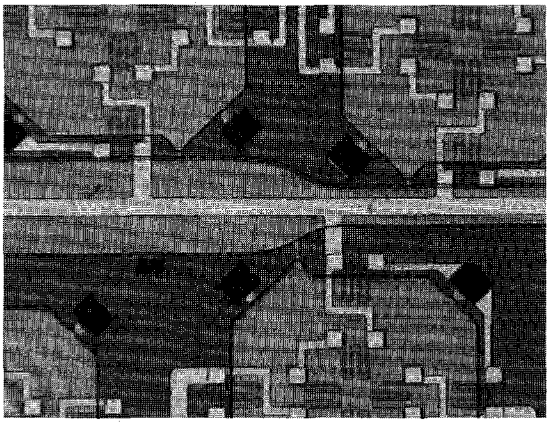
And there are a legion of applications beyond the biomedical. For example, self-contained laboratories on a chip could be used as process controls in industries from brewing beer to refining gasoline. Beyond the factory gates, such sensors could form the basis for rugged yet compact air- or water-pollution monitors.

Along with the usual M^3 problems of component integration, these projects are hampered by a lack of fundamental knowledge of what goes on in machinery of cellular dimensions. "There are so many promising applications that everybody has been spending their resources developing new devices," says Tai. "But we're neglecting the study of fundamental micromaterial properties, which we need in order to keep advancing. I can't overemphasize the importance of fundamental research, and I feel that academia, rather than industry, has an obligation to do it because it benefits everybody." Tai and Miu are therefore running a silicon microproperties lab, too.

As we saw in the case of the airplane-wing flaps, specks of silicon can behave quite differently than silicon in the large. One of the questions the microproperties project is trying to answer is just how small you can make, say, a hinge—at some point, there are simply going to be too few atoms to accommodate the bending force. The project is studying static properties such as tensile strength (how much you can stretch a sample) and fracture strength, as well as dynamic properties such as fracture propagation. The project is also looking at composites, in which the silicon has been coated with a metal, an alloy, a polymer, or even a ceramic. Most silicon microgadgets incorporate other materials, if only as the metal lead to an electrical connection. Says Tai, "Composite materials have been a big research topic in materials science, but microcomposites are relatively new and there's no general theory describing them. Microcomposite materials open up a whole new range of properties and behaviors that we can use in ways we can't even imagine because we don't know enough about them. We've already found a lot of interesting things we don't see on the macro scale." They've discovered, for example, that applying a layer of metal to the top of a silicon beam markedly alters its fracture behavior. Whenever you do a set of fracture experiments, there's always a certain amount of statistical scatter in the results. But the metal layer reduces that scatter—the results cluster more tightly around a single value. Furthermore, the alloy's exact composition strongly affects the clustering.

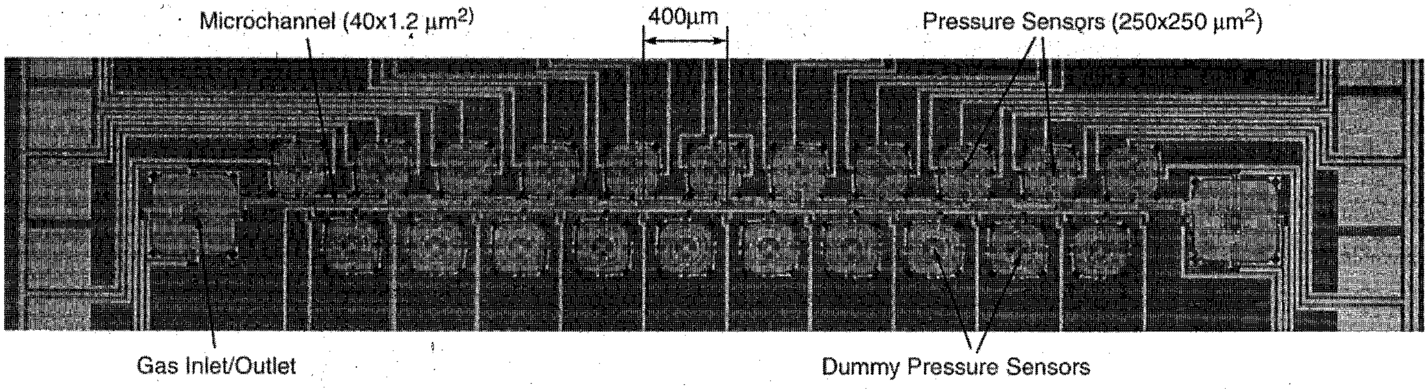
There's also a pressure spike at bottleneck number two, where the channel narrows to 40 microns. This could mean that the gas molecules pile up like a mob of Keystone Kops running full tilt at a narrow doorway.

And if the quintessence of rock-solid silicon changes with its bulk, it should come as no surprise that more evanescent phenomena are mutable as well. Take fluid—gas and liquid—flows, for example. The vast literature on fluids in enclosed channels (the sort of thing you use to design natural-gas pipelines or chemical plants) tends to streamline the calculations by concentrating on what's happening in the middle of the pipe and neglecting the complexities, called edge effects, that occur along the walls. But you can't do this in a microchannel, where the channel's height is comparable to the mean free path—the average distance a fluid molecule travels before colliding with another fluid molecule. At that scale, everything is edge effect. "If you don't have micromachining technology, it's very hard to do these experiments, and there's really no need for them. Now, suddenly, we have this technology, and people are showing that many useful microfluid devices can be made. But in order to properly design micropumps and so forth, you have to know how fluids behave on this scale."



So the micromachine lab and the UCLA engineers are building wind tunnels on chips. This has required developing a micro pressure sensor that can be integrated into a channel so that the ensemble can be built as one unit. The flow at various points in the channel—which is what you really want to know—is then derived from the pressure data.

The first wind tunnel looked at the simplest possible situation—a pure gas (helium or nitrogen) in a straight, rectangular channel. And, says Tai, "we found that no theory, even when we modified the famous Navier-Stokes equations, could explain the differences we saw between helium flow and nitrogen flow." These equations, which work very well at macro scales, say that the two gases will behave differently. The gases, however, didn't behave differently in the way that the equations said they would—they behaved differently in a completely different way. None of the fluid mechanists that Tai talked to were able to explain what was going on, so the group eventually just published the data in an article that said, "Here, theorists—what do you make of this?" The group also discovered that the pressure distribution in the channel was nonlinear. In a big pipe, like a gas main, the pressure is high at the inlet, drops at a steady rate—linearly—as the gas flows down the pipe, and reaches its lowest value at the outlet. This pressure drop forces the gas through the pipe, just as an elevation drop forces water down an aqueduct. But in the microchannel, the pressure



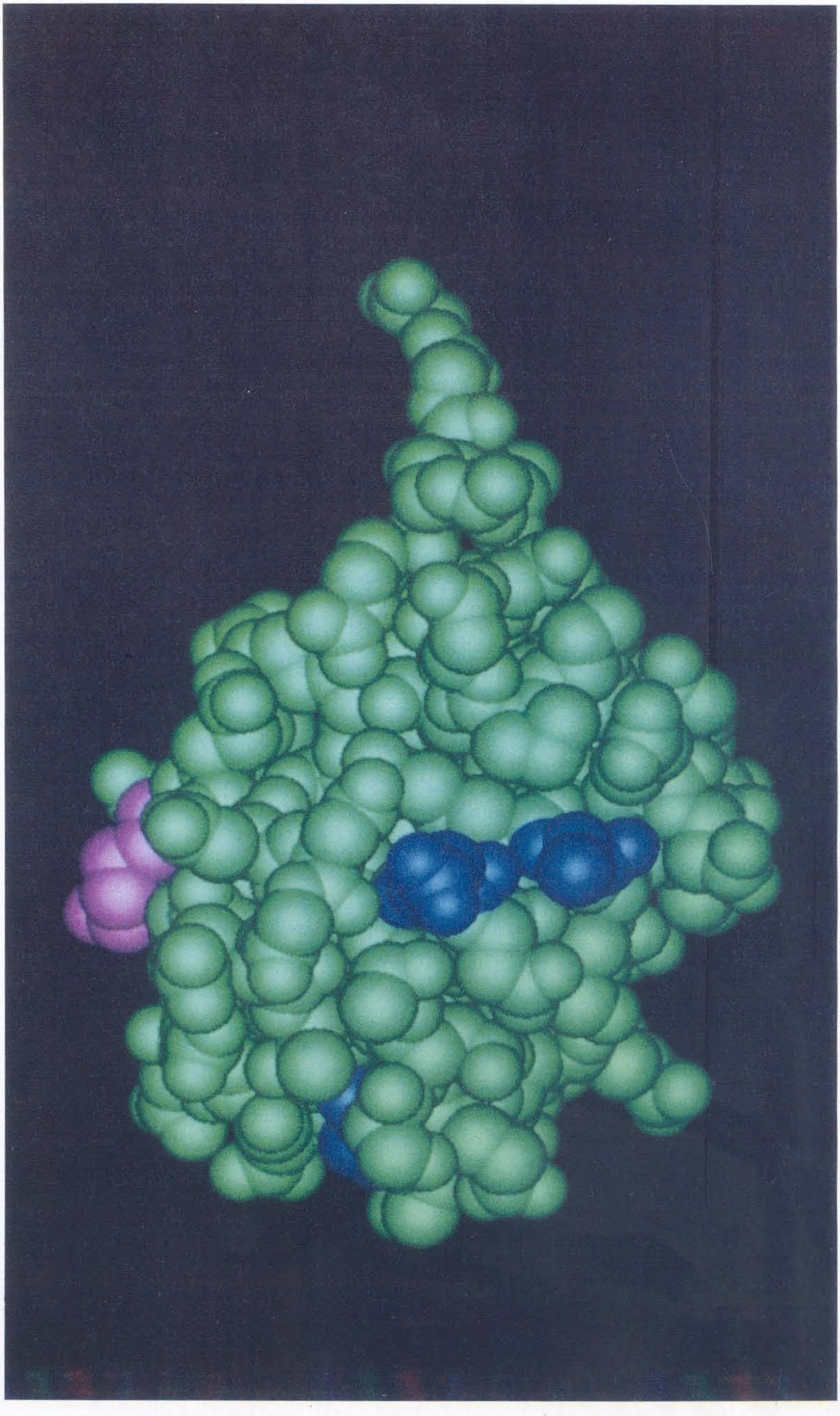
Above: The uniform—flow micro wind tunnel. The channel is 4.5 millimeters long by 40 microns wide by 2 microns deep. The structure at one end is labeled "Gas Inlet/ Outlet" because the tunnels are designed to accommodate flow in either direction. The "dummy" sensors along the bottom side of the tunnel provide for leak checks during the fabrication process.
Opposite page: The section where the newest wind tunnel narrows from 100 to 40 microns. Portions of three pressure sensors (two above and one below the tunnel) can be seen, as well as the very narrow channels that feed them. The sensors are cavities beneath thin diaphragms that flex as the pressure changes. These distortions are measured by the zigzag structures visible on the diaphragms. The thick, light-gray stripe down the center of the tunnel is an electrical lead.

didn't drop very fast in the first portion of the pipe, which may indicate that the gas molecules are clogging up the channel. "There are ideas as to why this should happen, but the bottom line is that we still don't understand the physics yet."

The lab's latest wind tunnel has three choke points in it. It starts off 100 microns wide, narrows to a 60-micron-wide throat, then expands back to 100 microns. Later on, the tunnel funnels down to 40 microns, and then, later still, there's an 18-micron-wide neck in the 40-micron channel. There's a micropressure sensor before, after, and near each choke point, as well as at the channel's inlet and outlet. "We see even stranger things in the nonuniform channel. We're more puzzled there than the day we started! That pretty much sums up the current state of our research." In addition to the high-pressure region in the early part of the channel, there's a pressure drop at the first and third bottleneck, which might confirm that the system doesn't have enough oomph to force many gas molecules toward the outlet. There's also a pressure spike at bottleneck number two, where the channel narrows to 40 microns. This could mean that the gas molecules pile up like a mob of Keystone Kops running full tilt at a narrow doorway.

