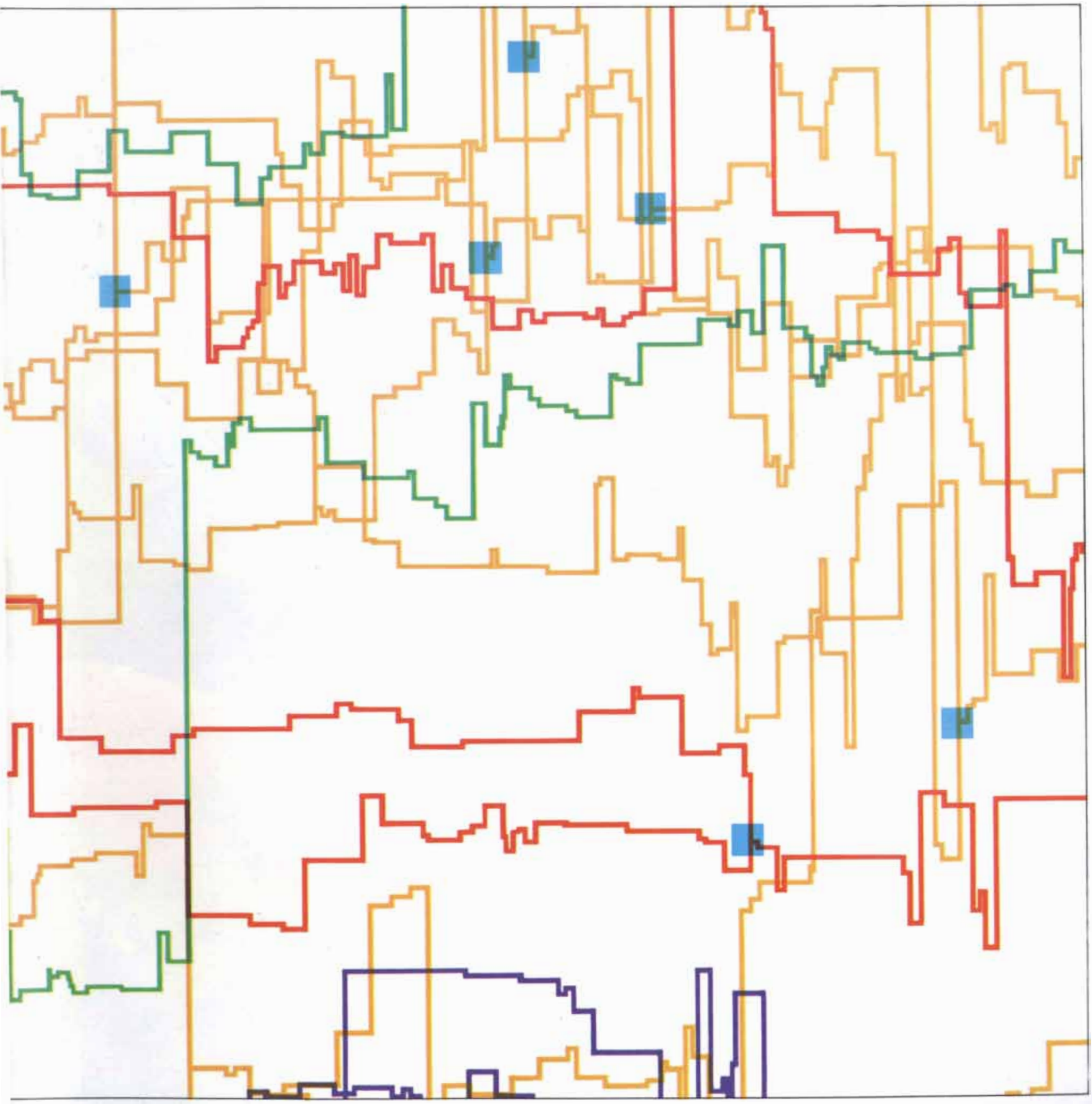
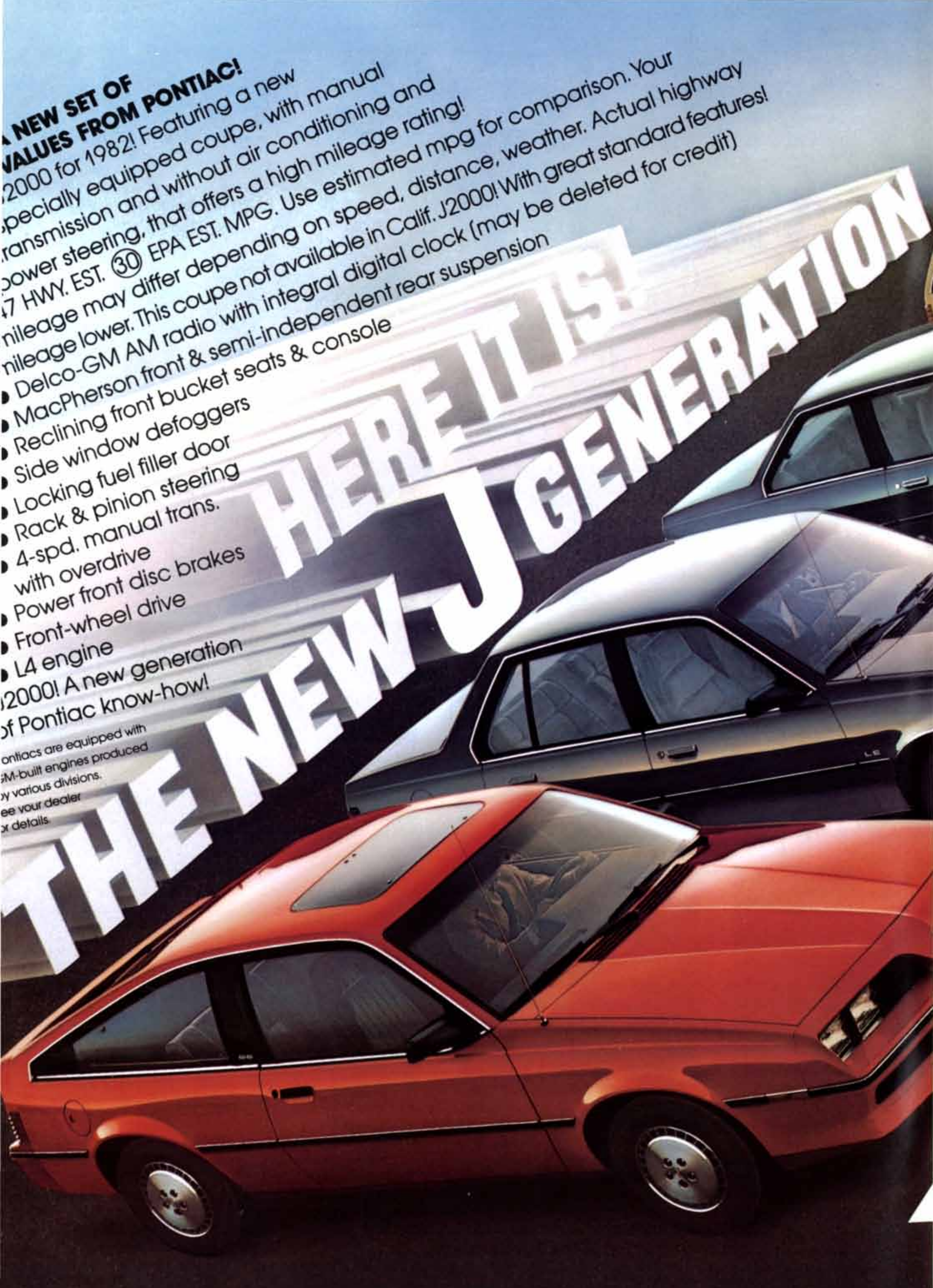
# SCIENTIFIC AMERICAN

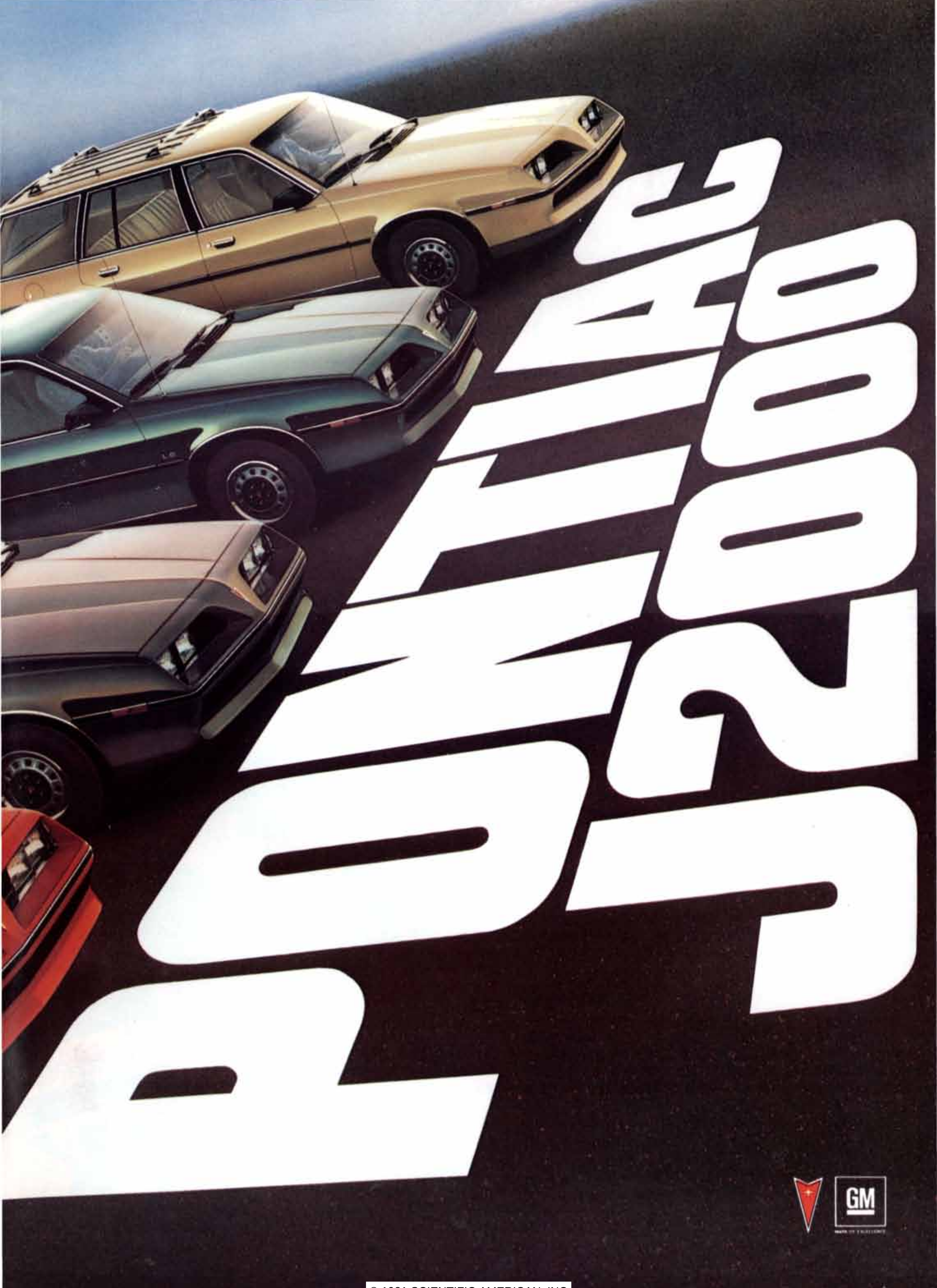


ACCRETION OF PLANETS

$2.00

*June 1981*


NEW SET OF
VALUES FROM PONTIAC!
2000 for 1982! Featuring a new
pecially equipped coupe, with manual
ransmission and without air conditioning and
power steering, that offers a high mileage rating!
7 HWY. EST. 30 EPA EST. MPG. Use estimated mpg for comparison. Your
nileage may differ depending on speed, distance, weather. Actual highway
nileage lower. This coupe not available in Calif. J2000! With great standard features!
Delco-GM AM radio with integral digital clock (may be deleted for credit)
MacPherson front & semi-independent rear suspension
Reclining front bucket seats & console
Side window defoggers
Locking fuel filler door
Rack & pinion steering
4-spd. manual trans. with overdrive
Power front disc brakes
Front-wheel drive
L4 engine
J2000! A new generation of Pontiac know-how!
ontiacs are equipped with
GM-built engines produced
by various divisions.
ee your dealer
or details.
HERE IT IS!
THE NEW J GENERATION


PONTIAC
GM

86 PROOF BLENDED SCOTCH WHISKY DISTILLED AND BOTTLED IN SCOTLAND. IMPORTED BY THE BUCKINGHAM CORPORATION, NEW YORK, N.Y
HERE'S TO GUT FEELINGS AND THOSE WHO STILL FOLLOW THEM.
Ted Turner does lots of things people advise him not to do. And he succeeds at them. He turned Atlanta's WTBS-TV into a "Superstation" using a communications satellite and recently founded Cable News Network, the world's first 24-hour TV news network. He bought the Atlanta Braves and moved them out of last place; won the 1977 America's Cup after being fired in the '74 races; and was named "Yachtsman of the Year" four times.
Ted Turner puts his feelings where his mouth is. He also puts a great scotch there: Cutty Sark. And while he's been called Captain Outrageous by some, one thing's sure: Ted Turner's enjoying himself.
CNN
Ted "Captain Outrageous" Turner
Lettick
The Scotch with a following of leaders.
CUTTY SARK®

Established 1845 SCIENTIFIC AMERICAN June 1981 Volume 244 Number 6

ARTICLES

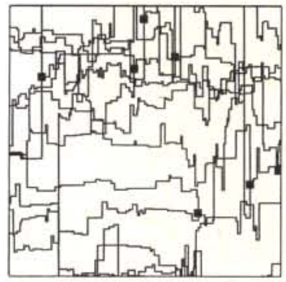


DEPARTMENTS





SCIENTIFIC AMERICAN (ISSN 0036-8733), PUBLISHED MONTHLY BY SCIENTIFIC AMERICAN, INC., 415 MADISON AVENUE, NEW YORK, N.Y. 10017. COPYRIGHT © 1981 BY SCIENTIFIC AMERICAN, INC. ALL RIGHTS RESERVED. PRINTED IN THE U.S.A. NO PART OF THIS ISSUE MAY BE REPRODUCED BY ANY MECHANICAL, PHOTOGRAPHIC OR ELECTRONIC PROCESS, OR IN THE FORM OF A PHONOGRAPHIC RECORDING, NOR MAY IT BE STORED IN A RETRIEVAL SYSTEM, TRANSMITTED OR OTHERWISE COPIED FOR PUBLIC OR PRIVATE USE WITHOUT WRITTEN PERMISSION OF THE PUBLISHER. CONTROLLED CIRCULATION POSTAGE PAID AT NEW YORK, N.Y. (USPS 509-530), AND AT CHICAGO, ILL. (USPS 508-030). AUTHORIZED AS SECOND-CLASS MAIL BY THE POST OFFICE DEPARTMENT, OTTAWA, CANADA, AND FOR PAYMENT OF POSTAGE IN CASH. SUBSCRIPTION RATE: $21 PER YEAR, U.S., ITS POSSESSIONS AND CANADA; $27 PER YEAR, ALL OTHER COUNTRIES. POSTMASTER: SEND ADDRESS CHANGES TO SCIENTIFIC AMERICAN, 415 MADISON AVENUE, NEW YORK, N.Y. 10017.



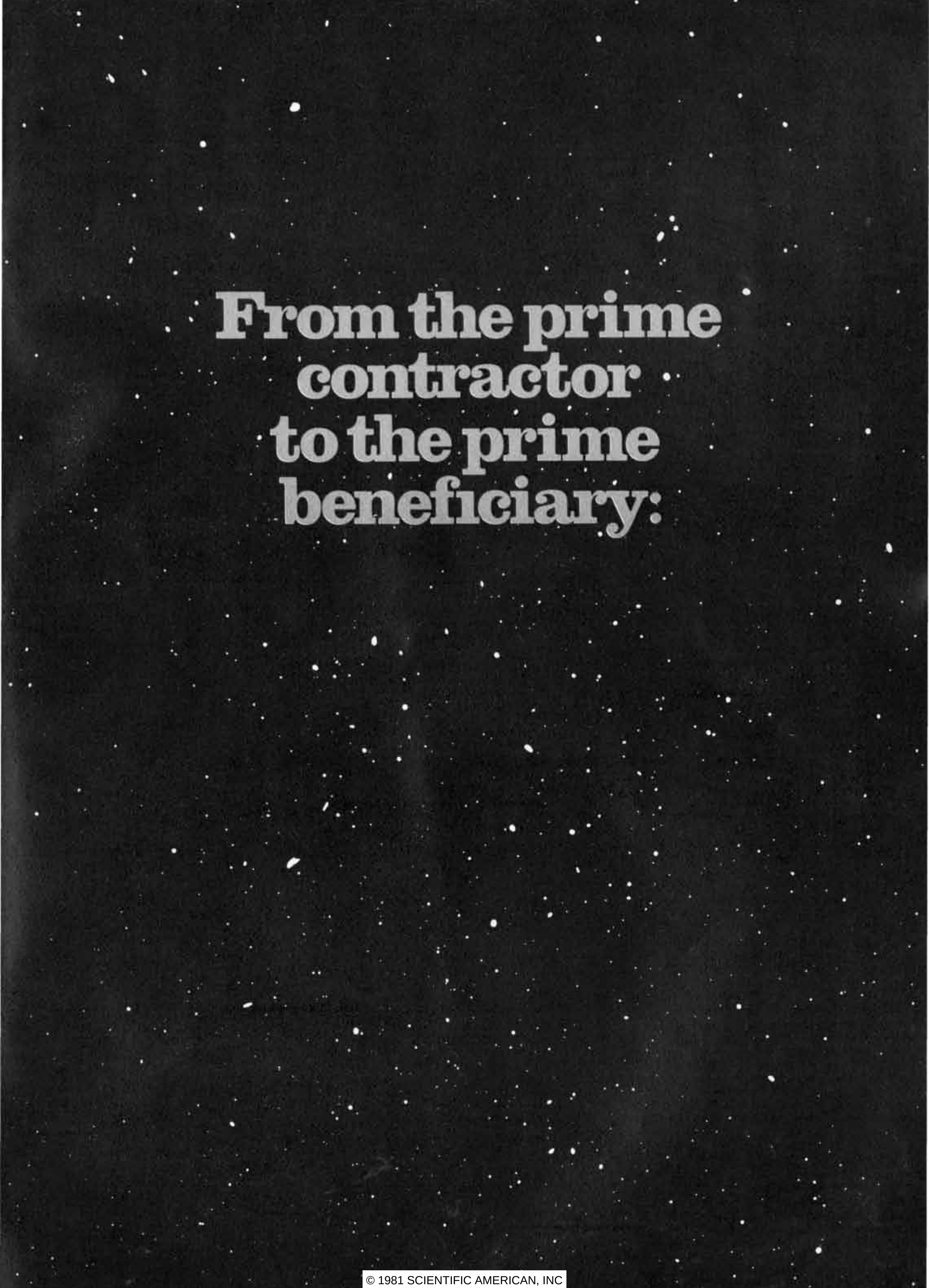
## THE COVER

The illustration on the cover shows part of a computer simulation in which collisions of bodies in elliptical orbits around the sun lead to the accretion of planets of the same rocky composition as Mercury, Venus, the earth and Mars (see "The Formation of the Earth from Planetesimals," by George W. Wetherill, page 162). Each path from left to right displays the career of one of the bodies. Specifically, the illustration plots the semimajor axis of each ellipse, a quantity that measures the orbital distance of the body from the sun; the vertical excursions in each path represent changes in that distance. The changes are caused by collisions and by near-misses with another body. Each blue square marks a collision, from which a single, more massive body emerges. The paths in yellow represent bodies that ultimately become part of a single planet much like the earth in both its size and its position in the solar system. The other colors represent bodies that will merge into three other rocky planets. The simulation was made by Larry P. Cox of the Massachusetts Institute of Technology.

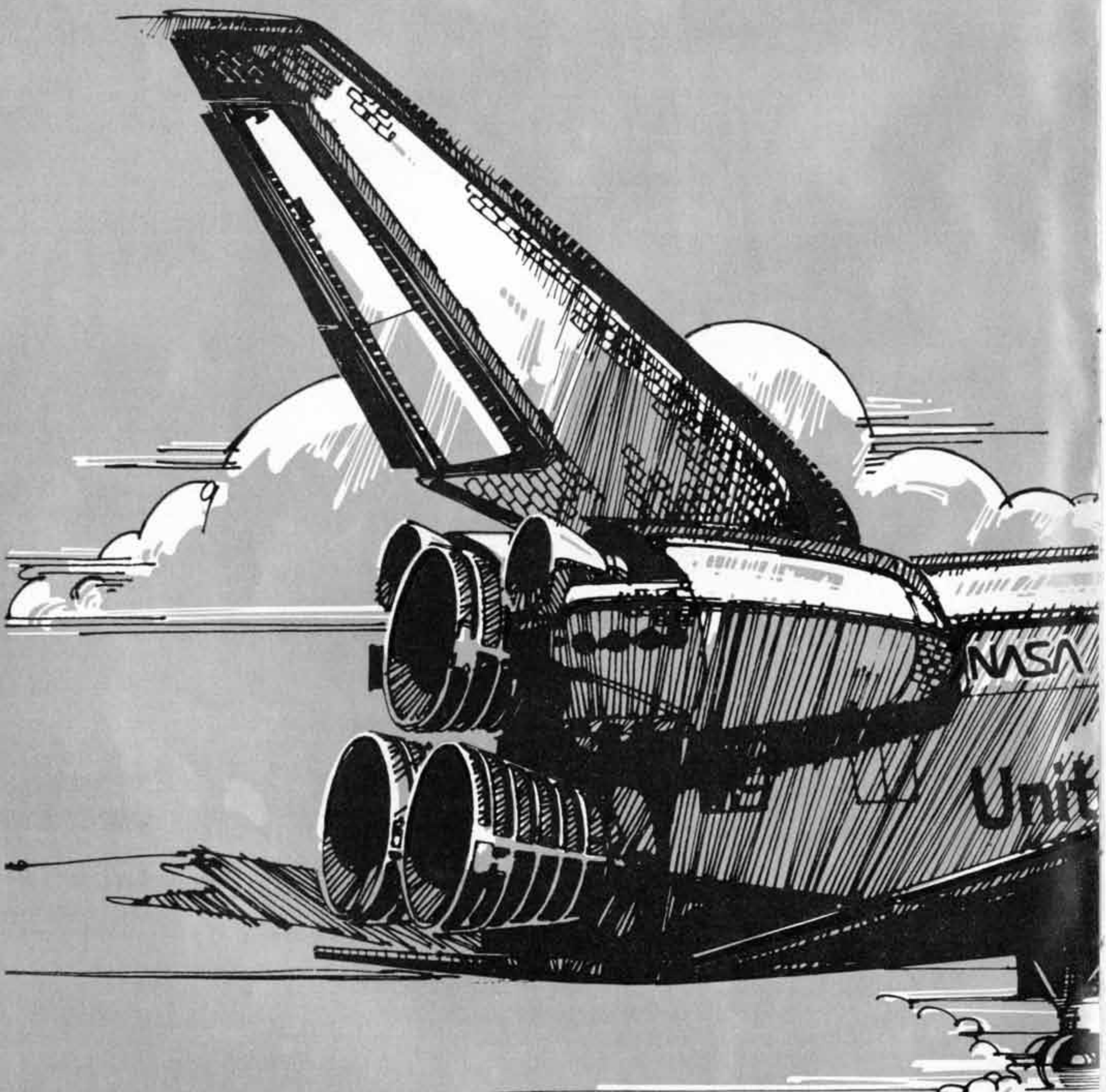
## THE ILLUSTRATIONS

Cover illustration by Allen Beechel

| Page | Source | Page | Source |
|---|---|---|---|
| 22–24 | Dana Nordhausen | 114 | Alexander J. B. Zehnder, Swiss Federal Institute of Technology (*top*); Frank Mayer, University of Göttingen (*bottom*) |
| 29–32 | Ilil Arbel | 116–121 | Alan D. Iselin |
| 56 | George V. Kelvin | 127–142 | Jerome Kuhl |
| 57 | General Atomic Company | 146 | Bernd Heinrich, University of Vermont |
| 58–59 | George V. Kelvin | 148–156 | Tom Prentiss |
| 60–61 | General Atomic Company | 158 | Tom Prentiss (*top*); Bernd Heinrich, University of Vermont (*bottom*) |
| 62–63 | George V. Kelvin | 162 | Larry P. Cox, Massachusetts Institute of Technology |
| 65 | Ralph Morse | 164–174 | Allen Beechel |
| 66 | Hans W. J. Courant, University of Minnesota | 177 | Noël H. Gale, University of Oxford |
| 68–75 | Gabor Kiss | 182–184 | Patricia J. Wynne |
| 77 | Milton R. J. Salton, New York University School of Medicine | 185–186 | Ashmolean Museum |
| 78–81 | Ilil Arbel | 188–191 | Ilil Arbel |
| 82 | Ilil Arbel (*top*); Brian G. Spratt, University of Sussex (*bottom*) | 192 | Noël H. Gale, University of Oxford |
| 83–85 | Ilil Arbel | 196 | Jearl Walker |
| 86 | Royal Society of London | 197–200 | Michael Goodman |
| 99–112 | Alan D. Iselin | | |


From the prime
contractor
to the prime
beneficiary:

# Congratulati



# ons, America.

Once again, America is leading the world into space.

Rockwell International is prime contractor for the Shuttle orbiter. Also, our Rocketdyne Division built the main engines. And we assist NASA in the integration of the Space Transportation System. Our achievements in space and aircraft development demonstrate the high technology which characterizes all the businesses of Rockwell International.

We join America in saluting NASA, the *Columbia* crew — John W. Young and Robert L. Crippen — and the 50,000 people in many companies who worked with us to build America's Space Shuttle.

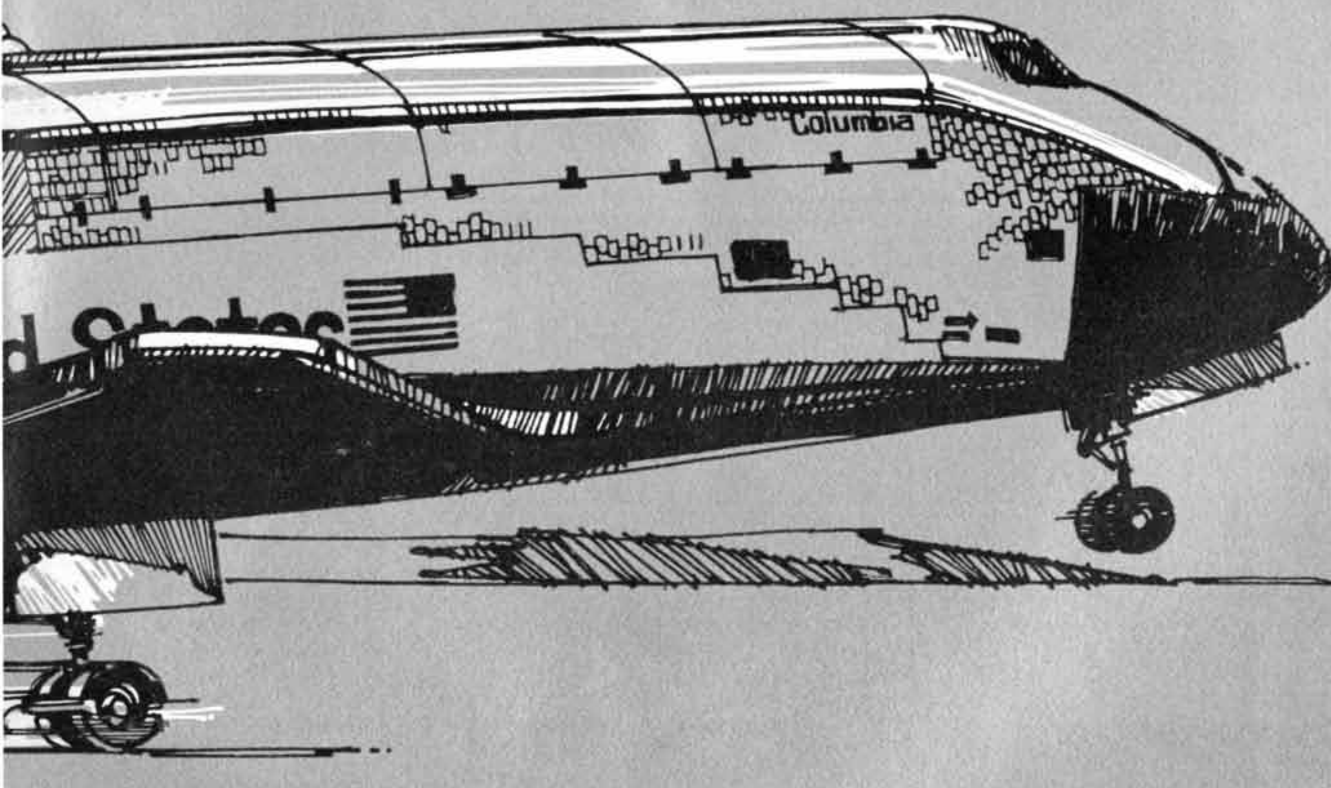
Congratulations, America. Through the Shuttle, designed for repeated flights into space, you have built a technology bridge to the benefits of this vast new frontier. It is a uniquely American achievement.

Good old American "know-how" is alive and well.

**Rockwell International**

...where science gets down to business

Automotive / Aerospace
Electronics / General Industries



# LETTERS

Sirs:

"The Origin of Genetic Information," by Manfred Eigen, William Gardiner, Peter Schuster and Ruthild Winkler-Oswatitsch [SCIENTIFIC AMERICAN, April], contains many illuminating ideas about the mechanisms of prebiotic evolution. I wish only to question one statement in the article, where the authors are discussing the compartmentation of life into cells. They say: "Cellular organization was needed because it was the only way to solve the one problem of information processing in evolution that self-replicative competition and hypercyclic cooperation were unable to address: the evaluation of the information in genetic messages." In other words, when a primitive gene is catalyzing the synthesis of a primitive protein, Darwinian selection cannot pick out and reward the superior gene for producing a superior protein unless the protein and the gene remain physically attached to each other. If the protein is free to float away from its parent gene into the primeval soup, the quality of the protein cannot influence the chance of survival of the gene.

This argument is faulty in only one respect. It is true that primitive genes and proteins must remain physically attached if the selection of genes is to be based on the performance of their offspring. But it is not true that cellular organization is the only way to attach a gene to its product. There are two ways in which a gene and its product may be topologically linked. One way is to enclose both molecules in a closed two-dimensional structure, i.e., a primitive cell membrane. The other way is to let both the gene and its product be closed one-dimensional loops. If a cyclic gene and a cyclic protein are linked to each other when the protein is synthesized, they will remain linked as long as they both survive.

I am not saying that cyclic RNA and cyclic proteins linked together were necessarily the first symbiotic structures. It is equally possible that cellular structures came first. But it is worthwhile at least to examine the hypothesis that the first great Darwinian struggle was fought between linked-cycle structures and cellular structures. It may have happened in this case, as in other cases subsequently, that the later arrivals were the winners. It may also have happened in this case, as in other cases, that the vanquished structures were not annihilated but survived as organelles or appendages of the victors.

We see in today's organisms many examples of cyclic RNA and DNA serving as carriers of genetic information. Cyclic proteins occur much more rarely. It is possible that some at least of the cyclic RNA may be a vestige of an early time when linked cycles were an essential step toward life's origin.

FREEMAN J. DYSON

Institute for Advanced Study
Princeton, N.J.

Sirs:

We entirely agree with Professor Dyson's view. The formulation in our article "cellular organization was... the only way" is indeed misleading. We have always emphasized that compartmentation may be realized in many different ways: specific structural linkage between genotype and phenotype, or by enclosing both into vesicular compartments, or even more simply in an inhomogeneous spatial distribution such as those distributions present in dissipative structures. Cellular organization was only the ultimate way nature found to evaluate genetic information. The initial compartmentation, in which gene duplication and compartment division were unsynchronized, may have taken a quite different form. We are grateful to Professor Dyson for emphasizing this point.

MANFRED EIGEN

WILLIAM GARDINER

PETER SCHUSTER

RUTHILD WINKLER-OSWATITSCH

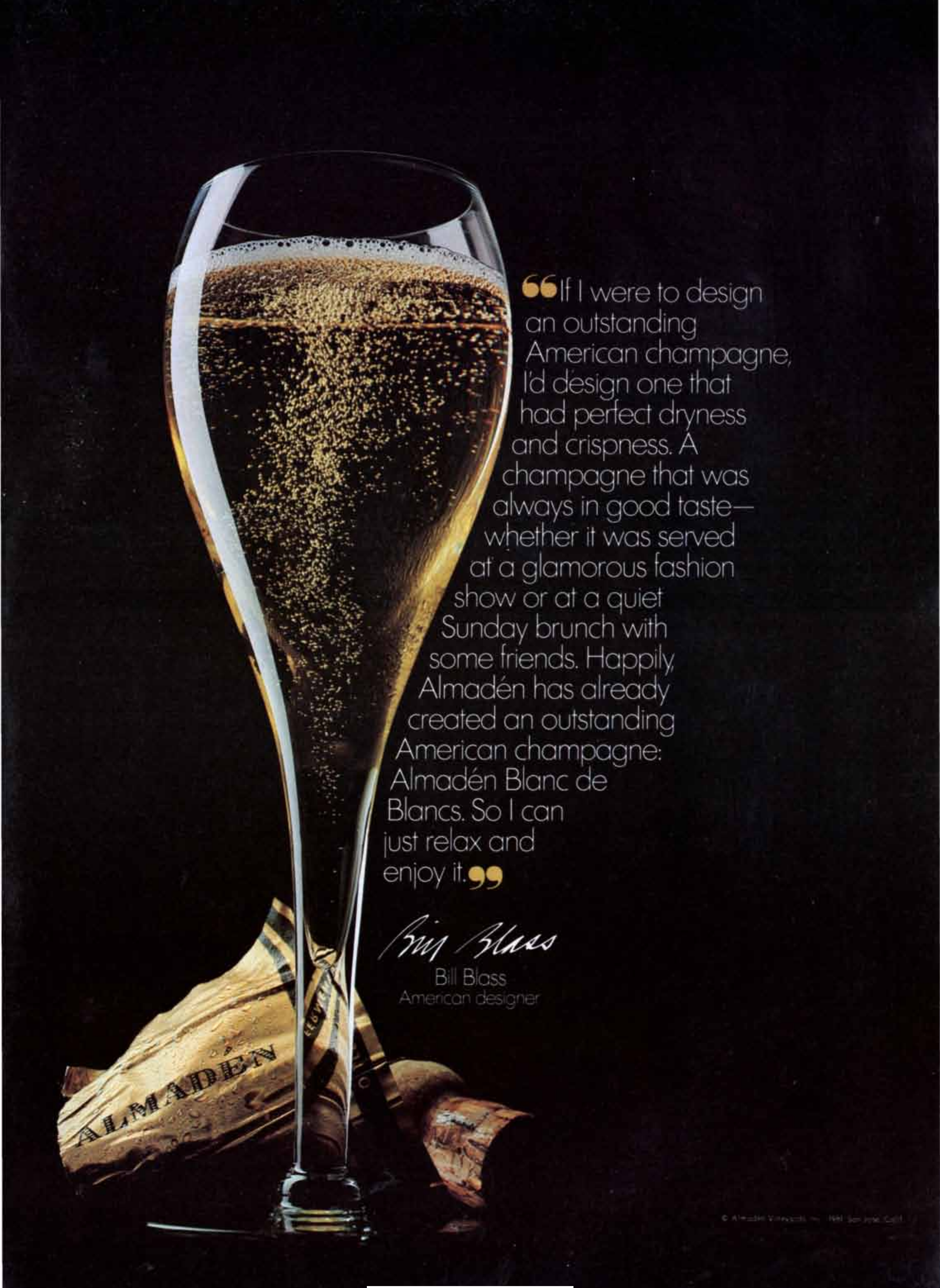
Max Planck Institute
for Biophysical Chemistry
Göttingen

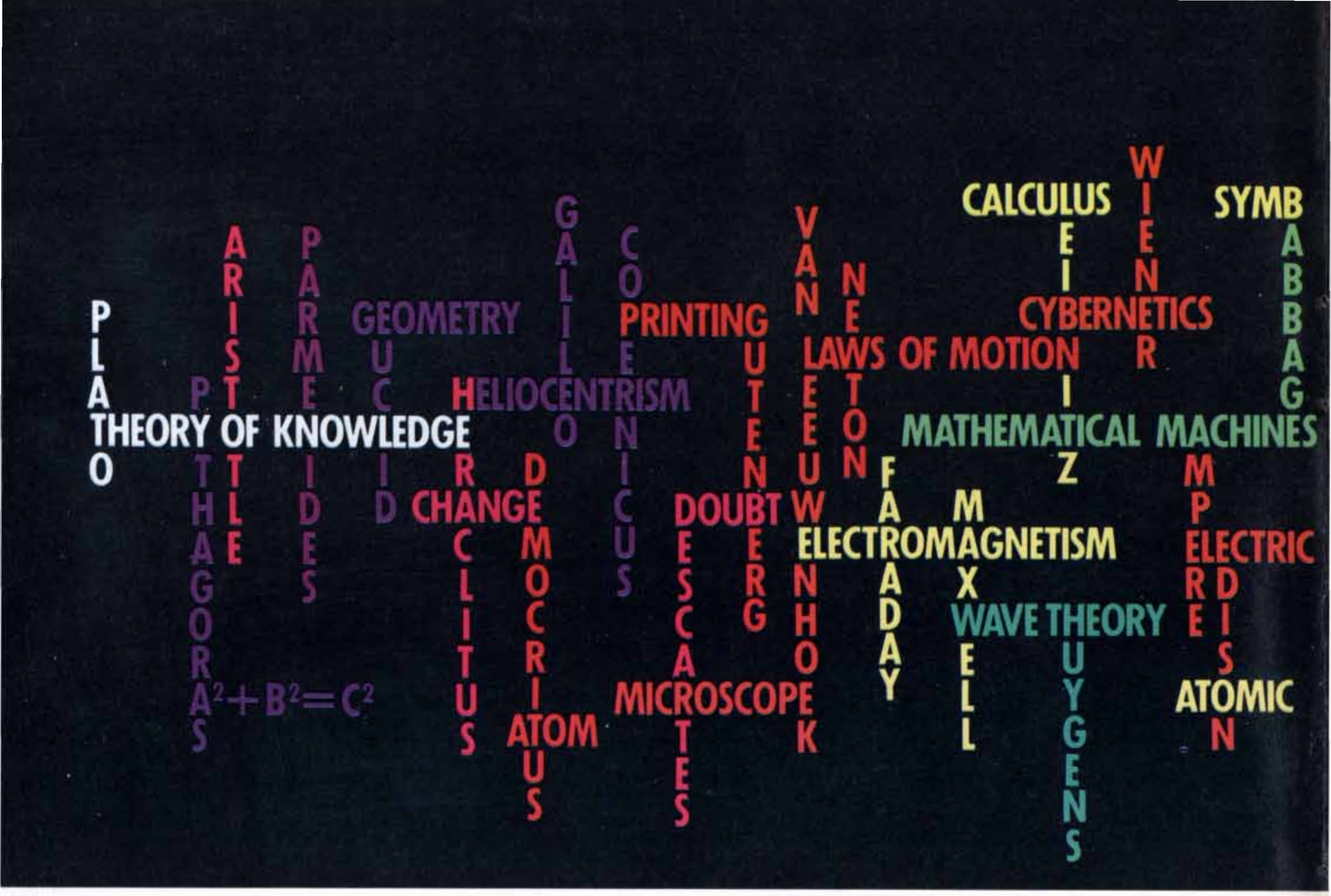
Sirs:

I should like to comment on the letter from Graham F. Lee in your January issue on "backward English."