Tai is now expanding these studies into liquid flows. Liquid flows will no doubt act even odder, because liquids are more viscous than gases. For one thing, it will be harder to force a liquid through a microchannel, which means that the channel will have to resist substantially greater stresses. Fortunately, we've already seen that microstructures actually get stronger as they get smaller. And if the knowledge needed to design sturdy, efficient microplumbing systems emerges from Tai's research, then hand-held blood-sample screening devices become more plausible, which brings us back to Dr. Fujimasa and to Lyndon Johnson's grandma.... □

Yu-Chong Tai earned his BS in electrical engineering from National Taiwan University in 1981, and his MS and PhD in electrical engineering from UC Berkeley in 1986 and 1989, respectively, inventing the first electrically spun micromotor along the way. He came to Caltech as an assistant professor in 1989. Numerous people have contributed to the work described herein: Amish Desai, Raanan Miller, Wei-Long Tang, Viktoria Temesvary, and Shu-Yun Wu to the microactuators; Charles Grosjean, Fu-Kang Jiang, Chang Liu, and Tom Tsao to the flap projects, with Bhusan Gupta and Sarah Bates of Goodman's lab; John Wright and Svetlana Tatic-Lucic (MS '90, PhD '94) to the waveguides, with JPL's Bruce Bumble, Henry LeDuc, and William McGrath; Wright and Tatic-Lucic to the neuron projects, with Hannah Dvorak and Michael Maher in Pine's lab; Michael Debar, Grosjean, and Wen Hsieh, to the microproperties studies; Jian-Qiang Liu (PhD '95) and Xing Yang to the wind tunnels. Tai's UCLA collaborators are Chih-Ming Ho, Jin-Biao Huang, T. S. Leu, John Mai, Kin-Cheok Pong, and Steve Tung. Tai's work is primarily funded by the Advanced Research Projects Agency, the Air Force Office of Scientific Research, the National Institutes of Health, the National Storage Industry Consortium, and Hewlett-Packard.



The World of Ubiquitin

The story of an old protein molecule is a tale of hazard and tear, of unceasing collisions with other molecules in the cell and assaults by a legion of highly reactive compounds that form in the process of metabolism.

by Alexander Varshavsky

The human ubiquitin molecule, shown here with its C-terminus at the top, differs from the yeast version by only the three amino acids rendered in blue. (The spheres represent individual atoms.) These three residues lie at positions 19, 24, and 28, as counted from the N-terminus. The pink atoms depict a lysine residue at position 48, through which another ubiquitin can attach itself to form a link in a multiubiquitin chain. (Ubiquitin's three-dimensional structure was determined by Senadhi Vijay-Kumar, Charles Bugg, and William Cook at the University of Alabama in Birmingham. Image courtesy of Michael Carson, Leigh Walter, and Cook.)

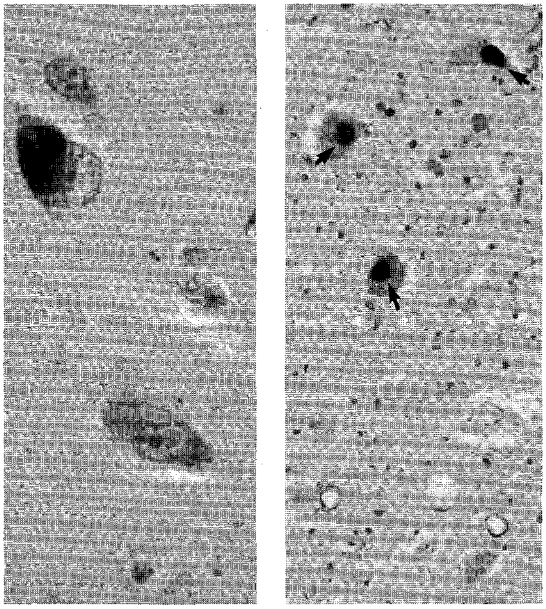
The pessimists have known it all along. Things of value in our eyes—fresh fruit, good weather, ourselves—tend to decay and fall apart. Proteins—the major constituents of living organisms—are no exception to this dreadful law. They are being destroyed inside and outside of cells, often in complicated ways, for a variety of reasons. The tale of protein degradation is a braid of interacting plots; in this article we focus on those that star a remarkable protein called ubiquitin.

But first, let's recall some basic molecular biology. Proteins are polymers, built from 20 different amino acids, which are assembled into linear chains according to instructions by segments of DNA called genes. The DNA's instructions are conveyed through messenger RNA to protein-making intracellular machines called ribosomes, which themselves are built from proteins and RNA. The protein's chain of amino acid residues (or simply residues) is called a polypeptide chain, and the residues are linked by chemical bonds called peptide bonds. The two distinct ends of a polypeptide chain are called the N-terminus and the C-terminus. The N-terminus bears a nitrogen-containing chemical group called the amino group, while the C-terminus bears the carbon-containing carboxyl group.

A newly formed protein, which emerges from the ribosome with its N-terminus first, faces a staggering variety of potential fates, one of which is degradation. Proteins are destroyed in a process called proteolysis, which may involve just a few cuts in a polypeptide chain, but can also result in the degradation of a protein all the way back to its constituent amino acids. Making proteins is an incredibly complex undertaking—why should they be destroyed at all? One reason for the existence of proteolysis is also kind of sad: proteins of a cell can be food for other cells, which often reside in a different organism. A lion dining on antelope looks utterly unlike a vegetarian munching a cucumber, but the strategy of both eaters is the same—to keep alive by subsisting on components of other living beings.

The enzymes (biological catalysts) that carry out proteolysis are a special class of proteins called proteases. Their size and complexity vary enormously—from relatively small proteases like trypsin and pepsin, which function outside of cells and digest proteins in food, to much larger ones called proteasomes, which consist of many protein subunits (polypeptide chains) and reside inside the cells.

Another function of proteolysis is the destruction of damaged or otherwise abnormal proteins. The story of an old protein molecule is a tale of hazard and tear, of unceasing collisions with other molecules in the cell and assaults by a legion of highly reactive compounds that form in the process of metabolism. Sometimes a protein molecule is abnormal from its very beginning, either because it is the product of a defective gene or because it failed to fold properly (folding properly is a complicated affair, assisted by special proteins). Yet another source of protein damage is environmental stress. Consider, for example, a yeast cell feeding on a grape at high noon. This cell has to cope, among other things,



Ubiquitin is a component of the intracellular protein aggregates characteristic of many human diseases, particularly those of the central nervous system. (The plaques that form in Alzheimer's disease are perhaps the best-known example.) An accumulation of these aggregates eventually kills the afflicted cell. Ubiquitin's role in aggregate formation (the aggregates also contain a variety of other proteins) remains unknown.

At left is a sample of brain cortex from a patient with Lewy Body dementia that has been stained with a ubiquitin-recognizing antibody (dark). A large ubiquitin-containing aggregate (called an inclusion body) is visible in one cell. The circular structures are cell nuclei.

At right, several ubiquitin-containing inclusion bodies (arrows) can be seen in the spinal cord neurons of a patient with amyotrophic lateral sclerosis, better known as Lou Gehrig's disease. Micrographs courtesy of John Mayer and colleagues at the University of Nottingham, England.

with the sun's heat—possibly a problem because the cell's temperature may become high enough to unfold and render inactive some of the yeast proteins.

Damaged proteins have to be repaired or eliminated. Protein repair systems (they do exist) are beyond the scope of our discussion. If repair fails or isn't attempted, a damaged protein has to be distinguished from its normal counterparts in the cell, singled out amidst the stir and bustle of other protein molecules, and then destroyed without perturbing nearby structures. We can now glimpse some of the reasons behind the complexity of the intracellular proteolytic machines—their task is vastly more subtle than the task of pepsin in the stomach, where every protein is fair game. The recognition and elimination of damaged proteins keeps a cell nearly, but not quite, free of them, because the surveillance mechanisms are blind to certain types of protein damage. In other cases, these mechanisms appear to recognize a damaged structure as such, but can't destroy it because it's protease-resistant or physically inaccessible—for example, by being a part of a huge protein aggregate, as happens in several chronic diseases. A damaged protein may also be difficult to reach in an otherwise normal structure. For example, the lenses of our eyes become more opaque with age, and often (if we live long enough) develop cataracts, in part because of a relatively inefficient protein turnover deep in the lenses, where the tightly packed lens proteins leave little room for anything else.

There exists yet another reason for a protein to be destroyed—if it *evolved* to be degraded quickly. Proteins like these often function as regulators—devices that control the activities of specific biological processes such as the transcription and replication of DNA, the life cycle of a virus inside its host, or the fluxes of specific compounds through the metabolic pathways of a cell. To understand the reason for making a regulator short-lived, imagine that a specific biochemical pathway, controlled by an activator protein, is required before but not after cell differentiation—a process in which a cell converts itself into a cell of another kind. Stopping the synthesis of the activator may not be a fast enough way to get rid of it: the activator would linger indefinitely in a nondividing cell (many differentiated cells no longer divide), and even a dividing cell would dilute the activator only twofold upon each division—too slowly for a good off-switch. But make the activator short-lived, and stopping its synthesis would result in a rapid decline in the activator's concentration, and therefore in a rapid shutoff of the no-longer-appropriate pathway.

Enter ubiquitin. Its saga began in 1975, when a group of scientists in New York reported the isolation of a 76-residue protein that was present in all tested organisms. The name proposed for the new molecule—"ubiquitin"—turned out to be remarkably apt, because later studies indicated that ubiquitin is one of the most highly conserved proteins among eukaryotes. (The eukaryotes include you, me, all other animals, plants, fungi, and everything else alive except bacteria. One characteristic feature of a eukaryotic cell is its nucleus—a membrane-enclosed compartment where the cell keeps most of its DNA in long, tightly coiled fibers called chromosomes.) "Highly conserved" means that the amino acid sequence (and hence the structure) of ubiquitin is nearly the same among different organisms. Since the sequences and, to a lesser extent, the structures of most proteins tend to change in the course of evolution, the sequence of a protein that performs a given function in one organism may be quite unlike the sequence of its functional "twin" in another organism. By contrast, the sequence of ubiquitin remained essentially unchanged in the course of roughly two billion years—the span of time since the nearest common ancestor of this writer and baker's yeast. This extraordinary evolutionary stability implies that almost the entire structure of the ubiquitin molecule participates in some extremely important cellular functions. But what those functions were was anybody's guess.

Two years later, scientists at the Baylor

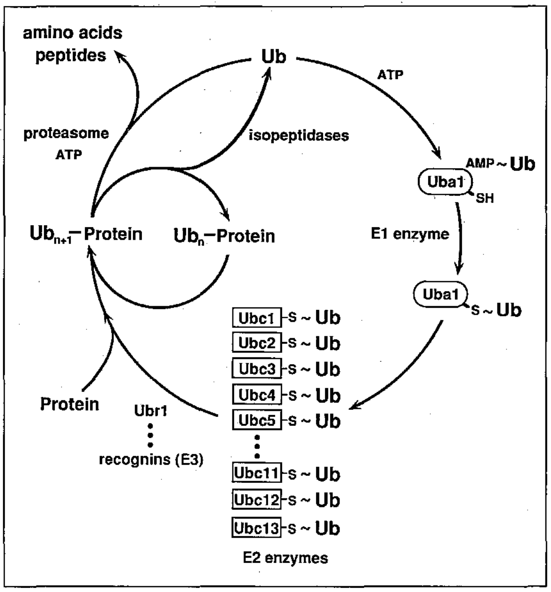
College of Medicine in Houston identified a mammalian protein of unusual structure, in which a chromosomal (DNA-bound) protein called H2A was linked to another protein—ubiquitin. In this "branched" protein, which they named ubiquitin-H2A or uH2A, ubiquitin was linked ("conjugated," as chemists say) to a lysine (an amino acid) within H2A, resulting in a protein with one C-terminus but two N-termini. The function of uH2A in chromosomes remains obscure to this day, but the branched structure of uH2A provided the first glimpse of a fundamental property of ubiquitin, soon to be encountered by scientists analyzing protein degradation.

The sequence of ubiquitin remained essentially unchanged in the course of roughly two billion years—the span of time since the nearest common ancestor of this writer and baker's yeast.

Many proteins that are slated for destruction meet their fate in specialized intracellular structures called lysosomes, but protein degradation also occurs elsewhere in a cell, including the cytosol and the nucleus. (Cytosol is the intracellular milieu outside of the many compartments that reside inside a cell. The nucleus is but one such compartment.) This extralysosomal protein degradation was found to require adenosine triphosphate (ATP), a universal source of chemical energy in living organisms. The ATP requirement for proteolysis was puzzling, because cleavage of the peptide bond between two residues in a polypeptide chain normally happens rapidly (and without a net input of energy) in the presence of a "simple" protease such as trypsin. (Try calling trypsin simple after you see its three-dimensional structure!) In 1978, Avram Hershko and his coworkers in Israel used an extract from reticulocytes (cells on their way to becoming red blood cells) in an attempt at understanding the mechanism of ATP-dependent protein degradation. They separated reticulocyte extract into two fractions that were inactive by themselves but became active when mixed together. The first fraction contained mostly hemoglobin and another, smaller protein, which was purified and shown to be the only factor that the second fraction required for ATP-dependent proteolysis; this protein was named APF ("ATP-dependent proteolysis factor").

At that time, it was unclear why some of the test proteins were degraded and some left intact in reticulocyte extract. So the strategy was simple—useful protein substrates were those that were short-lived in the extract, were degraded in an ATP-dependent manner, and were easy to obtain. Something unusual was happening to the short-lived proteins in these experiments: before disappearing, they temporarily became *larger.* A single species of the substrate—the protein about to be degraded—was observed in the extract samples that lacked ATP, whereas a set of larger substrate-containing molecules was formed in the presence of ATP. The researchers determined that these larger molecules were almost certainly those of the substrate conjugated to one or more APF molecules. The exploration of APF continued in Israel and the United States, and in 1980 APF was found to be—what else?—ubiquitin. This result brought together the study of ATP-dependent proteolysis and the earlier analysis of uH2A in chromosomes.

Meanwhile, my colleagues and I at MIT were studying chromosome replication and often discussed ubiquitin: what exactly is that branched protein, uH2A, doing in chromosomes? On a fateful day in 1981, I came across a paper from Tokyo University that described a mutant mouse cell line called ts85. The researchers showed that a specific nuclear protein disappeared at elevated temperatures from ts85 cells. They suggested that this protein might be uH2A. When I saw their data, I had to calm down to continue reading, because I *knew* that this protein *was* uH2A! If so, the ts85 mutant was a godsend to anyone who wanted to apply the power of genetics to the puzzle of ubiquitin. Like flipping a wall switch to see what lamp it controls, one could use ts85 cells to turn the conjugation of ubiquitin to other proteins on and off at will, and then observe what the cell did or didn't do. Daniel Finley (then a graduate student in my laboratory) and Aaron Ciechanover (then a postdoc at another MIT lab) started the analysis of ts85 and found that an extract from these mutant cells, in contrast to an extract from normal cells, produced ubiquitin-protein conjugates



The ubiquitin cycle. From the top, clockwise: In the presence of adenosine triphosphate (ATP), the last residue of a ubiquitin molecule (Ub) becomes joined through a high-energy bond (denoted by a ~) to a cysteine (an amino acid) of a ubiquitin-activating, or E1, enzyme (Uba1). This enzymatic reaction proceeds through an intermediate in which ubiquitin is joined to adenosine monophosphate (AMP). The activated ubiquitin is then transferred to another cysteine in one of several ubiquitin-conjugating, or E2, enzymes (Ubc1, etc.). An E2 enzyme, guided by an accessory protein called recognin, or E3, links the activated ubiquitin to its ultimate acceptor protein, whatever that may be. Many ubiquitin molecules can be linked, sequentially, to one molecule of the protein substrate, as shown by the subscript indicating the number of ubiquitins in a multiubiquitin chain. The substrate is then degraded, in yet another ATP-requiring step, by a protease called the proteasome. Ubiquitin molecules linked to the substrate are not degraded and reenter the free ubiquitin pool, after their liberation from the multiubiquitin chain by enzymes called isopeptidases.

The seemingly paradoxical idea—that ubiquitin may function as a protein stabilizer as well as a signal for protein degradation—was supported by other findings.

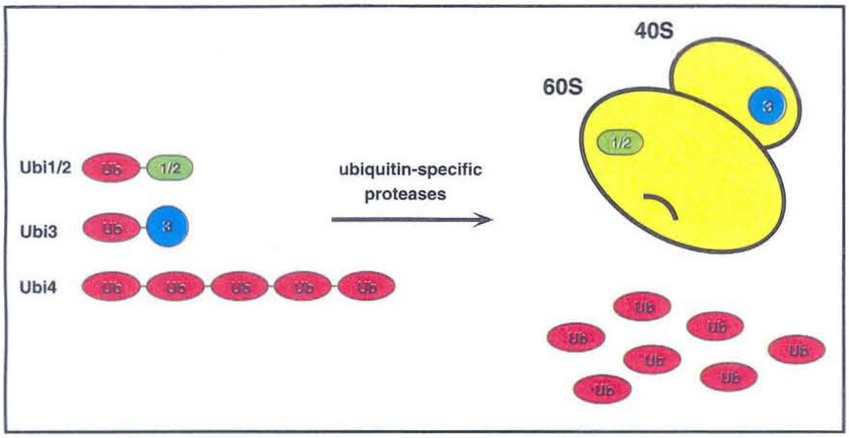
only at a relatively low temperature.

By then, the mammalian ubiquitin system had been resolved by other researchers into three components. The first of these was the ubiquitin-activating enzyme, or E1. This protein catalyzes an ATP-dependent reaction in which the C-terminal glycine residue of ubiquitin is joined to a specific cysteine residue in the E1 enzyme itself. The E1-ubiquitin complex then transfers this "activated" ubiquitin to a specific cysteine in another protein, called the ubiquitin-conjugating enzyme, or E2. The E2 enzyme, either by itself or in a complex with an accessory protein called recognin, or E3, forms ubiquitin-protein ligase—an enzyme that links ubiquitin to its ultimate acceptor proteins.

With this knowledge in mind, let us return to the adventure with ts85. We traced the heat sensitivity of ubiquitin conjugation in ts85 cells to the heat sensitivity of their mutant ubiquitin-activating (E1) enzyme. Since E1 is the first in the cascade of enzymes that prepare ubiquitin for its conjugation to other proteins, we could ask whether the ATP-dependent proteolysis I mentioned earlier also required E1. The results were striking: the degradation of short-lived proteins in ts85 cells was indistinguishable from that in normal cells at 30°C but nearly ceased at 39°C, whereas no inhibition of proteolysis was observed in normal cells at 39°C. These findings provided the first direct evidence that ubiquitin was required for protein degradation in living cells.

The study of ts85 cells was my first encounter with the power of approaches that bring together biochemical and genetic methods. But in the early eighties a sortie into mammalian genetics was still hampered by the impossibility of altering genes at will. (Things have improved greatly since then.) Therefore we embarked on a study of ubiquitin pathways in the species of yeast called *Saccharomyces cerevisiae.* This fungus was "domesticated" by humans eons ago for making bread and those mind-altering beverages called beer and wine. By 1983, when we started working with *S. cerevisiae,* it had already become a fair-haired eukaryote for genetic analysis, not only because of its rapid growth and simplicity (in comparison to plant and animal cells) but also because earlier work by geneticists had resulted in powerful techniques for manipulating yeast genes.

Our first target was the family of ubiquitin genes. Surprisingly, all of these genes were found to encode not the "mature" ubiquitin but precursor molecules that were enzymatically cleaved shortly after their synthesis, to yield ubiquitin and other proteins. One gene encoded a polyubiquitin, while the others encoded ubiquitin linked to unrelated ("tail") proteins. The mystery of the tails was solved in 1989, when Finley (by now a postdoc) and graduate student Bonnie Bartel in my lab, and Martin Rechsteiner's laboratory at the University of Utah, discovered that the free tails were components of the ribosome. We also showed that if the tail proteins were manufactured without ubiquitin, the assembly of ribosomes became inefficient, resulting in slowly growing cells. The likely explanation of this result stems from the fact that ubiquitin is an uncommonly stable and fast-folding protein.



Ubi1/2, Ubi3, and Ubi4 are ubiquitin precursor molecules in yeast. They are nearly identical to the precursors of human ubiquitin. These precursors are ubiquitin fusions—either to itself, as in the polyubiquitin protein Ubi4, or to other, "tail" proteins, as in Ubi1/2 and Ubi3. The precursors are cleaved by ubiquitin-specific processing proteases immediately after the ubiquitin's C-terminal residue, yielding free ubiquitin and (in the case of Ubi1/2 and Ubi3) the free tail proteins. These proteins were found to be components of the ribosome (shown in yellow). Tail 1/2 is a part of the small (40S) ribosomal subunit, whereas tail 3 resides in the large (60S) subunit. The tails' functions and exact locations within the ribosomal subunits are unknown.

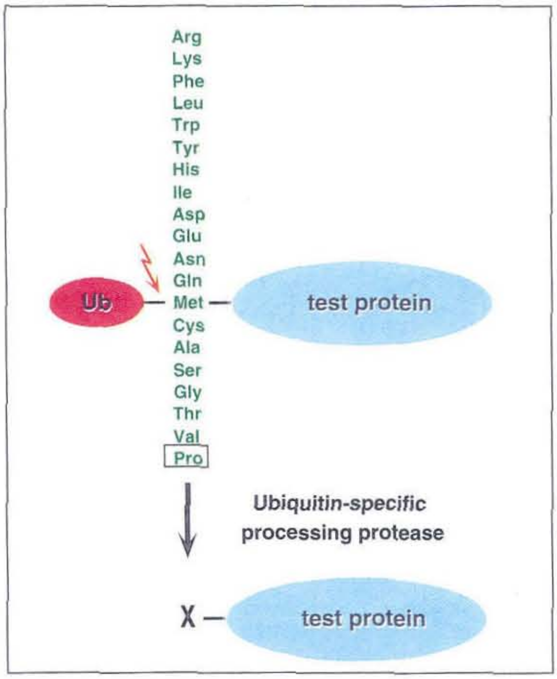
It may therefore protect the rest of a precursor protein from attack by the cell's ever-vigilant proteolytic systems. This protection is transient, because a newly formed ubiquitin precursor is cleaved at the junction of the ubiquitin and the tail. Since this cleavage is fast but not instantaneous, we suggested that ubiquitin's presence provides a partial protection to the ubiquitin-linked tail for the few fleeting seconds when the nascent tail is in gravest danger of being destroyed. As a result, a vulnerable tail-protein molecule may have a better chance of making it in one piece from the ribosome that produced it in the cytosol to an assembly site for ribosomes in the nucleus, where the tail is incorporated into a new ribosome.