I suggest that the "reverse English" mentioned by Mr. Lee as having been prevalent in Sydney in the 1920's is the English "back slang," or something resembling it. The difference is that Andrew Levine in the original "Science and the Citizen" article apparently articulates the sounds in reverse order, whereas in back slang the words are deliberately deformed to make them easier to say, and perhaps also to make them less intelligible to outsiders.

I was brought up in a poor district of inner London in the early 1920's and I can well remember back slang being spoken by the costermongers (street-market traders in fruit and vegetables). Some traces still remain today, for example "yob," now used to mean a young lout. These people were able to speak back slang so fluently that it was completely unintelligible to the ordinary person, although I should add that it

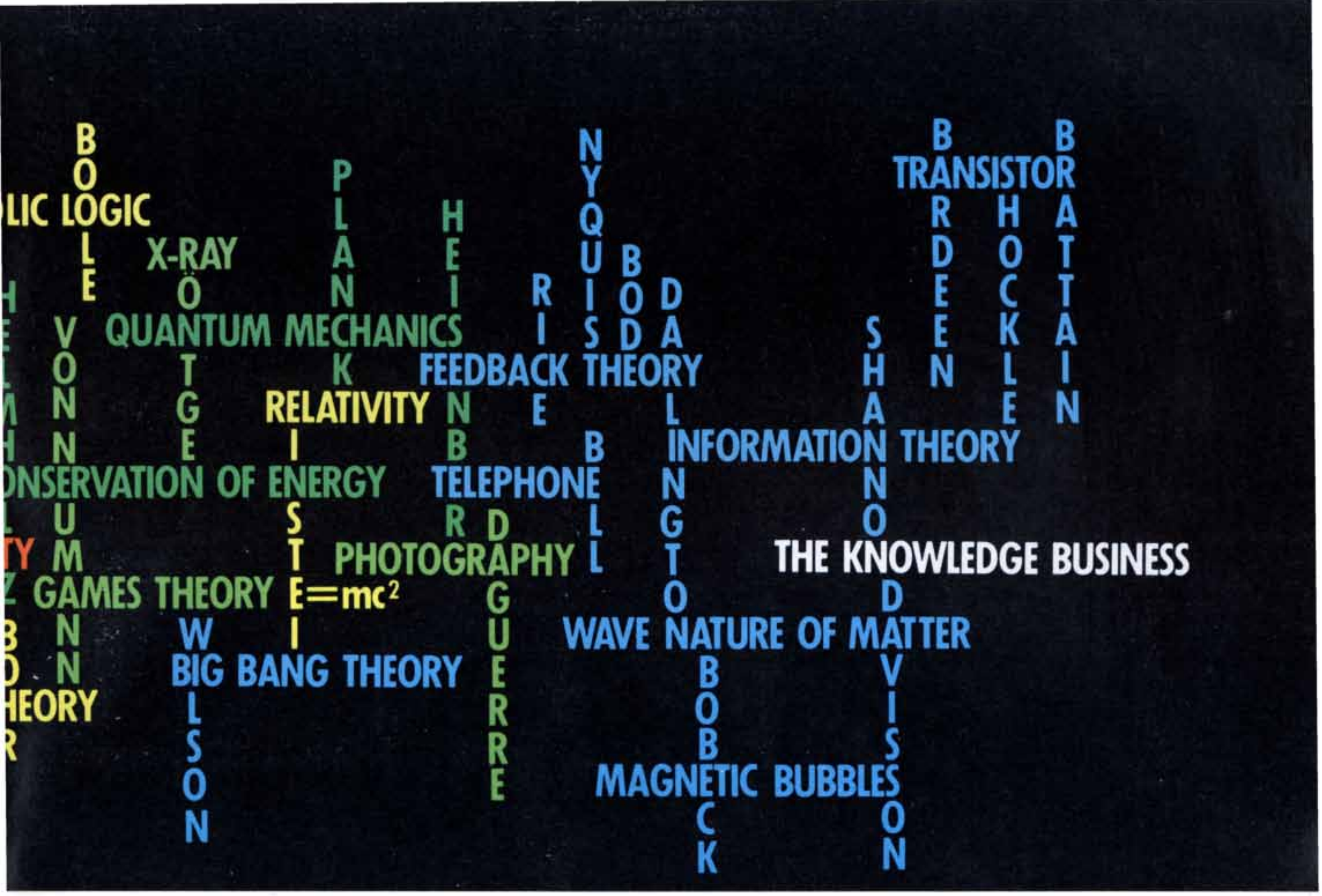

"If I were to design an outstanding American champagne, I'd design one that had perfect dryness and crispness. A champagne that was always in good taste—whether it was served at a glamorous fashion show or at a quiet Sunday brunch with some friends. Happily, Almadén has already created an outstanding American champagne: Almadén Blanc de Blancs. So I can just relax and enjoy it."
Bill Blass
Bill Blass
American designer
ALMADÉN




# Enter the information age.

From the beginning, knowledge has shaped the future.

Now, America is entering a new age, the Information Age. Bell scientists have used their knowledge to invent the tools for managing information in this new age: the transistor, magnetic bubbles and other breakthroughs shown above in blue. We apply this knowledge to advanced systems for business by drawing on our knowledge of information networks and of each industry's specific information management needs.




This knowledge can be applied to your business with the help of your local Bell Account Executive, an industry specialist, to produce a single, integrated system. This unique system can add to your profits by increasing productivity, saving energy, improving quality of life for your employees.

A typical example is our Dimension® PBX system for managing information in hospitals. In addition to updating voice communications, it distributes timely information on patient bills, medical records, prescribed diets. It reports bed status, staff availability, department inventories. A special feature transmits energy-saving information to heating, cooling and lighting units, reducing energy costs as much as 20 percent.

The system we design for you will be part of our nationwide network, the largest and most advanced system in the world for managing information. You can make your own changes as your business changes and grows. Your system will be flexible, to accommodate further advances in technology; it will be backed by and integrated with the Bell System.

The knowledge business can help you move into the Information Age.

## The knowledge business

was mainly the older costers who were fluent, and it was obviously dying out among the younger generation.

There is a quite detailed account of the use of back slang in London in the 1850's in Henry Mayhew's *London Labour and the London Poor,* which I can recommend as a completely fascinating, and I believe unique, social document. A typical example where the word is not simply reversed is "exisyenep" for "sixpence." Some changes clearly occurred with time. For instance, Mayhew gives the word for "police" as "esclop," whereas in my time the everyday form was "slop." In view of the large amount of emigration from the U.K. to Australia in the 19th and early 20th centuries, it does not seem unreasonable to suppose back slang was carried over with the emigrants.

Incidentally, I am often amazed at the way old slang words will suddenly appear in unexpected places with renewed vitality. I had believed the word "lush" for an alcoholic was modern American, but it is not. Mayhew quotes it as having been in common use in 1851 with the same meaning. Even stranger, perhaps, is "raspberry," which is actually London rhyming slang, having been shortened from "raspberry tart."

L. E. WEAVER

St. Leonards-on-Sea
East Sussex, England

**Editorial correspondence** should be addressed to The Editors, SCIENTIFIC AMERICAN, 415 Madison Avenue, New York, N.Y. 10017. Manuscripts are submitted at the author's risk and will not be returned unless accompanied by postage.

**Advertising correspondence** should be addressed to C. John Kirby, Advertising Director, SCIENTIFIC AMERICAN, 415 Madison Avenue, New York, N.Y. 10017.

**Offprint correspondence** and orders should be addressed to W. H. Freeman and Company, 660 Market Street, San Francisco, Calif. 94104. For each offprint ordered please enclose 60 cents.

**Subscription correspondence** should be addressed to Subscription Manager, SCIENTIFIC AMERICAN, 415 Madison Avenue, New York, N.Y. 10017. For change of address, notify us at least four weeks in advance. Send both old and new addresses and enclose an address imprint from a recent issue. (Date of last issue on your subscription is shown at upper right-hand corner of each month's mailing label.)

Name ______________________

New Address ______________________

______________________

______________________

Old Address ______________________

______________________

______________________

# Increase Your Productivity With Radio Shack's TRS-80® Model III Computer.

Radio Shack's TRS-80 Model III desktop computer is the key to getting more things done in much less time. It's a powerful tool that's ready to help with the professional's most demanding tasks.

**A Highly Versatile Problem Solver.** Model III has the memory capacity, disk storage and programming flexibility you need for many engineering, scientific and statistical applications. The powerful Model III BASIC language lets you design your own unique programs, and includes machine language programming, a program line editor, and much more.

Of course, Radio Shack also has ready-to-run software. For example, our VisiCalc™ planning and modeling program can help you set up and solve design equations and other complex computations.

**Use Your Time More Efficiently.** You'll have more time to use Model III for planning and problem solving when you also use it for your routine tasks. Let Model III keep your records, as well as store and analyze data with our Profile software. Or add an optional printer and our Scripsit™ Word Processing program for fast, error-free memos and documentation.

**Judge Their Extras by Our Standards.** You don't have to pay more for the features you want. Our $2495 Model III Desktop Computer System includes 32K RAM (expandable to 48K), Model III BASIC in 14K ROM, and two built-in 175K double-density disk drives. You get a full 64-character by 16-line screen format, with both upper and lower case characters. A 65-key keyboard with 12-key datapad. A 12" high-resolution monitor. An RS-232 serial interface *and* parallel printer interface. All are standard features — not expensive extras.

**Supported Coast-to-Coast.** When you're ready to expand your Model III with more disk storage, a modem for data communications, a printer, or other peripherals, Radio Shack is ready. With 6100 stores and dealers, 130 Computer Centers and 135 service centers.

Improve your productivity — get a personal demonstration of the TRS-80 Model III today, or write for more information.

# 50 AND 100 YEARS AGO

SCIENTIFIC AMERICAN

JUNE, 1931: "Aspiration to duplicate in the laboratory the powerful radiations that hitherto have been produced only by nature promises to lead to further discoveries by a trio of Carnegie Institution of Washington physicists. At present the vacuum tubes developed by M. A. Tuve, L. R. Hafstad and Odd Dahl have operated reliably at potentials up to two million volts. This potential produces the most powerful radiations yet made by man: artificial beta rays, which are high-speed electrons, and artificial gamma rays. Tuve and his associates are now at work on a new method of building up voltages. Thirty million volts are possible with a modification of what is called the Faraday cage. By using these voltages to accelerate the positively charged cores of hydrogen atoms these scientists will have by far the most powerful projectiles ever available to human beings, resembling the alpha rays from radium. They expect the artificial radiations to be extremely useful in probing deeper the fundamental structure of matter."

"One of the difficulties in television has been to find a wide enough wave band. Until a solution is found it will not be possible to transmit images containing much detail. Where television in theaters and other places of entertainment is concerned, however, this problem does not arise, as the television program can be transmitted over telephone lines. In the Baird laboratories in England a demonstration of such television was recently given. The apparatus used in the Baird zone televisor consists, at the transmitting end, of a large mirror drum with 30 mirrors arranged around its periphery, each mirror being tilted slightly differently from the next. The images cast by the mirror drum as it revolves pass over three apertures so arranged that each aperture explores one third of each image. Behind these apertures are three photoelectric cells, connected to the necessary amplifying equipment. At the receiving station the currents are received by three neon tubes. The light from them is focused by means of a lens upon a revolving drum, similar to that at the transmitting station, the light being thus reflected upon a translucent screen."

"The problem of drug addiction concerns the United States more than any other country. Our per capita use of narcotics, according to Dr. Walter E. Dixon, is 245 milligrams, as compared with 150 for Europe, eight for Asia and 24.5 for Africa. Numerous studies have been made as to the causes of narcotic addiction. The latest figures supplied by Dr. W. L. Treadway of the United States Public Health Service indicate that the main causes are previous use of drugs in medical treatment, self treatment for relief of pain, recourse to drugs during emotional distress, the influence of other addicts and the indulgence in drugs for the sake of curiosity, thrill or bravado. The National Research Council is attempting to solve the problem of narcotic addiction by promoting research with a view to finding some synthetic preparation derived from opium that will have all of the qualities of opium and morphine without their tendency to establish drug addiction."

SCIENTIFIC AMERICAN

JUNE, 1881: "In *Comptes Rendus de l'Académie des Sciences* M. Pasteur writes: 'Of the various results that I have had the honor of communicating to the Academy concerning the disease commonly called chicken cholera, I shall take the liberty of recalling the following: (1) Chicken cholera is a virulent disease of the highest order. (2) The virus is a microscopical parasite, which may be multiplied by cultivation outside of the body of an animal. (3) The virus presents various degrees of virulence. Sometimes the disease is followed by death; at other times, after giving rise to morbid symptoms of variable intensity, these are followed by cure. (4) The differences noted in the power of various viruses are not merely the results of observations made of natural phenomena, where the experimenter can give rise to them at will. (5) As generally happens with virulent diseases, chicken cholera does not recidivate, or rather it may be said that recidivation is in inverse ratio to the intensity of the first attack of the disease, and it is always possible to carry the preservative action far enough to prevent the most virulent virus from producing any effect. (6) Without wishing to affirm anything at this time on the relations of the virus of smallpox to that of vaccine, it appears from the foregoing that in the chicken cholera there are conditions of the virus which, relative to the most virulent virus, act as a human vaccine to the virus of smallpox.'"

"The researches of M. Gaston Planté on the polarization of voltameters led to his invention of the secondary cell, composed of two strips of lead immersed in acidulated water. These cells accumulate and, so to speak, store up the electricity passed into them from some outside generator. When the two electrodes are connected with any source of electricity, the surfaces of the two strips of lead undergo certain modifications. Thus the positive pole retains oxygen and becomes covered with a thin coating of peroxide of lead, while the negative pole becomes reduced to a clean metallic state. Now, if the secondary cell is separated from the primary one, we have a veritable voltaic battery, for the symmetry of the poles is upset and one is ready to give up oxygen and the other eager to receive it. When the poles are connected, an intense electric current is obtained, but it is of short duration. This difficulty has now been overcome by M. Faure by substituting for strips of lead masses of spongy lead. A battery composed of Faure's cells and weighing 150 lb. is capable of storing up a quantity of electricity equivalent to one horsepower during one hour."

"Captain W. de W. Abney, R.E., F.R.S., and Lieutenant-Colonel Festing, R.E., describe the influence of the molecular grouping in organic bodies on their absorption in the infra-red region of the spectrum. The source of light for obtaining a continuous spectrum was the incandescent positive pole of an electric arc. The light was passed through tubes containing the fluid, and the absorption spectra were photographed in the infra-red region. In the abstract of a paper read before the Royal Society the authors point out that each radical has its own definite absorption in the infra-red, and that such a radical can be detected in a more complex body."

"We cannot tell whether electricity is some peculiar kind of substance or some modification or motion of ordinary matter. No phenomena have thus far been discovered that absolutely negative the notion that electricity may be a subtle, imponderable fluid. Another view, however, seems to carry a greater weight of opinion in its favor—that, namely, of Maxwell. He regards an electric charge as the establishment of a peculiar state of strain among the atoms of the charged bodies and in the medium between them. A discharge consists in the sudden relief of this strain by a giving way of the intervening medium, without necessarily implying any transfer of substance through it. In its application the theory is mathematically difficult, but it opens the way for the establishment of relations between electricity and the other physical agents, especially light and heat. There is certainly great probability that some hypothesis will yet be found that will include in one general theory all the physical agents—light, heat, gravity and chemical affinity, as well as electricity and magnetism. But the hour and the man have not yet come."


# JETTA: THE GREATEST COLLECTION OF VOLKSWAGENS EVER ASSEMBLED UNDER ONE ROOF.

Want us to run that by again? It's really not as crazy as it sounds. Because the Jetta just happens to combine the best features of every other Volkswagen we make.

On the road, for example, the Jetta handles like a Scirocco.

Which is only natural. Since the Jetta *is* like a Scirocco.

With rack-and-pinion steering, front-wheel drive, independent 4-wheel suspension and a powerful CIS fuel-injected engine.

Inside, the Jetta is like our elegant Dasher.

With a luxurious amount of room, and a surprising amount of luxury.

And at the gas pump?

You guessed it.

The Jetta's like a Rabbit.

In fact, it gets an EPA estimated [25] mpg, 40 mpg highway estimate. (Use "estimated mpg" for comparisons. Your mileage may vary with weather, speed and trip length. Actual highway mileage will probably be less.)

Of course, the best thing about the Jetta isn't only that it's like a Scirocco, a Dasher or a Rabbit. The best thing about a Jetta is that it's also like, well, a Jetta.

With sleek European styling, sophisticated German engineering, and a trunk so well designed it can easily hold eight suitcases.

When you consider everything a Jetta has to offer, it's no wonder so many people are adding one to their car collections.

After all, how often can you acquire an entire collection in one car?

## ALL OUR BEST IDEAS IN ONE CAR. JETTA.


COURVOISIER
COGNAC
Admiré pour
sa perfection,
sa qualité a
conquis le
monde entier
VSOP
Fine Champagne
COURVOISIER.
The Brandy of Napoleon
COGNAC
The Brandy of Napoleon

# THE AUTHORS

HAROLD M. AGNEW ("Gas-cooled Nuclear Power Reactors") is president of the General Atomic Company, which he joined in 1979 after more than eight years as director of the Los Alamos Scientific Laboratory. His work on the development of nuclear energy dates from the early 1940's, when as a recent graduate of the University of Denver he joined the small group that worked with Enrico Fermi on the first nuclear-fission chain reaction. In 1943 Agnew joined the Los Alamos laboratory to participate in the development of the atomic bomb. From 1946 to 1949 he was again with Fermi at the University of Chicago, obtaining his Ph.D. there in 1949. Thereafter Agnew was at Los Alamos except for three years (1961–64) as scientific adviser to the Supreme Allied Commander in Europe.

STEVEN WEINBERG ("The Decay of the Proton") is Higgins Professor of Physics at Harvard University and senior scientist at the Smithsonian Astrophysical Observatory. He is currently visiting professor at the University of Texas at Austin. He did his undergraduate work at Cornell University, being graduated in 1954, studied for the next year at the Niels Bohr Institute in Copenhagen and received his Ph.D. from Princeton University in 1957. Thereafter he worked at Columbia University, the Lawrence Radiation Laboratory of the University of California, the University of California at Berkeley and the Massachusetts Institute of Technology before going to Harvard in 1973. He has written two books: *The First Three Minutes: A Modern View of the Origin of the Universe* and *Gravitation and Cosmology: Principles and Applications of the General Theory of Relativity.* He has received numerous honors for his work on the theory of elementary particles, including five honorary degrees and the 1979 Nobel prize in physics, which he shared with Sheldon Lee Glashow and Abdus Salam. Weinberg was recently elected a foreign member of the Royal Society.

E. P. ABRAHAM ("The Beta-Lactam Antibiotics") is professor emeritus of chemical pathology at the Sir William Dunn School of Pathology of the University of Oxford. He did his undergraduate work at Oxford and obtained his D.Phil. there in 1938. After two years as a Rockefeller Foundation traveling fellow in Stockholm he returned to Oxford to work on the isolation and chemistry of penicillin with Howard W. Florey, Ernst B. Chain and others. In 1940 he and Chain discovered the enzyme penicillinase; in 1953 he and G. G. F. Newton isolated cephalosporin *C* from an impure preparation of penicillin *N.* Subsequent work at Oxford and in pharmaceutical companies led to the introduction of cephalosporins into medicine. Abraham, a Fellow of the Royal Society, was knighted last year. His leisure interests include walking and skiing.

CARL R. WOESE ("Archaebacteria") is professor of microbiology and of genetics and development at the University of Illinois at Urbana-Champaign; from 1972 to 1979 he held a third appointment as professor of biophysics. His bachelor's degree (in mathematics and physics) was awarded by Amherst College in 1950 and his Ph.D. (in biophysics) by Yale University in 1953. He then spent two years at the University of Rochester School of Medicine and Dentistry before returning to Yale to do research in biophysics. From 1960 to 1963 he worked as a biophysicist at the General Electric Research Laboratory, joining the faculty of the University of Illinois in 1964. "My entire career," he writes, "has been a deepening venture into the recesses of evolution." Much of his work has been on the evolution of the mechanism whereby the genetic code is translated in the cell by the ribosomes. Now, Woese says, "it is time to press deeper, and my interest is turning to the evolution of the ribosome itself."

ROBERT G. BLAND ("The Allocation of Resources by Linear Programming") is assistant professor in the School of Operations Research and Industrial Engineering and the Center for Applied Mathematics at Cornell University. "I studied at Cornell," he writes, "and got my Ph.D. in operations research there in 1974. I was assistant professor of mathematical sciences at the State University of New York at Binghamton, research fellow at the Center for Operations Research and Econometrics in Louvain and professor of management at the European Institute for Advanced Studies in Management in Brussels before returning to Cornell in 1978. My research interests are in the theory and applications of graphs and networks, mathematical programming and discrete optimization."

BERND HEINRICH ("The Regulation of Temperature in the Honeybee Swarm") is professor of zoology at the University of Vermont. He writes: "I grew up in rural Maine after coming to this country (when I was 10 years old) from Germany. I received bachelor's and master's degrees at the University of Maine and a Ph.D. from the University of California at Los Angeles, all in zoology. For the next 10 years I was in the entomology department at the Uni-


Before this RCA
silicon image sensor,
these little stars were just a twinkle
in the astronomer's eye.
Stars that used to be no more than a dim blur in the telescope are now having their pictures taken with astonishing clarity and brilliance by a tiny silicon chip produced by RCA engineers and scientists.
The chip, called a CCD or Charge-Coupled Device, is built into a camera attached to a telescope. It makes today's most powerful telescopes 5 to 10 times more powerful.
The CCD is capable of reading light from stars too dim to be captured by conventional photography. It transforms the faintest visual image into thousands of electrical signals—digital information that can be displayed on a TV screen or analyzed by computer.
CCD technology allows the astronomer to see what could never be seen before, giving us a clearer picture of our universe than we have ever had in the past.
This CCD image sensor is a product of RCA research—the kind of dedicated research that has been a tradition at RCA for over 60 years. Today, many years after Nipper started listening to "His Master's Voice," research and development of electronic products are funded at the highest level in our company's history. Our commitment is, as it has always been, to the advancement of technology—to the creation of products that expand the human horizon. © 1981 RCA Corporation
RCA
A tradition on the move

versity of California at Berkeley. Last summer I came back East because I was homesick for the changing seasons and the forests and the streams that have always been my love, my inspiration and the source of my spiritual well-being." In addition to his scientific work Heinrich is a runner. "I got in shape to run the Boston Marathon last year," he says, "and celebrated turning 40 by winning the Masters (over 40) division, later also setting a new U.S. Masters record for 50 kilometers (3 : 03 : 56). Otherwise I rarely run competitively because it takes too much time from all the other things I am interested in."

GEORGE W. WETHERILL ("The Formation of the Earth from Planetesimals") is director of the Department of Terrestrial Magnetism of the Carnegie Institution of Washington. His degrees are all from the University of Chicago, concluding in 1953 with a Ph.D. (in physics). On finishing his studies he joined the staff of the department he now heads, remaining there until 1960. He then spent 15 years as professor of geophysics and geology at the University of California at Los Angeles. For four of those years (1968–72) he was chairman of the department of planetary and space science. Wetherill was appointed to his present post in 1975.

NOËL H. GALE and ZOFIA STOS-GALE ("Lead and Silver in the Ancient Aegean") are husband and wife; Gale is codirector of the Geological Age and Isotope Research Laboratory of the University of Oxford and Stos-Gale is attached to the Heberden Coin Room of the university's Ashmolean Museum. Gale studied physics at the Imperial College of Science and Technology in London and then specialized in nuclear physics, the subject in which he got his Ph.D. from the University of Manchester in 1963. His work in nuclear physics led to an increasing interest in astrophysics, the origin of the elements and studies in meteorites. He moved to Oxford in 1965 to pursue meteorite studies and was led eventually into studies of ancient mining and metallurgy. Gale writes that he retains "an active interest in cosmochemistry, the geochronology and geochemistry of the Indian Archaean and the calibration of the geologic time scale while also developing other aspects of the application of science to archaeological problems." Stos-Gale was born in Poland and studied archaeological history and the history of art at the Jagiellonian University in Cracow. A British Council fellowship brought her to Oxford in 1977 to work in the Research Laboratory for Archaeology and the History of Art. She met Gale through collaborative work on a project to establish the composition and provenance of the metal in Early Dynastic Egyptian gold and silver jewelry.

# MATHEMATICAL GAMES

## *The inspired geometrical symmetries of Scott Kim*

by Martin Gardner

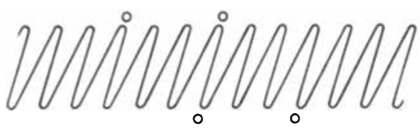
Scott Kim's *Inversions,* published this year by Byte Publications, is one of the most astonishing and delightful books ever printed. Over the years Kim has developed the magical ability to take just about any word or short phrase and letter it in such a way that it exhibits some kind of striking geometrical symmetry. Consider Kim's lettering of my name below. Turn it upside down and presto! It remains exactly the same.

Students of curious wordplay have long recognized that short words can be formed to display various types of geometrical symmetry. On the Rue Mozart in Paris a clothing shop called "New Man" has a large sign lettered "NeW MaN" with the *e* and the *a* identical except for their orientation. As a result the entire sign has upside-down symmetry. The names VISTA (the magazine of the United Nations Association), ZOONOOZ (the magazine of the San Diego zoo) and NISSIN (a Japanese manufacturer of camera flash equipment) are all cleverly designed so that they have upside-down symmetry.

BOO HOO, DIOXIDE, EXCEEDED and DICK COHEN DIED 10 DEC 1883 all have mirror symmetry about a horizontal axis. If you hold them upside down in front of a mirror, they appear unchanged. One day in a supermarket my sister was puzzled by the name on a box of crackers, "spep oop," until she realized that a box of "doo dads" was on the shelf upside down. Wallace Lee, a magician in North Carolina, liked to amuse friends by asking if they had ever eaten any "ittaybeds," a word he printed on a piece of paper like this:

After everyone said no, he would add: "Of course, they taste much better upside down."

Many short words in conventional typefaces turn into other words when they are inverted. MOM turns into WOW and "up" becomes the abbreviation "dn." SWIMS remains the same. Other words have mirror symmetry about a vertical axis, such as "bid" (and "pig" if the *g* is drawn as a mirror image of the *p*). Here is an amusing way to write "minimum" so that it is the same when it is rotated 180 degrees:



It is Kim who has carried this curious art of symmetrical calligraphy to heights not previously known to be possible. By ingeniously distorting letters, yet never so violently that one cannot recognize a word or phrase, Kim has produced incredibly fantastic patterns. His book is a collection of such wonders, interspersed with provocative observations on the nature of symmetry, its philosophical aspects and its embodiment in art and music as well as in wordplay.

Kim is no stranger to this column. He is a young man of Korean descent, born in the U.S., who is now doing graduate work in computer science at Stanford University. He was in his teens when he began to create highly original problems in recreational mathematics. Some that have been published here include his "lost-king tours" (April, 1977), the problem of placing chess knights on the corners of a hypercube (February, 1978), his solution to "boxing a box" (February, 1979) and his beautifully symmetrical "*m*-pire map" (February, 1980). In addition to a remarkable ability to think geometrically (not only in two and three dimensions but also in four-space and higher spaces) Kim is a classical pianist who for years could not decide between pursuing studies in mathematics and in music. At the moment he is intensely interested in the use of computers for designing typefaces, a field pioneered by his friend and mentor at Stanford, the computer scientist Donald E. Knuth.

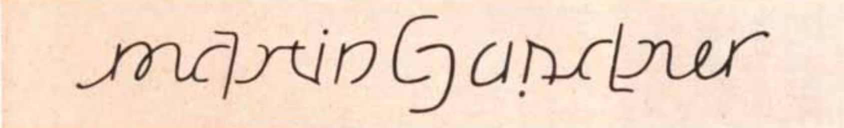
For several years Kim's talent for lettering words to give them unexpected symmetries was confined to amusing friends and designing family Christmas cards. He would meet a stranger at a party, learn his or her name, then vanish for a little while and return with the name neatly drawn so that it would be the same upside down. His 1977 Christmas card, with upside-down symmetry, is one of the illustrations on the opposite page. (Lester and Pearl are his father and mother; Grant and Gail are his brother and sister.) The following year he found a way to make "Merry Christmas, 1978," mirror-symmetrical about a horizontal axis, and in 1979 he made the mirror axis vertical.

For a wedding anniversary of his parents Kim designed a cake with chocolate and vanilla frosting in the pattern shown on the opposite page. ("Lester" is in black, "Pearl" is upside down in white.) This is Kim's "figure and ground" technique. You will find another example of it in *Gödel, Escher, Bach: An Eternal Golden Braid,* the Pulitzer-prize-winning book by Kim's good friend Douglas R. Hofstadter, with whom I am alternating columns this year. Speaking of Kurt Gödel, J. S. Bach and M. C. Escher, another illustration on the opposite page shows how Kim has given each name a lovely mirror symmetry. In an illustration on page 24 Kim has lettered the entire alphabet in such a way that the total pattern has left-right symmetry.

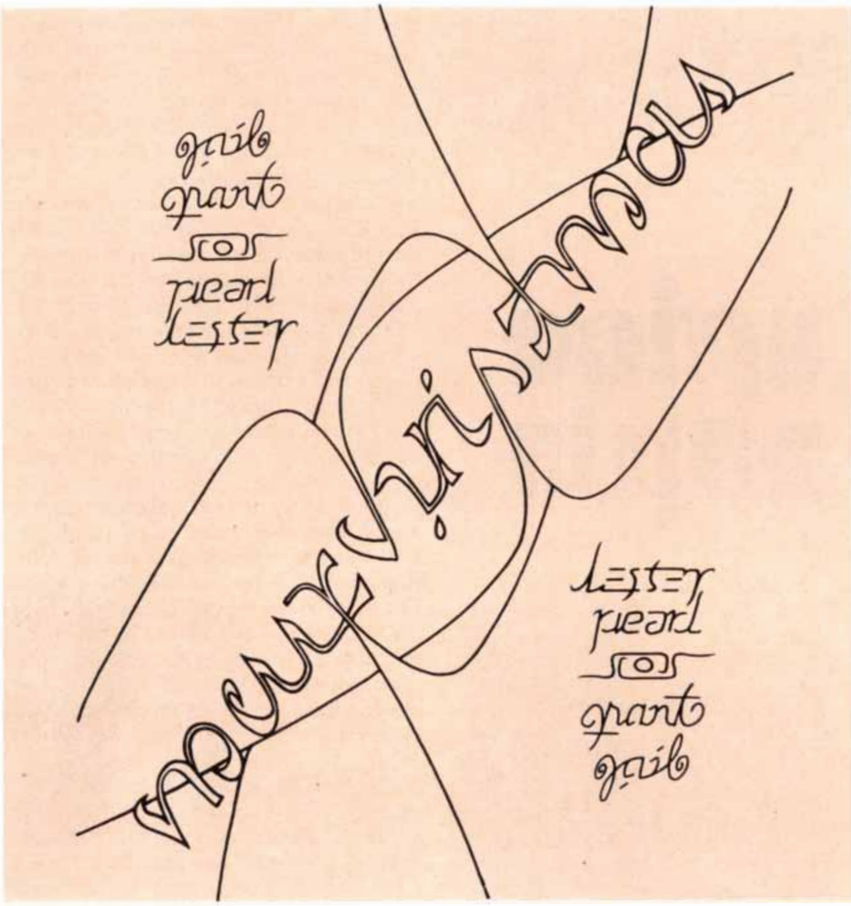
Kim's magic calligraphy came to the attention of Scot Morris, an editor at *Omni.* Morris devoted a page of his popular column on games to Kim's work in *Omni's* September 1979 issue, and he announced a readers' contest for similar patterns. Kim was hired to judge the thousands of entries that came in. You will find the beautiful prizewinners in *Omni's* April 1980 issue and close runners-up in Morris' columns for May and November of the same year.

All the patterns in Kim's book are his own. A small selection of a few more is given on page 24 to convey some notion of the amazing variety of visual tricks Kim has up his sleeve.