Many if not all of the ribosomal proteins are short-lived in vivo unless they associate with each other and the ribosomal RNA to form the ribosome. This way of running the assembly of a multiprotein structure assures that any of its components produced in excess won't end up lingering in the cell. But why were only two of the many ribosomal proteins "chosen" to be produced as ubiquitin fusions during evolution, and why has this arrangement persisted in the course of the two billion years that separate fungi and humans from their nearest common ancestor? Here is a partial answer: the presence of ubiquitin and a ribosomal protein within a single precursor seems to be, among other things, the way to establish a coupling between the numbers of newly made ubiquitin molecules and the numbers of newly assembled ribosomes. An interdependence of this sort may be a useful homeostatic (order-maintaining) arrangement, because ribosomes are in the business of making proteins, whereas the ubiquitin system is about protein destruction—it would be helpful to the cell if these systems were sensitive to each other's abundance and activity.

The seemingly paradoxical idea—that ubiquitin may function as a protein stabilizer as well as a signal for protein degradation—was supported by other findings, which showed that if the gene for a protein that had been difficult to produce because of its rapid intracellular destruction was extended by adding a region that encoded ubiquitin, the yield of the resulting fusion of ubiquitin and the protein was often much higher than the yield of the initial protein.

What about the gene encoding polyubiquitin? Finley and I found that this gene was activated by just about every stressful treatment we could think of. For example, heating cells beyond their normal temperature range, starving them of nutrients, or exposing them to toxic compounds like hydrogen peroxide all resulted in the overproduction of ubiquitin by the polyubiquitin gene. Furthermore, a yeast mutant lacking the polyubiquitin gene was hypersensitive to the stresses that activated this gene in wild-type (normal) yeast. The mutant grew well in the absence of hardships, and seemed normal in other respects as well—until the going got tough. We concluded that ubiquitin, in addition to whatever else it does in a cell, functions as a stress protein—a member of the large class of proteins that all organisms produce, sometimes in copious amounts, in response to adversity. Many of these proteins are also present, at lower concentrations, in cells that are doing just fine, suggesting that stress-specific roles of these proteins are but enhanced versions of their functions in the absence of stress.

Why should a cell under stress overproduce ubiquitin? An oxidative or heat injury increases the amount of damaged proteins in the cell and therefore increases the demand for ubiquitin, whose conjugation to damaged proteins is required for their degradation. Interestingly, an overproduction of ubiquitin in stressed cells doesn't increase their level of *free* ubiquitin, suggesting that the essential function of the polyubiquitin gene is to maintain the cell's free-ubiquitin level in the face of the increased rate at which free ubiquitin is depleted through the formation of ubiquitin-protein conjugates. This property of being distributed between free and tightly protein-bound states is also characteristic of many stress proteins other than ubiquitin. Finley and I proposed that a stress-induced



Right: The ubiquitin fusion technique. Linear fusions of ubiquitin to other proteins are rapidly cleaved in vivo at the ubiquitin-protein junction, irrespective of the identity of the residue at the N-terminal side of the cleavage site. This feature of ubiquitin-specific proteases makes possible the generation of otherwise identical proteins bearing any of the twenty amino acids in the N-terminal position. Proline (Pro) is the only residue that partially inhibits cleavage at the ubiquitin-protein junction. Below: The standard one- and three-letter abbreviations for the amino acids.

Alanine	Ala	A
Arginine	Arg	R
Asparagine	Asn	N
Aspartate	Asp	D
Cysteine	Cys	C
Glutamate	Glu	E
Glutamine	Gln	Q
Glycine	Gly	G
Histidine	His	H
Isoleucine	Ile	I
Leucine	Leu	L
Lysine	Lys	K
Methionine	Met	M
Phenylalanine	Phe	F
Proline	Pro	P
Serine	Ser	S
Threonine	Thr	T
Tryptophan	Trp	W
Tyrosine	Tyr	Y
Valine	Val	V

increase in the *total* level of a stress protein is mediated by a regulatory mechanism that acts to maintain the required level of a *free* stress protein. Examples of such "feedback" circuits have recently been described for several stress proteins.

In 1987, Stefan Jentsch (then a postdoc in my lab) found that one of the ubiquitin-conjugating (E2) enzymes was encoded by a gene called *RAD6*. This gene has been known for many years, because mutations in *RAD6* perturb a number of processes, from sporulation to DNA repair. (Sporulation is one of the stress responses in yeast: when out of food, yeast cells form spores—small, tough, dormant cells ready to outlast the bad times until a wind or whatever transfers them onto anything edible.) Subsequent work greatly expanded the list of known E2 functions; it now includes the ability of cells to resist poisoning by toxic metals, the regulation of the cell cycle, and the control of protein transport across membranes. These remarkably diverse functions are probably underlain by a common mechanism—the degradation of specific proteins tagged by E2 enzymes.

We are halfway through the story but quite a few things are still unexplained. For instance: why attach ubiquitin to a short-lived protein at all—why is this bulky and metabolically costly modification so necessary for the in vivo degradation of many proteins? And furthermore: what features of a protein make it a target of the ubiquitin system? Let us begin with the last problem.

There is no such thing as a totally nonspecific protease—a protease that can cleave any peptide bond with equal dexterity. Even "simple" extracellular proteases like trypsin or pepsin have their preferences, specific for each protease. Features of proteins that make them susceptible to proteolysis are called degradation signals, or degrons. In 1986, Andreas Bachmair and Finley (then postdocs in my lab) discovered the first intracellular degradation signal, and showed it to be recognized by a pathway that involves ubiquitin.

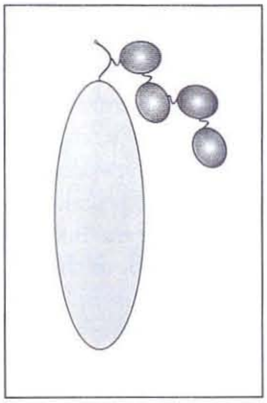
As often happens, the experiments that led to this insight were initially aimed at something else: we wanted to design a fusion protein whose ubiquitin component could not be removed by the ubiquitin-specific proteases that normally cleave a precursor protein at the junction between ubiquitin and a "downstream" polypeptide. To this end, a gene was constructed that encoded ubiquitin fused to an enzyme called β-galactosidase (βgal). (This enzyme was chosen because its fate in the cell could be followed in several convenient ways.) The gene was mutated to convert the methionine (Met) residue at the ubiquitin-βgal junction into a variety of other amino acids. Alas, the ubiquitin-specific proteases couldn't care less about these alterations of their substrate—they continued to cut ubiquitin off the ubiquitin-X-βgal fusion (X being the varied residue) as if nothing had happened.

This result proved to be good luck in disguise—we were thwarted, for a time, in making a fusion protein whose ubiquitin portion stays put, but the near indifference of the proteases to the identity of residue X yielded a method for producing, in a living cell, any residue at the N-terminus of any protein—until then an impossible feat. Why impossible? Because of the way the genetic code works: every mRNA molecule—the messenger that carries the protein's assembly instructions from the genes to the ribosomes, where the proteins are manufactured—is "read" starting from the codon (a unit of RNA encoding one amino acid) that specifies methionine. The ribosome needs some way of knowing where to begin, but why a methionine codon instead of a codon for another amino acid was chosen for this purpose at the dawn of earthly life is unclear. However, once this fundamental early choice had been made, it became "fixed" in the design of living cells. Thus all proteins produced in vivo start off with an N-terminal methionine. Lots of things can happen to this methionine later on—it's retained in many proteins, and it's chemically modified in others; it may even be removed by specific proteases, but the current understanding of these reactions is insufficient for their assured manipulation. Linking ubiquitin to the

Below: The N-end rule for yeast.
Top right: A comparison of the N-end rules in three organisms of increasing complexity. Open circles stand for stabilizing N-terminal residues; red circles are destabilizing ones. The N-end rule is actually more elaborate than is shown here, in that some destabilizing residues are recognized directly, whereas others undergo specific in vivo modifications before recognition.
Bottom right: This isn't a balloon animal, but a multiubiquitin chain containing four ubiquitins, drawn roughly to the scale of the X-βgal subunit to which the chain is attached. Multiubiquitin chains in vivo can contain more than 50 ubiquitins!

Residue **X** in **X-ßgal**	In vivo half-life of **X-ßgal**
Arg	2 min
Lys	3 min
Phe	3 min
Leu	3 min
Trp	3 min
Tyr	10 min
His	3 min
Ile	30 min
Asp	3 min
Glu	30 min
Asn	3 min
Gln	10 min
Cys	> 20 h
Ala	> 20 h
Ser	> 20 h
Thr	> 20 h
Gly	> 20 h
Val	> 20 h
Pro	> 20 h
Met	> 20 h

Amino acids	F	L	W	Y	R	K	H	I	N	Q	D	E	C	A	S	T	G	V	P	M
Bacterium (*E. coli*)	●	●	●	●	●	●	○	○	○	○	○	○	○	○	○	○	○	○	○	○
Yeast (*S. cerevisiae*)	●	●	●	●	●	●	●	●	●	●	●	●	○	○	○	○	○	○	○	○
Mammal (rabbit reticulocyte)	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	○	○	○	○

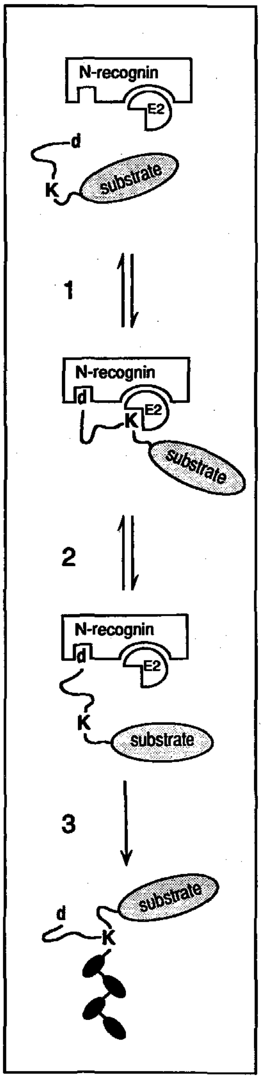


N-terminus of a protein bypasses this problem. The desired N-terminal residue can now be produced by the ubiquitin-specific proteases that cut the fusion protein after the last residue of ubiquitin—away from the initial N-terminus of a ubiquitin-protein fusion.

The new method in hand, we discovered something remarkable almost immediately: X-βgal proteins bearing different N-terminal residues had different in vivo half-lives. (The half-life of a protein is the time it takes for 50 percent of the protein molecules initially present to disappear.) For example, Met-βgal, (which bore N-terminal methionine) had a half-life of at least 30 hours—an eternity by the standards of short-lived proteins. In striking contrast, Arg-βgal (which bore N-terminal arginine) had a half-life of two minutes. One way to appreciate the fleetingness of this half-life is to consider that it takes the ribosome about two minutes to synthesize the approximately 1,100-residue Arg-βgal. In other words, a newly formed molecule of Arg-βgal is destroyed in about the time it took to make it in the first place!

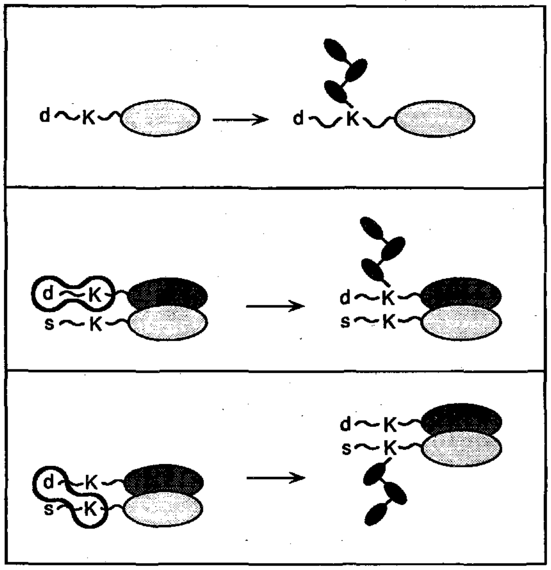
Measurements of degradation rates of X-βgal proteins in yeast yielded a relationship between the in vivo half-life of a protein and the identity of its N-terminal residue—a new, startlingly simple code. We named it the N-end rule and proceeded to explore the vistas opened up by this discovery. It was soon found that distinct versions of the N-end rule operated in all organisms examined, from bacteria to mammals. The three N-end rules in the illustration above are different but also hauntingly similar: the set of destabilizing residues in bacteria is a subset of the analogous set in yeast, and that, in turn, is a subset of the analogous set in mammalian reticulocytes—cells on their way to becoming red blood cells. We don't know the functional meaning of these differences, but it appears that the N-end rule book depends on the cell's physiological state. In other words, the N-end rule is a "soft-wired" code, in contrast, for example, to the genetic code, which is "hard-wired" in the sense that it is the same for all genes in all organisms. (There are, in fact, a few exceptions to the latter statement, as is the case with most statements in biology. Nearly every rule that can be broken in principle is actually violated somewhere in the world of living things, for evolution respects few constraints other than those imposed by physics.) The N-end rule is just beginning to yield its secrets—another story, to be described someday in an article of its own.

Central to understanding the N-end rule is the underlying degradation signal, which we named the N-degron. Is it actually as simple as a single residue at the N-terminus of a protein? What is the role of ubiquitin in the function of the N-degron? My colleagues and I addressed these questions by mutating N-end rule substrates (proteins that are degraded in accordance with the N-end rule) and determining their in vivo half-lives. By 1989, genetic analysis had shown that the N-degron consists of two components: a destabilizing N-terminal residue, and an amino acid residue called lysine at a specific position in the substrate. A parallel biochemical study indicated that multiple ubiquitin molecules



Left: The mechanism by which the N-end rule recognizes a substrate and prepares it for degradation. 1.) N-recognin binds to the substrate's destabilizing N-terminal residue (d). 2.) The relevant lysine (K) is captured by the ubiquitin-conjugating enzyme (E2) associated with the N-recognin. 3.) The lysine capture results in the synthesis of a lysine-linked multiubiquitin chain (black ovals) by the E2 enzyme.

Below: *Cis-trans* recognition and degradation of N-end rule substrates. The upper panel shows the single-subunit case, with d, K and the multiubiquitin chain as above. The middle panel illustrates *cis* recognition of a two-subunit protein, one subunit of which bears a stabilizing N-terminal residue (s). The bottom panel shows how the same protein can be recognized in *trans*. Note that the multiubiquitin chain is now linked to the other (lower) subunit.



become linked to an N-end rule substrate shortly before its degradation. Strikingly, all of these ubiquitin molecules were found to dangle from one lysine—the same one that had been pinpointed by genetic analysis. Thus, instead of being attached to several different lysines of a substrate such as Arg-βgal, the ubiquitin molecules formed a multiubiquitin chain.

What makes a lysine in an N-end rule substrate the site of ubiquitin conjugation? The relevant lysine must be located *spatially* close to the N-terminus—this requirement includes a proximity to the N-terminus *along* the substrate's polypeptide chain. The recognition of a short-lived protein by the proteolytic machinery starts with the binding of a protein, called N-recognin, to the substrate's N-terminal residue. This binding is reversible, and unless the E2 enzyme (which exists in a complex with N-recognin) binds rapidly to the proper internal lysine of the same substrate, the N-recognin-E2 complex "falls off" the substrate and has to start again. A critical lysine should be easy to find if it's positioned for a nearly simultaneous capture of both it and the substrate's N-terminal residue by the complex's two binding sites. Alternatively, the relevant lysine could be a part of a mobile region of the substrate that doesn't fold up into one preferred structure (or conformation, as we say in the trade). While flopping around, the substrate's lysine may approach the bound N-recognin-E2 complex often enough for the E2 enzyme to catch it before the entire complex dissociates from the substrate.

Now that we have gotten sophisticated about the recognition system, let's push it a little further. Thus far, the N-degron's two components have been assumed to reside in the same polypeptide; they are said, in this case, to be recognized in *cis*. But there's also an arrangement called *trans,* in which a destabilizing N-terminal residue and the relevant lysine are in two different subunits (polypeptide chains) of a multisubunit protein. Would such a "split" N-degron work? In 1990, Erica Johnson (then a graduate student in my lab) showed that it would. This discovery revealed a previously unsuspected ability of the N-end rule pathway: of the two subunits bearing the split N-degron, only one subunit—the one containing the relevant lysine—was degraded, whereas the other subunit was left unharmed. In other words, the destruction of a multisubunit N-end rule substrate is confined to those subunits that can be linked to a multiubiquitin chain.

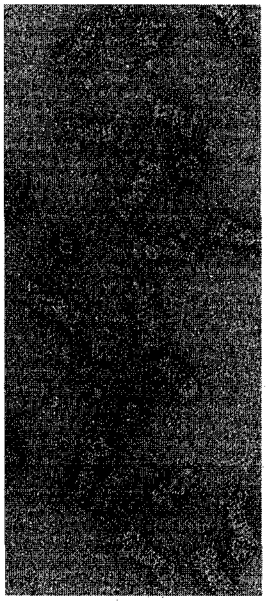
How many distinct degrons (recognized by different recognins) are there in a cell? We don't know, but "at least three" is a safe answer. One class contains the N-degrons I've already discussed. Another distinct class of degradation signals is present in proteins called cyclins, which function as subunits of cyclin-dependent kinases—enzymes that control cell growth and division. Several studies have shown that cyclin degradation is ubiquitin-dependent; moreover, a stretch of nine residues is conserved among many cyclins and is required for their destruction. Yet another class of degradation signals has been described by Martin Rechsteiner and coworkers at the University of Utah, who noticed that many short-lived proteins (including certain cyclins) contain sequences that are unusually rich in the amino acids proline, glutamate, serine, and threonine. Rechsteiner has suggested that some of these sequences may act as degrons. Indeed, deleting such a region from a short-lived protein often stabilizes the protein. And the end of the list of degrons is not yet in sight: for example, Mark Hochstrasser (then a postdoc in my lab) and I have described two distinct degradation signals in a single protein called Matα2—a repressor of RNA synthesis that regulates sexual differentiation in *S. cerevisiae* (yes, fungi have sex, but this story is about ubiquitin).

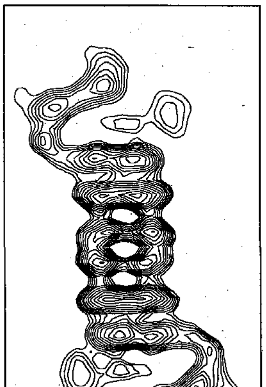
The two-component design established for the N-degron appears to be characteristic of other degradation signals as well. The first component of these signals is an internal region of a protein (instead of its N-terminus) that is specific for each degron, while the second component is likely to be a conformationally mobile lysine (or lysines). If so, these other, still dimly understood degradation signals may also exist in versions analogous to the "split" N-degron.

Indeed, Peter Howley and his colleagues at NIH and Harvard Medical School have shown that a protein called p53 can be marked for destruction as a result of its binding by a protein known as E6—a product of an oncogenic (cancer-causing) human papilloma virus. (Names such as "p53," "E6," and many others are often little more than labels used to distinguish one protein among the multitude of its fellows, which are often discovered before their functions are known. For instance, p53 means "a protein with a molecular mass of about 53,000 atomic mass units.") Oncogenic papilloma viruses, whose sexual transmission among humans increases the risk of certain cancers, are able to induce the proliferation of infected cells. The viruses achieve this in part by decreasing, through ubiquitin--dependent degradation, the level of the cellular regulatory protein p53, whose binding by the viral protein E6 destabilizes p53 *without* destabilizing E6 itself. There is a striking analogy between this effect (mediated by an unknown degradation signal in p53) and the mechanics of a split N-degron.

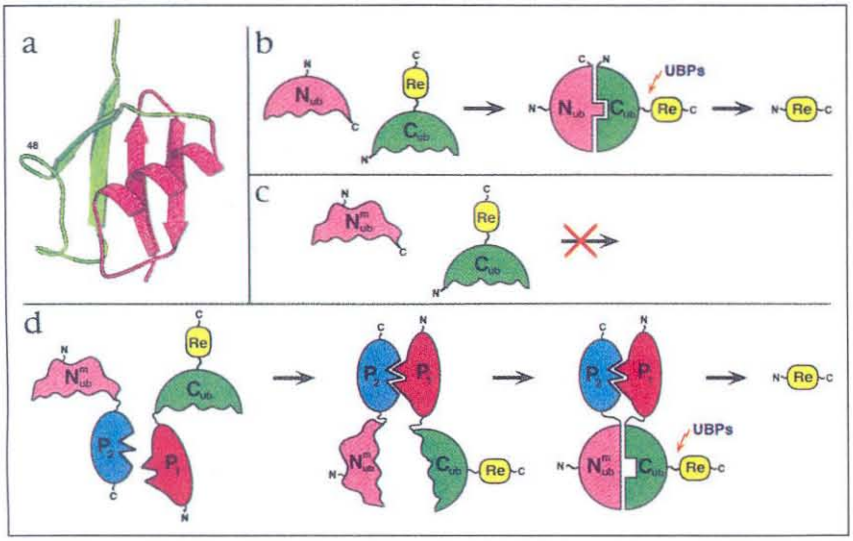
The protease that degrades ubiquitin-linked proteins is called the 26S proteasome; "26S" (26 Svedberg units) is shorthand for how rapidly this large particle sediments in a centrifuge. The 26S proteasome attacks a protein that bears a multiubiquitin chain in a reaction that requires ATP and the multiubiquitin itself. Thus ATP is used at least twice in ubiquitin-dependent proteolysis: first at the step of ubiquitin attachment (or rather ubiquitin activation), and then at a poorly understood step during the actual degradation of a substrate. The 26S proteasome contains some 40 distinct subunits and is unstable in the absence of ATP, dissociating into several components. One of them is called the 20S proteasome, a particle that can cleave a variety of peptide bonds in a reaction that doesn't require ATP. Biochemical studies of the 20S proteasome, and electron-microscopic observations of an analogous protease from bacteria, suggest that the proteasome destroys a protein substrate in a process that involves "threading" the substrate's polypeptide chain through a channel inside the cylinder-shaped proteasome.

Now that we have a nodding acquaintance with the gadgetry of the ubiquitin system, let us attempt an answer to the central question: what exactly is ubiquitin's function? One possibility is that the formation of a multiubiquitin chain linked to a substrate produces additional binding sites for the proteasome's components. As a result, the probability of the proteasome "falling off" the substrate would decrease, and *that* could facilitate the substrate's destruction. Here's why: the proteasome must at least partially unfold the protein in order to thread it through the channel where the proteolysis actually occurs. A folded protein molecule is not a static structure: its polypeptide chain moves about a bit, and sometimes quite a bit, as it gets kicked by packets of water molecules. If the proteasome can "catch" a mobile, relatively unstructured region that becomes exposed during these occasional partial unfoldings (called fluctuations), the substrate's conformation might be destabilized strongly enough for the proteasome to start its work. This implies that the proteasome is "waiting" for a fluctuation; the longer the wait, the greater the probability of a suitable unfolding event. And if the formation of a multiubiquitin chain retards the dissociation of the substrate from the proteasome, the allowed waiting time becomes longer, increasing in turn the probability of





Top: An electron-microscopic image of a crowd of 26S proteasome particles, magnified 300,000 times.
Bottom: A computer-enhanced image of a single 26S proteasome, magnified 1,800,000 times. Electron micrographs courtesy of Wolfgang Baumeister and colleagues at the Max Planck Institute in Martinsried, Germany.