I turn now to two unusual mathematical problems originated by Kim, both of which are still only partly solved. In 1975, when Kim was in high school, he thought of the following generalization of the old problem of placing eight

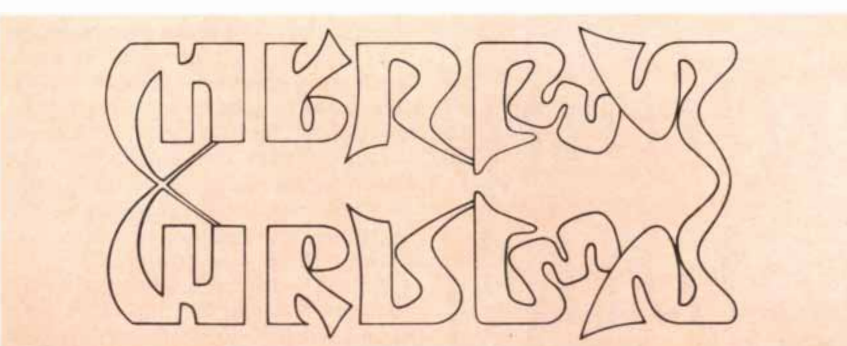


*The name of the columnist reads the same upside down. Try it!*



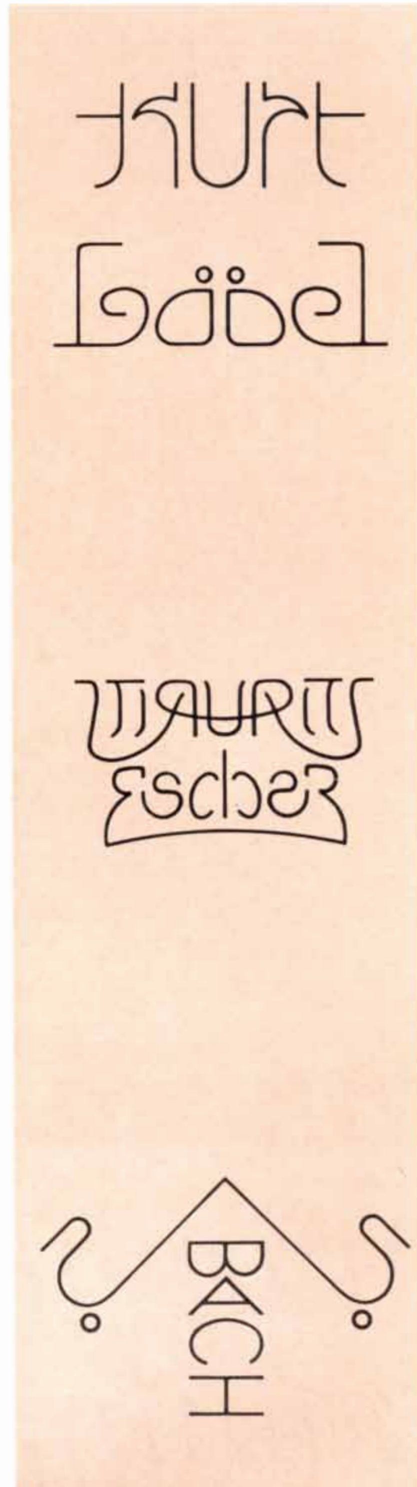

*Kim's Christmas card also reads the same upside down*


*"Merry Christmas" is mirror-symmetrical about a vertical axis*



*"Merry Christmas" is mirror-symmetrical about a horizontal axis*

*"Lester" and "Pearl" are figure and ground*




*Gödel, Escher and Bach in mirror symmetry*

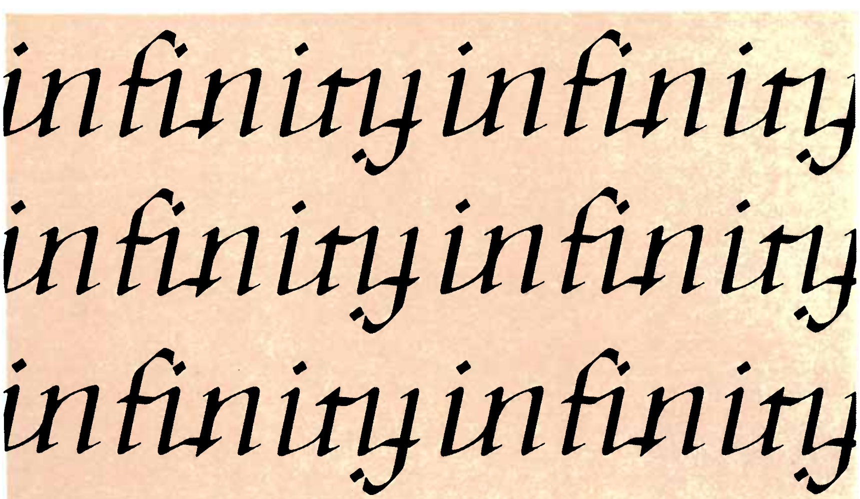
A mirror-symmetrical alphabet

"True" is embedded in "false"

"Upside" is the same upside down

A divided "communication" can be inverted


Invertible "man" and "woman"




"Infinity" keeps going to infinity

queens on a chessboard so that no queen attacks another. Let us ask, said Kim, for the maximum number of queens that can be put on the board so that each queen attacks exactly $n$ other queens. As in chess, we assume that a queen cannot attack through another queen.

When $n$ is 0, we have the classic problem. Kim was able to prove that when $n$ is 1, 10 queens is the maximum number. (A proof is in *Journal of Recreational Mathematics,* Vol. 13, No. 1, page 61; 1980–81.) A pleasing solution is shown in the top illustration on the opposite page. The middle illustration shows a maximal solution of 14 queens when $n$ is 2, a pattern Kim described in a letter as being "so horribly asymmetric that it has no right to exist." There are only conjectures for the maximum when $n$ is 3 or 4. Kim's best result of 16 queens for $n = 3$ has the ridiculously simple solution shown in the bottom illustration, but there is no known proof that 16 is the maximum. For $n = 4$ Kim's best result is 20 queens. Can you place 20 queens on a chessboard so that each queen attacks exactly four other queens before I give Kim's solution in my next column?

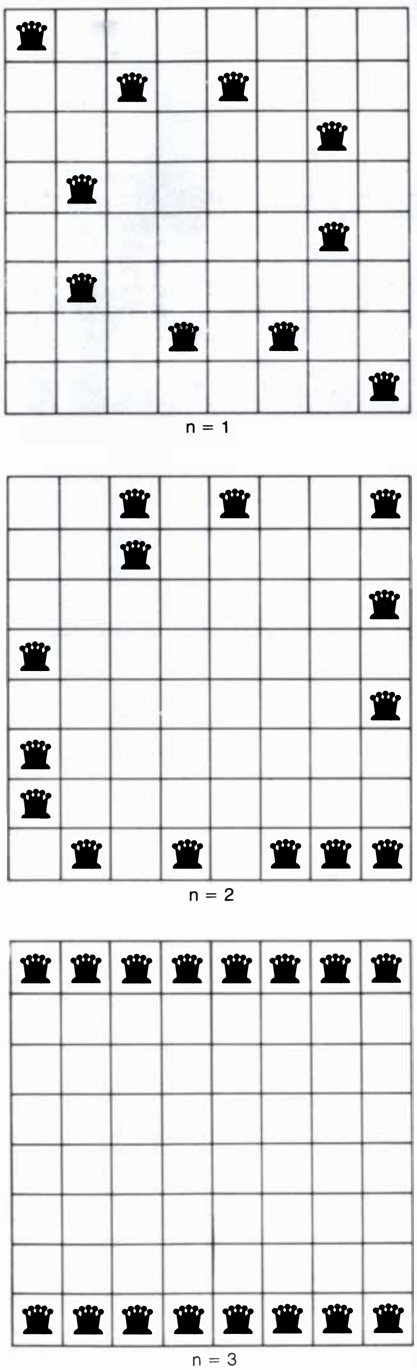
The problem can of course be generalized to boards of any size, but Kim has a simple proof based on graph theory that on no board, however large, can $n$ have a value greater than 4. For $n = 1$ Kim has shown that the maximum number of queens cannot exceed the largest integer less than or equal to $4k/3$, where $k$ is the number of squares along an edge of the board. For $n = 2$ he has a more difficult proof that the maximum number of queens cannot exceed $2k - 2$, and that this maximum is obtainable on all even-order boards.

Kim's problem concerning polycube snakes has not previously been published, and he and I would welcome any light that readers can throw on it. First we must define a snake. It is a single connected chain of identical unit cubes joined at their faces in such a way that each cube (except for a cube at the end of a chain) is attached face to face to exactly two other cubes. The snake may twist in any possible direction, provided no internal cube abuts the face of any cube other than its two immediate neighbors. The snake may, however, twist so that any number of its cubes touch along edges or at corners. A polycube snake may be finite in length, having two end cubes that are each fastened to only one cube, or it may be finite and closed so that it has no ends. A snake may also have just one end and be infinite in length, or it may be infinite and endless in both directions.

We now ask a deceptively simple question. What is the smallest number of snakes needed to fill all space? We can put it another way. Imagine space to be completely packed with an infinite number of unit cubes. What is the smallest number of snakes into which it can be dissected by cutting along the planes that define the cubes?

If we consider the two-dimensional analogue of the problem (snakes made of unit squares), it is easy to see that the answer is two. We simply intertwine two spirals of infinite one-ended flat snakes, one gray, one white, as is shown at the top left on page 29.

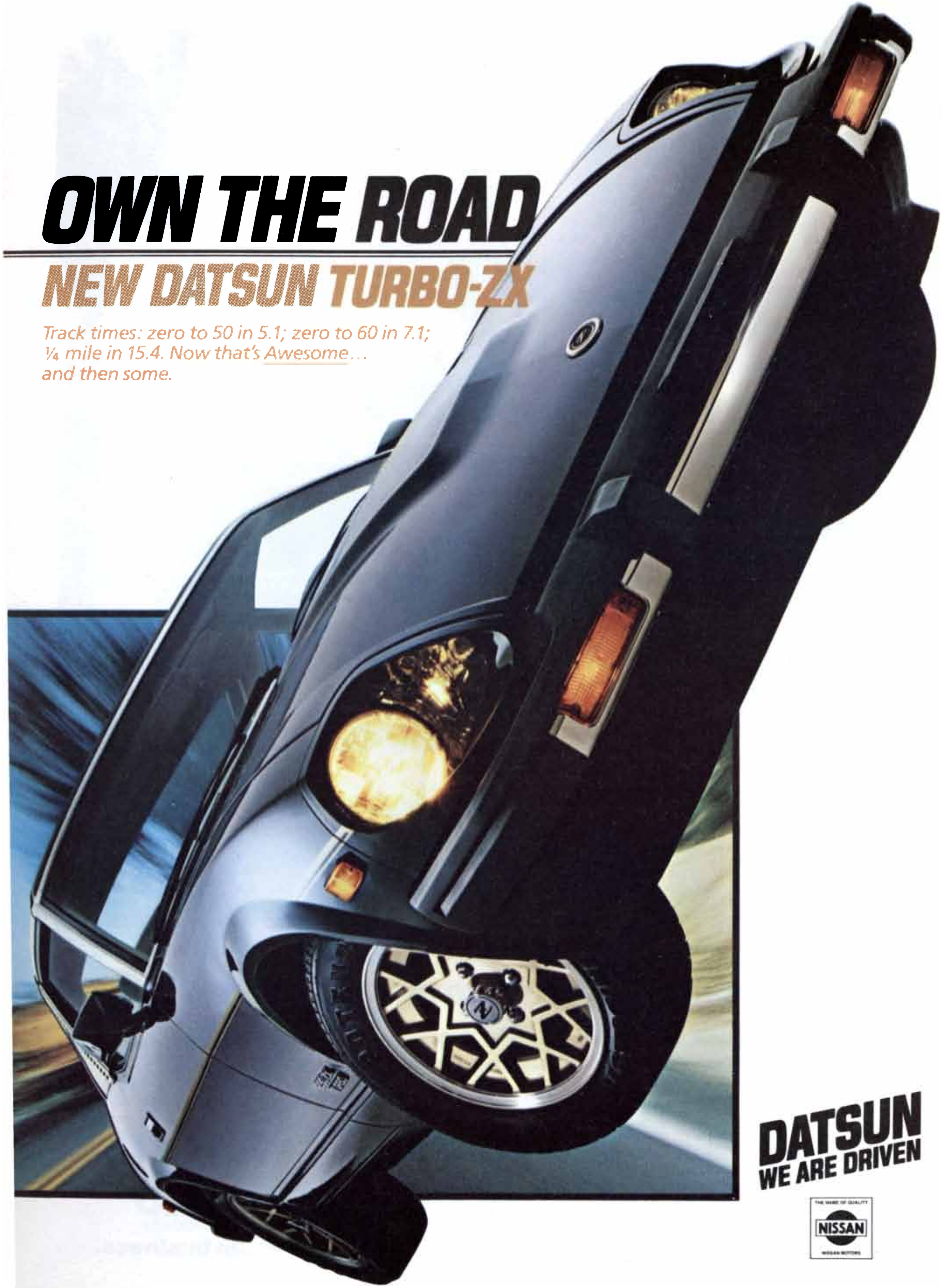
The question of how to fill three-dimensional space with polycube snakes is not so easily answered. Kim has found a way of twisting four infinitely long one-ended snakes (it is convenient to think




*Kim's chess-queens problem*


Allied Chemical Corp.


Allied Corp.


OWN THE ROAD
NEW DATSUN TURBO-ZX
Track times: zero to 50 in 5.1; zero to 60 in 7.1; ¼ mile in 15.4. Now that's Awesome... and then some.
DATSUN
WE ARE DRIVEN
NISSAN

of them as being each a different color) into a structure of interlocked helical shapes that fill all space. The method is too complicated to explain in a limited space; you will have to take my word that it can be done.

Can it be done with three snakes? Not only is this an unanswered question but also Kim has been unable to prove that it cannot be done with two! "A solution with only two snakes," he wrote in a letter, "would constitute a sort of infinite three-dimensional yin-yang symbol: the negative space left by one snake would be the other snake. It is the beauty of such an entwining, and the possibility of building a model large enough to crawl through, that keeps me searching for a solution."

The problem can of course be generalized to snakes made of unit cubes in any number of dimensions. Kim has conjectured that in a space of $n$ dimensions the minimum number of snakes that completely fill it is $2(n - 1)$, but the guess is still a shaky one.

A few years ago I had the pleasure of explaining the polycube-snake problem to John Horton Conway, the Cambridge mathematician. When I concluded by saying Kim had not yet shown that two snakes could not tile three-dimensional space, Conway instantly said, "But it's obvious that—" He checked himself in mid-sentence, stared into three-space for a minute or two, then exclaimed, "It's *not* obvious!"

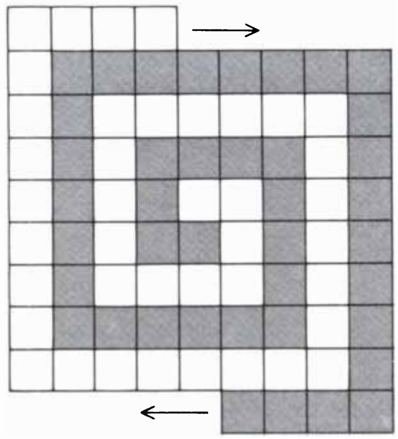
I have no idea what passed through Conway's mind. I can only say that if the impossibility of filling three-space with two snakes is not obvious to Conway or to Kim, it probably is not obvious to anyone else.

Here are the solutions to April's short problems:

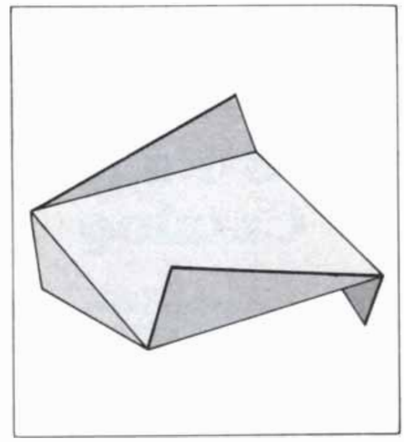
1. Consider the two outermost spots, $A$ and $K$. A spot $L$ on the line anywhere between $A$ and $K$ (inclusive) will have the same sum of its distances to $A$ and $K$. Clearly this is smaller than the sum if $L$ is not between $A$ and $K$. Now consider $B$ and $J$, the next pair of spots as you move inward. As before, to minimize the sum of $L$'s distances from $B$ and $J$, $L$ must be between $B$ and $J$. Since $L$ is then also between $A$ and $K$, its position will minimize the sum of its distances to $A$, $K$, $B$ and $J$.

Continue in this way, taking the spots by pairs as you move inward to new nested intervals along the line. The last pair of spots is $E$ and $G$. Between them is the single spot $F$. Any spot between $E$ and $G$ will minimize the distances to all spots except $F$. Obviously if you want also to minimize the distance to $F$, the spot must be exactly at $F$. In brief, Lavinia should move into the same building where her friend Frank lives.

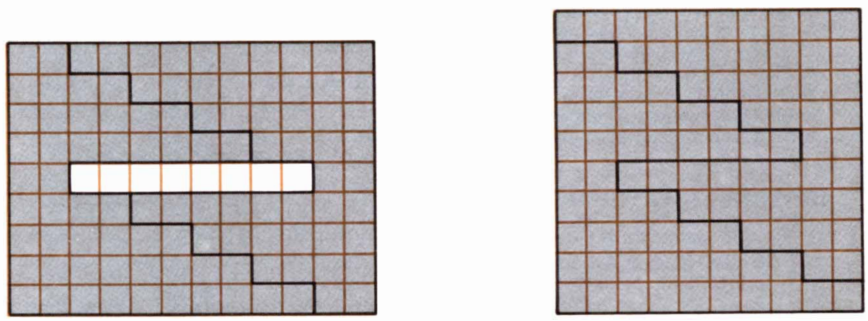
To generalize, for any even number of spots on a straight line, any spot between the two middle spots will have a mini-



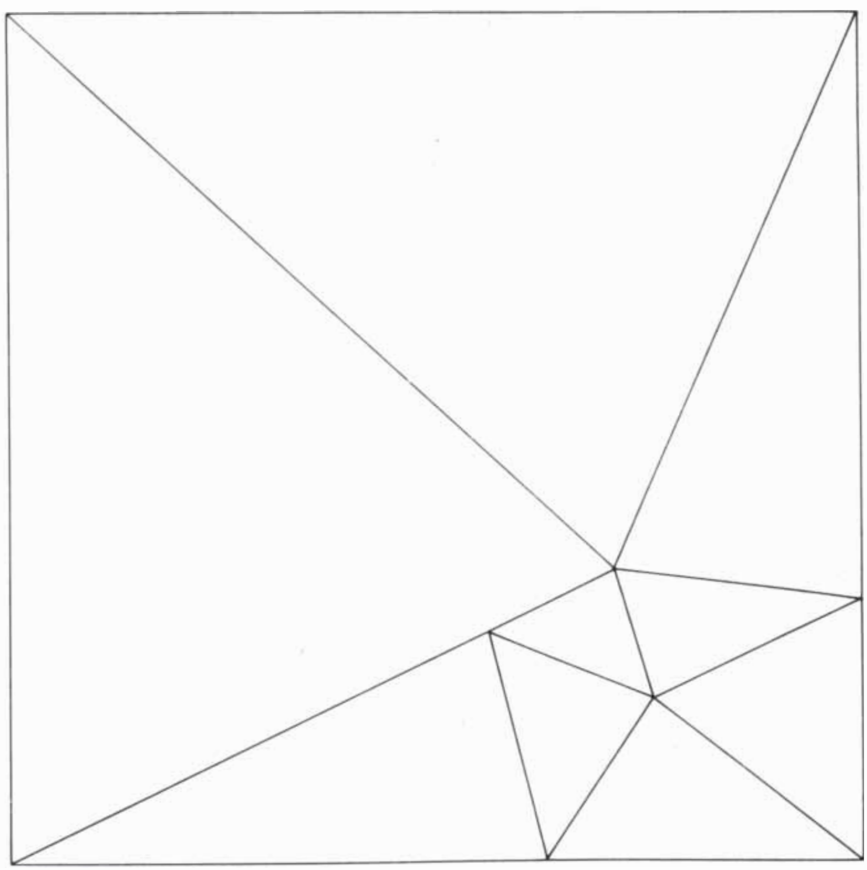
*How two flat snakes tile the plane*



*A fold with mirror-rotation symmetry*



*A damaged patchwork quilt is sewn together to make a square quilt without a hole*



*How to cut a square into nine acute triangles* mal sum of distances to all spots. For any odd number of spots the center spot is the desired location. The problem appeared in "No Calculus, Please," by J. H. Butchart and Leo Moser (*Scripta Mathematica,* Vol. 18, Nos. 3–4, pages 221–236; September–December, 1952).

2. The illustration at the top right on the preceding page shows how to fold a square of paper into a shape that has no plane of symmetry yet can be superposed on its mirror reflection. The figure is said to have "mirror-rotation symmetry," a type of symmetry of great importance in crystallography. I took this example from page 42 of *Symmetry in Science and Art,* by A. V. Shubnikov and V. A. Koptsik (Plenum Press, 1974).

3. The middle illustration on the preceding page shows how the damaged patchwork quilt is cut into two parts that can be sewn together to make a square quilt without a hole. The problem is No. 215 in Henry Ernest Dudeney's *Puzzles and Curious Problems* (Thomas Nelson and Sons, Ltd., 1931).

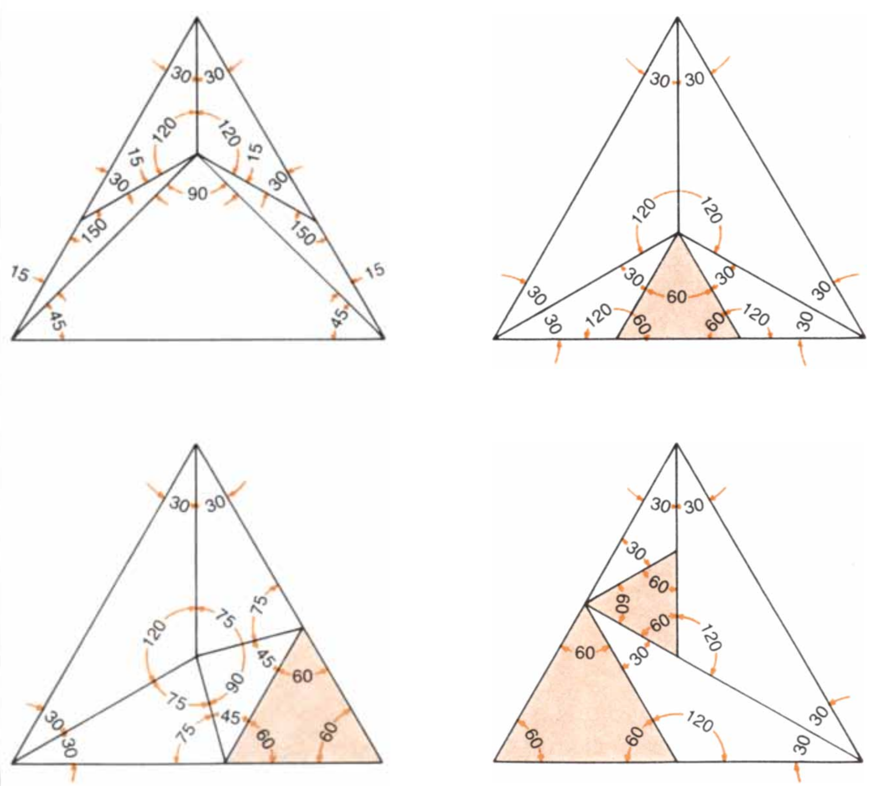
4. The bottom illustration on the preceding page shows how to cut a square into nine acute triangles. The solution is probably unique. If triangulation is taken in the topological sense, so that a vertex is not allowed to be on the side of a triangle, then there is no solution with nine triangles, although there is a solution for eight triangles, for 10 and for all higher numbers. This curious result has been proved in an unpublished paper by Charles Cassidy and Graham Lord of Laval University in Quebec.

The illustration below shows four ways to cut an equilateral triangle into five isosceles triangles. The first pattern has no equilateral triangle among the five, the second and third patterns both have one equilateral triangle and the fourth pattern has two equilateral triangles. The four patterns, devised by Robert S. Johnson, appear in *Crux Mathematicorum* (Vol. 4, No. 2, page 53; February, 1978). A proof by Harry L. Nelson that there cannot be more than two equilateral triangles is in the same volume of the journal (No. 4, pages 102–104; April, 1978).

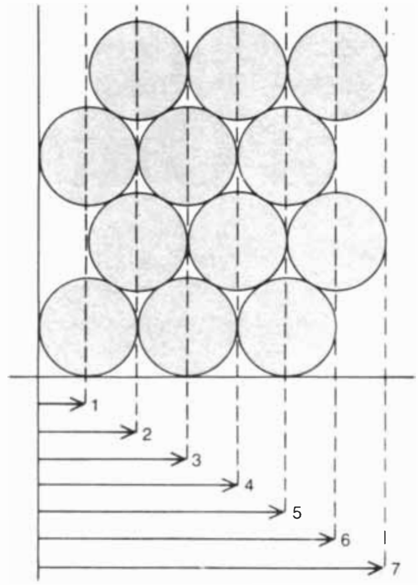
5. Integral distances in centimeters along a line can be measured by one-yen Japanese coins, each with a radius of one centimeter, in the manner shown in the top illustration on page 32.

6. The map in the bottom illustration on page 32, found by Lloyd Shapley, a mathematician at the Rand Corporation, proves that when five colors are used in playing Steven J. Brams's map-coloring game, there is a map on which the maximizer can always win.

The map is a projection of the skeleton of a dodecahedron, with an outside region (*A*) that represents the "back" face of the solid. The maximizer's strategy is always to play on the face of the dodecahedron opposite the face where the opponent last played, using the same color. (In the illustration regions representing pairs of opposite faces are given the same letter.) As you can readily see, this strategy eliminates successive colors from further use, forcing the game




*Four ways to cut an equilateral triangle into five isosceles triangles*

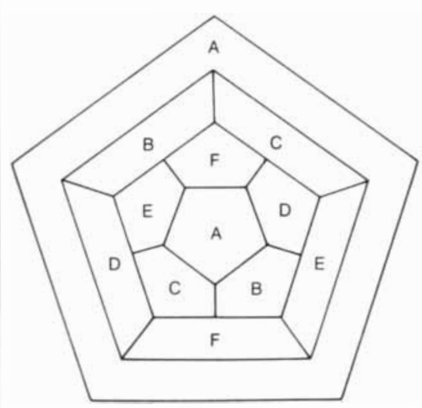


*One-yen coins measure integral distances*

to the point where the minimizer cannot play without using a sixth color.

Is there a map on which the maximizer can force the use of a seventh color when the opponent plays optimally? This remains an unsolved problem.

7. John Horton Conway's strategy for his game of whim is as follows. Treat the whim move as if it were another pile consisting of one counter if there is a pile of four or more counters, and as if it were another pile consisting of two counters if there is no pile of four or more. Until someone makes the whim move the invisible whim pile remains. Making the whim move removes the whim pile. A whim position is a winning position for the player who would win in ordinary nim if the whim pile were actually there. The winning strategy, therefore, is simply to imagine the whim is present until a player removes it and to play the strategy of ordinary nim. If a move changes a position from one in which at least one pile has four or more counters to one in which there is no such pile, the whim pile acquires its second counter after the move, not before.



*A proof of how to win a map-coloring game*

# EUROPEAN AEROSPACE: COLLABORATION AND COMPETITION

**This survey of European Aerospace, which includes technical reports on recent industrial advances and innovations, was coordinated by Joel Stratte-McClure following visits to Europe's major aerospace companies.**

(Airbus Industrie)

(ESA)

(British Aerospace)

*Three of Europe's success stories: the Airbus A300, the Ariane launcher and the Tornado multi-role combat aircraft.*

Airbus. Adour. Alpha Jet. Ariane. Euromissile. Eurosatellite. Lynx. Spacelab. Tornado. Turbo-Union. These and other collaborative programs among European aerospace companies have created a technologically imaginative, commercially sound competitor in every sector of the aerospace industry – from avionics and simulators to satellites, helicopters and military trainers to civil aircraft, ballistic missiles and engines to rocket launchers.

Although the United States remains the leader in aerospace productivity, Europe, which represents 28 per cent of the Western world's aerospace market, has impressively continued to increase its importance and export its expertise. Today European aerospace is an $18,000 million industry, stretches from Greece to Northern Ireland and Sweden to Spain, and employs over 450,000 persons. During the 1970's, total European Economic Community aerospace sales increased approximately 55 per cent while U.S. sales declined 15 per cent.

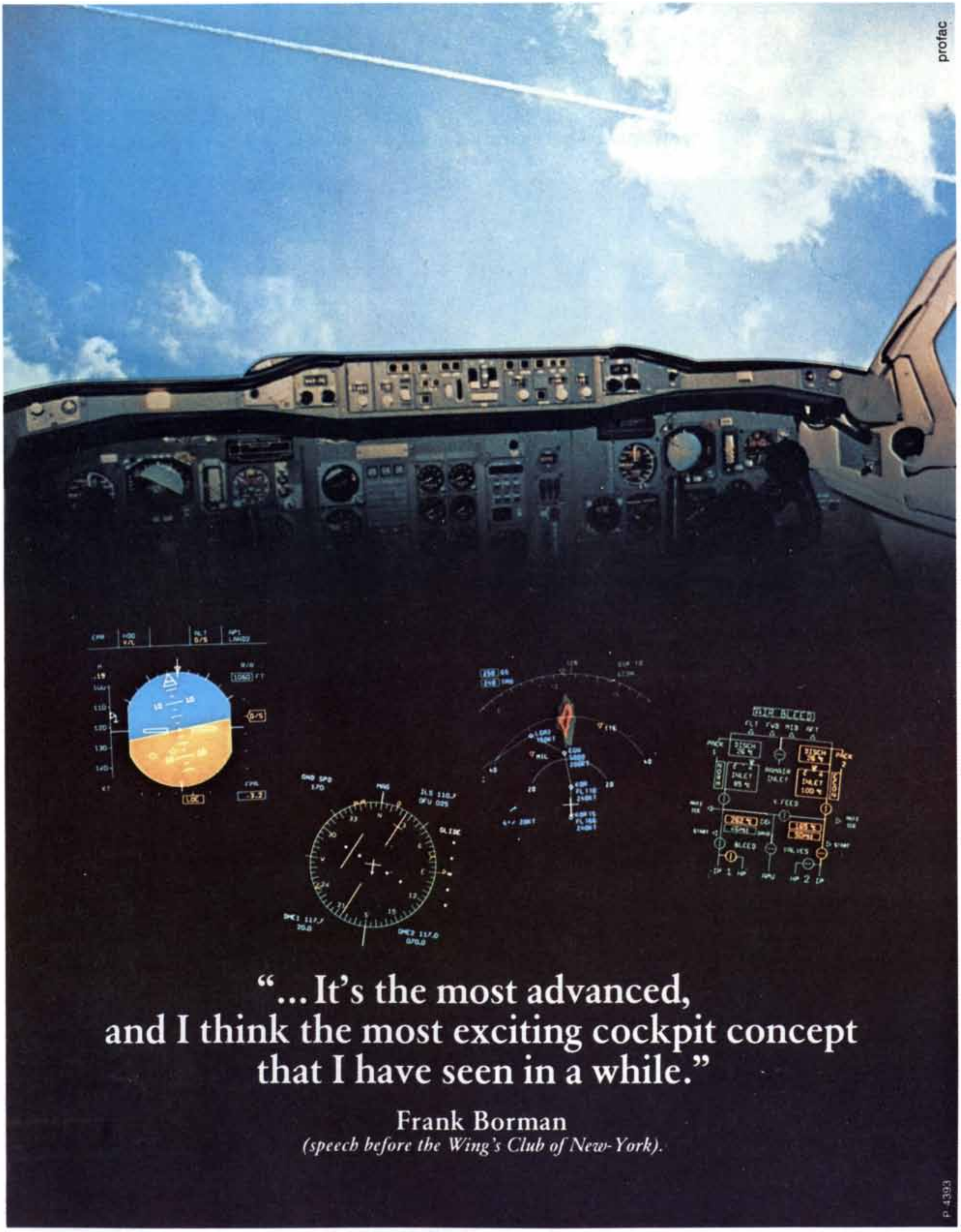
Throughout Europe, aerospace R & D has assumed a substantial share of total expenditure as governments and the private sector attempt to augment their industrial productivity, establish specialized technological niches and become commercially independent. In addition, there has been successful collaboration on numerous intra-European projects for financial, technical, industrial, political, military and commercial reasons.

Collaboration continues to be the predominant theme of the European aerospace industry. It has permitted participants to share prohibitively high R & D costs for major projects while pooling different technical resources and expertise. It has led to manufacturing of products with standardized features and produced significantly larger markets. Most importantly, it has resulted in a mastering of management and organizational methods for complex, multinational projects which are becoming the norm in the European aerospace industry.

## EUROPEAN COLLABORATIVE PROGRAMS: AIRCRAFT, ENGINES, MISSILES AND SPACE

| | | Belgium | France | W. Germany | Italy | Netherlands | Spain | United Kingdom |
|---|---|---|---|---|---|---|---|---|
| **Aircraft:** | | | | | | | | |
| A300/A310 Airbus | Twin jet wide-body transport | ● | ● | ● | | ● | ● | ● |
| Alpha Jet | Twin jet trainer and strike aircraft | ● | ● | ● | | | | |
| Atlantic | Twin turboprop maritime patrol aircraft | ● | ● | ● | ● | ● | | |
| Concorde | Four jet supersonic transport | | ● | | | | | ● |
| F.27 Friendship | Twin turboprop transport | ● | ● | | | ● | | |
| F.28 Friendship | Twin jet transport | | | ● | | ● | | ● |
| Jaguar | Twin jet tactical support and trainer aircraft | | ● | | | | | ● |
| Mercure | Twin turbofan transport | ● | ● | | ● | | | |
| Tornado | Variable geometry twin-jet multi-role combat aircraft | | | ● | ● | | | ● |
| Transall C160 | Twin turboprop transport | ● | ● | ● | | | | ● |
| VFW 614 | Twin turbofan transport | ● | ● | ● | | ● | | ● |
| Gazelle | Single turbine multi-purpose helicopter | | ● | | | | | ● |
| Lynx | Twin turbine general purpose helicopter | | ● | | | | | ● |
| Puma | Twin turbine logistic helicopter | | ● | | | | | ● |
| Argus | Battlefield surveillance system | | ● | ● | | | | |
| **Aero Engines:** | | | | | | | | |
| Adour | Turbofan, for Jaguar | | ● | | | | | ● |
| Astazou IIIN | Turbine, for Gazelle | | ● | | | | | ● |
| Gem | Turbine for Lynx | | ● | | | | | ● |
| Larzac | Turbofan, for Alpha Jet, etc. | ● | ● | ● | | | | ● |
| Olympus 593 | Turbojet, for Concorde | | ● | | | | | ● |
| RB 199 | Military turbofan | | | ● | ● | | | ● |
| Turmo IIIC4 | Turbine, for Puma | | ● | | | | | ● |
| Tyne | Turboprop, for Atlantic and Transall | ● | ● | ● | | | | ● |
| Viper | Light turbojet | | | | ● | | | ● |
| **Missiles:** | | | | | | | | |
| Milan | Anti-tank | | ● | ● | | | | |
| HOT | Anti-tank | | ● | ● | | | | |
| Roland | Surface-to-air | | ● | ● | | | | |
| Otomat | Ship-to-ship, ground-to-surface | | ● | | ● | | | |
| **Space:** | | | | | | | | |
| Ariane | Three-stage rocket vehicle | ● | ● | ● | ● | ● | ● | ● |
| ECS | European communication satellite | ● | ● | ● | ● | ● | ● | ● |
| GEOS | Geostationary satellite to measure electric and magnetic fields | ● | ● | ● | ● | ● | ● | ● |
| MARECS | Maritime European Communications satellite | ● | ● | ● | ● | ● | ● | ● |
| Meteosat | Weather satellite | ● | ● | ● | ● | | | ● |
| OTS | Orbital test satellite for TV relay and communication | ● | ● | ● | ● | ● | | ● |
| Spacelab | Re-usable space laboratory | ● | ● | ● | ● | ● | ● | ● |
| Exosat | Scientific satellite for X-ray research | ● | ● | ● | ● | ● | ● | ● |

Source: AECMA


profac
"... It's the most advanced,
and I think the most exciting cockpit concept
that I have seen in a while."
Frank Borman
(speech before the Wing's Club of New-York).
P 4393

"Large programs in Western Europe today must be undertaken on a collaborative basis," says Jean Charles Poggi, corporate vice president of France's Aerospatiale, Europe's largest aerospace company. "They may cost five to ten per cent more but the technical and commercial benefits easily outweigh the obstacles."