How to detect a protein interaction in vivo as it occurs. (a) This diagram illustrates the folding pattern of ubiquitin's polypeptide chain, without detailing the amino acids. The N- and C-terminal halves are shown in pink and green, respectively. (The 48 marks the lysine where other ubiquitins can attach.) (b) If a "reporter" protein (Re) is fused to a free C-terminal half (C_{ub}), ubiquitin-specific proteases (UBPs) won't cleave the fusion until the C-half associates with an N-half (N_{ub}) to form a nearly normal ubiquitin molecule. Once liberated, the reporter protein can be detected in several ways. (c) If the N-half is mutated (N^m_{ub}) in a way that weakens its interaction with the C-half, the reporter is not cleaved off. (d) But if the C-half and the mutant N-half are linked to proteins that interact in vivo (P_1 and P_2), the interaction will bring the two halves so close together that their residual affinity will be sufficient to form a functional ubiquitin anyhow, causing the reporter protein to be cut free.

catching a partially unfolded substrate.

Two results indicate that the unfolding of a protein substrate is indeed a prerequisite for its destruction by the proteasome, and that a multiubiquitin chain plays a role in the process. Jennifer Johnston (a postdoc in my lab) has found that the ubiquitin-dependent degradation of a protein—dihydrofolate reductase, or DHFR—by the N-end rule pathway can be inhibited by methotrexate, a small molecule that specifically binds to DHFR. This finding—that a modest increase in the conformational stability of DHFR as a result of its binding to methotrexate is sufficient to stop the proteasome juggernaut in its tracks—is consistent with the idea that a substrate's conformation is one major barrier faced by the proteasome. In addition, Tillmann Rümenapf (then a postdoc in my lab), James Strauss (PhD '67, Caltech's Bowles Professor of Biology), and I have found that the formation of a substrate-linked multiubiquitin chain is actually unnecessary for the substrate's degradation by the N-end rule pathway, provided that the substrate is conformationally unstable to start with. These findings are consistent with the model discussed above, but they are also consistent with another idea—that the substrate-linked multiubiquitin chain, by virtue of being in close proximity to the substrate, may interact with it and thereby play a direct role in destabilizing its conformation.

The above models are specific enough to make testable predictions, but barely begin to address the true range and subtlety of reactions at the proteasome. For example, we've discussed multiubiquitin chains as if they simply hang there—conjugated to a substrate, bound to the proteasome. In fact, a multiubiquitin chain has a life of its own: it folds in certain preferred ways; it also grows through the activity of E2 enzymes and shrinks through cuts made by ubiquitin-specific proteases, at least one of which is a component of the proteasome. These and other complexities are trying to speak to us and will be understood someday, when even a popular yarn about ubiquitin shall require a book to be told.

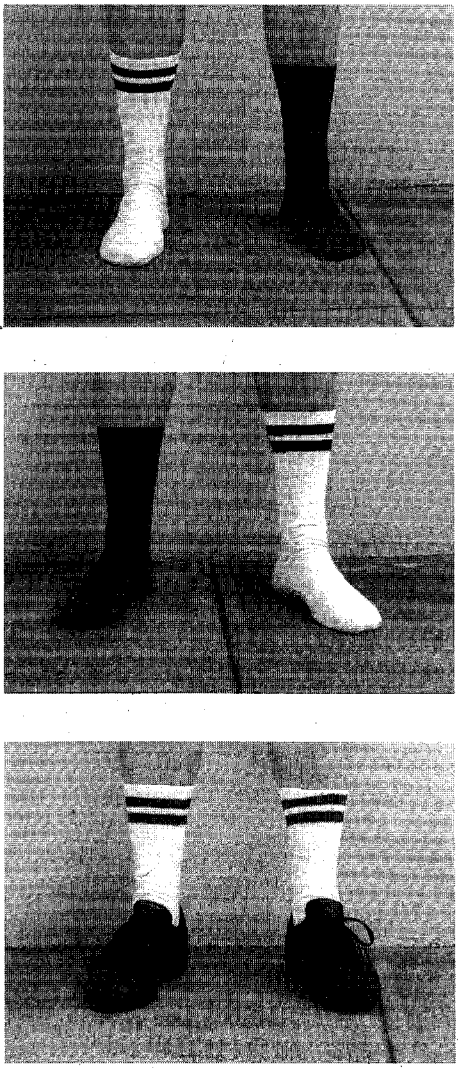
In the meantime, I shall mention just one instance of research on ubiquitin bearing fruit in other fields. Nils Johnsson, a postdoc in my laboratory, has found that the compact organization of ubiquitin belies a subtlety: the ubiquitin's N-terminal "half" retains elements of its folded structure even in the absence of the rest of the molecule. Moreover, the N-terminal half can bind in vivo to a *separately produced* C-terminal half, forming a nearly normal ubiquitin. This discovery has led to a new method for detecting protein interactions in living cells.

The growing understanding of intracellular proteolysis is providing us with powerful tools for manipulating the in vivo half-lives of intracellular proteins, including those whose malfunction or overproduction leads to cancer and other illnesses. Most drugs of today are incapable of altering the in vivo stability of a protein target. But this is likely to change, and when it does, an entirely new class of therapeutic agents will emerge, with exciting implications for the cure of currently intractable diseases. □

Alexander Varshavsky is the Smits Professor of Cell Biology at Caltech. He is also a member of the National Academy of Sciences and the American Academy of Arts and Sciences. Varshavsky was born and educated in Moscow, Russia. In 1977, he joined the faculty at the Massachusetts Institute of Technology in Cambridge and stayed there until 1992. Varshavsky and coworkers discovered the exposed regions in chromosomes (which form at the beginnings of active genes, at the origins of DNA replication, and at other sites of localized activity in the chromosomes), deciphered the mechanism of separation of intertwined sister DNA molecules during chromosome replication, and described the phenomenon of induced gene amplification that contributes to rapid evolution of cancer cells within a tumor. Varshavsky's initial interest in ubiquitin stemmed from its presence in chromosomes. His laboratory produced the first direct evidence that ubiquitin is required for protein degradation in living cells, and in 1986 discovered the first intracellular degradation signal.

Lab Notes

Feet have handedness (or chirality), too—your left foot and your right foot are not identical. On the other hand—or, rather, foot—socks are achiral. The black sock will go on either foot with equal ease. Shoes, however, are chiral—each shoe fits only one foot.



When One Hand Is Better Than Two

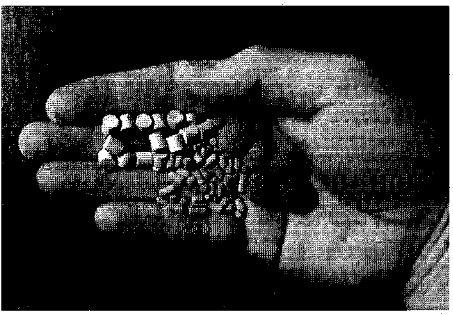
Almost invariably, only one enantiomer of the drug is good for what ails you. The other one is, at best, inert.

When you're gulping a couple of tablets of your favorite analgesic to soothe your pounding skull, it probably wouldn't cheer you any to reflect that more than 50 percent of the pill is binders, buffers, and other non-pain-relievers. Well, here's some more good news: in many nonprescription drugs, fully one-half of the active ingredient isn't. That's because biologically active chemicals generally contain a chiral center. "Chiral" comes from the Greek word for hand, and just as we have left and right hands, molecules can have left- and right-handed forms called enantiomers. ("Enantios" is Greek for "opposite.") "Your shoes are also chiral," notes Mark Davis, the Schlinger Professor of Chemical Engineering. "Your left shoe has to go on your left foot, and your right shoe on your right foot. Unless you have children..." And if the kids haven't been playing in your closet, a quick inventory should reveal an equal number of right and left shoes—what a chemist would call a racemic mixture of shoes.

While racemic shoes in the closet are desirable, racemic molecules in a medicine aren't, because almost invariably, only one enantiomer of the drug is good for what ails you. The other one is, at best, inert. Ibuprofen, for example, is sold racemically in Advil and Motrin, but only the left-hand variety does anything for your headache. However, both versions cause stomach irritation, so taking the racemic mixture gives you twice the queasiness per unit of aaahhhh. Sometimes the wrong enantiomer has serious side effects—for example, one enantiomer of Ventolin, the generic anti-asthmatic inhalant, dilates your bronchial passages, while the other form causes high blood pressure in a small percentage of patients. And then there's thalidomide. This drug, sold in Europe to pregnant women for morning sickness in the early 1960s, caused some 3,000 malformed infants to be born before the drug was pulled from the market. It turned out that while one enantiomer was, in fact, a powerful and specific sedative, the other caused massive birth defects.

Unfortunately, it's very difficult to synthesize one enantiomer exclusively. (Nature does it routinely by using enzymes, but doesn't supply enzymes

"If you're making tons of a compound, you can't have a thousand people sitting in your factory picking crystals."



Solid-state catalysts like these are an industrial mainstay.

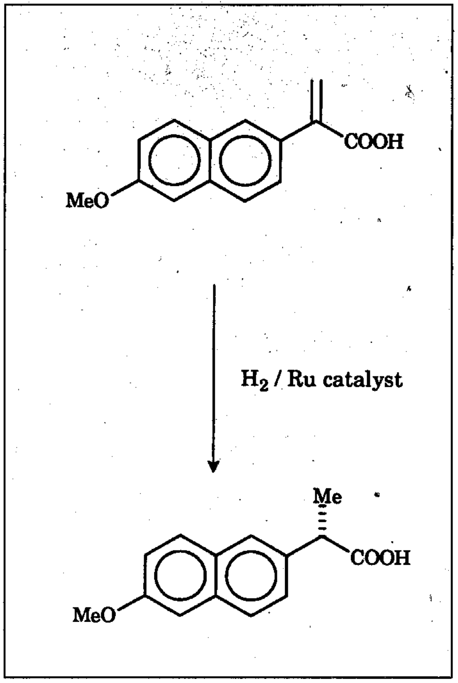
for many of the compounds we wish to make.) Recognizing this, the Food and Drug Administration (FDA) until recently allowed racemic drugs to be sold, provided that testing showed that the other enantiomer had no untoward effects. In 1992, however, the FDA revised its guidelines to recommend that new drugs should be enantiomerically pure, unless the manufacturer can prove that the racemic mixture is actually more beneficial.

The traditional path to enantiomeric purity, says Davis, "was to perform a racemic synthesis that made both hands, and then do what chemical engineers do well—design a separation process and throw half of your product away. That's been done for many drugs. That's what Pasteur did when he discovered enantiomers—he saw two different crystals in a sample of tartaric acid, and he picked one out from the other. But if you're making tons of a compound, you can't have a thousand people sitting in your factory picking crystals." Of course, pharmaceutical companies use much more sophisticated separation techniques to meet the FDA's exacting purity standards.

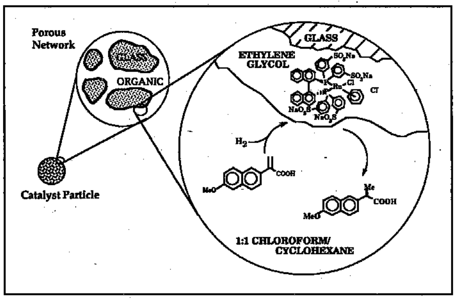
In the late 1970s, chemists finally succeeded in copying Nature's strategy by developing catalysts that themselves had a handedness, and imparted it to their products. Unlike the enzymes, these catalysts were relatively simple—metal ions bedecked with chiral organic shrubbery that held the ingredients in such a way that only the correct enantiomer could result from their reaction. But the chemists weren't home free yet—these catalysts had to be dissolved in the reaction medium to do their job, and once in solution, they often proved as difficult to remove as the wrong enantiomer had been. And leaving the catalyst in the drug is no better than leaving the wrong enantiomer.

Now if the catalyst were a solid, it could simply be filtered out once the reaction was finished. (Achiral catalysts that *are* solids are widely used industrially.) Many people have tried to solidify these chiral catalysts, but the problems inherent in having a catalyst that is at once a filterable solid *and* soluble in the reaction medium are obvious. The most promising approach was to form a chemical bond between the catalyst and some insoluble substance, allowing the catalyst to stick out into the reaction medium while still being tethered to something retrievable. But the tethered catalysts generally proved to be less active (and most often less selective in their output!) than their free-swimming counterparts, an effect that can probably be blamed on the nearby solid's preventing the catalyst's organic shrubbery from springing into its proper positions, just as a rose bush planted too near the house winds up growing flat against the wall.

Davis realized that there was a way to make the catalyst stick to a solid without having to tie the two so closely together. Simply coat the solid (in this case, porous glass beads so tiny that they look like powder) with a solvent that the catalyst will dissolve in but the reaction medium won't. And if the catalyst is more soluble in your solvent than in the reaction medium, when you mix all the ingredients together the catalyst should migrate into the solvent, while at the same time the solvent and the reaction medium separate like oil and water. And if the solvent has a greater affinity for the glass beads than the reaction



Above: The final step in synthesizing naproxen. The planar precursor molecule (top) is achiral. The chiral naproxen molecule (bottom) has its methyl group (Me) behind the plane of the page in the "good" form; in the toxic form the methyl group sticks out in front of the page.
Below: The catalytic system. Porous glass beads are coated with the catalyst dissolved in ethylene glycol. Ru stands for ruthenium.



medium does, a little brisk stirring coats the beads with a very thin layer of the catalyst-containing solvent. Since the catalyst is fully dissolved in a liquid, the organic shrubbery is free to take its preferred shape, and as the solvent layer is very thin, the catalyst is close enough to the reaction medium to slurp up the starting ingredients, run the reaction, and spit the finished products back into the reaction medium.

Davis chose to try this approach on naproxen, the active ingredient in the prescription painkiller Naprosyn and its weaker over-the-counter cousin, Aleve. Molecule for molecule, right-handed naproxen is 55 times more potent than aspirin. (Left-handed naproxen is a liver toxin.) A ruthenium-containing catalyst for synthesizing left-handed naproxen had already been developed, making it an ideal test case. Davis's group chose ethylene glycol as their solvent and a mixture of chloroform and cyclohexane as their reaction medium, and were then faced with the task of trying to modify the catalyst so that it would dissolve in ethylene glycol and adhere to the beads. Recalls Davis, "This was the hardest part—it took about a year to synthesize this catalyst with the right kind of stickers on it without destroying its chirality." With the right stickers, "we threw everything into a bucket, and the whole thing self-assembled. As a comparison, we didn't add the solid, and, in fact, it didn't react." With the beads, they got 96 percent yield of the correct enantiomer—good enough for the FDA's new guidelines—and 100 percent removal of the catalyst after filtration. The solid catalyst is about one-third as fast as the soluble version, Davis says, but the ease of separation is more than worth it from the manufacturing standpoint.

Proving that this approach works in one particular case is a far cry from codifying it into a recipe that one could use to stock an entire pharmacy, but Davis expects to see a lot of other people applying this method. "The wave of the future is not through separating compounds, because you're wasting half of what you make, but in never synthesizing the wrong compound in the first place." □ *—DS*

Letters

Editor:

I just read your excellent article on Linus Pauling in the most recent *E&S*. You are probably overloaded with stories about him, but here's another one.

In 1972, I was teaching a chemistry appreciation course at the University (or whatever it was called then) of Wisconsin in Stevens Point. These "science for poets" courses were very popular in those days. The objective was to show the wonder and excitement of chemistry and its applications to our daily lives, and not to bore or confuse the students with a lot of theory (chemistry majors were forbidden). Part of the course involved "case studies"—for example, we read *The Double Helix*, as much for its insight into the personalities and politics of science as for its importance to what is now molecular biology. And, of course, that brought us back to Linus: vitamin C was hot, Vietnam and the peace movement was hotter, and we had already run into Pauling in Crick's race for the structure of DNA.

Sensing that the students might like to meet him, I wrote him, saying that I was sure he wouldn't remember me from Atom (sorry), even though I was in his freshman chemistry course in 1958–59. Explaining the situation, I took a long

Letters continued

Pauling visits Caltech in 1970.

shot and asked for an hour of his time for a two-way conversation with my class (we had just acquired a speaker-phone, the high-tech pinnacle of communications technology of the day). Linus, of course, was more than happy to oblige.

The students asked about *all* of it—they were most interested in the vitamin C controversy, especially since the speaker was not part of the establishment. They also asked about nuclear testing (although they were too young to remember it). As I recall, they didn't care too much about molecular modeling and structure. And the students got to ask the questions.

They were entralled. Anyone who ever heard a Pauling lecture didn't forget it. I suspect that Linus Pauling is all that many of these students ever learned or remember about chemistry, and I think that may have been more important than the rest of it. November 15, 1972. I still remember it.

Oh yes, I still remember him as a chemistry professor. That might be why I went on for degrees in chemistry, although they are now fully depreciated and I've had several different careers since. I learned much from Linus about the actual practice of science, and about having values and acting on them, and this has done more for me that the technology and science itself.

I've seen other academic institutions treat their free radicals much as Caltech treated Pauling, and they are much the worse for it. Thanks, Caltech, for finally giving him the recognition he so richly deserved.

D.E.I.

Bob Rouda, '62 Ch

Editor:

I owe my Caltech career to Linus Pauling. In the fall of 1952 I came to Caltech instead of MIT. My teacher of geology and chemistry had said, "Sam, you have to go to Caltech. Linus Pauling's there, and he's the greatest crystallographer in the world!" End of argument.

This compelling logic and *Facts About Caltech* must have worked their wonders on that 16-year-old boy, but they ill prepared me for the shock of my first day at Caltech. Monday, 8:00 a.m., Freshman Chemistry: "Good morning, boys. My name is Linus Pauling." Those were the last words I understood all morning. When he spoke about the Bragg equation and read his five-inch slide rule to seven places, all the valedictorians around me nodded as if they understood. I did not. Afterward, everyone raced to the nearest calculator to confirm the slide-rule answer. Of course (thanks probably to small writing and Scotch tape) it was correct.

Flash forward nearly 40 years. I had written Dr. Pauling at his institute, and he invited me to visit. Although he couldn't have known me from Adam's off ox, he was extremely courteous and friendly. His appearance was energetic, and his voice retained the uniquely clear enunciation that I had remembered. (A biochemist friend had recently opined that Dr. Pauling was slowing down, becoming only half as sharp as previously and therefore only five times as sharp as anyone else.)

The conversation turned to how his interest in chemistry began. As a boy he was a forester. He became interested in the minerals he found, then metallurgy, forging, and welding—particulary interesting to me as a mechanical engineer. His interest then expanded to crystals and then to all of chemistry. At his Big Sur ranch, he still used his geology hammer until it was inadvertently left in a car that was sold. When I sent him a replacement, he responded with copies of his books on vitamin C and nuclear testing; I began taking the vitamin C and have never felt better.

His knowledge was extraordinarily broad. When I mentioined that my wife worked with a great-ape language-acquisition project, he began a discussion of the 40 differences between the fetal bloods of humans and gorillas. Like Edison's, Dr. Pauling's career was remarkably productive, and for a similar reason. Not only did he put out a great deal every day, he worked more days, still productive at an advanced age.

A wonderful afternoon with a great man—one of the many benefits of attending Caltech.

Samuel R. Phillips, '56 Eng, MS '57 ME

Six Easy Pieces

Essentials of Physics Explained by Its Most Brilliant Teacher

by Richard P. Feynman
Addison-Wesley Publishing Company, Reading, Massachusetts, 1994

The six easiest of Feynman's *Lectures on Physics* (actually five easy ones and one hard one) may not provide much food for thought for Caltech graduates who have tasted the real thing in two years of the famous three red books. The editors intended this to be a physics primer for a wider nontechnical audience and to introduce the nonscientific public to Feynman's genius as a teacher. But the book comes with an added bonus: six tapes or CDs of Feynman himself, originally recorded on reel-to-reel tape in 201 East Bridge when Feynman began the course. The old tapes, which have languished in Caltech's Archives for 30-something years, have been digitally remastered; the sound quality leaves something to be desired by today's standards, but Feynman's unique style (and his Brooklyn accent) come through loud and clear.

The five "easy" lectures (atoms in motion, basic physics, the relation of physics to other sciences, conservation of energy, and the theory of gravitation) were recorded in September and October 1961. Then it's fast-forward to April 1962 for quantum behavior, which he describes to his class as an "entertainment lecture." He admits in his preface to the original edition of *Lectures on Physics*, which is included in this volume, that his experiment to describe the principles of quantum mechanics in a way that did not require partial differential equations was not entirely successful. But it *is* entertaining.

In addition to Feynman's own original preface, the book comes with an introduction by Paul Davies, and a special preface, by David Goodstein and Gerry Neugebauer, to a commemorative edition of *Feynman's Lectures on Physics* published in 1989. Goodstein and Neugebauer call Feynman "a truly great teacher, perhaps the greatest of his era and ours," and also "an extraordinary teacher of teachers." They note that in 1961–62 students began dreading the class (it was not known as "easy"); as their numbers dropped off, their seats were taken by more and more faculty and grad students. If you want to relive Freshman Physics with Feynman for yourself, the set can be ordered from the Caltech Bookstore (with tapes, $49.95; with CDs, $59.95; the book alone is $22.00).

Braving the Elements

by Harry B. Gray, John D. Simon, and William C. Trogler
University Science Books, Sausalito, California, 1995

After the nonscientific public has mastered physics with Feynman, it can take on chemistry with Harry Gray, Caltech's Beckman Professor of Chemistry and director of the Beckman Institute, and his two coauthors from UC San Diego. Ostensibly a textbook for nonchemists, something with the title *Braving the Elements* has to be—you would think— livelier than an ordinary textbook. And indeed it is. Anyone "interested in learning about modern chemistry and how it relates to the environment, energy, health, and other areas of human concern" should find it readable. This includes, according to the authors, lawyers, media people, "and even physicists."

Under the chapter heading "Newsworthy Molecules" the reader can discover the chemical structure of, among many others, ibuprofen, sunscreen, vitamin C, testosterone, AZT, LSD, caffeine, TNT, and sarin (but this isn't a how-to book; it doesn't tell you how to make them). You can learn the chemistry of indigestion and of book decay; read about the chemical industry, including titanium alloy bike frames and composite tennis rackets, in a chapter called "Wall Street Chemistry"; and discover everything a potential juror should know about DNA; not to mention the chemistry of nuclear power, ozone depletion, global warming, smog, cancer treatment, and just about everything else an informed citizen, who doesn't happen to be a chemist, might just be curious about.

The book is briskly and entertainingly written, sprinkled with historical sketches of great moments in modern chemistry—the first controlled nuclear fission reaction, the invention of nylon, the cleanup of the Love Canal. Chemistry is alive and well, say the authors, and to prove it they have written what might almost qualify as a page-turner.