European industry has been rationalized to meet the collaborative challenge. When the two principal German manufacturers officially complete a government-induced merger, most European countries will have only one main, often government-owned, airframe manufacturer – Aeritalia in Italy, Aerospatiale in France, British Aerospace in the United Kingdom, Messerschmitt-Bölkow-Blohm (MBB) in the Federal Republic of Germany, Construcciones Aeronauticas S.A. (CASA) in Spain, Fokker in Holland, and Saab-Scania in Sweden.

The result: a pan-European industry about one-third the size of that in the United States.

"We have realized our individual strength depends on being European," says Michael Goldsmith, managing director of the Hatfield-Chester Division of British Aerospace. "And the more successful we are in collaborative efforts, the more European we become."

There is a feeling among most Europeans that, despite some failures in the 1960's and 1970's, collaborative success is creating an improved technical and commercial reputation.

"Cooperative programs have put us on a technical par with the Americans in numerous areas," says Dr. Gustav Bittner, vice president at MBB.

Contends Michel Schmit, director of international operations for France's MATRA group, "Successful collaborative programs enable us to achieve technical and commercial independence from manufacturers in the U.S."

Rene Collette, head of communication satellite programs department for the European Space Agency (ESA), the coordinating body for most space activities in Europe, adds "Europe used to have the reputation of being technically competent but commercially inept. Collaborative programs have enabled us to be strong in both areas."

In fact, yesterday's collaborative failures are partly responsible for today's successes. "Europe's come a long way since the technical excellence but commercial failure of the Concorde," says Tore Gullstrand, general manager of Saab-Scania's Aerospace Division. "We've realized technical perfection is a dubious achievement without commercial benefits."

Naturally there are collaborative pitfalls which every European project must resolve. These include: program coordination; engineering approaches and design standards; ownership of technology; language differences; exchange rates; specifications and quality control; varying labor laws and different employment conditions. But, in most cases, European ventures have managed to overcome the problems. "We don't really have a choice when it comes to collaboration," concludes Claes-Ulrik Winberg, president of Sweden's Bofors Ordnance. "The alternative is economic stagnation."

Perceptions of European aerospace have altered considerably with the technical and commercial success of various collaborative ventures, as discussed in this assessment by one of America's top aviation and space technology writers.

## EUROPEAN AEROSPACE: CHALLENGING AMERICAN INDUSTRY

by Robert R. Ropelewski

Now recovered from the devastation of two major wars and slowly mastering the delicate art of international collaboration, Europe's leading aviation enterprises are investing heavily in design, engineering and manufacturing capabilities to compete with U.S. industry. Those investments are beginning to pay dividends.

Airbus A300 jumbo jetliners can be seen at most major airports around the world, including many in the United States. With over 500 sold or on option, they have captured a sizable chunk of the world market for commercial airliners in recent years and are expected to continue doing so throughout the 1980's. Built from components manufactured in several European countries, the A300 and its successive derivatives incorporate advanced flight controls and automatic landing systems that are at the forefront of technology in this field.

On the military side, dozens of air forces around the world are equipped with combat aircraft supplied from Europe. Not a single new market arises that doesn't have one or more European contenders along with traditional U.S. competitors.

Growing demand for helicopters—both civil and military—has spawned a thriving helicopter industry in Europe. American manufacturers are battling fiercely for new and traditional customers, not only in foreign markets but in the U.S. The U.S. Coast Guard, for example, is preparing to take delivery this year of the first of 90 Aerospatiale Dauphin helicopters it is producing for search and rescue missions.

Britain and France have made solid inroads in the business jet field with more than 1000 aircraft sold worldwide. About two-thirds of those sales have been in the U.S., against formidable competition from domestic American manufacturers. The U.S. Coast Guard, after evaluating both domestic and European candidates, has ordered 41 French Falcon jets for its maritime patrol responsibilities. Ironically, problems with the American engines for the Falcon jets have delayed deliveries of those aircraft.

From nuclear ballistic to close-in air-to-air missiles, Europe has fielded a complete line of strategic and tactical guided weapons comparing favorably in capability and performance with similar U.S. equipment. The U.S. Army is preparing to take delivery of its first Franco-German Roland Air Defense Missiles, manufactured by Euromissile members Aerospatiale and MBB, to protect army ground units against hostile, low flying aircraft.

Europe's most recent challenge to U.S. dominance of world aerospace markets has come in the space sector, where a new line of European satellites is under development to compete with U.S. technology in markets around the world. To unshackle itself from current dependence on U.S rocket boosters to launch its spacecraft Western European nations — through ESA — are developing the Ariane launch vehicle. With development costs exceeding $1 billion, it is Europe's biggest space program to date and will compete with the NASA Space Shuttle for the launching of future U.S., European and Third World spacecraft.

Though funding has been a problem, there has been no shortage of good ideas or technological innovation in Western Europe. European jetliners were making routine automatic landings in poor weather long before the necessary systems were incorporated in U.S.-built airliners. European helicopter manufacturers were installing high-performance, longer-life composite rotor blades several years before American manufacturers adopted them.

European spacecraft designers may also be striding ahead of the U.S. in the development of commercial broadcasting and communications satellites that can be marketed to many Third World countries.

U.S. aerospace leaders are warning that similar gains may soon be made by European industry in the field of remote-sensing satellites with both civil and military applications. All of these could be sold and launched by Europe without any dependence or political interaction with the U.S.

Western Europe has always had a solid foundation in aeronautical technology, but this has been fragmented by the multiple nationalities involved and the lack of funds to pursue promising technologies. A greater European cohesion in recent years has encouraged a consolidation of R & D funds, most noticeable in the Airbus program and various European space programs. This, in turn, has allowed Western Europe to explore new and promising technology much more cost-effectively.

The European industry also learned its lessons in terms of aggressively marketing the technology it has developed. There is a big difference between the sales effort of the Concorde supersonic transport in the late 1960's and early 1970's and the more sophisticated, aggressive multi-national marketing effort for the Airbus. The same aggressiveness is being shown by European teams marketing the Ariane space launcher to potential customers in the U.S. and elsewhere.

Additionally, European government and industry officials have shown an interest in capitalizing on the momentum of the Airbus sales record to launch other new commercial jetliner programs, but it remains to be seen if the traditional arguments over the division of work between the various national partners can be resolved.

Meanwhile, Britain, Northern Ireland, West Germany, France, Italy and Spain are all studying or already developing—independently for the most part—a new line of smaller, commuter aircraft that will be competing for a growing market, particularly in the United States. U.S. manufacturers have yet to respond to the needs of this market.

One thing is clear, U.S. industry dominance of the world aerospace markets is being seriously challenged by Western Europe–and Western Europe's industries and governments are showing a strong determination to sustain this trend. ■

## AIRBUS: THE FLAGSHIP OF COLLABORATION

Airbus Industrie, the consortium of risk-sharing aerospace manufacturers producing a family of wide-bodied fan jet aircraft, is the best known European collaborative venture today.

It offers the strongest competition to U.S. dominance in the air transport market and is making commercial headway throughout the world with the A300 aircraft and its variations.

Partners in the project include Aerospatiale (37.9 per cent); Deutsche Airbus, comprising MBB and VFW, (37.9 per cent); British Aerospace (20 per cent); CASA (4.2 per cent); associate members Fokker and Belairbus as well as suppliers and sub-contractors from numerous other countries, including the United States. "The combination of different technical talents has produced an exceptional cross-fertilization of development ideas and engineering techniques," says Airbus Industrie president Bernard Lathiere. "The continual technical scrutiny by the partners creates a near fail-safe, cost effective development."

Airbus has brought the second member of its family, the A310, into production and the partners are currently expanding their facilities to increase output from a current four to eight aircraft per month by 1984. Airbus, attempting to obtain between 25-30 per cent of the world market, will soon decide which other aircraft variations it will produce. "Airbus has put Europe in the big leagues," says Frank Borman, chairman and president of Eastern Airlines. "It is a technical success because it is elegant in its simplicity, has thoughtful systems, low maintenance costs and tremendous reliability."

Future Airbus products will continue to exploit this technical expertise as discussed in the following report.

## AERODYNAMIC DEVELOPMENT OF THE A300 AND A310

by Roger Beteille, Executive Vice President, and Jean-Pierre Saint Girons, Deputy Chief Engineer A310, Airbus Industrie

The market penetration achieved by the A300 and A310 is largely due to their substantial fuel economy in a period of rising fuel costs. Advanced aerodynamics, most notably in the wing design, has contributed to this success and there is a continual evolution of key aerodynamic features.

### The A300 concept

The A300 was conceived in the late 1960's as a high capacity twin turbofan transport with the best possible economic performance on short and medium range routes. The wing was designed around two main requirements:

- Flight at high subsonic Mach numbers with low drag.
- Field performance compatible with routes flown by major European operators.

Many configurations were investigated in detail to ensure that, while meeting these requirements, there was:

- Adequate volume for fuel and undercarriage.
- An optimum installation of high lift devices and other wing movable surfaces.
- No stability problems at high or low speed within the design flight envelope of the aircraft.

Experience with over 100 aircraft in service has established the validity of the original concept. In addition, developments of the original design have produced versions with passenger capacity varying from 250 to over 340 and range increased from 2300 km to more than 5900 km.

### From the A300 to A310

The A310 is being developed to make the maximum use of this A300 experience. It uses the same wide body cross-section selected for the A300 and the fuselage will be built in the same jigs. The well-proven layout for wing movable surfaces, tailplane position and engine location has also been retained, giving the A310 a strong family resemblance to the A300. However, there are significant differences between the two aircraft. Most notably the A310 has:

- An all-new wing to take advantage of the latest advances in aerodynamics.
- A new rear fuselage permitting additional seating capacity without increasing fuselage length.
- A new reduced area tailplane.
- Digital systems and an advanced flight deck concept.
- Increased application of advanced materials.
- New engines.

Many of these improvements will be employed in future A300s, most notably the series 600 version.

### A300—High speed wing design

The main objective was to improve on contemporary design standards by increasing lift at a given Mach number (or vice versa) without drag penalty. This was achieved by close attention to leading edge design by developing the 'peaky' concept already proved on VC10 and Trident, but without compromising low speed performance; and further development of the rear loading concept building on previous experience with the Trident and HS 125 and subsequent extensive research.

However, the desirable supercritical flow characteristics on the wing upper surface were obtained by a process based on subcritical theoretical methods combined with extensive wind tunnel investigations. The standard achieved was more advanced than in any previous or contemporary design.

### Evolving aerodynamics

Since the A300 was designed there has been an exciting evolution in theoretical aerodynamics based on the availability of powerful general purpose computers with advanced numerical techniques.

The result has been that mixed subsonic/supersonic flows, including shock waves, can be treated theoretically. Available methods permit the analysis of three dimensional flow around a wing-fuselage combination, with the inclusion of a representative approximation to the boundary layer flow.

These new methods allowed the A300 supercritical flow and rear loading characteristics to be developed to a much greater extent for the A310. They also enabled the basic wing to be designed with much less dependence on wind tunnel development than the A300.

However, wind tunnel testing has been conducted from the earliest stage of the A310 design to confirm the theoretical results. Particular attention has been paid to delaying the onset of buffet and avoiding pitchup at high angles of attack.

All interference aspects have been treated very carefully, including the wing-fuselage junction and the flap track fairings. Emphasis has been put on the optimization of the engine installation, including the influence of the jet. Intensive use of live engine simulators has been made including both turbine driven simulators and blown nacelles.

The similarity of the configurations and the fact that the A300 already has a very aerodynamically advanced wing permits back to back wind tunnel tests on A300 and A310 models. This allows a safer extrapolation to full scale backed up by actual A300 flights.

These investigations have resulted in:

- An increase in design lift coefficient of about 20 per cent relative to the A300.
- An increase in wing thickness relative to A300, particularly on the inboard wing.
- A reduction in wing area of 16 per cent relative to the A300 while retaining adequate volume for fuel and undercarriage stowage and maintaining good structural efficiency.
- An increase of 14 per cent in aspect ratio while maintaining the span within 2 per cent of the A300, thus improving induced drag and hence fuel efficiency.

Wind tunnel work has confirmed the soundness of the A310 wing design which includes:

- An increase in cruise lift-to-drag ratio of more than five per cent relative to the the A300.
- Buffet boundaries increased by more than 12 per cent giving the A310 high altitude cruise capability despite its small wing area.
- High speed stability at high lift as smooth as on the A300.

### Low speed high lift design

To achieve the A300 field performance requirements, very efficient high lift devices were needed.

Full span leading edge slats associated with tabbed fowler flaps at the trailing edge were finally selected. This track mounted system permitted take-off and landing configurations to be aerodynamically optimized. The system was examined in extensive wind tunnel tests guided by theoretical methods.

The A300 proved to be more effective than expected after flight testing and maximum lift coefficients up to three have been demonstrated in the landing configuration.

A similar approach has been used to optimize A310 high lift devices and the experience from the A300 has been used to improve the system.

A flexible leading edge krüger was looked at for the A310 but rejected in favor of the A300 type slat system.

The flap tabs were replaced by a forward expanding vane which offered a cleaner cruise configuration than the tabbed flap when combined with the advanced airfoil. Early wind tunnel tests showed that the vane on the outboard flap could be eliminated with a substantial saving in weight and complexity.

In conclusion, the advantages offered by designing a family of aircraft which can incorporate progressive advances in technology are well illustrated by the progress achieved in the A300 and A310 programs. Each aircraft and its developments can benefit from past experience and a closed loop is established which allows proved benefits to be introduced with minimum technical and financial risk. ■

Another technical evolution at Airbus Industrie involves a new forward facing crew cockpit (ffcc) which displays data in a new manner to improve crew efficiency and minimize the workload. "The updating in technology and control of information has allowed us to develop a rational system which reduces the pilots' workload and maximizes safety in the aircraft," says Bernard Ziegler, Airbus Industrie's senior vice president of flight and support.

The ffcc controls and displays are logically organized and include clear color coding; co-locations of selectors, controls and indicators; groups of inter-dependent systems; simplification of the man-machine interface through cathode ray tube (CRT) displays; and careful distribution of working areas.

"The Airbus ffcc is at least one generation ahead of any of the American producers, particularly in terms of software and presentation," says Eastern's Borman.

The Airbus project has enabled a variety of companies in Europe to expand their own operations while contributing to the collaborative effort, which has cost $3 billion to date. "We've been able to increase our R & D budget and will hopefully boost our overall industrial production because of participation in the Airbus program," says Marcel Claisse, managing director of SONACA in Belgium.

In addition, the Airbus is usually sold with General Electric or Pratt and Whitney engines, which provide a slice of business for American firms. "American engines and components comprise 25 per cent of every Airbus," says Lathiere. "Each Airbus sold is a boost for American industry." General Electric manufactures its Airbus engines in conjunction with SNECMA in France. Pratt and Whitney became a major collaborator when it joined forces with Airbus in 1977 to certify the JT9D engine on the A300. Since then, the Pratt and Whitney improved JT9D-7R4 series engine has been selected to power the A310 and the increased capacity A300 B4-600 jet.

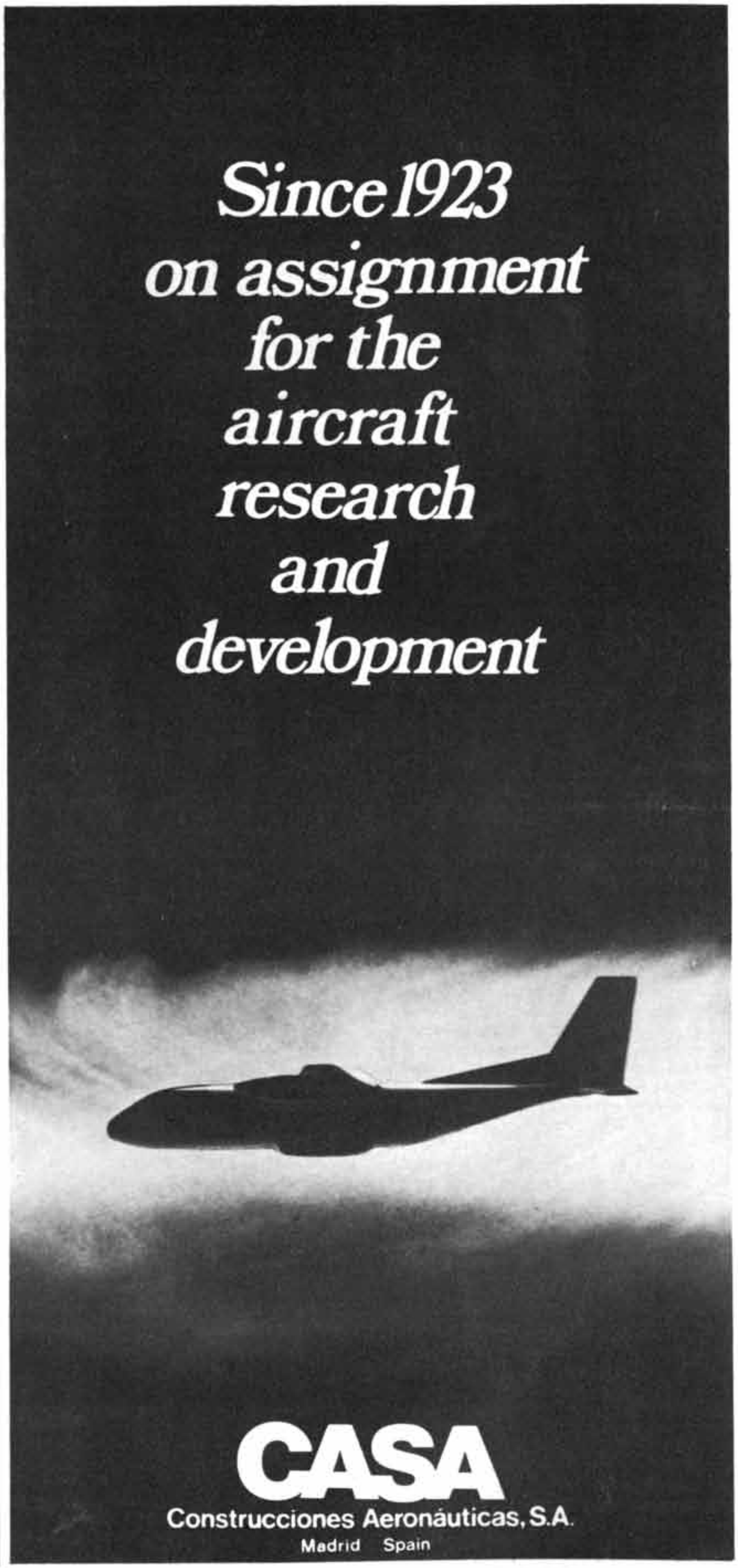
## TORNADO: MERGING MILITARY EXPERTISE

Another European collaborative project, the Panavia Tornado swing-wing all weather multi-role combat aircraft (MRCA), is being manufactured by a tri-national company including British Aerospace (42.5 per cent), MBB (42.5 per cent) and Aeritialia (15 per cent). The Munich-based Panavia, which began delivering the Tornado to the national services in the three countries in 1979, has total design responsibilities for the project and directs a program involving 500 companies.

The three-shaft RB 199 turbo fan engines for the Tornado are manufactured by Turbo-Union, a joint company established by Britain's Rolls-Royce, West Germany's Motoren und Turbinen Union and Italy's Fiat Aviazione.

"We have developed a military aircraft which forms a solid base for European defense and a structure for future projects in this field," says Panavia managing director Carl-Peter Fichtmüller.

Like the Airbus and most other collaborative projects in Europe, Panavia participants hope the program will lead to other cooperative ventures. "We would be wasting a great deal of development, management and manufacturing expertise if we did not give serious consideration to expanding the Panavia program," says



Renato Bonifacio, managing director of Aeritalia. "You don't go through the teething operations of a multinational project on this scale without thinking of future variants."

In addition, the Turbo-Union group is seeking engine contracts in different European countries, such as Sweden. "There's no reason Turbo-Union shouldn't be the springboard for a range of engines manufactured by the three participating companies," says Gian Carlo Boffetta, director general of Fiat Aviazione and chairman of Turbo-Union.

While the Tornado is primarily a military aircraft, some of its technical achievements, like the system described below, could have applications in the civil area.

## THE TORNADO SELF-TEST SYSTEM

by Friedrich Muller-Kraenner
Avionics Coordination Manager
Panavia Aircraft

The Tornado was designed in the early 1970's to far more stringent requirements than those applied to other operational weapons systems in Europe. Considerable emphasis was placed on reducing the life cycle cost to the user while developing a complex system for assured missile effectiveness. Requirements included continuation of mission despite defects, fast turnaround time and optimal maintenance.

The mission success of Tornado is dependent to a large degree on the correct performance of a fully integrated and predominantly digital avionic system which has an inherent operational flexibility through redundancies and fail-safe design techniques. On-board sensors are combined with a soft-dependent main computer which coordinates system operations for modes including navigation, weapon aiming and terrain following.

The integrity of all Tornado systems is protected by a self-test system known as On Board Check Out and Monitoring System (OCAMS), which represents the best possible technical solution to detect, identify and locate failures at an early stage. In the event of a safety-critical or mission-critical failure, this system provides the flying crew with a cockpit indication enabling an instant decision. The generated failure data, indicating the affected area, are stored for subsequent representation. This contributes considerably to operation of the aircraft with a minimum dependence on aerospace ground equipment.

The OCAMS is based on three major engineering concepts:

- The provision of Built-in Test Facilities (BIT) within the equipment which operate continuously or interruptively to detect and locate 80 per cent of all faults occurring over a statistically significant period of time down to line replaceable unit (LRU) level.
- The provision of software packages located within the aircraft's 64k main computer using existing interfaces with equipment and operational displays to provide a defect detection and location capability supplementary to the BIT.
- The utilization of dedicated recording facilities for maintenance and defect data for post-flight behavioral analysis.

Through this design it is possible to achieve an effective failure indication and location whereby 80 per cent are directly indicated on the aircraft's maintenance panel and operational displays, down to LRU level. About 15 per cent must be located indirectly through system behavior while the remainder can only be found by visual inspection.

A typical example of the mechanization of the BIT circuitry for a piece of avionic equipment can be illustrated by the Doppler Radar Equipment. Weight and volume for installation do not exceed eight per cent and five per cent respectively for the hardware and the equipment mean time between defects, due to additional complexity, does not increase by more than five per cent. The continuous monitoring of the equipment during all operational modes detects 80 per cent of all defects. By using interruptive BIT (manual initiation of test inputs which conducts a functional test of the installed equipment), a 90 per cent detection rate of all defects is possible. The mean time to test with BIT takes two minutes. The time to repair the aircraft, which includes five minutes for replacement of the defective equipment, is nine minutes. The time to locate and exchange the defective module in the Doppler Radar itself is, with the aid of an Automatic Test System, a further 170 minutes.

The test software operates within the main computer and provides defect detection and location capability for as many links and input/output interfaces as possible.

An integral part of the operational software enables the monitoring of avionic system parameters during flight and stores specific parameters for subsequent replay and analysis should they become invalid. It further displays status information on the navigator's video display and continuously performs a computer self-test. On the ground it can be used to test various autopilot modes. In addition, a dedicated test program replaces the operational flight program and provides the ground operator with a tool for checking interfaces and mode control through indication of their functions on the video display.

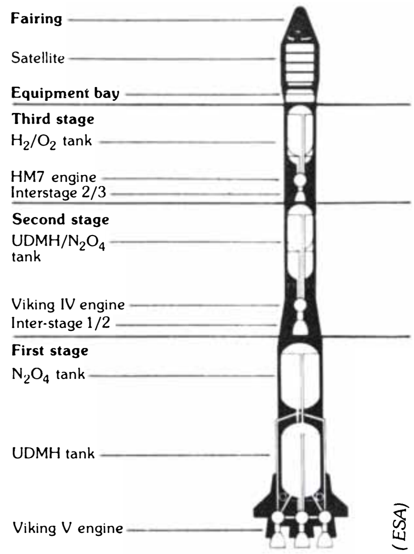
Engine and structural parameters are recorded by the incidence and maintenance recorders for subsequent replay and analysis on ground equipment.

The interface between the Tornado and its operators is designed as simply as possible. The aircrew are warned of any malfunction in the aircraft's system, which affects safety or may force them to alter their mission, by means of a central warning panel in each cockpit and by information displayed on the video display in the navigator's cockpit. These malfunction data are stored on board the aircraft in the central computer and maintenance panel. This allows the ground crew to easily locate and replace the defect LRU. The OCAMS subsequently reconfirms the integrity of the aircraft prior to its new mission. ■

## SPACE: EUROPE'S TECHNICAL FUTURE

European companies, on a joint and competitive basis, are rushing to take advantage of commercial opportunities in space because they believe they can make substantial strides against the Americans who have slackened their efforts since the Apollo program. "Space is Europe's future," says Pierre Morel, associate managing director of the French National Center of Space Studies (CNES). "We've been pushed out of some traditional markets, like automobiles, and the experience has taught us to attempt to take a technical lead in future technologies."

To establish this lead Europe is forging ahead in all sectors of space activity. For example, it is developing a three stage launcher designed to place space systems in geostationary and polar heliosynchronous orbits, and a third test launch for Ariane is scheduled this summer.



(ESA)

*Ariane: Europe's three stage launcher.*

Frederic d'Allest, president of Arianespace, a company responsible for the production, marketing and operating of Ariane, hopes the launcher will obtain a 25-30 per cent share of the world market and lead to a launcher family which will rival the capabilities of the U.S.-built space shuttle.

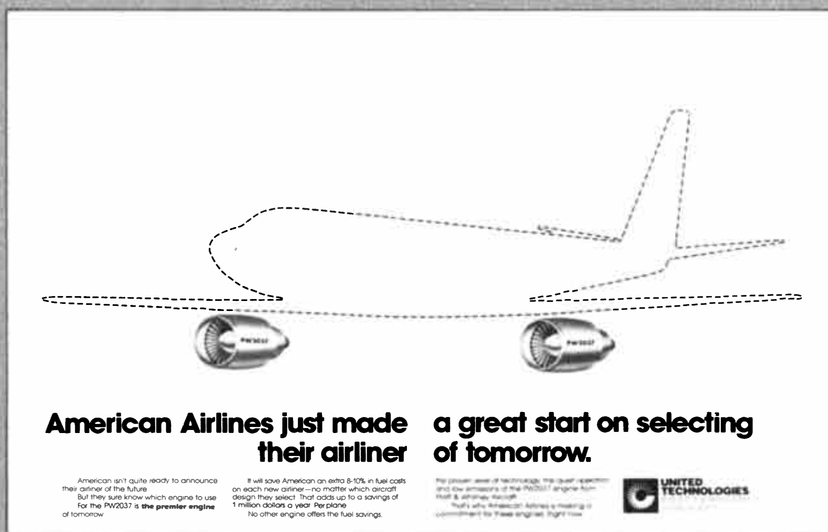
"Two hundred civil, scientific, commercial and weather satellites will be launched during the 1980's and we will have our slice of that market," d'Allest predicts.

The Ariane program, a European Space Agency project with CNES as the prime contractor, involves 50 companies from ten participating nations. The six principal contractors for the development phase are Aerospatiale, the industrial integrator with responsibility for technical management and physical integration of the launcher elements, SEP, Air Liquide and MATRA in France; ETCA in Belgium and Aeritalia in Italy. Arianespace has formed an agreement with Grumman Aerospace in the U.S. to market the launch service and provide technical support to American companies.

The project, like Airbus, enabled companies to develop technical expertise through participation. "The Ariane program, and other space ventures under the auspices of ESA, have enabled us to become specialized in specific areas rather than trying to tackle every aspect of a particular job," says Colin Wearmouth, executive director communications and science, space and communications division, British Aerospace Dynamics Group (BADG).

In fact, most participants agree the European space program has produced a rationalized and concentrated R & D effort, defined common European programs, led to the definition of


Wall Street Journal, December 22, 1980




Wall Street Journal, December 23, 1980

# MTU and Fiat were nearly as happy as we were.

The introduction of the fuel-efficient PW2037 is not just our success. It's the success of a three-way international partnership.

For Motoren-und Turbinen-Union will design and manufacture the low pressure turbine for every PW2037 engine. Fiat will design and manufacture gear boxes and other external engine hardware.

At Pratt & Whitney, we're convinced that international partnerships are the key to the future of the business. And there's no better evidence than the success of the PW2037.

Maybe we should talk.

UNITED TECHNOLOGIES
PRATT & WHITNEY AIRCRAFT

integrated industrial structures and identified areas of commercial profitability. "Ariane gave the Europeans a program with a healthy project leadership, innovative technical requirements and a strong commercial organization," says Hans Hoffmann, managing technical director of West Germany's ERNO. "It's helped us define our own R & D goals: bring down the cost of a specific technology rather than leap from one technology to another." Michel Lasalle, director of space affairs for Thomson-CSF, the French electronics and communications group, adds "In Europe, specialization has become as important as competition."

The second most important European collaborative effort in space is the Spacelab, a multipurpose laboratory which will be carried in the payload bay of the American-built space shuttle and used for a variety of scientific, industrial and research applications. "It is the pilot project for all future manned research in space," says Hans Hoffmann of ERNO, prime contractor to ESA for Spacelab. "On the first flight there will be experiments under weightlessness dealing with metallurgy, composite materials, semiconductor crystals, surface forces and the dynamics of fluids."

A number of satellites, both scientific and commercial, are being developed on a pan-European, bilateral and national level. ESA is controlling a variety of programs including the European Communications Satellite for television relay, the Maritime European Communications Satellite, the Exosat designed for X-ray research in outer space, and L-Sat, a multi-role communications satellite with direct broadcast capabilities. On a bilateral level, the Germans and French have formed Eurosatellite to create a direct broadcasting satellite network. It will be developed and manufactured by the four Eurosatellite participants: MBB and AEG-Telefunken in West Germany, Aerospatiale and Thomson-CSF in France.

"Space technology in Europe has reached the stage where some countries will go it alone," says ESA's Collette. "In this case collaboration has spawned competition."

In any event, Europe's ambitions, from a technical and commercial viewpoint, are fairly clear. Concludes Pierre Morel of CNES, "We must attempt to close the European space market to Americans, we must participate in all international projects which affect us and we must obtain a fair share of the market in developing countries."

CNES will be doubling its R & D budget during the next five years concentrating on high frequency traveling wave tubes and transistors, platforms and control systems, digital electronics, new structures and remote manipulation.

The ESA budget for 1981 totals approximately $850 million from eleven member countries and three other participants. The funds are distributed as follows: communication programs 27.7 per cent, Spacelab 19.1 per cent, Ariane 17.2 per cent, scientific programs 15.9 per cent, and earth observation 5.3 per cent, with the remainder allocated to the general budget.

## HELICOPTERS: EXPORTING TECHNICAL EXPERTISE

Although there is still no pan-European helicopter project underway, the principal manufacturers (Aerospatiale, MBB, Agusta in Italy and Westland Aircraft in the United Kingdom) have agreed to establish a joint R & D program while the four governments plan to rationalize their future helicopter requirements.

In addition, there is a great deal of bilateral discussion between the different manufacturers. Westland and Agusta have created EH Industries to develop an anti-submarine helicopter for the navies of Italy and Britain. Aerospatiale and MBB are collaborating on an anti-tank helicopter program.

"The idea of a European helicopter program is not a dream and we certainly coordinate with each other," says Agusta president Pietro Fascione. "The overlapping R & D programs and the fact that we all need large export markets could prompt further cooperation, "

Aerospatiale, which exports 80 per cent of its helicopters, describes the type of R & D occurring at its facilities near Marseilles.

## IMPROVING THE HELICOPTER

by Rene Mouille
Associate Director of Helicopter Studies, Aerospatiale

Aerospatiale's Helicopter Division has established four main objectives to further improve the efficiency of its military and civilian helicopters. These include R & D in the following areas:

- Aerodynamic characteristics of rotors and fuselages will be improved to have higher take-off weights, higher speeds and lower fuel consumption.
- Components will be improved technologically to obtain reduced weights and costs, improved noise and vibration level, and reduced vulnerability of military aircraft.
- Operating costs will be reduced by a cut in production costs, simple maintenance routine and high reliability.
- Systems will be better integrated in military and civilian helicopters.

These long term programs follow breakthroughs in other areas which play a major role in helicopter performance. They include:

- The aerodynamic characteristics of the blades where research has led to the definition of a wide range of blade profiles —thin profiles for blade tips, thicker profiles for blade roots — permitting optimized aerodynamic characteristics.
- The aerodynamic characteristics of the fuselage where fuselage shapes have been streamlined, landing-gear made retractable and interference drags (rotor-fuselage interference drag especially) reduced.

The most dramatic improvement, however, has probably been achieved in the field of helicopter component technology.

Composite non-metallic blades, made of glass or carbon fibers coated with epoxy resin, offer numerous advantages including: the best aerodynamic optimization achieved through molding operations; the best dynamic optimization achieved through possible evolutions in weight and stiffness; an exceptional fatigue strength, sensitivity to notches and corrosion which is practically nil; an intrinsically fail-safe nature of the blade, due to the fibrous materials used, resulting in a high reliability; and greatly reduced fabrication costs.

Glass, kevlar or carbon composite materials are used widely for secondary structures. In primary structures the headway made by composite materials is slower. Although they do not always represent the best alternative, they are of great interest because of their weight and price for very variable shapes.

Progress has also been made in the field of helicopter power-plants due to the advent of new turboshaft engines with better performance, lower fuel consumption, smaller overall dimensions and lower noise level. Moreover, installing them on aircraft led to fruitful research results in the field of inevitable power loss in air intakes and at the exhaust.

Finally, there has been a progressive improvement in the integration of on-board systems: electric, hydraulic or automatic stabilization systems with a possible coupling with radionavigation.

## ENGINEERING THE AERO ENGINES

The only European engine manufacturer capable of competing with the Americans is Rolls-Royce in the United Kingdom. It has collaborated with German and Italian companies in Turbo-Union, but is struggling to maintain its role as the largest civil engine manufacturer outside the United States. Although 50 per cent of its sales are military engines, Rolls-Royce has had a difficult time in the civil market. It has not, for example, placed one of its engines on an Airbus.

But the company is continuing its R & D effort in numerous areas, from basic improvement in component technology to engine demonstrator programs. Rolls-Royce engineers are working on high temperature demonstrator units (hot end technology), quiet engine demonstrators, mixed engine demonstrators and advanced core engine technology.

"If we add up all of the improvements, we can anticipate a 13 per cent fuel saving in our engines by the middle of 1988," says John Bush, Rolls-Royce chief engineer, advanced engineering.

In addition, Rolls-Royce is contemplating one major development program.

## THE SUPERSONIC VERTICAL AND SHORT TAKE-OFF AND LANDING (V/STOL) AIRCRAFT

by W. J. Lewis
Assistant Chief Engineer, Advanced Projects, V/STOL, Rolls-Royce Limited

Jet lift was first proposed in the United Kingdom in 1941 and consequently is almost as old as the jet engine itself. Exploratory development of the concept was underway in the late 1940's and the 1950's produced a number of research vehicles which led to various operational prototypes in the early 1960's.

Two notable examples of these prototypes were the German EWR/Sud VF101 and the French Mirage IIIV—both of which flew at supersonic speeds, as well as demonstrating the ability to take-off and land vertically.

Despite the interest that supersonic V/STOL has long generated within the aeronautical industry, no-one has succeeded in configuring a vehicle which combines a short field performance with supersonic flight capability—at a

sufficiently low cost and performance penalty relative to a conventional aircraft to convince the military customer that V/STOL is worthwhile.

As a result, the subsonic Harrier built by British Aerospace and powered by the Rolls-Royce Pegasus vectored thrust engine is the only operational V/STOL aircraft in the Western world today.

Although the Harrier/AV8A has been in operational service for over 10 years, the payload/range performance is still regarded to be poor in some quarters, despite the special circumstances for which both the Royal Air Force and the U.S. Marine Corps need it. This payload/range penalty—real or imagined — is probably one of the major reasons why the AV8B has not yet been fully funded.