Books continued

The Art of Alessandro Magnasco: An Essay in the Recovery of Meaning

by Oscar Mandel
Leo S. Olschki Editore,
Florence, Italy, 1994

The immediate subject of Professor of Literature Oscar Mandel's monograph is a rather peculiar painting by Magnasco (1667–1749) that hangs in Pasadena's Norton Simon Museum. Labeled *Calefactorium with friars*, the painting depicts a ragtag bunch of gaunt, hooded Capuchin friars untidily, and unreligiously, warming themselves around a monastery fireplace; the disorderly—some would say decadent—scene was described, even in Magnasco's time, as "bizarre." Mandel begins his search for the painting's meaning by querying present-day museumgoers on their perception of the painter's attitude toward his subjects; although opinions varied widely, a clear majority thought it hostile or at least uncomplimentary. After comparing these responses to the opinions of "experts" (i.e., art critics), the majority of whom found the painter either sympathetic to his Capuchins or morally neutral (only a few thought it scornful), Mandel reveals that he himself lines up with those who consider the painter neutral or uncommitted. He then procedes to marshall the historical and textual evidence for his view. He explores Magnasco's own and his contemporaries' attitude toward the church and compares the painting with traditional representations of monks and friars in Italian art, concluding that Magnasco's painting, while perhaps eccentric, is devoid of any moral or ideological viewpoint and represents no negative propaganda.

Why bother to go to such lengths to recover the meaning of a work of art? Mandel approaches this question from an aesthetic point of view: perceiving the meaning of a work adds to the pleasure of viewing it. But he's also using Magnasco's friars to illustrate a larger point about art (and, one presumes, literature). A work's meaning often "spreads out" during the intervening centuries; why is it important to recover the artist's original intent rather than to adopt an interpretation that speaks to our own times? Mandel maintains that we normally dislike separating the work of art from the "hand" that gives it to us. We labor to recover original meanings because aesthetic pleasure is embedded in the larger pleasure of grasping the whole human act of creation: the creation and the creator bound together.

Nano

by Ed Regis
Little, Brown and Company,
Boston, 1995

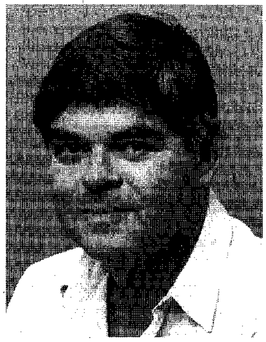
This is not another book about Richard Feynman, although his ghost hovers protectively over most of the story. Subtitled "The Emerging Science of Nanotechnology: Remaking the World —Molecule by Molecule," it's mostly about K. Eric Drexler, who, as an MIT undergraduate in the seventies, conceived the visionary idea of a molecular nanocomputer and, ultimately, a molecular manufacturing machine: a little black box that "will make for you, atom by atom, everything you ever wanted." He was chagrined to discover in 1979 that Feynman had thought it all up first—two decades earlier. In "There's Plenty of Room at the Bottom," Feynman's talk to the American Physical Society in December 1959 (and reprinted in the February 1960 *E&S*, where, over the past 35 years, it has become our most requested article), he prophesied building on an atomic scale: "I am not afraid to consider the final question as to whether, ultimately—in the grand future—we can arrange the atoms the way we want; the very *atoms*, all the way down! . . .The principles of physics, as far as I can see, do not speak against the possibility of maneuvering things atom by atom."

The grand future was not so very far off. Feynman never bothered to think up a way to use his atomic machines, but Eric Drexler did. He started by designing atomic bearings and gears. Working scientists greeted his work with some skepticism—atoms, after all, aren't marbles. He also had to fight the science fiction label and the ridicule of a "Captain Future" image. By the beginning of the nineties, however, which was coming to be known as the nanotechnology decade, Drexler had written a book full of equations. He then was pronounced sane and even testified before Congress. Nanotechnology *is* the future, it is now assumed, and all that remain are the philosophical questions: Will nanomachines take over the world? And what will people do when work becomes unnecessary?

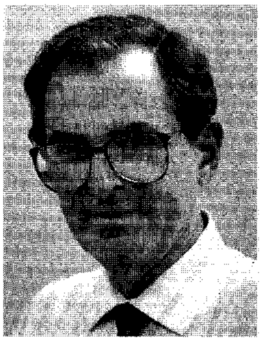
Ed Regis, the author of *Who Got Einstein's Office*, has previously written about the weirder fringes of science in *Great Mambo Chicken and the Transhuman Condition*, in which Drexler also appears. He writes with humor but treats his subject seriously at the same time. It may sound like science fiction, but it isn't anymore.

Random Walk

R. Stanton Avery (left), chair emeritus of the Board of Trustees, admires a drawing of Avery House with President Tom Everhart at the structure's May 9 groundbreaking. The innovative residence, which will open its doors in the fall of 1996, will house undergraduates, grad students, and faculty, and will provide a meeting place for the exchange of ideas with visiting entrepreneurs.



Paul Dimotakis



David Goodstein

Bradford Sturtevant

Philip Saffman

New Professorships

Appointments to three new professorships and one older one were announced this spring.

Paul Dimotakis has been appointed the John K. Northrop Professor of Aeronautics, a chair established with funds from the dissolution of Northrop University. Dimotakis, who has been a member of the faculty since 1973 and whose research is on turbulent flows, will also remain professor of applied physics.

Vice Provost and Professor of Physics and Applied Physics David Goodstein will become the first Frank J. Gilloon Distinguished Teaching and Service Professor, a chair endowed by the estate of Gilloon, who taught civil engineering at Caltech in 1919–20 and who died recently at the age of 99. Goodstein joined the Caltech faculty in 1966 and works in condensed-matter physics.

The first Hans W. Liepmann Professor of Aeronautics will be Bradford Sturtevant, who specializes in the study of shock waves. Sturtevant, who earned his MS and PhD from Caltech, has been a member of the faculty for 35 years. The chair honors Liepmann, the Theodore von Kármán Professor of Aeronautics, Emeritus.

Philip Saffman will succeed Anatol Roshko, who succeeded Hans Liepmann in the von Kármán chair. Saffman, whose title will be the Theodore von Kármán Professor of Applied Mathematics and Aeronautics, has conducted pioneering research on various types of fluid interactions. He came to Caltech as professor of fluid mechanics in 1964.

Honors and Awards

Three Caltech faculty members were elected fellows of the American Academy of Arts and Sciences: Tom Ahrens, professor of geophysics; Paul Jennings, acting vice president for business and finance and professor of civil engineering and applied mechanics; and Anthony Readhead, professor of astronomy. Ahrens has also been named the recipient of the 1995 Arthur L. Day Medal and of a life fellowship in the Geological Society of America.

Clarence Allen, professor of geology and geophysics, emeritus, will receive the 1996 Medal of the Seismological Society of America.

Lew Allen Jr., senior faculty associate

Random Walk continued

and former director of JPL, has received the 1995 Goddard Astronautics Award.

Felix Boehm, the William L. Valentine Professor of Physics, has been awarded the 1994 Tom W. Bonner Prize in Nuclear Physics.

Jehoshua Bruck, associate professor of computation and neural systems and electrical engineering, has been selected to receive a Sloan Research Fellowship.

Thomas Caughey, the Richard L. and Dorothy M. Hayman Professor of Mechanical Engineering, has been elected a fellow of the American Association for the Advancement of Science.

Ray Deshaies, assistant professor of biology, has been named a 1995 Searle Scholar.

Jeffrey Dubin, associate professor of economics, has been awarded a 1995 Haynes Foundation Faculty Fellowship.

Sam Epstein, the William E. Leonhard Professor of Geology, Emeritus, and Hugh Taylor, the Robert P. Sharp Professor of Geology, with alumnus Robert Clayton, PhD '55, professor of cosmochemistry at the University of Chicago, have received the Urey Medal of the European Association of Geochemistry.

Harry Gray, the Arnold O. Beckman Professor of Chemistry and director of the Beckman Institute, has been elected a foreign member of the Royal Society of Arts and Sciences in Sweden.

Roy Gould, the Simon Ramo Professor of Engineering, has received the James Clerk Maxwell Prize of the American Physical Society.

Gregory Hjorth, the Bateman Research Instructor in Mathematics, has been awarded the Sacks Prize in Mathematical Logic.

Wolfgang Knauss, professor of aeronautics and applied mechanics, has received the 1995 Murray Medal of the Society for Experimental Mechanics.

Rudy Marcus, Nobel laureate and the Arthur Amos Noyes Professor of Chemistry, has received the Honorary Professorship at Fudan University in the People's Republic of China, and the Lavoisier Medal of the French Chemical Society. He has also been named an honorary fellow of the Chemical Institute of Canada, an honorary member of the International Society of Electrochemistry, and an honorary fellow of University College, Oxford.

Three Caltech professors were elected to the National Academy of Sciences this year: Elliot Meyerowitz, professor of biology; Anthony Readhead, professor of astronomy, and Alexander Varshavsky, the Howard and Gwen Laurie Smits Professor of Cell Biology (see page 26).

Clair Patterson, professor of geochemistry, emeritus, has been awarded the $150,000 Tyler Prize, the world's highest honor in environmental science.

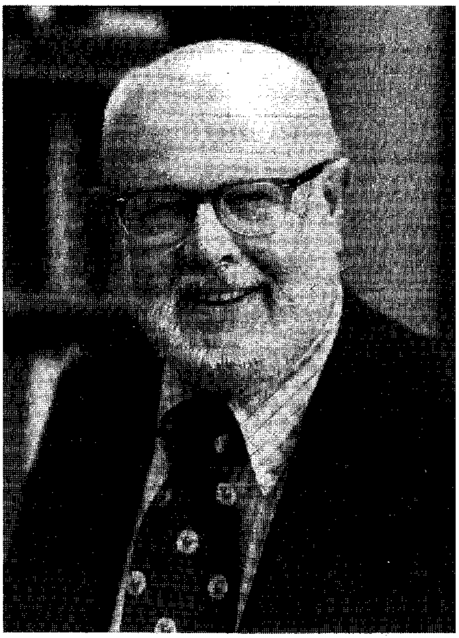
E. Sterl Phinney, associate professor of theoretical astrophysics, has received the Helen B. Warner Prize for Astronomy from the American Astronomical Society.

Douglas Rees, professor of chemistry, has been awarded the 1994 Johnson Foundation Prize by the Johnson Research Foundation of the University of Pennsylvania Medical School.

Ares Rosakis, professor of aeronautics and applied mechanics, has been named a Fellow of the American Society of Mechanical Engineers.

Philip Saffman, the Theodore von Kármán Professor of Applied Mathematics and Aeronautics, has received the American Institute of Aeronautics and Astronautics Fluid Dynamics Award.

Ahmed Zewail, the Linus Pauling Professor of Chemical Physics, is the recipient of the 1995 Herbert P. Broida Prize of the American Physical Society.



William A. Fowler 1911–1995

William A. (Willy) Fowler, Nobel laureate and Institute Professor of Physics, Emeritus, died March 14 in Pasadena at the age of 83. He first came to Caltech as a graduate student in 1933 to work with Charles Lauritsen; he earned his PhD in 1936, whereupon he joined the Caltech faculty, which he never left. His work in the Kellogg Radiation Laboratory put Fowler and his collaborators at the forefront of some of the central issues in modern physics and cosmology. Fowler was primarily concerned with studies of fusion reactions—how the nuclei of lighter chemical elements fuse to create the heavier ones in a process known as nucleosynthesis. It was for this work that he was awarded the Nobel Prize in 1983.

Gerald Wasserburg, Crafoord laureate and the John D. MacArthur Professor of Geology and Geophysics, heads a committee that is planning a symposium on "Nuclear Astrophysics/A Celebration of Willy Fowler," to be held December 14 through the morning of December 16.

Computer models of the ocean can now, thanks to an increased understanding of ocean physics and the development of computer technology, produce a very good approximation of ocean circulation. Compare this simulation (by the Miami Isopycnal Coordinate Ocean Model) of sea surface temperature in the North Atlantic (bluish colors are cold and reddish ones are warm) with the actual satellite infrared image on page 2.