Today, interest in supersonic V/STOL combat aircraft is at a higher level than it has been for some time. Some of the reasons for this are the existence of runway denial weapons and the high in-flight performance requirements for modern combat aircraft, which eliminate the need for oversized or additional engines solely for take-off and landing.

The supersonic conventional combat aircraft, without exception, has a compact, high specific thrust power unit achieved by reheating a turbojet or turbofan engine. The supersonic V/STOL aircraft has the same powerplant requirements in-flight–but needs two additional features. These include a deflecting of the thrust to provide lift and an increasing of the thrust of the basic engine for the take-off and landing phases of flight.

This thrust increase can be provided in a number of ways, in addition to the obvious one of deflecting the reheated thrust of the basic engine, thereby eliminating the complication and cost of an additional system and minimizing the volume of the powerplant.

The different methods of providing this thrust boost all have to make use of either heat addition—that is, increasing jet temperature and velocity, air flow addition or a combination of heat and flow addition.

A wide variety of engineering methods has been studied with varying degrees of success.

While the heat addition method offers the lower penalty in powerplant volume, the perhaps weight, the disadvantages are a high fuel flow during take-off and landing and possibly some operational restrictions due to the hot, high pressure exhaust jets impinging on the ground. The flow addition method has a low fuel flow—no increase over that of the basic engine—and a much less adverse ground jet footprint. However, because very large airflows at low velocity have to be handled, it is a high volume system and is very sensitive to pressure losses.

Current thinking at Rolls-Royce is in favor of thrust augmentation for take-off and landing by heat addition, and builds on the existing experience of the vectored thrust Pegasus engine.

This has led to the concept of augmenting the thrust of the bypass flow by burning fuel in the ducts leading to the front vectoring nozzles. Such a combustion system has been labeled Plenum Chamber Burning (PCB) and has the advantages of being viable in flight as well as take-off. This minimizes powerplant volume and avoids too great an increase in complexity over the existing system in the Harrier.

Opponents of the concept claim it is not compatible with good supersonic performance. These criticisms arise because performance estimates in the past have been based on the assumption that PCB would be added to the existing Pegasus engine.

Although this is completely feasible, it must be said that the Pegasus is a subsonic engine rated to give the maximum possible thrust at take-off and International Safety Association conditions. Project studies have indicated that a new engine of the same airflow as the Pegasus could increase the thrust by over 30 per cent at take-off and by more than 60 per cent at M=1.6.

Technology studies into the problems associated with the use of PCB have been carried out for several years by Rolls-Royce and solutions have been found to all the critical questions.

The stage has been reached where the real need now is for a demonstrator aircraft to study the operational aspect associated with high temperature exhaust flows.

After a successful demonstration, the way should be clear to achieve the ambition of many engineers in the aeronautical industry and meet the requirements of the services—an operational supersonic V/STOL combat aircraft.

## THE TECHNICAL FUTURE

Throughout Europe companies are attempting to reduce the weight and fuel consumption of their aircraft through innovations in avionics, propulsion, aerodynamics, structures and materials. A few examples which illustrate the scope of the activity include:

- A Control Configured Vehicle program at MBB, near Munich, where artificial stability is created in an aircraft by special control systems. This technology will make an essential contribution to increased performance in future military and civil high-performance aircraft.
- Lucas Aerospace in the United Kingdom is continuing the development of digital fuel controls which are being applied in the military area as quickly as they are developed.
- A Dutch-German built versatile low speed wind tunnel located in the Netherlands.
- A consultation system for technical documentation developed by the French company Sogitec, permitting technical publications, including illustrations, to be stored in a computer and remotely consulted in real time.
- The creation of Hellenic Aerospace Industry in Greece with the goal of making that country the major aerospace maintenance and overhaul center for civil and military aircraft in the eastern Mediterranean.
- The design evaluation in Sweden of a new light strike fighter which will replace the Saab Viggen and Draken aircaft in the 1990's.
- The development of a new technology wing by Dornier in West Germany to increase the performance and improve the economy of a new utility and commuter aircraft which goes into production later this year.
- The formation of Italian companies under the umbrella of Aeritalia to carry out subcontract work on the A310. These include Aeronautica Macchi, Agusta, Caproni Vizzola Fiat Aviazione, Piaggio and Siaimarchetti.
- A new short-to-medium range aircraft seating around 140 persons designed by Fokker using an increasing amount of composite materials.
- A new version of CASA's C101 advance trainer will be designed in Spain.

## THE FUTURE OF COLLABORATION

Naturally not every European collaborative project works as smoothly as Airbus, Ariane or Tornado.

One example of a collaborative problem is the Jaguar, the low level strike aircraft manufactured by the French (Dassault-Breguet) and British Aerospace. The aircraft competes with Dassault's Mirage aircraft and the French company obviously prefers to sell the latter and receive 100 per cent of the commercial benefits.

In addition, some Europeans are admittedly cool about collaboration and believe it has become as much a religion as a solution. Says Jack Pateman, managing director of Marconi Avionics, Europe's largest aircraft electronics company and a major participant in the United States aircraft products market, "We've been involved in most European cooperative projects and our basic conclusion is that when possible we should go it alone."

But there is little question that the success and necessity of collaboration has sparked Europe's commercial and technical spirit. And Europeans now anticipate increased American participation in some of their programs. "Americans are coming to Europe because they need our money, technical expertise and production facilities," says Bengt Eriksson, executive vice president of Sweden's Volvo Flygmotor. "We anticipate more transatlantic joint development and production programs." Says Carlos Marin, CASA's deputy chairman, "We are devoted to all forms of international cooperation and would like to see the formation of a pan-European organization which could initiate transatlantic collaborative programs."

Adds Aeritalia's Renato Bonifacio, whose company is participating with Boeing on the 767, "European collaboration is important but we must do business where we can. Although it may take a while, the future in most areas is going to be transatlantic."

Europeans do not, on the whole believe equal collaboration is yet possible with their American counterparts and they envisage a coordination of R & D activities on a scientific level prior to large-scale commercial ventures.

"We waste valuable funds by duplicating efforts on space research," says Morel of CNES. "There has been some good bilateral work done between the Americans and European nations but it's time to seriously evaluate joint research projects."

In the end, it could be commercial, rather than scientific, factors which spark increased American interest in Europe.

Says Alan Buley, chief executive officer of the Saab-Fairchild 340 project, a Swedish-U.S. venture to manufacture a twin-engine 34 seat pressurized turbo prop, "Americans should realize they can benefit by collaborating with Europeans to obtain a slice of this market. They should realize it's time for the two way street."

# How can you consult a million pages of documentation without actually handling it?

The technical literature on today's sophisticated equipment represents large continuously-growing and changing volumes of information.

To create, process and recall this information, extremely powerful data and graphic processing systems are required.

Sogitec has again pioneered in this field with the development of the SI 3000, the world's first system that permits local or remote real time access of textual or illustrated technical literature stored in a central computerized data bank.

The unique graphic and textual capability of the SI 3000 makes possible on-line access to the complete range of an equipment support documentation such as, spare parts catalogs, maintenance manuals which can be customized with user data ie; work cards, operating vouchers, etc.

Two pages of text and/or illustration appear simultaneously on the display screen. Should you require a hard copy, a high-resolution graphic printer copies out the desired pages on paper.

The SI 3000 has obvious advantages over paper or microfilm:
– extremely easy access to information
– updates available immediately throughout the world
– smaller volumes of paper or micro-supports to publish, distribute and handle
– reduced operating costs.

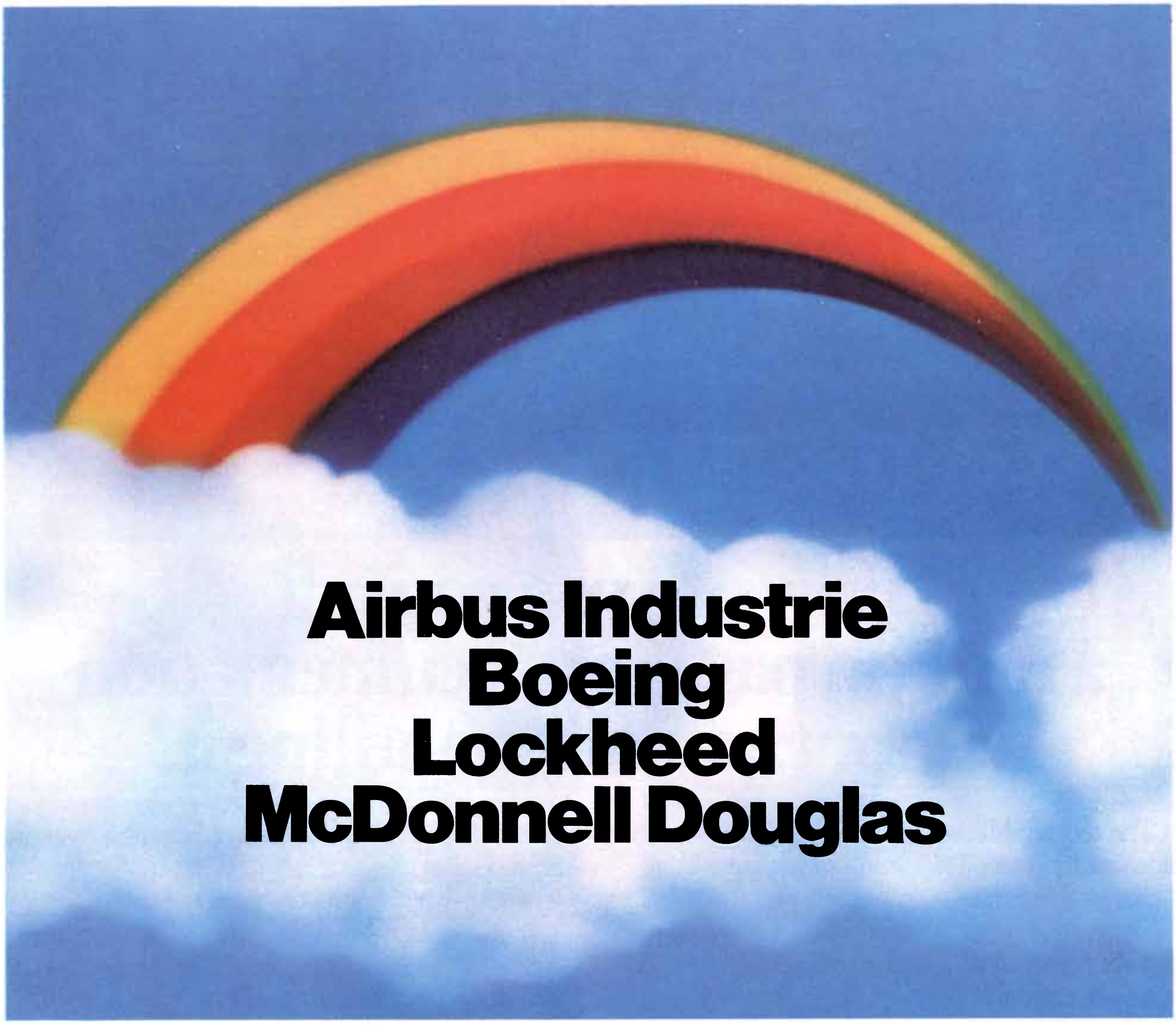
Advanced techniques of processing of technical data is only one of Sogitec's activities.
In addition, we design and produce audiovisual teaching aids to facilitate the transfer of technology. We also design and manufacture simulators for training personnel in the use of complex aviation equipment.

Logistic support to industry.
We have been doing this job for over 16 years. It is one to which we are passionately devoted and in which we have always been innovators. In this way, we seek to justify the confidence that manufacturers have placed in us, and to provide even better service to users.

SOGITEC

Product Support Services

SOGITEC S.A. 27, rue de Vanves, 92100 Boulogne France. Tél.: (1) 609 91 01.
SOGITEC INC. 1801 Dove Street, Newport Beach, California 92660. Tel.: (714) 955-3432.
SOGITEC GmbH. Bonn Center HI 1104, Bundeskanzlerplatz, 5300 Bonn 1. Tel.: (228) 21 50 17.




# The order is more than just alphabetical

In the last two years, Airbus Industrie's market share of short/medium-range twin aisle aircraft surpassed their competitors.

Total sales of the A300 and A310 now exceed 450 aircraft to 40 airlines. These airlines have chosen the Airbus concept for its true wide body comfort, exceptional cargo capability, remarkable fuel efficiency and sheer profitability. In short, a proven formula already hard at work for over 20 airlines worldwide, backed by the combined strength of Europe's entire aerospace industry.

Little wonder that the A300 and A310 are the number one twin aisle twins.


**Airbus Industrie**

# BOOKS

## *Fantastic tales of old China, the ailments of urban man and avant-garde telescopes*

by Philip Morrison

THE JOURNEY TO THE WEST, translated and edited by Anthony C. Yu. Volumes 1, 2 and 3; Volume 4 in preparation. The University of Chicago Press ($25 per volume). Fifty years ago Sir Aurel Stein loaded the saddlebags of his camels with drinking ice for a winter trip across the arid Gobi. His adventures on central Asian tracks from India to China became famous, but long before him other men of learning had undertaken the perils of that trip. The annals name a hundred earlier pilgrims: Chinese monks who traveled similar routes both ways, seeking the dharma in the West, in that India from which the lessons of the Compassionate Buddha had reached out to engage Chinese faith. Their travails spanned four centuries, the last pilgrim returning as an old man to China in A.D. 789.

One such scripture pilgrim, neither the first nor the last, was the monk Hsüan-tsang, called Tripitaka. He set out under the second T'ang emperor, to return years later to Ch'ang-an in A.D. 645, bearing nearly 700 items, "three baskets of Mahāyāna scriptures." But his story alone, not without the quality of myth, was fused into a new substance, an intoxicating elixir of fantasy, audacity, satire and overreaching adventure within the crucible of the Chinese mind over almost 1,000 years. These folktales and popular texts were newly worked in about 1580 (probably by Wu Ch'êng-ên, an otherwise obscure scholar of Kiangsu) into a 100-chapter narrative, half epic and half Ming novel, poetry embedded in prose, with the high art of an unchallenged master.

It is those 100 chapters we are now being given for the first time in English. (Several truncated versions have appeared, one of them still available in paperback, a single volume by the gifted Arthur Waley with the title *Monkey, Folk Novel of China.*) But never before have we had a chance to approach the real thing, "one of the four or five lasting monuments of traditional Chinese fiction." This *Journey* is alive today both in China and overseas, wherever inked characters are read. The skillful paper cutouts offered in China for a few coins often show the silhouettes of the four companions of the journey. The Peking Opera plays their story regularly; the action drawn out in modern cartoons fills four thick paperback volumes in comic-book format for sale in China and Hong Kong. An issue of colorful stamps of the People's Republic celebrated the *Journey* in 1979.

How can an ocean of tales be sampled? The arduous trip is turned into a magical epic, a combat not only against mundane thirst and endless rocky paths but also against a diverse multitude of monster spirits and demon kings. No human heart, not even that of the good, if timid, Tripitaka, could survive alone. He is accompanied by three supernatural bodyguards—disciples and intercessors—with a white horse that is an errant dragon in disguise. Elder Brother among the three reprobate immortals is the half-repentant Monkey King, Sun Wu-k'ung, born of a stone and self-styled as a Great Sage, Equal to Heaven. His all but boundless powers, his ceaselessly fertile invention, his daring irreverence, his cocky self-assurance and his feisty combativeness make him at once the hero rebel, the sly, ugly-faced trickster and the invincible warrior. He, like the others, bears his allegorical burden lightly; he is the Monkey of the Mind, at once the glory and the peril of our cruel and marvelous species.

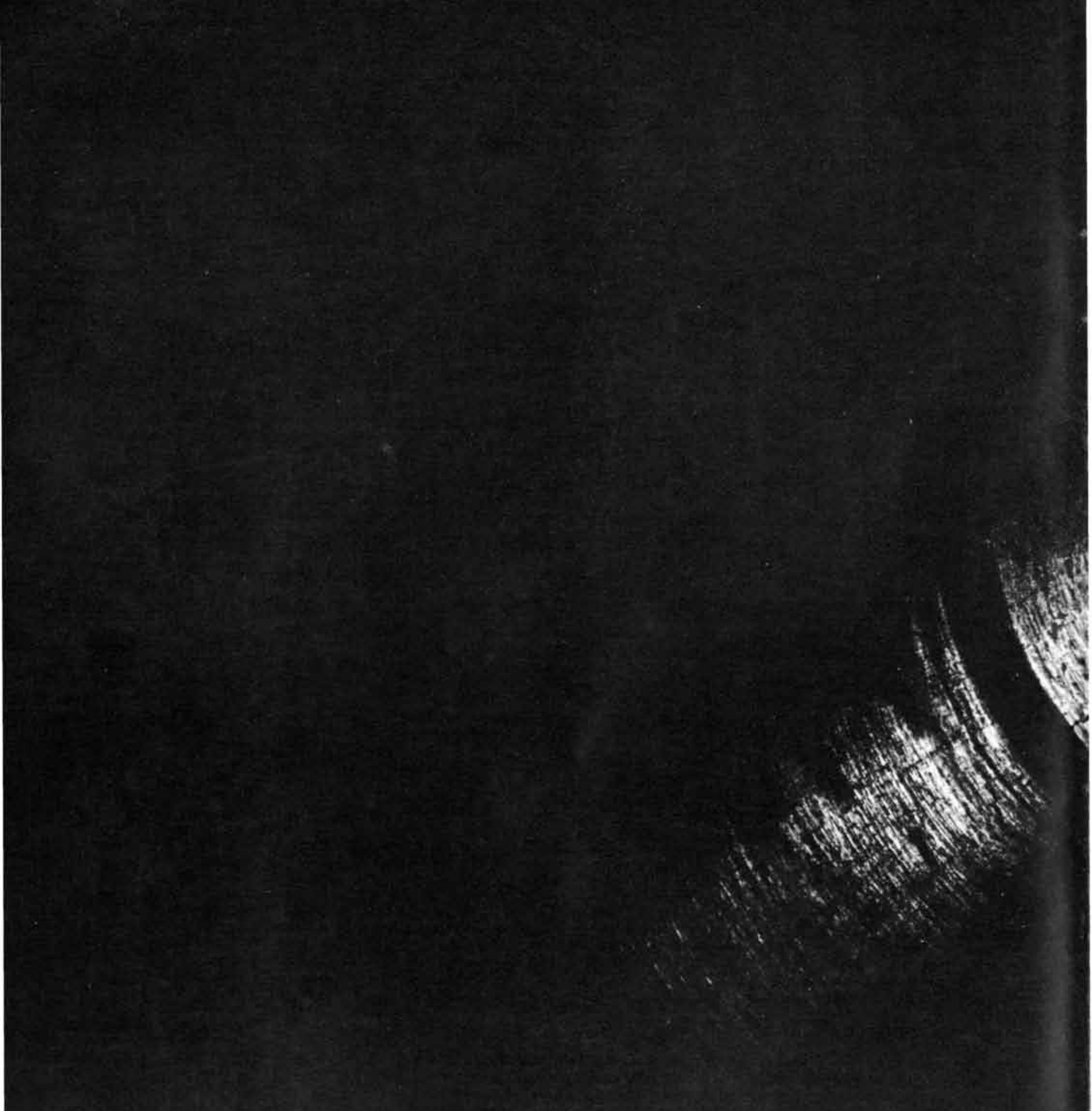
Would you read of strange instruments? Try the imp-reflecting mirror, the pineal phoenix eye, the Compliant golden-hooped iron rod, which can grow from an embroidery needle concealed in Monkey's ear into a weapon as thick as a rice bowl or even to the height of a mountain. (There is evidence that once it was the axis of the earth.) Would you find a surprising turn of events? The good monk and his second disciple Piggy drink of the cool, clean river. Half an hour later, what a stomachache! "Their bellies began to swell.... Inside their abdomens there seemed to be a clot of blood or a lump of flesh... kicking and jumping wildly about." It had been the Child-and-Mother River in the Nation of Women, and the men were pregnant, although they were "without any birth canal." "That water your master drank was not the best," says an old woman in some amusement.

Is it mortal combat and all the martial arts you seek? This story is the very fount of kung fu; the play of sword and rod, fire and wind, cloud somersaults and wild stratagems is unending. Magical transformations? A single hair of Monkey is often turned into a fighting little monkey ally, and in need an entire handful can be torn out, at once a loyal army. Cormorants, crabs and flies are familiar guises for Monkey; indeed, such efforts are no newer than the old Sumerian tales. But to become a pangolin? Or what of the occasion when old Monkey changed "into a little temple for the local spirit. His wide-open mouth became the entrance, his teeth the doors, his tongue the Bodhisattva and his eyes the windows. Only his tail he found to be troublesome, so he... changed it into a flagpole." There never was such a temple with a flagpole behind it; that architectural eccentricity gave him away.

The scenes of ever fresh quandary and conflict are played out under the engaged gods above. Those rulers include the magical theocracy of Buddhism, the stars in their constellations and the masters of the deep lore of Taoism. There is a well-staffed heavenly bureaucracy in T'ang style, from grooms and gardeners to the examiners, magistrates and ministers of the Celestial Court.

Professor Yu, in a crisp vernacular, lively and lighthearted, never affected or cute, sets it all down for a flood of pleasurable reading. Hundreds of passages of poetry, mainly scene descriptions or fast-paced chronicles of exalted combat, are here to be enjoyed. The personages we encounter span this world and the next; we meet, among a few dozen other demons and monsters of character, the terrible young demon king Red Boy (the master of true fire), a passel of thunder squires, a false Monkey (really a magical six-eared macaque), Lao Tzu and Kuan-yin, Mars (the Star of Fiery Virtue), the Divine King Water Lord of the Yellow River (ready to drop on any adversary from his chalice the waters of the entire stream), the monster Clean Iguana and the comic little fish spirit Busy Bubble. There are many more, friends or foes of Pilgrim Sun Wu-k'ung.

Plenty of learned footnotes explicate the allusions that tumble past modern readers, particularly those unfamiliar with the culture; without pedantry the notes often cite the sources in Chinese. Alchemy, astrology and medicine are all sketched out as Monkey uses them or deals with their adepts. Great Sage undertakes to treat the Western monarch, King of Scarlet-Purple Kingdom, whose grave illness has defied all local thera-

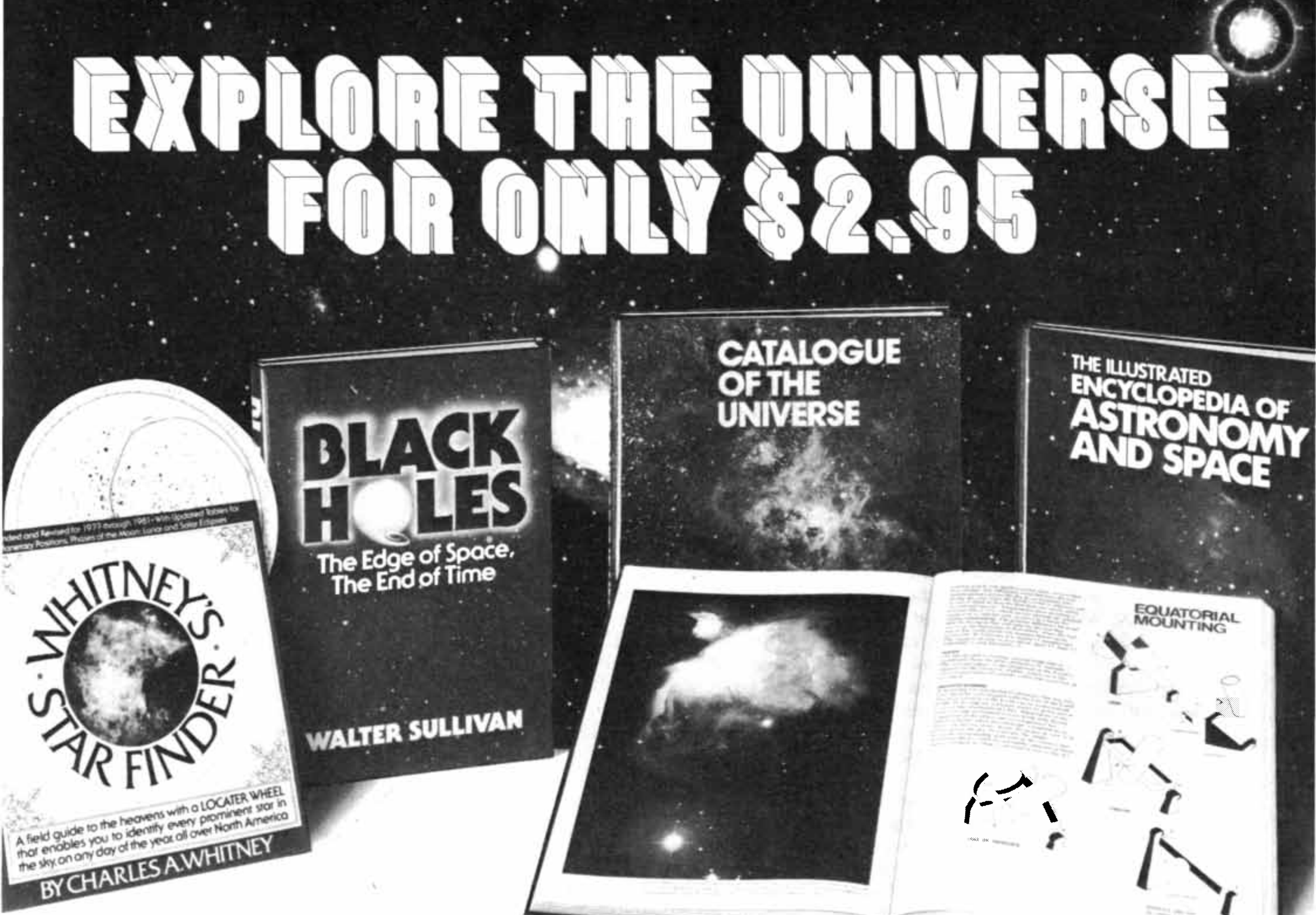


pies. He invokes the *Classic of Medical Problems* and applies the rare technique of taking the pulse by the feel of three long threads of gold, each tied to a point on the king's pulsing wrist, as the unapproachable patient sits up in his dragon bed. Pilgrim Sun studies the revealing signals with meticulous care, making "his own breathing regular." The diagnosis is full, and Great Sage's pharmacology offers the cure. Sun orders three pounds each of all the 808 flavors of medicine, both raw and cooked, along with all the mortars and pestles he may need. This is a professional ruse, to guard his prescription by misleading the imperial college. Actually only two flavors and soot are used to roll three large pills. The active ingredient is horse urine, to be sure, taken from a loud-talking horse that was "originally a flying dragon." Sourceless water is required for taking the pills, and a dragon rain obliges. In a little while the king is literally purged of his illness: it is a complete cure and a hyperbolic look at ancient medical ritual.

One volume is still to come; it will end the pilgrimage in victory, surely freed from illusion. Dear Great Sage! His diamond pupils see all; his spirit is marked by an irrepressible "heroic gall." We last hear him, at the end of Chapter 75, speaking from within a demon who has swallowed him. Undaunted, the Monkey inside threatens to use his portable frying pan from Canton to grill liver, chitterlings, stomach and lungs, so preparing the authentic southern dish *tsa-sui:* a chopped-up miscellany. By sea change four centuries later it is our American chop suey!

DISEASE AND URBANIZATION, edited by E. J. Clegg and J. P. Garlick. Symposia of the Society for the Study of Human Biology, Volume 20. Distributed in the U.S. by the Humanities Press, Atlantic Highlands, N.J. ($22.25). The mosquito *Aedes aegypti* was once an African forest species; it has now followed man everywhere in the Tropics. Only recently has it struck inland from the coasts of Southeast Asia. Adapted to life near its numerous mammalian prey, it has come to prefer human blood and to place its eggs in tin cans, old tires and household water jars instead of the "ancestral tree-hole"; an old adaptation to survive the dry season in the egg serves still in the new domain. *A. aegypti* has long been caught up with our own fate: it is the chief vector of yellow fever. That virulence, however, has never been endemic in tropical Asia, perhaps because no monkey host is there in the jungles nearby. What the mosquito carries instead is the virus of a less mortal disease, dengue hemorrhagic fever, which affects mainly children.

Throughout bustling Bangkok about 800,000 containers of water support the




# Take this 4-volume Basic Astronomy Library for $2.95 (a $60.80 value)

if you will join the **Astronomy Book Club**. You simply agree to buy 3 more books—at handsome discounts—within the next 12 months.

Black holes, quasars, pulsars, supernovae, white dwarfs—explore the fantastic discoveries being made about the universe with this outstanding selection of four critically acclaimed astronomy books. Together they form a marvelously comprehensive Basic Astronomy Library—a library which will allow you to tour our cosmos—from the Red Planet to the Seyfert galaxies . . . from the Earth's core to a black holes horizon . . . from how the universe was born to how, when, and where it will die.

Here is what your library will include:

**WHITNEY'S STAR FINDER.** *Charles A. Whitney.* Top-notch field guide to the heavens. Enables you to identify every major star in the sky over North America. Comes with a Star Finder-Locater Wheel. Publisher's price $6.95.

**BLACK HOLES: The Edge of Space, The End of Time.** *Walter Sullivan.* Bottomless pits into which whole stars can disappear . . . *Black Holes* takes you on an incredible journey that will change your picture of the universe. Hardbound. Publisher's price $17.95.

**CATALOGUE OF THE UNIVERSE.** *Paul Murdin et al.* Three-volumes-in-one: an atlas, encyclopedia and book of descriptive astronomy. Surveys over 220 astronomical objects. More than 300 photographs. Publisher's price $17.95.

**THE ILLUSTRATED ENCYCLOPEDIA OF ASTRONOMY AND SPACE.** Revised edition. *Edited by Ian Ridpath.* Your window to the universe. Over 1,000 entries and 150 photographs and illustrations. Hardbound. Publisher's price $17.95.

. . . quite a trip for only $2.95—and that's just the beginning!

### What is the Astronomy Book Club?

**Astronomy Book Club** is a very special club—a book club operated exclusively for astronomy enthusiasts. It's devoted to making the finest astronomy books available conveniently and inexpensively. Recent offerings include titles by such eminent authorities as Carl Sagan, Robert Jastrow, John Gribbin, Paolo Maffei, and Isaac Asimov.

As an **Astronomy Book Club** member, you'll receive books of your choice at savings up to *30% off publishers' prices.* In addition, you'll immediately become eligible for our Bonus Book Plan, with savings of 70% off the publisher's price

Discover the phenomena of space—join **Astronomy Book Club** today.

If reply card has been removed, please write to: **Astronomy Book Club**, Dept L-AJ5. Riverside, N.J. 08075 to obtain membership information and an application.

mosquito's larval stage. The plates holding flowerpots and the ant traps on the legs of every kitchen cupboard harbor a few larvae, but almost three-fourths of all the mosquitoes were reared in the same big earthenware jars that store household water. In Jakarta indoor concrete tubs for bathwater serve the same role. These containers are domestic necessities: piped water is in irregular supply in the marginally equipped, fast-growing cities. Lids do not fit tightly enough; regular emptying is too much to ask; the householder must be persuaded against common sense to contaminate his clean water with an insecticide.

It seems, however, that first of all the people must accept responsibility. In Singapore it was enforced on them by law; it worked, with a small, homogeneous population and area and an efficient staff. Enforcement of such a law seems far away in most of the rest of Southeast Asia. The lesson of this remarkable piece of ecology is coolly drawn; we have many volumes of relevant studies of mosquitoes but not much on their human partners. Dengue will spread until "studies of vector management are supplemented with studies of man management."

Take a second example, close to the volume's title. Coronary heart disease is by popular consensus a disease of the affluent city dweller, particularly if that person is sedentary, a smoker and over-indulgent in a steak-and-butter diet, rich in saturated fats. Maybe; there is some evidence in the indictment, but it is not yet a true bill. An up-to-date survey in England and Wales shows quite the reverse: among five occupational classes the most affluent suffer the least mortality from the coronary-disease complex. Among British civil servants the same trend holds: men in the lowest grade die of coronaries four times oftener than those in the highest grade. The dwellers in cities large and small in England and Wales show no trend with city size, although overall there is some relief from the condition in rural areas. In transitional Puerto Rico and in Yugoslavia the city folk are at twice as much coronary risk. But in largely rural China, as in some studies of Polynesian migrants to New Zealand cities, the urban tendency to hypertension rises sharply: it is almost four times higher in the city than it is in the country.

There is good reason to infer that the pattern of coronary disease is changing with time; people simply adapt. Those most newly exposed to urbanization are at the highest risk. Just what this adaptation is remains a puzzle; not more than half of the effect can be blamed on blood pressure, fatty diet, smoking, cholesterol or obesity—which is not at all to exculpate those well-supported hazards. There are no national exceptions to the rule that high coronary disease goes with high saturated-fat diet. The converse, however, is not true; coconut eaters on an isolated atoll show low blood cholesterol in spite of a high intake of saturated fat. There is more going on; one idea is that not the higher fat intake but an inadequate fiber intake from rich foods is the key, but data remain scant.

We do not yet understand what it is that has let the affluent begin to cut their coronary risks. It may somehow be the effect of widespread exercise, stopping the cigarettes and cutting down on eating. Or it may be quite other, subtler processes. Until we do learn we can expect a wave of coronary disease to pass through all the classes of our population. One graph here strikingly, although the data are not new, plots the rate of lung-cancer mortality in 1950 against national cigarette use per capita in 1930. Eight countries supply the data points, from Canada and Switzerland to the four Scandinavian lands. A straight line fits the data with a coefficient of regression of a neat 95 percent. We do know something.

This tight, small volume presents 11 such review papers, nearly all by British authors, each paper examining some health issue in its full context, with the data drawn from vital statistics all the way from poor Ugandan villages to Copenhagen. The diseases range from acute infections to stress-induced hormone levels, from multiple sclerosis to a wide variety of cancers. The main method of the epidemiologist is contrast; the symposium, diverse but not synoptic, was a series of contrasts: poverty and affluence, urban and rural, tropical and temperate. The conclusion is not hard to draw: the easy ecological successes of the past—the new fly screen, the removal of old tires from the backyards—will have fewer counterparts. The problems we understand now are two-sided; for example, the more affluent expect more gratification, and so the incontrovertible hazards of smoking wane very slowly, outside the force of public law.

European infants long ago began to receive steel-milled cereals, a richer, more nutritious diet, accompanied by dairy products and sugar. These high-energy foods carried them through the infectious attacks of infancy. The Newcastle two-year-old, even 30 years ago, averaged only two illnesses in the second year; the frail Ugandan village child in 1970 had five times more episodes of illness (quite apart from malaria, which by itself was present more often than all the English infections). Immunization, cleanliness, good water and even protein are not likely to explain the contrast. It is the refined, energy-rich foods that have replaced the coarse paps and porridges of the village that may make most of the difference. Perhaps we pay later in the coin of appendicitis, colonic cancer and coronary disease for the early ability to fight off infection. Yet economic development has overall clearly brought a longer life span, if we are not overtaken by our worst potential social epidemic, nuclear war at long range.

OPTICAL AND INFRARED TELESCOPES FOR THE 1990s, edited by Adelaide Hewitt. Proceedings of the Kitt Peak National Observatory Conference, 1980. Kitt Peak National Observatory, Tucson, Ariz. ($35). This century, the century of the galaxies, opened with our biggest telescopes about a meter in aperture. Collecting area increased as the mountaintops were occupied until 1948, when the Palomar five-meter started operation, its conception then 20 years old. Even now only one larger telescope exists, the Russian six-meter in the Caucasus, still in its break-in stages. Photon-collection area has thus gone up some 30 times in the 20th century, and new detector techniques have by themselves improved performance by an even larger factor. In 1984 we hope to see the first good-size general-purpose telescope in space, the Space Telescope, a shuttle-launched free-flier in earth orbit, its elegant mirror 2.4 meters across.

Our National Observatory on Kitt Peak was host to a conference early last year that gathered optical and infrared astronomers and engineers from many lands to seek out the future of bigger ground-based telescopes. The optical and infrared bands are the regime of atomic and molecular energies, still the channels central for the analysis of motion and composition even in this age of radio, X-ray and gamma-ray astronomy. The bulky report is swollen by the quick-publication format, chiefly double-spaced typescript reports of some 75 papers, together with verbatim discussion and plenty of graphs and pictures. One learns of experience with the newest telescopes, new techniques of construction and of detection, the scientific problems appropriate for the new instruments, the overall technical issues in the interplay of scale, form and cost, and a good deal of detailed and ingenious technology, both at the level of methods and materials and at that of optical design. Unanimity is not at hand.

The central dream for a decade hence is a mountaintop telescope some 20 meters in diameter. Such an instrument will handle the crucial spectroscopy of faint objects much faster than the space telescope, and it will image a wider field as well. The cost of the big telescopes scales pretty proportionately with their weight; early experience gave a still steeper curve. Pound for pound optical telescopes cost almost three orders of magnitude more than their big radio counterparts; that is the price of optical

# Mothers ought to eat their spinach too.

As people get older, even eating properly can become something of a trial.

Foods that used to be favorites don't sit well anymore.

It's a hassle just going shopping for food.

And somehow, the idea of preparing a homecooked meal when there's nobody else at home isn't very appetizing.

With more people living longer in America today, we can't ignore problems like these.

That's why the nutritional needs of our senior citizens will be on the agenda at the White House Conference on Aging.

And it's why the food people at ITT began a major three-year program on the subject.

We're funding research to learn more about the diet that older people need to stay healthy.

And we're working on new ways to get the latest information out to them. Where it can be useful.

When you think about all they did to make us eat right, we have to do something to make sure they eat right.

We have to make their later years better years.

**The best ideas are the ideas that help people.** ITT precision. Today's techniques, however, much lighten optical telescopes. Computation makes it possible to design mechanical supports—perhaps air pads—to maintain the figure of mirrors much thinner than the heavy disk of Palomar. The focal length is kept short, so that the tube dwindles; a thin mirror can be sagged into the deep-dish form. The lightweight mirror keeps the mount light; complex computer controls track the moving sky more easily with one vertical and one horizontal axis than can the geometrically simple but structurally demanding tilted axes of the classical clock-driven analogue mounts. The dome is kept small too; indeed, such a dome, rather like a gun turret with an open face, is working well with the new multiple-mirror telescope of the Smithsonian Institution and the University of Arizona. Around such a lightweight mirror a cheaper observatory might grow, its cost intermediate between the experience with radio and classical optical observatories, say $100 million for a 20-meter telescope, still low compared with placing a big mirror in orbit.

The newest detectors are marvels. In space and on the ground the photons are caught in a pure silicon disk, where every photon between wide color limits frees an electron for conduction. (An ordinary photoelectric surface is never that efficient; if an electron starts the wrong way, it never leaves the surface.) The usual microingenuity of the silicon world has found more than one way to sort out these electrons; in one device clock pulses drive the electrons along conduction channels in the disk by a kind of electric peristalsis, taking them in order from more than half a million distinct little picture elements almost without loss.

It is air that is the enemy of the ground-based telescope. Clouds and upper-air motions are obvious, but in addition the detail of the image is severely degraded not only by the air of "the dome and its contents" but also by the motions of the air at the mirror surface itself. Rigorous uniformity of temperature must be sought. Half a dozen renderings in these two volumes show a varied set of conceptual designs for big telescopes. There is much meat here for the browser, and more of course for those who would follow the arguments closely. The volumes end with a lively discussion that involves all astronomy, its aims and its social structure. Is one large, powerful, costly device worth more than many smaller ones, three-meter telescopes "for the people"? The answers are not easy.

If anything seems lacking in this rich stew of ingenuity, experience and (more doubtful) projection, it is enough attention to the possibility of combining the output of many small telescopes by novel electro-optical techniques of the future. Optical observatories might come increasingly to resemble radio arrays. It does look as though a large, enduring observatory in orbit, a few pounds of thin aluminum film exquisitely backed, controlled and arranged over an acre or so, will mark the maturity of optical science, although perhaps the 1990's are too soon for such ultimates.

CORRIDORS OF TIME: 1,700,000,000 YEARS OF EARTH AT GRAND CANYON. Panoramic photography and text by Ron Redfern, illustrations by Gary Hincks, introduction by Carl Sagan. Times Books ($55). A view of the Grand Canyon is above all else a grasp of space, of volume, of empty air—perhaps it holds a soaring hawk, or even a plane lost in the scene—given boundaries by the strong and colorful architecture of the layered rock. A photograph can, if it is artful enough, suggest that depth. It cannot, however, stretch to the sides, filling the vision, rewarding every turn of the gaze with new richness. This remarkable book seeks to transcend the camera's limits.

The pictures are genuinely wide as well as deep. Some of the sharp color photographs stretch across an entire four-page gatefold, a view almost 20 times the area covered by a standard 35-millimeter lens. They are not distorted by wide-angle optics or made by a lens revolving about an axis. Rather, they are meticulously built up from many shots, the mosaic all but seamless, the color balance and exposures remarkably uniform, the fruit of years of the most painstaking work in the field. Redfern, a man of intensity, ingenuity and taste who does not quite tell all about his techniques of camera mount and computation, has carried it off in triumph. Some of his photographs appear in eight-by-36-foot color blowups as the focus of a canyonlands exhibition now traveling among museums in the U.S. and Britain.

The chief theme is geology, the geology of the entire uplifted Colorado Plateau. Bryce Canyon, Zion and the Little Colorado are all treated with the same wide view and clarifying text. Redfern is of course moved by geologic time, as the title conveys, but he does not lose in all the plate tectonics, the paleontology and the erosional unconformities and ancient peneplains the remarkable story of the recent canyon cutting itself. His account of that inference is more up-to-date than other lay sources offer, with strong new field support for reversed flow through the present Little Colorado channel, which five million years ago was great Lake Bidahochi.

Around the bright core of these extraordinary photographs, in sun and shade, in snow and summer heat, the author has built quite a comprehensive volume. Widely seen are not only the geology but also the local fauna and flora, the archaeology and the people of today, both the indigenous and the intrusive. The history of exploration is here, with the feel of the boatmen old and new shooting the rapids. Only the elaborated tourism of the canyon rims, the cars, the hotels and the crowds is left in shadow.

The freshest of all the discussion is the author's sensitive account of his predecessors of just a century ago, the English romantic painter Thomas Moran, who was a follower of J. M. W. Turner, and W. H. Holmes, whose marvelous drawings evoke the very panoramas Redfern has captured in his mosaics. "It is extraordinary that Moran was selected by scientists to convey the scenery. [He] made no pretense that he even tried to convey accurately geological detail, and [John Wesley] Powell . . . must have known it. But what a stroke of genius it was to capture the spirit of the canyonlands through one artist and convey the strict and necessary detail of them through another." It is a true delight to compare the intricate Holmes vista from Point Sublime, reproduced as the endpapers of this big volume, with comparable views from the modern panoramic camera.

VOYAGER 1 ENCOUNTERS SATURN. Jet Propulsion Laboratory, National Aeronautics and Space Administration. Government Printing Office ($3). Just the size and gloss of the souvenir program you are offered in the lobby of most theatrical spectacles, this 40-page brochure (authorship corporate) commemorates the gala performance of November, 1980, with a certain panache. The magazines showed us some of the color plates on a still shorter deadline, but the Jet Propulsion Laboratory has put together a rich Saturnian gallery. The departing view of Saturn, the black shadow thrown unprecedentedly toward the viewer across the bright ring plane, the astonishing complexity of the rings, the big crater that nearly splits the little icy moon Mimas, plenty of atmospheric features—the memorable shots are all here.

JPL wit and erudition are manifest in the epigraph, just about the aptest science-fiction forecast ever to enrich a NASA publication. It is from Voltaire's philosophical novel *Micromégas,* published in 1752. The text is given in two languages; the English reads: "Our voyager knew marvelously the laws of gravitation, and all attractive and repulsive forces. He used them in such a timely way that, once with the help of a ray of sunshine, another time thanks to a cooperative comet, he went from globe to globe, he and his kin, as a bird flutters from branch to branch." Kincraft *Voyager 2* will reach Saturn in late summer; thence it will flutter out to the globe Uranus by 1986.

CHIVAS REGAL • 12 YEARS OLD WORLDWIDE • BLENDED SCOTCH WHISKY • 86 PROOF • GENERAL WINE & SPIRITS CO., N.Y., N.Y.

Father's Day 1974

Father's Day 1975

Father's Day 1976

Father's Day 1977

Father's Day 1978

Father's Day 1979

Father's Day 1980


Father's Day 1981

# Patience does have its rewards.

Established 1845 SCIENTIFIC AMERICAN June 1981 Volume 244 Number 6

# Gas-cooled Nuclear Power Reactors

*Although the U.S. has only one such reactor, they have served well overseas. They have an attractive safety feature: a loss-of-coolant accident such as the one at Three Mile Island is all but impossible*

by Harold M. Agnew

In March, 1979, the nuclear power industry suffered a shock from which it has not yet recovered: the accident that disabled one of the nuclear reactors at Three Mile Island. It is ironic that an event that caused no discernible physical harm to anyone crippled the prospect for expanding nuclear power at the very time the nation was becoming generally aware of the need for new domestic sources of energy.

Although the experience at Three Mile Island demonstrated to the satisfaction of technically qualified people that present-day water-cooled nuclear reactors offer no significant threat to the health and safety of the general public, it also showed that such accidents and equipment failures can jeopardize the operability of the plant and place at risk the heavy capital investment it represents. In the extreme case an accident such as the one at Three Mile Island can threaten the financial survival of the operating utility.

Perhaps the principal lesson of Three Mile Island is that the current generation of nuclear power plants is vulnerable to certain rare events that can lead to a condition where the time available for responding correctly can be less than a minute. In such low-probability events if the appropriate actions are not undertaken immediately, the consequences can be extremely costly even when public safety is not at issue. It is reasonable to ask: Do we need to be content with nuclear reactors of a design such that operators must react correctly within a minute in order to prevent damage to the reactor? The answer is no.

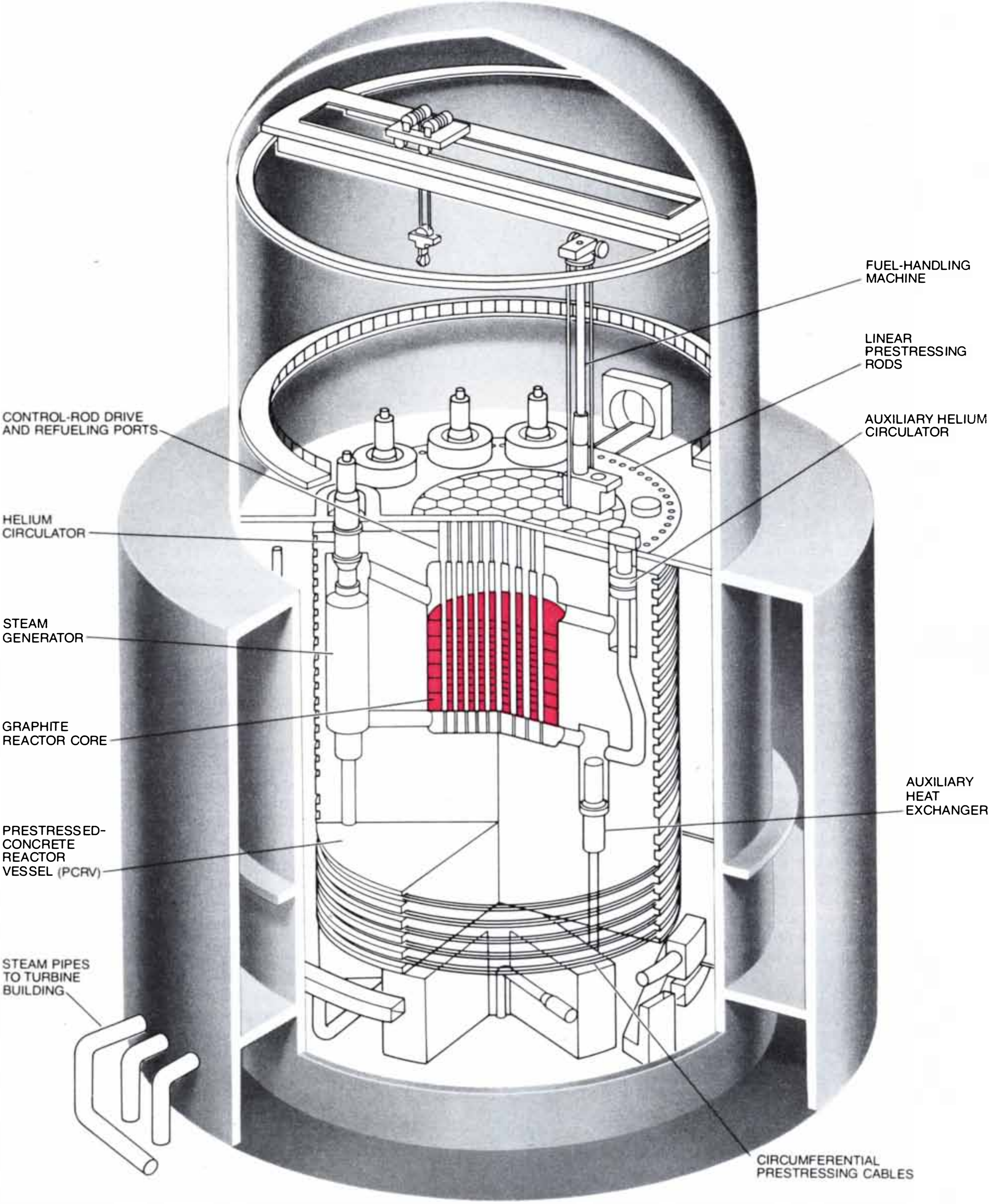
That being the case, how did the U.S. nuclear power industry come to follow the path it did? The dominance in the U.S. of the light-water reactor has a simple explanation. The pressurized-light-water reactor is a straightforward adaptation of the highly compact reactor designed to propel the first nuclear-powered submarine, the U.S.S. *Nautilus,* launched in 1954. An electric-power version of the submarine reactor, built by the Westinghouse Electric Company, went into service at Shippingport, Pa., three years later. The General Electric Company soon introduced a reactor design of its own, the boiling-water reactor, in which the heat generated by nuclear fission was carried away from the core by steam rather than by pressurized hot water.

In both types of reactor it is essential that the reactor core not be uncovered, even briefly, lest the temperature in the core quickly rise and melt the metal jackets around the fuel pellets, as indeed probably happened at Three Mile Island. Light-water reactors are equipped with redundant safety features to cope with a "loss of coolant" accident. In such accidents the emergency equipment is designed to flood the core with water from a plentiful and assured source. When the normal coolant flow was interrupted at Three Mile Island, a sequence of improbable events, including apparent operator error, interrupted the delivery of the emergency cascade of water for too long a time.

All but one of the 71 commercially licensed and operating nuclear power plants in the U.S., which currently supply about 11 percent of the nation's electric power, are light-water reactors. The exception is a helium-cooled reactor, the Fort St. Vrain Nuclear Generating Station, which was accepted for service in the summer of 1979 by the Public Service Company of Colorado. The plant's rated capacity is 330 megawatts of electric power, or MWe, which is about a third the output of a standard commercial power plant. The reactor has been operating at up to 70 percent of rated power and has recently been released by the Nuclear Regulatory Commission for testing at up to full power.

The Fort St. Vrain demonstration plant was designed and built for the Public Service Company of Colorado by the General Atomic Company as a part of the Atomic Energy Commission's Power Reactor Demonstration Program. This followed the successful operation of a 40-MWe prototype, Peach Bottom Atomic Power Station No. 1, on the system of the Philadelphia Electric Company. In its seven and a half years of operation, from 1967 through 1974, the Peach Bottom reactor was available for service 86 percent of the time (except for scheduled shutdowns related to the research and development objectives of the reactor itself). The comparable figure for all U.S. nuclear reactors is about 66 percent.

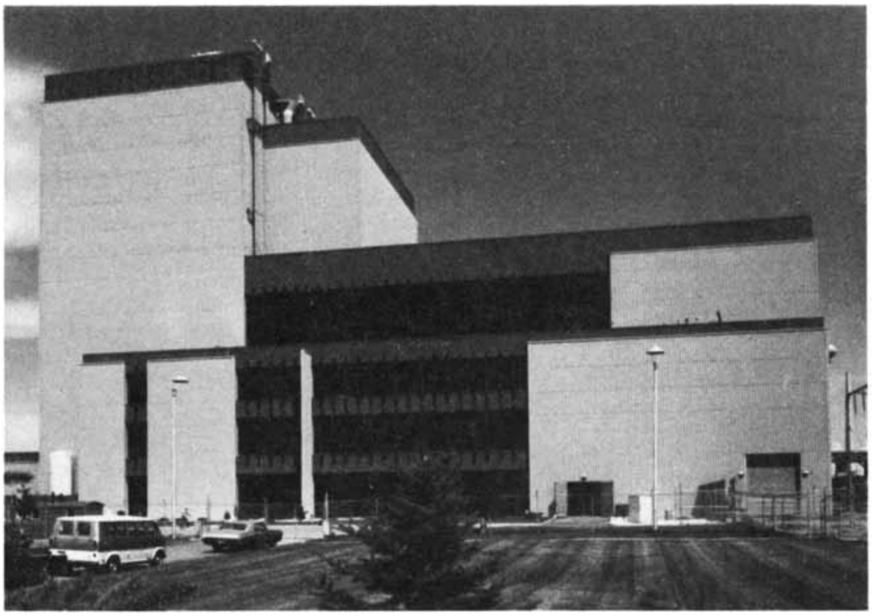
The key safety features that differentiate the helium-cooled reactor from water-cooled reactors are two. First, since the reactor core is cooled by a circulating gas completely confined within a massive reactor vessel, the reactor cannot lose its primary coolant because of a rupture of pipes outside the vessel. Second, if the circulation of the gas is interrupted by some mishap to all of the main helium-circulation system, the temperature within the reactor core rises only slowly because the fuel elements are embedded in a massive matrix of graphite, which serves as the moderator for slowing down neutrons and which can absorb the heat released by fission products after the nuclear chain reac-




**HIGH-TEMPERATURE GAS-COOLED REACTOR (HTGR) is a second-generation system more efficient than the 71 light-water power reactors that now supply about 11 percent of U.S. electricity. In this HTGR designed by the General Atomic Company the moderator (the material that slows neutrons in the reactor core) is graphite and the coolant is helium. In light-water reactors ordinary (but demineralized and conditioned) water serves both as the moderator and as the coolant. The HTGR shown would have an output of 860 megawatts of electricity (MWe), slightly less than that of the largest power plants, which generate 1,000 MWe or more. The HTGR has a thermal efficiency of 38.5 percent, which is comparable to the efficiency of the best fossil-fuel plants and is higher than the 32 to 33 percent attained by current light-water reactors. Because the core of the HTGR contains nearly 1,500 tons of graphite, which has a high capacity for absorbing heat, an HTGR is much less likely to be damaged than a light-water reactor if there is an interruption in the flow of coolant or a loss of coolant. It was such an interruption that caused the accident at the Three Mile Island nuclear power station near Harrisburg, Pa. The reactor core and steam-generating system of the HTGR are housed in a prestressed-concrete vessel with walls some 15 feet thick.**

tion itself has been halted. In a helium-cooled graphite-moderated reactor the nuclear reaction is halted by the insertion of control rods, similar to those in all reactors, or alternatively by the injection of small boron-containing balls that "poison" the reaction. In water-cooled reactors the loss of the coolant, which also acts as the moderator, stops the reaction.

If the emergency cooling systems in a light-water reactor should fail to function, the temperature in the reactor core would rise even after the reaction had been shut off because the fission products accumulated in the fuel elements would continue to release energy at a high rate. At the instant of shutdown, decay heat amounts to about 7 percent of the rated thermal output of the reactor, or about 210 megawatts in a water-cooled plant with a thermal rating of 3,000 megawatts (equivalent to an electric output of 1,000 MW). It is estimated that in such a loss-of-coolant accident the temperature of the cladding around the fuel elements would reach 3,000 degrees Fahrenheit and fuel failure would begin in as little as 50 seconds in a pressurized-water reactor and in less than two minutes in a boiling-water reactor. With a helium-cooled reactor, in a comparable event involving a system depressurization and the total failure of the helium-circulation system, more than an hour would be required for the temperature inside the core to reach 3,000 degrees F. At that temperature both the coated fuel particles and graphite fuel elements in a helium-cooled reactor would not be affected. The fuel particles and graphite can readily withstand temperatures of up to 4,000 degrees F., which would not be reached until at least 10 hours had elapsed. In short, there is ample time to institute a variety of reasoned emergency measures for restoring the flow of helium coolant.



**ONLY HTGR IN THE U.S. is near Denver, Colo. It is the Fort St. Vrain Nuclear Generating Station, designed by General Atomic and owned and operated by the Public Service Company of Colorado. The plant, which has a capacity of 330 MWe, was placed in operation three years ago and has since supplied more than two billion kilowatt-hours of electricity. On several occasions the forced circulation of coolant in the reactor core has been interrupted for periods of as much as 15 minutes without doing any detectable harm to the core or to the fuel elements.**

The virtues of gas-cooled graphite reactors have been widely recognized elsewhere in the world. In the 1950's and 1960's, when the U.S. had committed itself to light-water reactors, Britain and France developed gas-cooled graphite-moderated reactors, in which the coolant was carbon dioxide rather than helium. Britain now has more than 40 gas-cooled reactors in operation or under construction, France has seven and Italy, Spain and Japan have one each. More than 600 reactor-years of operating experience has been acquired with the European gas-cooled reactors. Such reactors have accounted for nearly a fifth of the total nuclear power generated in western Europe, Japan and the U.S. so far.

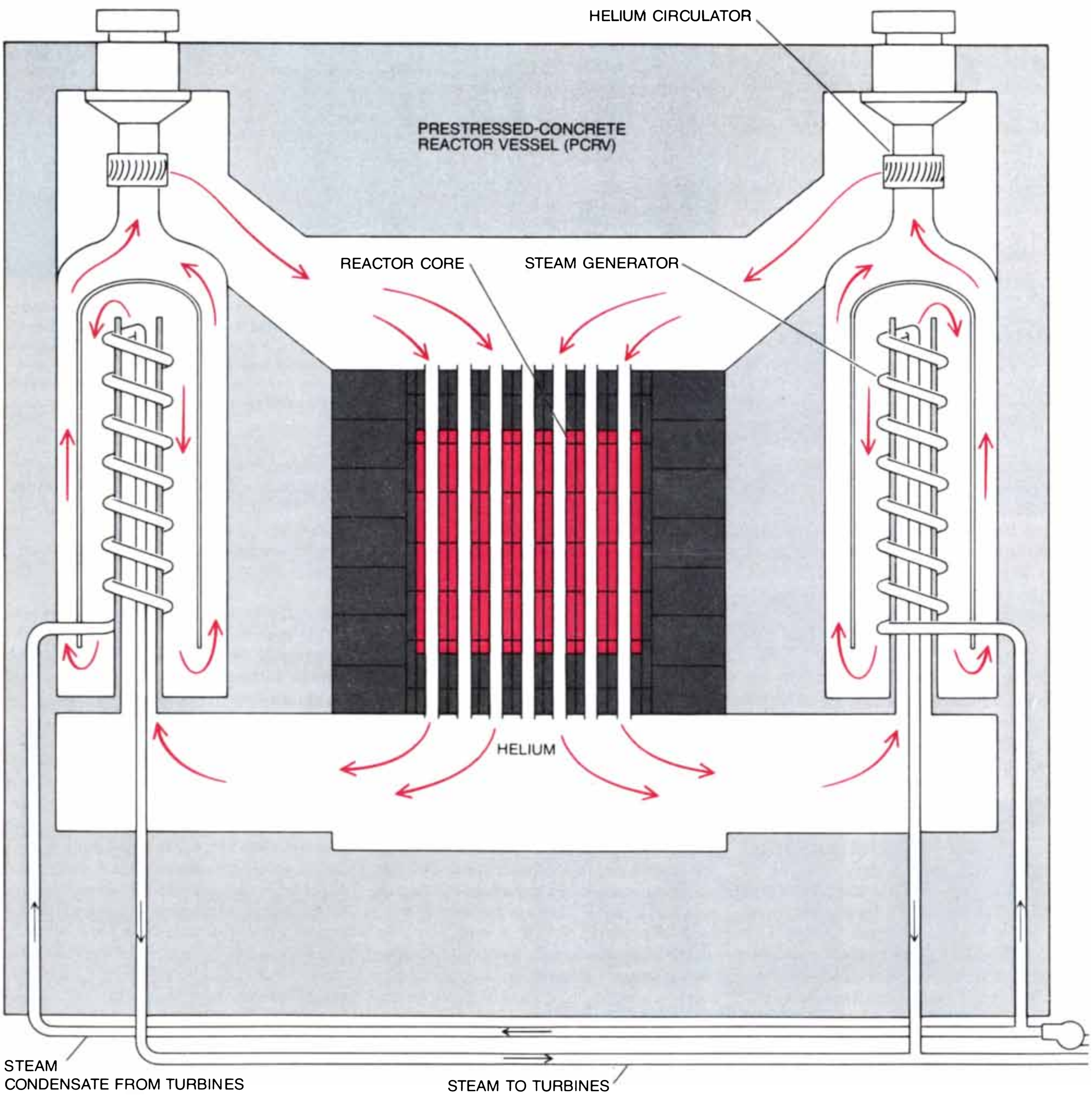
The British and French efforts were at an early stage in 1956 when a group of physicists, many of them with experience at the Los Alamos Scientific Laboratory, gathered at La Jolla, Calif., to consider the problem of designing a reactor that would be both more efficient and inherently more "forgiving" than the reactors then available. Among those present were H. A. Bethe of Cornell University, Freeman J. Dyson of the Institute for Advanced Study, Peter Fortescue of the Atomic Energy Research Establishment at Harwell in England and Frederic de Hoffmann, who was then president of General Atomic. Out of these early deliberations, aided by concepts from Britain and France, evolved the concept of the high-temperature gas-cooled reactor, or HTGR, tested at Peach Bottom and on a larger scale at Fort St. Vrain. Because the U.S. has plentiful supplies of helium, that gas could be selected as a coolant instead of carbon dioxide. Helium has the important advantage that it is stable to the high radiation flux in the reactor, does not become radioactive, is chemically inert and has excellent heat-transfer characteristics.

The attractive features of HTGR's were summarized by Joseph M. Hendrie, chairman of the Nuclear Regulatory Commission, in testimony before a congressional subcommittee in March, 1980. Such reactors, he said, "have efficiencies as good as the best fossil-fuel plants and are substantially more efficient than the water-cooled reactors. They not only get better thermal efficiency but also get better energy utilization out of each pound of uranium that is mined, better, in fact, by probably 15 or 20 percent than the best estimates for advanced light-water fuels." He added that HTGR's "have some safety advantages. They are machines in which you don't have to do a lot of things in a hurry if something goes wrong because the core structure is a great massive pile of graphite, a very high-temperature and stable material, so that if you get a power dropoff or the plant circulators go out, [you have time] to sit down and think about what to do."

The first series of gas-cooled reactors built in Britain were called Magnox reactors because the fuel rods, which contained natural unenriched uranium, were clad in a magnesium alloy. The reactor core, incorporating many tons of graphite, was housed in a large and expensive steel pressure vessel many times bigger than the pressure vessels needed for light-water reactors. Then in 1958 French engineers showed that the steel vessel could be replaced with a vessel of prestressed concrete that could be constructed in sizes large enough to house the entire reactor system, including the steam generators. The prestressed-concrete reactor vessel, or PCRV, is kept in compression at all times by a network of redundant, tensioned steel tendons that can be monitored and retensioned or even replaced if necessary. Tightness against leaks is ensured by a steel liner affixed to the inside of the PCRV, which acts only as a membrane seal to contain the coolant. The liner and the walls of the PCRV are cooled by water circulating through tubes that are welded to the outer surface of the vessel.

PCRV's were subsequently adopted for all French and British gas-reactor systems. The high degree of safety afforded by the concrete vessel contributed to the British decision to construct a second generation of reactors known as advanced gas reactors (AGR's) near urban sites. In this second generation the fuel was uranium oxide, a ceramic, clad in stainless steel, a change made possible by the adoption of slightly enriched uranium. With the new fuel AGR's could operate at higher temperatures than the Magnox-fueled reactors and were able to "burn" more of the uranium 235 in the fuel before refueling became necessary. With higher temperatures the efficiency of electric-power generation was raised from about 30 percent to a little more than 40 percent.

In the U.S. the Atomic Energy Commission (a predecessor agency of the Department of Energy) nurtured interest in gas-cooled reactors in the 1950's and 1960's by supporting the study of several advanced reactor concepts. One of the AEC's main objectives was to reduce the amount of uranium required




**REACTOR CORE AND STEAM GENERATORS of the HTGR are enclosed in a massive prestressed-concrete reactor vessel (PCRV). For a reactor designed to generate 860 MWe the PCRV would be 102 feet in diameter and 95 feet high. (A pressurized-light-water reactor of slightly larger capacity is shown at the same scale in the illustration on the opposite page.) The graphite core of the 860-MWe HTGR fills a cylindrical volume 26 feet in diameter and 21 feet high. Helium at a pressure of 1,050 pounds per square inch is circulated through some 27,000 vertical channels in the core by four primary circulators, of which only two appear in this cross section. The helium emerges from the reactor core at a temperature of 1,266 degrees Fahrenheit and enters the base of the steam generators, where it makes two passes over an array of helical and straight steam coils. Water boils upward through the coils and is further heated as it passes downward to emerge as superheated steam with a temperature of 1,000 degrees F. and a pressure of 2,500 pounds per square inch. Not shown are three coolant loops in which water-cooled heat exchangers can remove heat from circulating helium when the steam-generating loops are out of service. After a reactor shutdown fission products in the core release heat at a rate that is high initially but declines exponentially.**

per unit of electric power; at that time uranium resources appeared scarce in relation to the projected needs. As a result the study emphasized reactor concepts that were either breeders or advanced converters. A breeder creates at least one atom of new fuel for each atom of fuel consumed. Advanced converters generally create from .7 to one atom of fuel for each atom consumed. Light-water reactors yield between .5 and .6 atom of fuel for each atom consumed. The high-temperature gas-cooled graphite-moderated reactor qualifies as an advanced converter. It was one of the designs that survived the inevitable weeding out. The HTGR had strong support from the utility industry because it is competitive in capital costs with light-water reactors and because it exploits a uranium-thorium fuel cycle with a low uranium consumption and therefore low fuel costs.

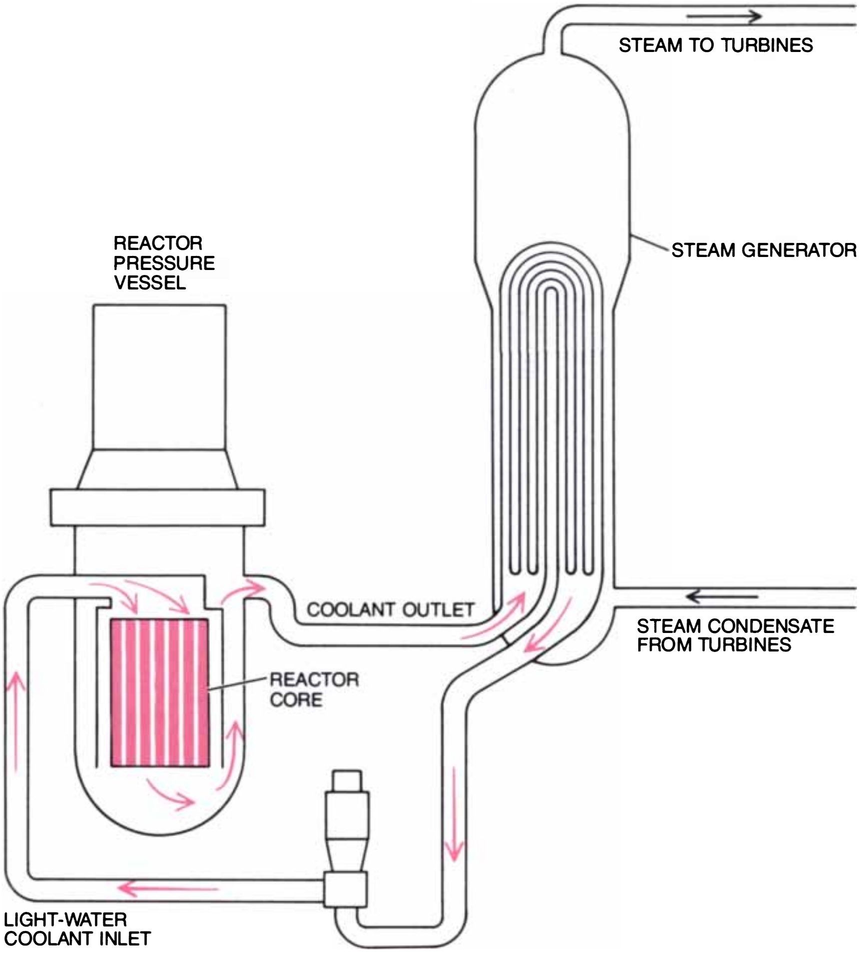
The continuing evolution of gas-reactor technology in Europe and the U.S. has led to a convergence in at least two important particulars for the next stage in the development of gas-cooled reactors. Helium replaces carbon dioxide as a coolant and the reactor core is charged with nuclear fuel in a unique system that dispenses with the need for a metal cladding. The two features have been demonstrated not only at Peach Bottom and Fort St. Vrain but also in two European reactors. The British operated a helium-cooled 20-MW thermal test reactor in southern England from 1965 to 1976. In Germany an HTGR of 15 MWe (called the AVR) has been generating electric power since 1967, with the outlet gas temperature being as high as 950 degrees Celsius. (The temperature of water leaving the core of a pressurized-light-water reactor is about 610 degrees F., or 321 degrees C.) A 300-MWe plant based on the AVR experience is now under construction in Germany and is scheduled for start-up in 1984 or 1985. In the U.S. the Fort St. Vrain reactor of 330 MWe has provided more than two billion kilowatt-hours of power since 1978 and has demonstrated the fuel performance and safety characteristics of a contemporary HTGR design. The reactor has been subjected to test transients up to and including the complete loss of forced-coolant circulation with no adverse effects on the reactor core or on other primary components of the system.

On the basis of the Fort St. Vrain experience General Atomic, in cooperation with Gas-Cooled Reactor Associates (an organization of U.S. utility companies) and the Department of Energy, has developed a reference design for an HTGR of 860 MWe. The goal has been a design that is simple and conservative and that places high emphasis on the safety and protection of capital investment. The reactor core is contained within a multicavity prestressed-concrete reactor vessel. Helium leaves the core at 1,266 degrees F. (reduced from 1,494 degrees F. at Fort St. Vrain) and passes through four primary coolant loops, where steam is generated at a temperature of 1,000 degrees F. and a pressure of 2,500 pounds per square inch.

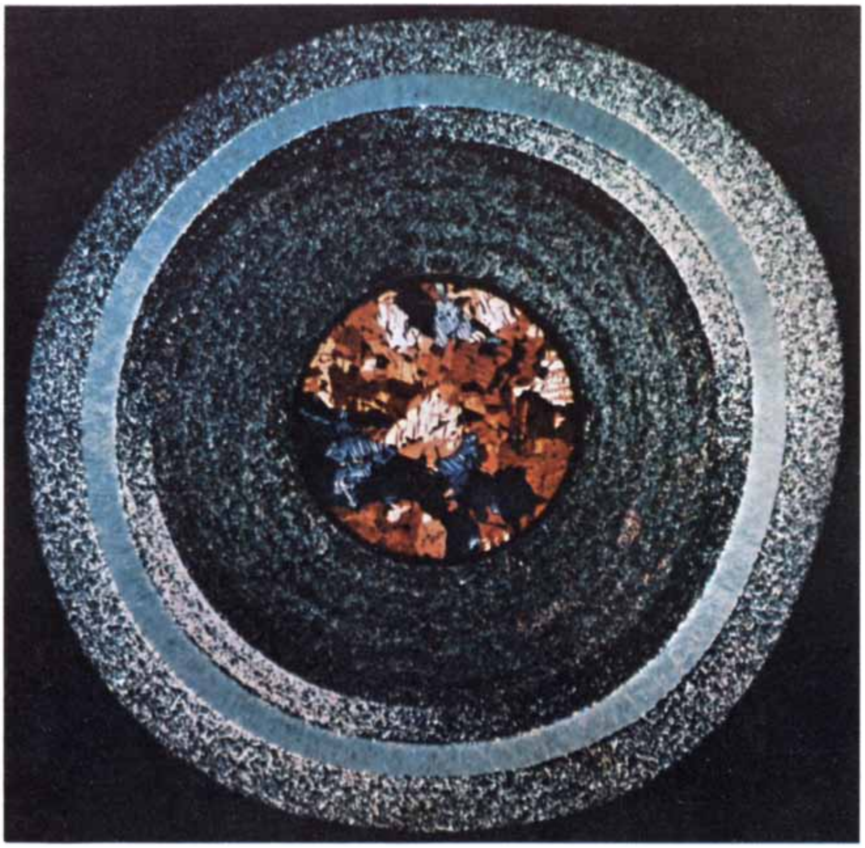
Helium is forced through each coolant loop by a circulator driven by an electric motor. (The Fort St. Vrain circulators are driven by steam.) The core also is provided with an auxiliary cooling system consisting of three loops, each sufficient by itself to deliver 100 percent of the required cooling when the helium in the reactor vessel is at the normal working pressure of 1,050 pounds per square inch, or 50 percent of the cooling when the vessel is depressurized. The helium that passes through the auxiliary cooling system is cooled with water circulated by electrically driven pumps that can be powered, if need be, by diesel generators.

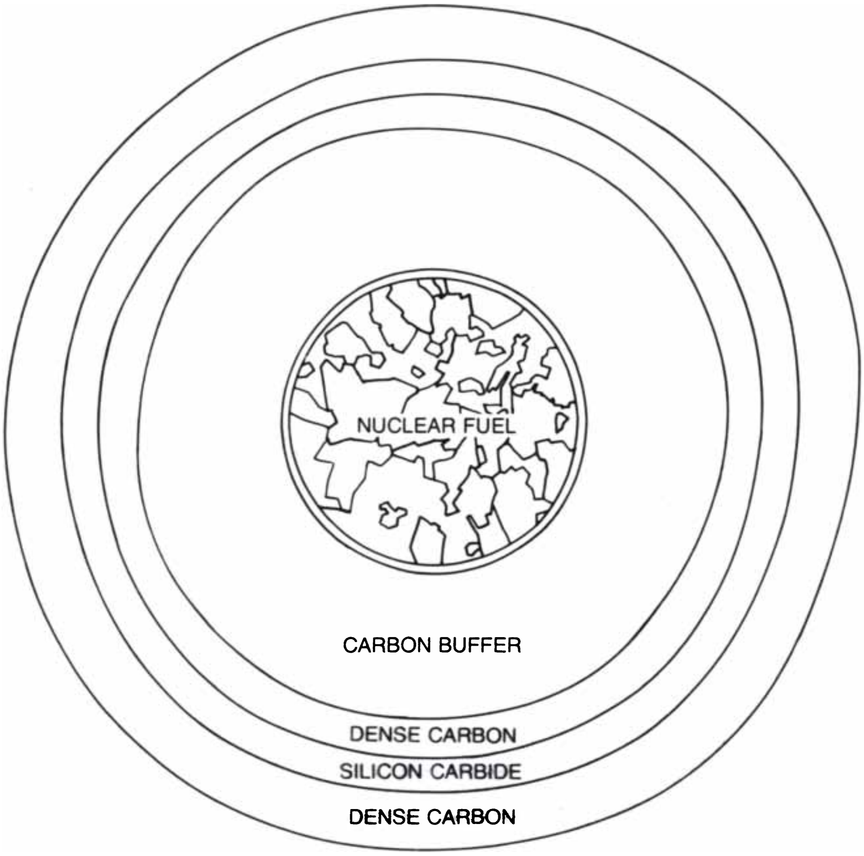
The combination of a stable, inert gas for a reactor coolant and a highly temperature-resistant graphite core structure allows steam to be generated at the high temperatures and pressures found in the modern electric-power plants that burn fossil fuel. The net electric-generating efficiency of the HTGR reference design is 38.5 percent, slightly below the 39.2 percent achieved at Fort St. Vrain. The small reduction was made in the interest of simplifying the steam-generating system and to furnish still further operating and safety margins.

A fundamental property of the helium coolant, a confined gas that cannot possibly condense to liquid form in the system, is that it follows a linear temperature-pressure relation; therefore instru-




**PRESSURIZED-LIGHT-WATER REACTOR has been the commonest type of nuclear power reactor in the U.S., with 44 reactors now in operation. Another 24 reactors are of the boiling-water type, in which heat is carried off from the core by steam rather than by pressurized heated water. The core of a pressurized-light-water reactor rated at 1,100 MWe is shown here. It is housed in a steel pressure vessel about 15 feet in diameter, 40 feet high and from six to 11 inches thick; the vessel is designed to operate with an internal pressure of 2,250 pounds per square inch. The coolant water leaves the reactor at 610 degrees F. and passes to four steam-generating loops, only one of which is shown here. Steam emerges from the generator at 540 degrees F. and a pressure of 1,000 pounds. At this temperature and pressure the system's thermal efficiency is only 32 to 33 percent, compared with 38.5 percent for an HTGR system.**






**FUEL PARTICLE developed for HTGR systems is .03 inch in diameter, about the size of a grain of sand. A cross section of the particle is enlarged 150 diameters at the top. The nuclear fuel itself is the crystalline-like material in the center. It consists of uranium oxycarbide in which for best performance the content of the fissionable isotope uranium 235 is enriched to 93 percent. Layers of carbon and silicon carbide are built up by a high-temperature process.**

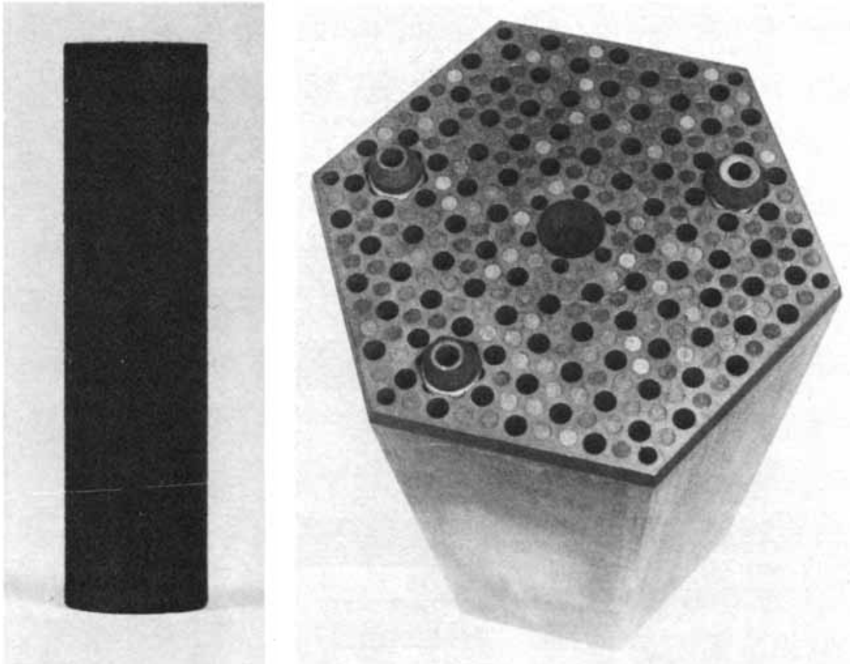
ment readings of temperature and pressure can provide independent checks on each other. Because there is no liquid-gas interface, as there is in boiling-water reactors (and in pressurized-water reactors under certain emergency conditions), a single unambiguous signal—pressure—always indicates the presence and physical condition of the coolant. Rapid depressurization of the primary cooling system can be tolerated without concern that voids have formed and left part of the core uncovered, as can happen when pressure is released from water that is above its atmospheric boiling point.

The Fort St. Vrain experience has verified several important safety and operating advantages of HTGR's. Operating and maintenance personnel have received exposures to radiation far below the limits established for nuclear plants. Fewer than 10 workers out of a total of several hundred have received amounts of radiation that were even measurable.

The Fort St. Vrain system has responded smoothly and gracefully to load changes caused either by transient excursions in the power-generating cycle or by the temporary shutoff of equipment within the plant. Because the core of the HTGR is large and releases less heat per unit volume than light-water reactors do and because the massive core, incorporating some 1,500 tons of graphite, has a large capacity to absorb heat if coolant flow is reduced or interrupted, the reactor responds slowly to an unexpected operational upset, allowing the operators enough time to take appropriate action: hours rather than seconds.

At Fort St. Vrain five such upsets have interrupted the forced circulation of helium for extended periods without giving rise to a measurable increase in the temperature of the core or harming the plant or the fuel in any way. The risk of damaging the reactor or the reactor core through operator error is virtually eliminated. Thanks to the HTGR's thermal stability the system for bringing the activity of the reactor to a halt by the insertion of neutron absorbers and the systems for emergency cooling can be of simple design. There is also ample time for such systems to be actuated manually if it is allowed by regulation. One consequence of the Three Mile Island accident is that the Nuclear Regulatory Commission now requires the full-time presence of an on-site expert, called a shift technical adviser, at nuclear power plants. Fort St. Vrain is the only reactor exempted from this rule; an expert is not required to remain on the site but is on call to report within an hour.

The prestressed-concrete reactor vessel is incorporated in the design as a major safety feature. First, a catastrophic rupture of the PCRV is such a remote possibility that risk analysts characterize it as being incredible. Every steel tendon that gives the PCRV its strength is independent and redundant; the vessel is in a constant state of compression. Second, the PCRV is designed to withstand an ultimate pressure of more than twice the normal operating pressure, or some 2,400 pounds per square inch. Any crack in the steel liner that might result from excessive pressure can do no more than give rise to a slow gas leak; such leaks tend to seal themselves when the pressure is reduced slightly. Third, total depressurization can result only if there is a failure of one of the pipe penetrations or small service lines that pass through the wall of the PCRV. Such a hypothetical failure is an extremely low-probability event. Moreover, at each penetration site the vessel is equipped with flow limiters that prevent the rapid release of gas that could cause structural damage to the core or to the cooling system.



**FUEL ROD AND FUEL BLOCK for an HTGR are shown respectively at the left and the right. The fuel rod, about 2.5 inches in length, consists of tens of thousands of fuel particles bound in a graphite matrix. Each fuel block, which is approximately 14 inches across and 31 inches high, holds 1,656 fuel rods packed in hexagonal arrays. The numerous empty channels in the block are paths for the flow of helium. The large central hole accommodates a mechanism for inserting the fuel blocks in the core of the reactor. The core of an 860-MWe reactor will require 3,512 blocks. Each 270-pound fuel block contains on the average 1.54 pounds of U-235 and 35 pounds of thorium 232. In its four-year residence in the reactor such a block would yield energy equivalent to 2,500 tons of coal or 12,000 barrels of fuel oil. If the unburned U-235 and the U-233 created from thorium were recovered and recycled, the energy equivalent of the original nuclear fuel would rise to some 11,000 tons of coal or 54,000 barrels of oil.**

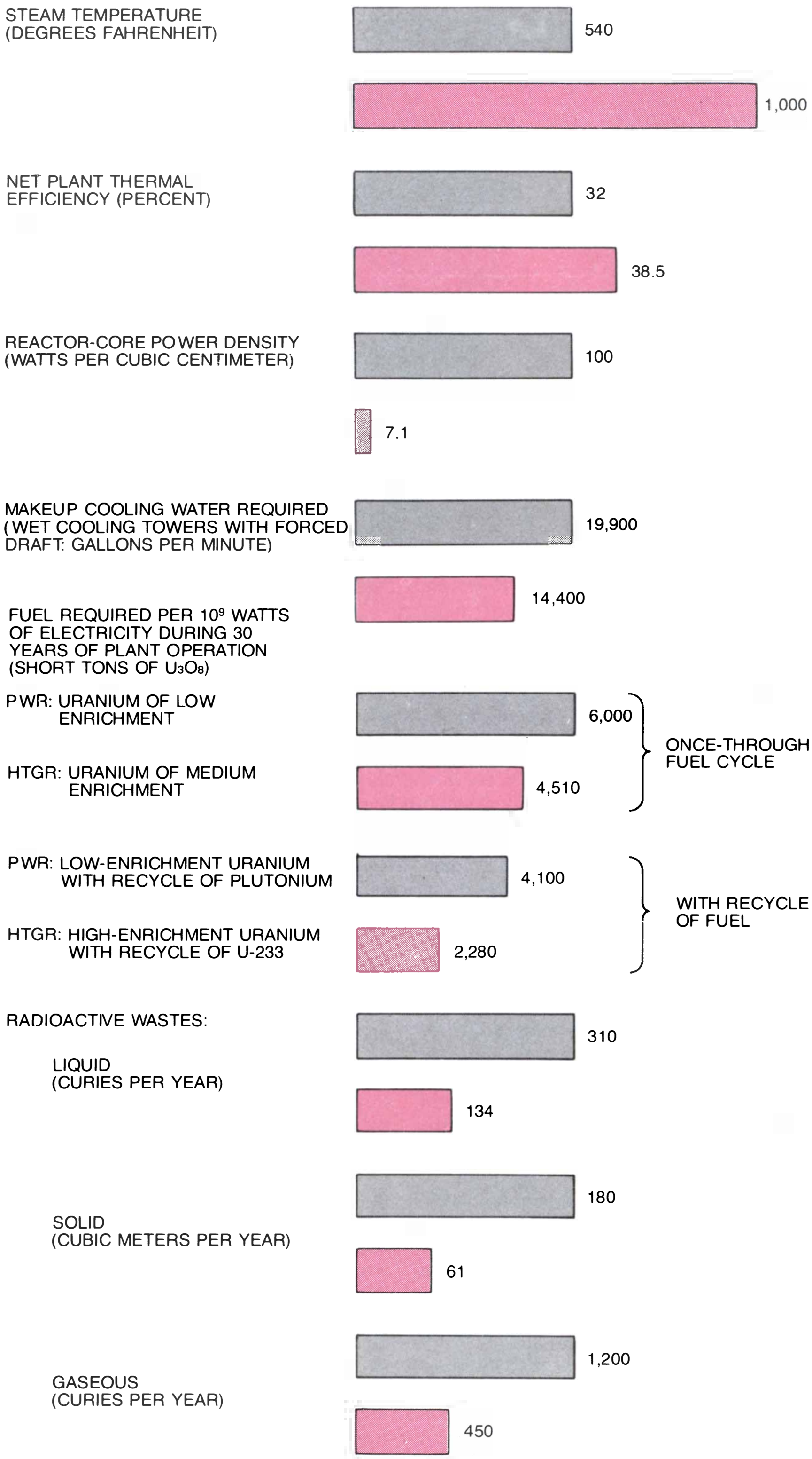
The improved performance characteristics of the HTGR also offer several environmental advantages over the current generation of reactors. Because an HTGR operates at an efficiency of about 39 percent compared with an efficiency of about 33 percent for light-water reactors, an HTGR releases about 25 percent less waste heat to be dissipated into the surrounding environment. If the heat, in the form of hot water, is rejected into a nearby lake or river, concern about raising temperatures to a point harmful to the aquatic ecosystem is reduced proportionately. If cooling towers are used to dissipate the heat, they consume less water and can be smaller and less expensive. If cooling ponds are used, an HTGR plant with about a third more megawatts of capacity than a light-water plant can be sited on a pond of a given size without exceeding a specified pond temperature. Where dry cooling towers must be adopted to meet environmental regulations or fit available water supplies, the loss of plant capacity in hot weather will be only about half as great with an HTGR as it is with existing nuclear power plants. As a result an HTGR plant can be situated at a remote arid or semiarid site with a smaller penalty in cost.

The level of radioactivity in normal discharges from all nuclear power plants is carefully monitored. An HTGR plant inherently releases into the plant process streams less radioactivity and at lower concentrations than a light-water reactor does. In addition an HTGR incorporates features that will ensure that releases of radioactivity from the plant to the environment are essentially zero. Routine decontamination procedures can be expected to produce small volumes of low-level liquid wastes (less than 2,000 gallons per year with a total activity of less than 150 curies). Such small volumes can be shipped off-site with little difficulty or retained on-site. The solid wastes produced by an HTGR should total less than 2,000 cubic feet per year. Some 80 percent will consist of low-level waste (such as paper, filter elements and spent resins) that is only slightly contaminated and can be shipped off-site in drums for burial or burning with virtually no effect on the environment. The remaining 20 percent will be intermediate-level waste, consisting chiefly of reflector blocks, which must be periodically replaced. Such waste can be shipped off-site in shielded 55-gallon transport casks for long-term safe disposal.

Helium-purification and gas-recovery systems incorporated in the standard HTGR plant should reduce the radioactive levels in released gases to several orders of magnitude below the current Government regulation of five millirems per year. Tritium (the radioactive isotope of hydrogen) generated within the primary system of an HTGR is removed in the helium-purification system by an oxidizer that converts the tritium into tritiated water, which is subsequently solidified and handled as solid waste that can be readily isolated as the tritium decays. (The half-life of tritium is 12.26 years.)

The HTGR has evolved a number of features that simplify operation, maintenance and refueling. For example, the entire primary coolant, helium, is confined within the prestressed-concrete reactor vessel. The PCRV itself provides all the necessary shielding for personnel, so that maintenance work can be done throughout the reactor building while the plant is in operation. Because the entire secondary steam system is essentially free of radioactivity all equipment in the steam cycle outside the PCRV, including the turbine-plant equipment, can be operated and maintained as it would be in a plant fired with fossil fuel. Because the amount of steam flowing to the turbines in an HTGR plant is only about 60 percent as large as that flowing to the turbines in a light-water power plant of the same output, all the equipment associated with the steam and feed-water cycles of an HTGR plant is small and therefore easier to maintain. In general, maintenance, repair and handling costs are lower in an HTGR plant than they are in light-water plants because helium, unlike water, is inert, nonradioactive and noncorrosive.

One big advantage of gas as a coolant is its transparency, which makes it possible to inspect many areas within the PCRV visually. The radiation shielding inherent in the design of the PCRV makes it possible to carry out many inspection and maintenance tasks while




**OPERATING CHARACTERISTICS of the 860-MWe HTGR (*color*) are compared with those of a pressurized-water reactor of the same generating capacity (*gray*). The lower fuel consumption of the HTGR can be attributed in part to higher thermal efficiency and in part to the fact that for each atom of U-235 consumed in the HTGR about .7 atom of new fuel is created. The pressurized-water reactor creates less than .5 atom of new fuel for each atom consumed. With a once-through fuel cycle both systems convert a certain fraction of U-238 or Th-232 atoms into isotopes of plutonium or uranium, some of which are beneficially consumed before the fuel needs replacing. If the spent fuel could be recycled (which was contrary to the policy of the last Administration), it would be preferable to fuel an HTGR with a mixture of highly enriched uranium (about 93 percent U-235) and thorium. Some of the thorium would be converted into fissionable U-233, which could be recovered and recycled to replace U-235 in subsequent fuel charges. Smaller volume of radioactive wastes from an HTGR results partly from its higher efficiency and partly from advantages of helium over water as a coolant.**

the reactor is running, which reduces the time the reactor must be taken out of service for such purposes.

Essentially all structural members of the PCRV, such as the vertical tendons and the circumferential cable wrapping, can be inspected visually while the reactor is operating. Selected members are continuously monitored for changes in tension or strain that would indicate a deterioration in performance. If necessary, any structural member can be replaced. All external concrete surfaces, except those immediately surrounding the ports for the control rods, can be inspected visually while the plant is running. The control-rod ports and the surfaces surrounding the site where the control-rod drives penetrate the PCRV can be readily inspected in the course of refueling.

Recent refueling experience at Fort St. Vrain has demonstrated the ease of handling the HTGR's block-type fuel elements. About 240, or a sixth, of the fuel elements were removed from the core and replaced with fresh fuel; the other 1,240 elements were left in place. The refueling crew was exposed to such low levels of radiation that measuring them called for a microrem meter. By extrapolating from existing data one can calculate that the sum of the integrated man-rem exposure for the entire refueling operation following on the operation of the reactor at full power will be less than five man-rem. Federal regulations currently limit individual workers to five rem over a period of a year.

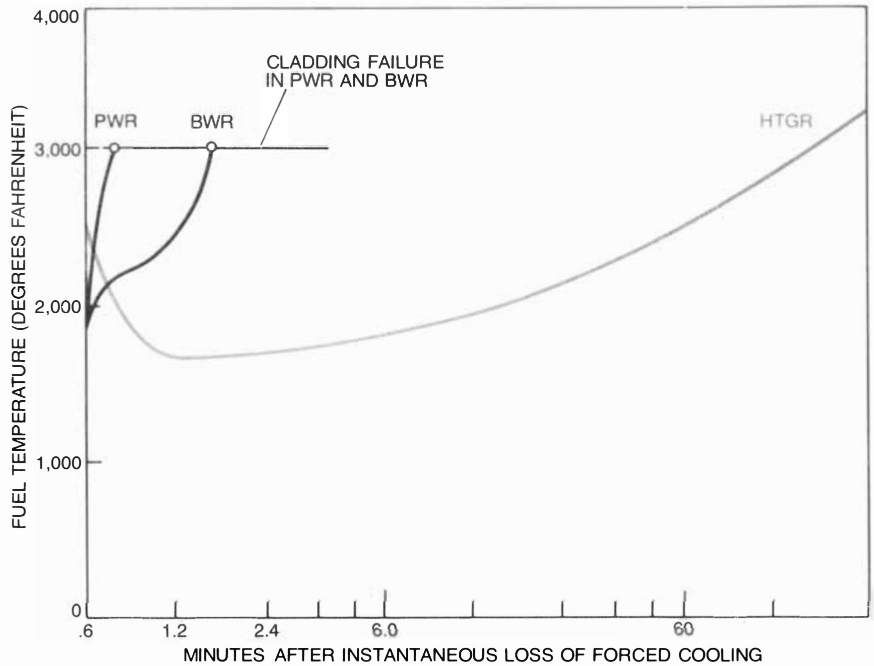
Each HTGR fuel element is a graphite block, hexagonal in cross section, 14 inches wide and 31 inches long. The block is perforated lengthwise with 72 coolant channels and 138 blind holes for fuel. Graphite is an ideal choice as a moderator and a structural material because its strength actually increases with temperature. In the reference design the graphite fuel blocks are stacked in columns of eight. This axially segmented arrangement facilitates fabrication, handling and refueling.

The convenient block configuration has been made possible by the development of a specially coated fuel particle. The kernel of each particle is a microsphere of uranium oxycarbide (suitably enriched in uranium 235) about .01 inch in diameter. Around each kernel thin layers of carbon, pyrolytic carbon and silicon carbide are applied at high temperature, yielding a tightly encased particle with a total diameter of about .03 inch. A similar form of encapsulation is used for the thorium particles. The technique ensures the containment of the fission products. The tiny spheres are tested in batches of 2,000 for structural integrity when they are exposed to a radiation flux that simulates the internal environment of the reactor. The particle-production process, which is semiautomatic, and the rigorous testing procedure work together to achieve a close control of quality.

Although severe and unforeseen service conditions in one region of the reactor core might cause the particle coating to fail and release fission products, the failure would be limited to the area directly involved. In most reactors, where the cladding of the fuel elements runs the entire length of the reactor core, an operating upset that ruptures a small section of cladding could release fission products from the entire length of the fuel rod. The performance of the fuel elements at Fort St. Vrain has fully met design expectations. Indeed, the release of fission products has been well below the predicted levels. In sum, the fission-product barriers in the HTGR fuel element have been demonstrated to have a high degree of reliability.

The properties of the HTGR make it possible to exploit a wide variety of nuclear fuel cycles with it. The cycle that has been most intensively studied and tested is the uranium-thorium one, in which fully enriched uranium (93 percent U-235) serves as the primary fissile material and thorium (Th-232) serves as a "fertile" material. In the reactor the thorium absorbs neutrons and is ultimately converted into the fissile isotope uranium 233, which can be recycled in subsequent fuel reloadings. The Fort St. Vrain reactor is fueled with uranium enriched to 93.5 percent U-235, in combination with thorium. The design of the plant allows the use of either fully enriched or medium-enrichment uranium (about 20 percent U-235). The HTGR fuel-cycle costs, under the current restraints on fuel reprocessing and recycling, are essentially equivalent to those of other commercial plants. Unless the policy is changed by the Administration spent fuel is to be stored indefinitely, without the recovery either of the unspent U-235 or of the U-233 or plutonium created during the operation of the reactor. This fuel cycle is commonly called the stowaway cycle.

If an HTGR were operated on a stowaway uranium-thorium cycle with fully enriched uranium, it would consume about 20 percent less uranium over its 40-year life than a light-water reactor would. If both types of reactor could be operated with a full recovery of their uranium and plutonium, the HTGR would consume about 50 percent less uranium. The HTGR therefore offers the opportunity of saving substantial amounts of uranium with either a stowaway policy or a full-recycle one, provided the reactor is designed to accept fully enriched fuel. The significance of the potential uranium saving can be appreciated when one considers that the total fuel cost over the life of a nuclear power plant is roughly equal to the total initial cost of the plant.




**INHERENT SAFETY OF AN HTGR is shown in graphs that compare the temperature in the core of an HTGR, of a pressurized-water reactor and of a boiling-water reactor following a hypothetical loss-of-coolant or loss-of-forced-circulation accident. In the water-cooled reactors the nuclear reaction is halted automatically by the loss of water, which serves as a moderator. In the HTGR the reaction must be stopped by the insertion of control rods that absorb neutrons. At the moment of shutdown decaying fission products in the fuel release heat at a rate equivalent to 7 percent of the thermal output of the reactor. The heat release falls to 1 percent in about two hours and to .5 percent in 24 hours. In the water-cooled reactors, in the absence of emergency cooling, the temperature of the cladding of the fuel would rise in less than two minutes to 3,000 degrees F., causing the cladding to fail. In the HTGR the mass of the graphite moderator would absorb the heat released by fission products, so that 3,000 degrees would not be reached for at least an hour. A temperature high enough to damage the graphite core (about 4,000 degrees) would be attained only after at least 10 hours without forced cooling.**

Over the past six years orders for about 55 nuclear power plants have been canceled. Only six years ago U.S. utilities had demonstrated interest in constructing 10 HTGR plants. Once the Fort St. Vrain reactor has been brought up to full power, which is scheduled for this summer, and has demonstrated the exceptional safety and reliability that its designers confidently predict, it is reasonable to assume that U.S. utilities will look favorably on the HTGR when they are again ready to place orders for nuclear power plants.

| YEAR | AVAILABILITY OF NUCLEAR STEAM SUPPLY SYSTEM (PERCENT) | PLANT AVAILABILITY (PERCENT) | OVERALL PLANT CAPACITY FACTOR (PERCENT) |
|---|---|---|---|
| 1967 | 81 | 78 | 69 |
| 1968 | 88 | 88 | 82 |
| 1969 | 86 | 84 | 67 |
| 1970 | 95 | 95 | 88 |
| 1971 | 90 | 87 | 78 |
| 1972 | 71 | 71 | 58 |
| 1973 | 95 | 94 | 78 |
| 1974 | 96 | 95 | 70 |
| CORE 1 AVERAGE | 85 | 83 | 73 |
| CORE 2 AVERAGE | 89 | 88 | 74 |
| TOTAL LIFETIME STATION AVERAGE | 88 | 86 | 74 |

**RELIABILITY OF FIRST HTGR PLANT designed by General Atomic, the Peach Bottom Atomic Power Station No. 1, is attested to by the statistics shown here. Apart from scheduled down time or time lost for reasons unrelated to the reactor, the HTGR was available for supplying steam for power generation 88 percent of the time. In achieving 74 percent of its rated electric-generating capacity over its seven-and-a-half-year lifetime the Peach Bottom reactor exceeded the typical figure of 66 percent achieved by light-water reactors operated by utilities.**

# The Decay of the Proton

*The proton is known to have a lifetime at least $10^{20}$ times the age of the universe, but theory indicates that it may not live forever. If it is not immortal, all ordinary matter will ultimately disintegrate*

by Steven Weinberg

The discovery of radioactivity by Antoine Henri Becquerel in 1896 dispelled the belief that all atoms are permanent and immutable. The energetic particles that had been detected by Becquerel were later understood as being emitted when the nuclei of the atoms of a radioactive substance decay spontaneously into other atomic nuclei. Interesting as this nuclear instability was, it seemed to be a rarity, a property only of certain heavy elements such as uranium and radium. The nuclei of common elements such as hydrogen and oxygen were thought to be absolutely stable.

There are now several theoretical reasons to suspect that all atomic nuclei ultimately decay and hence that all matter is in some small degree radioactive. In a decay of this kind one of the two types of particle in the atomic nucleus, a proton or a neutron, would be spontaneously transformed into energetic particles very different from the particles that make up ordinary atoms. Even the lightest nucleus, that of hydrogen, which consists of a single proton, would be subject to decay.

Much evidence, beginning with the great age of the earth, indicates that matter cannot be highly evanescent. If ordinary matter decays, it does so only very slowly, so slowly that experiments of an extraordinarily large scale will be needed to detect the decay. Becquerel discovered the radioactive disintegration of uranium nuclei in a crystal of uranium salts that weighed perhaps a few grams; in order to observe the feebler radioactivity associated with the decay of the proton, it will be necessary to monitor many tons of material. Nevertheless, experimental searches for the decay of the proton are now under way.

To see what is at stake in these experiments, it is useful first to ask why anything in the world should last forever. The electron, for example, is still thought to be absolutely stable. What physical principles prevent it from decaying into other particles? By understanding the stability of particles such as the electron one can judge whether there are any physical principles that prevent the decay of ordinary atomic nuclei.

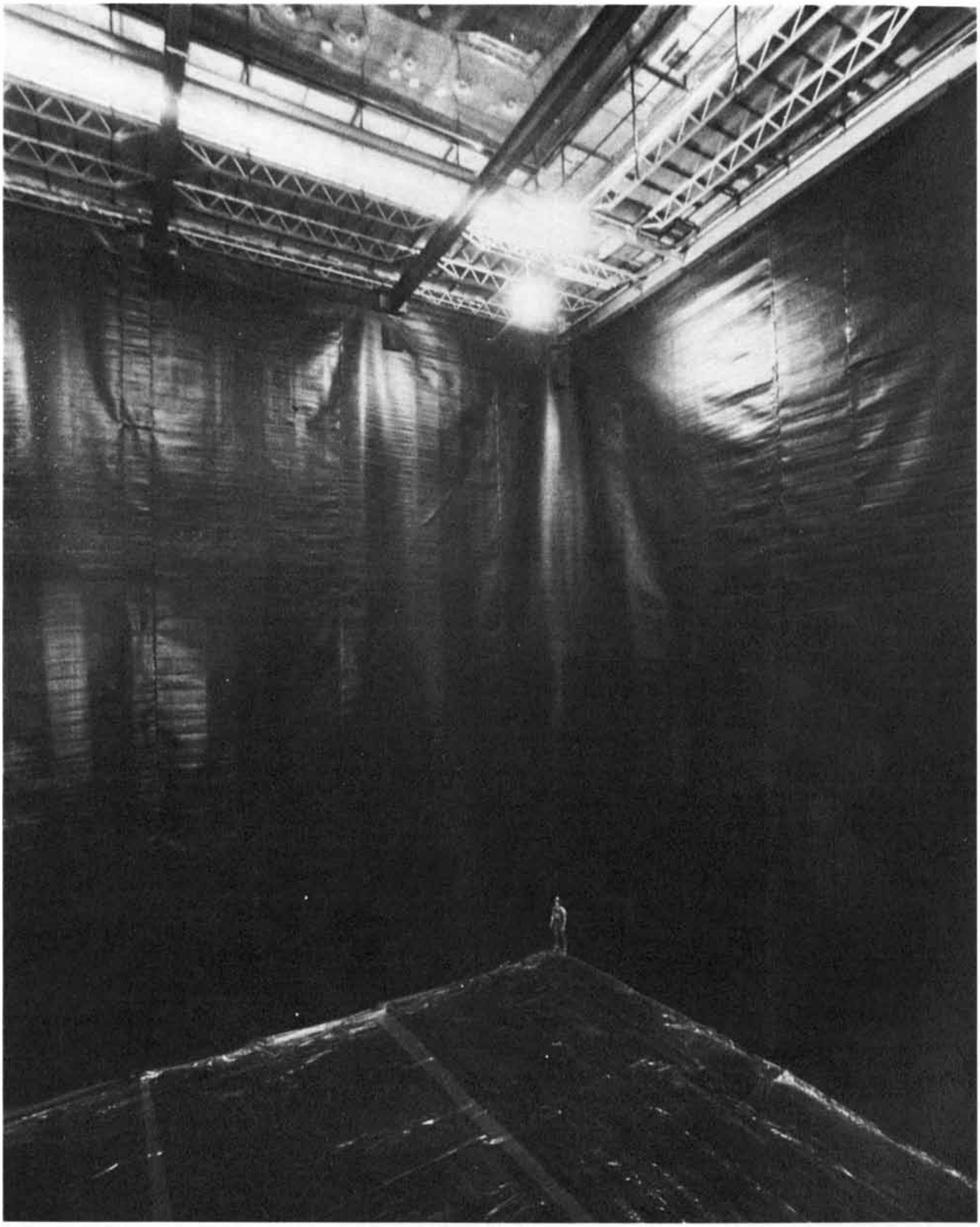
Experience in the physics of elementary particles teaches that any decay process one can imagine will occur spontaneously unless it is forbidden by one of the conservation laws of physics. A conservation law states that the total value of some quantity, such as energy or electric charge, can never change. Even if a decay process is not produced directly by one of the fundamental interactions of elementary particles, if it is not forbidden by a conservation law, it will be produced by some more or less complicated sequence of emissions and absorptions of particles. Thus in considering whether any particle is stable one has to ask whether its decay would violate any conservation law.

The law of conservation of energy is easy to apply. It simply requires that the mass of the decaying particle (or the energy equivalent of the mass) be greater than the total mass of the decay products. (It is not enough for the masses to be equal because some mass must be converted into the kinetic energy of the decay products.) Therefore a good way to start in judging the stability of any particle is to list all the less massive particles into which it might conceivably decay.

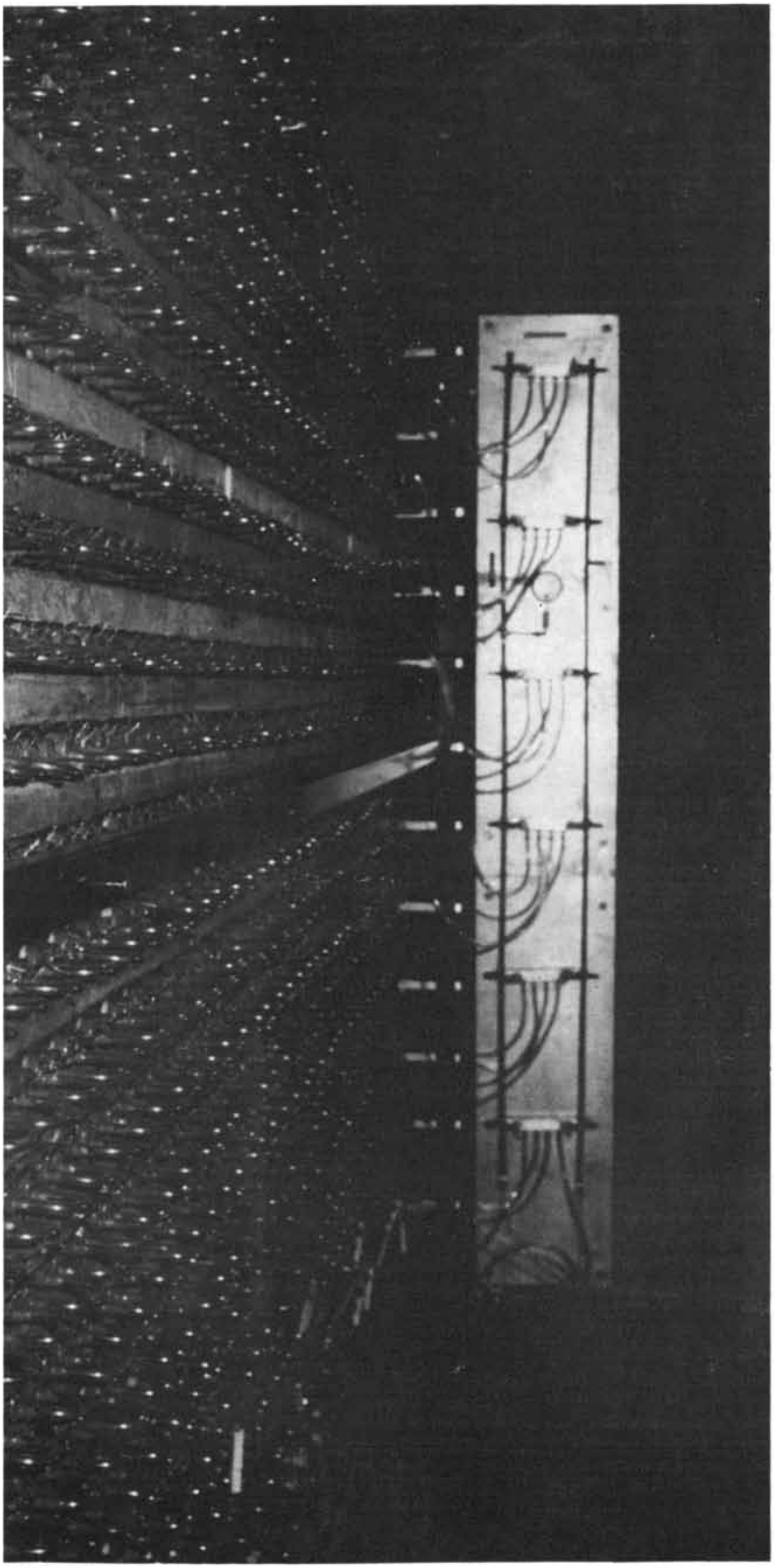
Consider the electron. As far as is known, there are only a few kinds of particle with a mass less than that of an electron. The most familiar of them is the photon, the quantum of light, whose mass is thought to be exactly zero. There are strong theoretical grounds for thinking there is also a quantum of gravitational radiation, the graviton, again with zero mass. Finally there are various species of particles called neutrinos, which are similar in some respects to the electron; they are emitted in the familiar kind of radioactivity known as beta decay, the kind that was discovered by Becquerel in 1896. Neutrinos have generally been thought to have zero mass, but the determination of their mass is at present an object of intense theoretical and experimental effort. Nevertheless, there is no doubt that at least one of the species of neutrino has a mass less than about a ten-thousandth of the mass of the electron.

Why then does the electron not decay into, say, neutrinos and photons? The answer is that although such a decay would satisfy the law of conservation of energy, it would violate another conservation law, that of electric charge. Benjamin Franklin was the first to recognize that the net quantity of electric charge (positive minus negative) never increases or decreases, although charges of opposite sign can be separated or recombined. Electrons carry a definite negative electric charge, but all the lighter particles into which the electron might decay (the photon, the graviton and the neutrinos) happen to carry zero electric charge. The decay of an electron would entail the destruction of a definite negative quantity of electric charge and is therefore strictly forbidden.

Now consider how these conservation laws might apply to the decay of the two kinds of particle that make up the atomic nucleus. For the moment consider only the lighter of the two particles, the proton, and return to the neutron later on. The proton carries a positive electric charge, equal in magnitude but opposite in sign to that of the electron, and so it too cannot decay into neutrinos, photons or gravitons. The proton, however, is about 1,820 times as heavy as the electron, and there are several particles of lesser mass that also have positive charge. The proton could decay into these other particles without violating the conservation of either energy or electric charge. For example, the electron has an antiparticle called the positron, with the same mass as the electron but with a positive electric charge equal to that of the proton. (For every kind of particle there is an antiparticle with the same mass but opposite values of other properties, such as electric charge. Incidentally, the positron is stable for the same reason that the electron is.) There is nothing in the laws of energy or charge conservation that would forbid a proton from decaying into a posi-



**LARGE CAVITY** some 1,950 feet underground in the Morton salt mine east of Cleveland will be filled in the next few months with 10,000 tons of water in preparation for a search for the decay of a proton or of a neutron bound in an atomic nucleus. The cavity is 60 by 80 by 70 feet. The decay of any one of the $2.5 \times 10^{33}$ protons and neutrons in the central region of the water will give rise to high-velocity particles. When a charged particle moves through a transparent medium faster than the speed of light in that medium, it emits a cone of blue light called Cerenkov radiation. The radiation is an optical boom analogous to the sonic boom generated by an airplane moving through the air faster than the speed of sound in air. The Cerenkov light will be detected by 2,400 photomultiplier tubes that will be installed along the walls of the cavity. The experiment is done underground to reduce the number of high-energy particles arriving from space that might be mistaken for the decay products of a proton or of a bound neutron. The consortium that will operate the detector includes investigators from the University of California at Irvine, the University of Michigan and the Brookhaven National Laboratory.



**SLAB OF CONCRETE in the Soudan mine in Minnesota is being monitored for events signaling the decay of the proton. Some 3,450 particle counters are embedded in the slab, whose mass is 31 tons. The counters will detect directly the products of proton decay. The experiment is being done by workers from the University of Minnesota and the Argonne National Laboratory.**

tron and any number of photons and neutrinos.

Another candidate for a decay product of the proton is the antimuon. The muon is a particle similar in many respects to the electron and with the same charge, but it is 210 times as massive. (The muon does decay into an electron and neutrinos.) The antimuon has the same charge as the proton but only about one-ninth the mass. A proton might therefore decay into an antimuon plus light neutral particles such as photons and neutrinos.

Still another possible product of the proton's decay is a meson, a member of the group of unstable particles intermediate in mass between the electron and the proton. The conservation laws of energy and charge would allow the proton to decay into, say, a positively charged meson and a neutrino or into a neutral meson and a positron. Any of these decay processes would lead to the total disruption of the hydrogen atom. In a heavier element they would change the chemical nature of the element and would release energy in amounts much greater than those released in ordinary radioactivity.

Why is matter everywhere not observed to be disintegrating as a result of such decay processes? This problem seems to have been first addressed by Hermann Weyl in 1929. Positrons, muons and mesons were unknown then, and so the conjectural proton-decay schemes outlined above could not have been imagined. Weyl was nonetheless puzzled about the stability of matter; he may have wondered why the protons in an atom do not absorb the orbiting electrons, leading for example to the decay of a hydrogen atom into a shower of photons. Weyl suggested that the stability of matter might be explained if there were two kinds of electric charge, one carried by the electron and the other by the proton. If each kind of charge were conserved separately, the mutual annihilation of a proton and an electron would be forbidden. Weyl's proposal did not attract much attention at the time.

The question was taken up again by E. C. G. Stueckelberg in 1938 and by Eugene P. Wigner (in a footnote) in 1949. They proposed what became the conventional view, namely that in addition to energy and electric charge there is another conserved property of matter, which has since come to be called baryon number. The baryons (from the Greek *barys,* heavy) are a family of particles that includes the proton and many particles heavier than the proton, such as the neutron and the highly unstable particles called hyperons. All baryons are assigned a baryon number of +1, and all lighter particles, including the photon, the electron, the positron, the graviton, the neutrino, the muon and the mesons, have a baryon number of zero. For an atom or another composite system of particles the baryon number is the sum of the baryon numbers of the constituent particles. It follows that any collection of particles lighter than the proton has a baryon number of zero. The law of baryon-number conservation is the assertion that the total baryon number cannot change. The decay of a proton into a collection of lighter particles would entail the conversion of a state whose baryon number is +1 into a state whose baryon number is zero, and so the decay is forbidden.

An antiparticle has a baryon number opposite to that of the corresponding particle. The antiproton, for example, has a baryon number of −1; it is an antibaryon. A proton and an antiproton can annihilate each other without violating the conservation of baryon number; the proton and the antiproton have a total baryon number of +1 plus −1, or zero, and so they can turn into a shower of mesons or photons. Thus baryon-number conservation does not require that each proton be immortal but rather requires that protons not decay spontaneously in ordinary matter, where there are no antiprotons.

So far I have discussed only the decay of the proton, but of course the nucleus of most atoms is made up not only of protons but also of neutrons. What about the possibility of the neutron's decaying? The neutron is a baryon with an electric charge of zero and a mass slightly larger than that of the proton. To be more precise, the mass of the neutron is a little greater than the mass of the proton plus the mass of the electron. This relation suggests one possible mode of decay for the neutron: it could give rise to a proton, an electron and some massless neutral particles. Energy can evidently be conserved in this process. So can electric charge, since the charges of the proton and the electron cancel each other. Baryon number is also conserved, since the neutron and the proton each have a baryon number of +1 and the other particles have a baryon number of zero.

A free neutron (one that is not bound in an atomic nucleus) does decay in exactly this manner: it yields a proton, an electron and an antineutrino. The half-life of the free neutron, which is the time required for half of the neutrons in any large sample to decay, is roughly 10 minutes. Neutrons in certain atomic nuclei, such as the nucleus of tritium (the heavy isotope of hydrogen with one proton and two neutrons), can also decay into protons; this is the process of beta decay. In most nuclei, however, neutrons do not decay because too much energy would be required to create a proton amid the repulsive electrostatic forces generated by the other protons in the nucleus. In such nuclei the neutrons are as stable as the protons.

The possibility remains that a neutron bound in a nucleus might decay in some other manner that does not conserve baryon number. For example, it might give rise to a positron and a negatively charged meson or to an electron and a positively charged meson. The discovery of such a neutron decay in an otherwise stable nucleus would be as significant as the discovery of the decay of a proton. Indeed, the experiments that are searching for proton decay are also looking for the decay of bound neutrons. Since the decay of free neutrons is already well known, however, the experimental tests of baryon-number conservation have come to be known as proton-decay experiments.

In recent years it has become widely accepted that the baryons and the mesons are made up of the more fundamental particles called quarks. A baryon consists of three quarks, an antibaryon consists of three antiquarks and a meson consists of a quark and an antiquark. The electron, the muon and the neutrinos belong to the family of particles called leptons, which are not made up of quarks and indeed give no sign of an inner structure. On this basis the baryon number of any system of particles is just one-third the net quark number, that is, one-third the difference between the number of quarks and the number of antiquarks. The conservation of net quark number is equivalent to the conservation of baryon number.

The skeptical reader may feel somewhat dissatisfied with baryon-number conservation as an explanation of the stability of the proton and the bound neutron. In my view he would be justified in this feeling. Baryon number was invented as a bookkeeping device, in order to explain the nonobservation of proton decays and related decays; it has no other known significance. In this respect baryon number is very different from electric charge, which has a direct dynamical significance: an electric charge creates electric and magnetic fields and the charge in turn is acted on by such fields, which have observable effects on the motion of the charge. The theory of electricity and magnetism would make no sense if electric charge were not conserved, but no such dynamical argument is known for baryon-number conservation.

Indeed, there is empirical evidence against the existence of any kind of field (call it a barytropic field) that might bear the same relation to baryon number as the electromagnetic field bears to electric charge. The earth includes some $4 \times 10^{51}$ protons and neutrons, and so it has a huge baryon number. If the earth were a source of a barytropic field, one would expect the field to attract or repel the protons and neutrons in ordinary bodies on the earth's surface. A barytropic force could be distinguished from the gravitational force because the gravitational force the earth exerts on a body is proportional to the mass of the body, whereas the barytropic force would be proportional to the baryon number. Bodies of equal mass that are composed of different elements can have baryon numbers that differ by almost 1 percent. Several highly accurate experiments (starting with those of Roland von Eötvös in 1889) show that the attraction of bodies to the earth is in fact closely proportional to their mass, not to their baryon number. In 1955 T. D. Lee of Columbia University and C. N. Yang of the Institute for Advanced Study in Princeton showed from an analysis of these experiments that any barytropic force between two nuclear particles would have to be much weaker than the gravitational force, which is itself almost 40 orders of magnitude weaker than the electromagnetic force. It cannot be absolutely ruled out that baryon number plays a dynamical role like that of electric charge, but the argument of Lee and Yang makes such a role appear quite unlikely.

The conclusion that baryon number has no dynamical role does not immediately imply that baryon number is not conserved. Indeed, since the mid-1930's physicists have become familiar with a number of other quantities that do not appear to have a dynamical significance like that of electric charge and yet are conserved, at least in certain contexts. Among these quantities are the ones called strangeness, isospin and charge conjugation. To take one example, protons and neutrons are assigned a strangeness of zero, some of the hyperons are assigned a strangeness of −1 and some of the mesons called *K* mesons are assigned a strangeness of +1. The conservation of strangeness was introduced as a bookkeeping rule to explain the observation that a *K* meson or a hyperon cannot be produced singly in collisions of ordinary atomic nuclei but that they can be produced in association, because one *K* meson and one hyperon have a net strangeness of zero. For years after the idea of baryon number was introduced it did not seem implausible that baryon-number conservation was another of these nondynamical bookkeeping rules, which happens to be universally obeyed.

This view of conservation laws has been radically changed by the development of the modern theories of elementary-particle interactions. The theories describe all known forces among elementary particles (apart from gravitation) in a way very similar to the way electromagnetism was described in the older theory of purely electromagnetic interactions; the latter theory, called quantum electrodynamics, was developed in the 1930's and 1940's. There are now thought to be 12 fields similar to the

| PARTICLE | MASS (MeV) | ELECTRIC CHARGE | BARYON NUMBER | PRINCIPAL DECAY MODE |
|---|---|---|---|---|
| PHOTON ($\gamma$) | 0 | 0 | 0 | NONE KNOWN |
| NEUTRINO ($\nu$) | 0? | 0 | 0 | NONE KNOWN |
| ANTINEUTRINO ($\bar{\nu}$) | 0? | 0 | 0 | NONE KNOWN |
| ELECTRON ($e^-$) | .511 | −1 | 0 | NONE KNOWN |
| POSITRON ($e^+$) | .511 | +1 | 0 | NONE KNOWN |
| MUON ($\mu^-$) | 105.7 | −1 | 0 | $\mu^- \rightarrow e^- + \nu + \bar{\nu}$<br>MASS: 105.7 → .511 + 0 + 0<br>CHARGE: −1 → −1 + 0 + 0<br>BARYON NUMBER: 0 → 0 + 0 + 0 |
| ANTIMUON ($\mu^+$) | 105.7 | +1 | 0 | $\mu^+ \rightarrow e^+ + \nu + \bar{\nu}$<br>MASS: 105.7 → .511 + 0 + 0<br>CHARGE: +1 → +1 + 0 + 0<br>BARYON NUMBER: 0 → 0 + 0 + 0 |
| PI MESONS ($\pi^+$) | 139.6 | +1 | 0 | $\pi^+ \rightarrow \mu^+ + \nu$<br>MASS: 139.6 → 105.7 + 0<br>CHARGE: +1 → +1 + 0<br>BARYON NUMBER: 0 → 0 + 0 |
| ($\pi^0$) | 135 | 0 | 0 | $\pi^0 \rightarrow \gamma + \gamma$<br>MASS: 135 → 0 + 0<br>CHARGE: 0 → 0 + 0<br>BARYON NUMBER: 0 → 0 + 0 |
| ($\pi^-$) | 139.6 | −1 | 0 | $\pi^- \rightarrow \mu^- + \bar{\nu}$<br>MASS: 139.6 → 105.7 + 0<br>CHARGE: −1 → −1 + 0<br>BARYON NUMBER: 0 → 0 + 0 |
| *K* MESONS ($K^+$) | 493.7 | +1 | 0 | $K^+ \rightarrow \mu^+ + \nu$<br>MASS: 493.7 → 105.7 + 0<br>CHARGE: +1 → + 1+ 0<br>BARYON NUMBER: 0 → 0 + 0 |
| ($K^0_S$) | 497.7 | 0 | 0 | $K^0_S \rightarrow \pi^+ + \pi^-$<br>MASS: 497.7 → 139.6 + 139.6<br>CHARGE: 0 → +1 + −1<br>BARYON NUMBER: 0 → 0 + 0 |
| ($K^0_L$) | 497.7 | 0 | 0 | $K^0_L \rightarrow \pi^0 + \pi^0 + \pi^0$<br>MASS: 497.7 → 135 + 135 + 135<br>CHARGE: 0 → 0 + 0+ 0<br>BARYON NUMBER: 0 → 0 + 0 + 0 |
| ($K^-$) | 493.7 | −1 | 0 | $K^- \rightarrow \mu^- + \bar{\nu}$<br>MASS: 493.7 → 105.7+ 0<br>CHARGE: −1 → −1 + 0<br>BARYON NUMBER: 0 → 0 + 0 |
| PROTON ($p$) | 938.3 | +1 | +1 | NONE KNOWN |
| ANTIPROTON ($\bar{p}$) | 938.3 | −1 | −1 | NONE KNOWN |
| NEUTRON ($n$) | 939.6 | 0 | +1 | $n \rightarrow p + e^- + \bar{\nu}$<br>MASS: 939.6 → 938.3 + .511 + 0<br>CHARGE: 0 →+1 + −1 + 0<br>BARYON NUMBER: +1 → +1 + 0 + 0 |
| ANTINEUTRON ($\bar{n}$) | 939.6 | 0 | −1 | $\bar{n} \rightarrow \bar{p} + e^+ + \nu$<br>MASS: 939.6 → 938.3 + .511 + 0<br>CHARGE: 0 → −1 + +1 + 0<br>BARYON NUMBER: −1 → −1 + 0+ 0 |
| Λ HYPERON ($\Lambda^0$) | 1115.6 | 0 | +1 | $\Lambda^0 \rightarrow p + \pi^-$<br>MASS:1115.6 → 938.3 + 139.6<br>CHARGE: 0 → +1 + −1<br>BARYON NUMBER: +1 → +1 + 0 |
| Λ ANTIHYPERON ($\bar{\Lambda}^0$) | 1115.6 | 0 | −1 | $\bar{\Lambda}^0 \rightarrow \bar{p} + \pi^+$<br>MASS:1115.6 → 938.3 + 139.6<br>CHARGE: 0 → −1 + +1<br>BARYON NUMBER: −1 → −1 + 0 |

**PRINCIPAL DECAY MODES of particles are governed by fundamental conservation laws. The law of conservation of energy requires that the mass of the decaying particle be at least as great as the total mass of the decay products. The law of conservation of electric charge requires that the charge of the decaying particle be equal to the total charge of the decay products. The law of conservation of baryon number states that the baryon number of the decaying particle must equal the sum of the baryon numbers of the decay products. The proton and many particles heavier than the proton have a baryon number of +1. All particles lighter than the proton have a baryon number of zero. The particles are listed in order of increasing mass. The mass is given in terms of the equivalent energy, in units of a million electron volts (MeV). The charge is given in units of the charge of the proton. Under each decay mode is an accounting of the three conserved quantities. The decay of a proton into a collection of lighter particles would entail a transformation of a state whose baryon number is +1 into a state whose baryon number is zero, and so the decay is forbidden by the conservation of baryon number. If the proton is found to decay, the conservation of baryon number is not universally true.**

electromagnetic field of quantum electrodynamics. They are the eight gluon fields that provide the strong nuclear forces that hold together the quarks inside baryons and mesons and the four electroweak fields that in a unified manner provide both the weak nuclear forces responsible for beta decay and electromagnetism itself. There are 12 corresponding conservation laws, similar to the conservation of electric charge, for quantities designated color, electroweak isospin and electroweak hypercharge. (Color is a property of quarks that has nothing to do with visual color; electric charge is a particular weighted combination of electroweak hypercharge and electroweak isospin.) Unlike baryon number, these conserved quantities have a direct physical significance: it is the particles carrying these quantities that give rise to the gluon fields and the electroweak fields, and the fields in turn exert a force on any such particle. The force depends on the values of the 12 quantities carried by the particle.

At the same time that new conservation laws have appeared the old nondynamical conservation laws have in a sense been demoted. For example, the modern theory of strong nuclear interactions is so tightly constrained by color conservation (and other principles) that there is no way it could include the kinds of complications that would be needed to violate the conservation of strangeness. One could try to introduce fundamental interactions that do not conserve strangeness, but it always turns out that one can redefine what is meant by strangeness in such a way that it is still conserved. Thus strangeness conservation is now understood to be not a fundamental principle like energy conservation or charge conservation but a consequence of the detailed theory of the strong interactions, and in particular of the genuinely fundamental law of color conservation. Since strangeness conservation is not a fundamental principle of physics, there is no general reason for it to be respected outside the realm of the strong forces. Indeed, it has been known since the discovery of strangeness that strangeness is not conserved by the weak nuclear forces.

The other nondynamical conservation laws have suffered a similar demotion in status; they are no longer seen as fundamental conservation laws, on the level of energy or charge conservation, but rather as mere mathematical consequences of the structure of present theories of elementary-particle interactions. A list of the conservation laws that now seem to be fundamental would include the conservation of the 12 quantities associated with the strong and the electroweak forces, the conservation of quantities such as energy and momentum that are associated in a similar way with gravitational forces, and the conservation of baryon number, which is not known to be associated with any force.

This fact alone should make one suspicious about baryon-number conservation: baryon number does not need to be conserved in the way that energy, charge, color and similar quantities need to be conserved in order to have sensible theories of elementary-particle interactions. Moreover, there are positive hints that baryon-number conservation is not exact. One of the hints is provided by the modern theory of electroweak interactions. Gerard 't Hooft of the University of Utrecht has shown that in this theory certain subtle effects that cannot be represented by any finite number of emissions and absorptions of elementary particles lead to baryon-nonconserving processes, but they are processes with an extraordinarily low rate. The processes are much too slow to be detected, but it is interesting that they arise precisely because baryon-number conservation is not related to any kind of barytropic field; no such effects could produce a nonconservation of quantities, such as electric charge, that are related to fields of force.

Another hint that baryon number may not be conserved comes from cosmology. One might have supposed, if only on aesthetic grounds, that the universe began with equal amounts of matter and antimatter and therefore with equal numbers of baryons and antibaryons. On this hypothesis the universe would have started with a total baryon number equal to zero. If baryon number were conserved, the total baryon number would have remained zero. Almost all protons and neutrons would have been annihilated through collisions with antiprotons and antineutrons and the universe today would contain only a thin gruel of photons and neutrinos, with no stars or planets or scientists.

It is possible the universe started with an excess of matter over antimatter, so that something would be left over after the annihilation of particles and antiparticles. It is also possible (although it is generally regarded as unlikely) that matter and antimatter have somehow become segregated and that we live in a patch of positive baryon number in a universe with a total baryon number of zero. If baryon number is not conserved, however, there is a more appealing possibility, namely that the universe did start with equal amounts of matter and antimatter and that the present excess of particles with positive baryon number is due to physical processes that have violated the conservation of baryon number. (It has been known since the 1964 experiment of James H. Christenson, James W. Cronin, Val L. Fitch and René Turlay of Princeton University that there is no exact matter-antimatter symmetry that would require processes in which antibaryons are created to go on at the same rate as those that create baryons.) These considerations, together with the absence of barytropic forces as shown by the arguments of Lee and Yang, led some theorists (including the Russian physicist Andrei D. Sakharov and me) to suggest in the 1960's that baryon number may not be exactly conserved. Cosmological considerations also instigated at least one of the proton-decay experiments done in this period, that of T. Alväger, I. Martinson and H. Ryde of the University of Stockholm and the Nobel Institute. In recent years many theorists have worked out schemes for the production of baryons in the very early universe.

Any suggestion of possible baryon-number nonconservation has to immediately confront the fact that ordinary matter is very stable. Maurice Goldhaber of the Brookhaven National Laboratory has remarked that "we know in our bones" the average lifetime of the proton is longer than about $10^{16}$ years. If the lifetime were any shorter, the $10^{28}$ or so protons in the human body would be decaying at an average rate of more than $10^{12}$ protons per year, or 30,000 decays per second, and we would be a health hazard to ourselves.

Of course, one can set a more stringent limit on the proton lifetime by actively searching for proton decay. The first experiment of this kind was carried out in 1954 by Frederick Reines and Clyde L. Cowan, Jr., who were then at the Los Alamos Scientific Laboratory, and Goldhaber. They used about 300 liters of a hydrocarbon scintillator, a material in which the energetic charged particles produced by a proton decay would generate a detectable flash of light. As in all subsequent proton-decay experiments, the apparatus was placed underground to shield it from cosmic rays. (The energetic particles of these rays can cause events that might be mistaken for proton decay.) With this precaution they observed only a few scintillations per second, almost all of which could be attributed to cosmic rays that penetrated deep underground. Reines, Cowan and Goldhaber concluded that the average lifetime of a proton or of a bound neutron must be greater than about $10^{22}$ years.

Subsequent experiments done by a number of physicists have gradually increased the empirical lower bound on the lifetime of the proton. The most elaborate search so far whose results have been published was undertaken by a consortium of investigators from Case Western Reserve University, the University of the Witwatersrand and the University of California at Irvine. They monitored 20 tons of a hydrocarbon scintillator at a depth of 3.2 kilometers in a South African gold mine from 1964 to 1971. A recent analysis of their data gave the result that the average lifetime of a proton or a bound neutron is longer than about $10^{30}$ years.

That is truly a long lifetime. For the purpose of comparison the present age of the universe is estimated to be a mere $10^{10}$ years. One can hope to observe the decay of particles with such long lifetimes only because radioactive-decay processes operate statistically: a sample of particles with an average lifetime of $t$ years will not all survive for $t$ years and then decay in unison; rather, a fraction $1/t$ of the total number of particles will decay in the first year, $1/t$ of the remaining particles will decay in the next year, and so on. The lower bound on the proton lifetime is set not by watching one proton for a long time and waiting for it to decay but by watching the $10^{31}$ protons and neutrons in 20 tons of scintillator for several years and waiting for a few dozen to decay.

It is the long lifetime of the proton that led to the idea of baryon-number conservation. How could the proton live so long if there were not some conservation law that makes it live forever? In the past few years an answer has emerged.

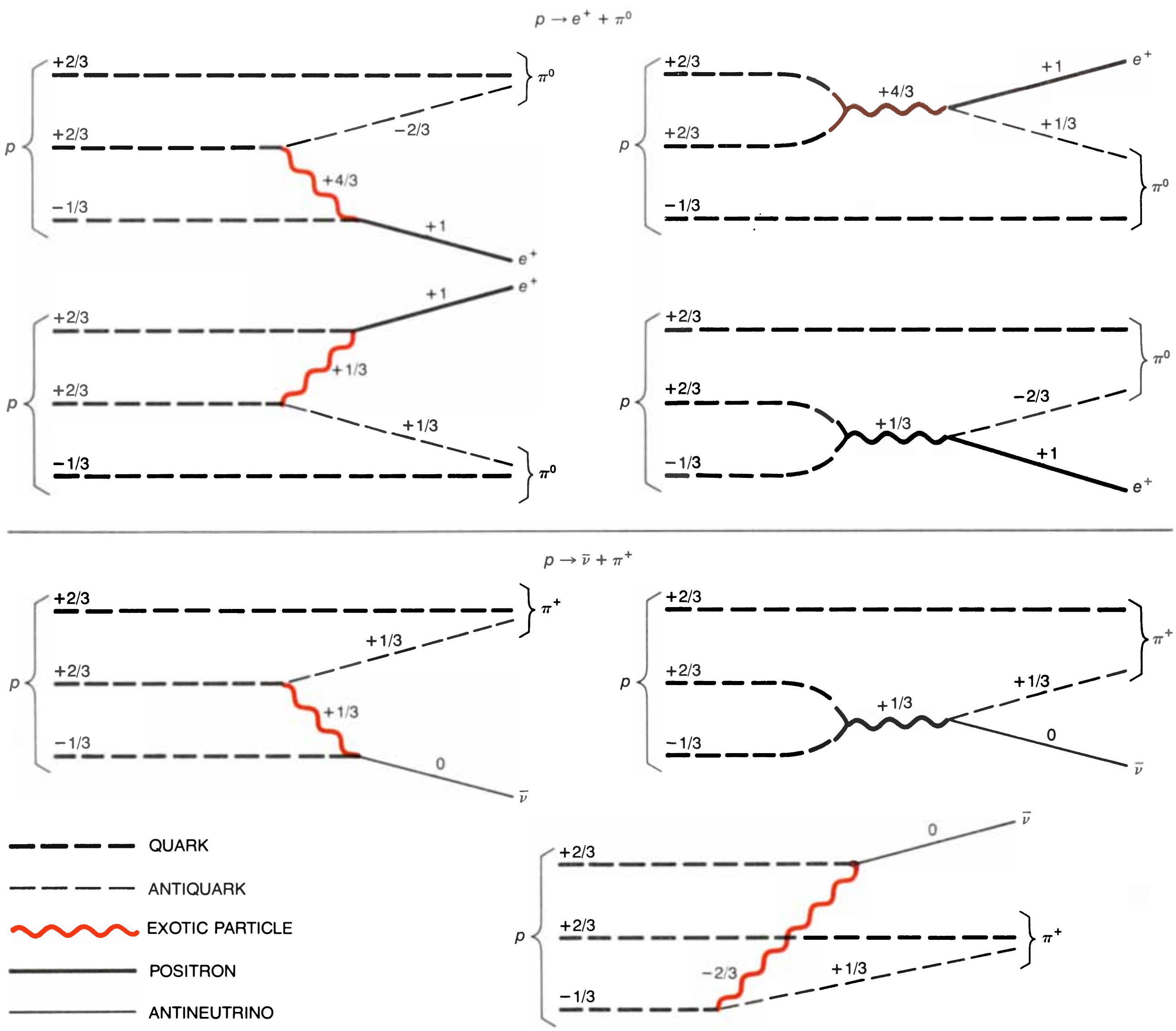
Remember that the modern theory of the weak, the electromagnetic and the strong interactions is highly constrained, so much so, for example, that it is impossible for the strong interactions to violate the conservation of the quantity called strangeness. It happens the theory is so constrained that it cannot be complicated enough (apart from the tiny 't Hooft effect) to allow for any violation of baryon-number conservation unless one introduces new kinds of particles with exotic values of charge, color and so on. Such particles would have to be qualitatively different from any particles now known.

If exotic particles of the right kind are admitted, proton decay becomes a possibility. The familiar conservation laws for charge, color and so on indicate that what is needed is a particle with an electric charge of +4/3, +1/3 or −2/3 times the charge of the proton; the particle must also have an intrinsic spin angular momentum equal to 0 or 1 and a color identical with that of the antiquark. For example, such an exotic particle could be produced when a quark turns into an antiquark and could then be destroyed when another quark turns into an antilepton (a positron, an antimuon or an antineutrino); in this way the three quarks that make up the proton could decay into an antilepton and a meson formed from the leftover quark and the antiquark.

Any such exotic particles would have to be very heavy or they would already have been detected. If they were heavy enough, they would be emitted and reabsorbed only with difficulty and hence would induce only a very small rate of proton decay. Thus it is now possible to explain the long lifetime of the proton without assuming any fundamental conservation law that would ensure it lives forever, and this has opened up the possibility that it does not live forever.

How heavy would the exotic particles have to be in order to explain the long lifetime of the proton? Assuming that the exotic particles interact more or less like photons, one can roughly estimate that a proton lifetime longer than $10^{30}$ years requires an exotic particle whose mass is greater than about $10^{14}$ proton masses. This is a stunningly large mass, larger than anyone can hope to produce with any accelerator that can now be envisioned. Yet there are at least two reasons for suspecting such enormously heavy particles may actually exist.

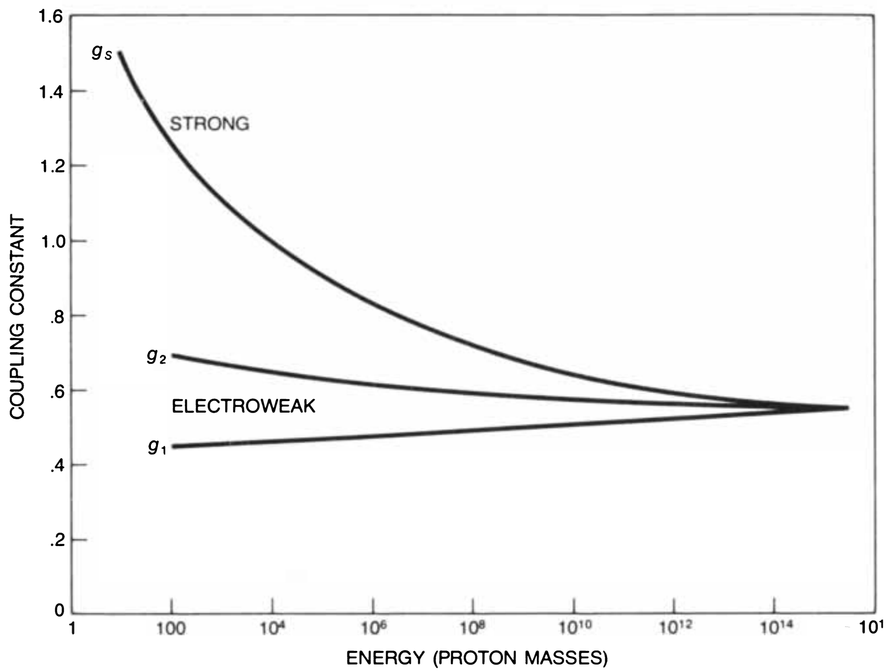
The first reason has to do with the phenomenon of gravitation, which has so far been left out of these considera-




**DECAY OF THE PROTON** by the emission and absorption of a heavy exotic particle would most likely give rise to a positron ($e^+$) and a neutral meson ($\pi^0$) or to an antineutrino ($\nu$) and a positive meson ($\pi^+$). The proton is thought to be a composite particle made up of three constituents called quarks; a meson consists of a quark and an antiquark. The emission and absorption of an exotic particle has the effect of transmuting two quarks into an antiquark and a positron (*upper four reactions*) or into an antiquark and an antineutrino (*lower three reactions*). The quarks, the positrons, the antineutrinos and the exotic particles are labeled by their electric charge. Charge is conserved in the seven reactions, which in essence represent all the ways the emission and absorption of a single exotic particle could lead to the proton's decaying into a meson and a positron or into a meson and an antineutrino. Several kinds of exotic particles could lead to proton decay. They are distinguished by their electric charge (plus or minus 1/3, plus or minus 2/3 and plus or minus 4/3) and by their intrinsic spin angular momentum. The spin can be equal to zero (as it is in the pi meson) or it can be equal to 1 (as it is in the photon).

tions. Einstein's general theory of relativity provides a satisfactory theory of gravitational interactions of particles at all experimentally accessible energies. Because of quantum fluctuations, however, the theory breaks down at very high energies, on the order of $10^{19}$ proton masses. This is known as the Planck mass, after Max Planck, who noted in 1900 that some such mass would appear naturally in any attempt to combine his quantum theory with the theory of gravitation. The Planck mass is roughly the energy at which the gravitational force between particles becomes stronger than the electroweak or the strong forces. In order to avoid an inconsistency between quantum mechanics and general relativity, some new features must enter physics at some energy at or below $10^{19}$ proton masses.

The other reason for expecting new degrees of freedom to appear at superlarge energies has to do with the electroweak and the strong interactions. The modern theory of these interactions involves three parameters, which are known as coupling constants. One of the coupling constants, called $g_s$, describes the strength with which the gluon fields of the strong force interact with particles that carry the conserved quantities designated color; the other two coupling constants, called $g_1$ and $g_2$, describe the strength with which the electroweak fields interact with particles carrying the corresponding quantities, electroweak hypercharge and electroweak isospin. One would like to believe all these interactions have some common origin, in which case the coupling constants should all have the same order of magnitude. But this is in apparent disagreement with the obvious fact that the strong interactions are strong; measurements yield a value for $g_s$ that is much larger than the value of $g_1$ or $g_2$.




**STRENGTH OF THREE FUNDAMENTAL FORCES (the strong force, the electromagnetic force and the weak force) that govern the interactions of elementary particles varies with the energy of the processes in which the forces are measured. The electromagnetic force acts between particles with an electric charge, the strong force binds quarks together to form protons and certain other particles and the weak force is responsible for certain radioactive decays. The electromagnetic force and the weak force have been subsumed under one theory called the electroweak theory, which posits two kinds of electroweak force. The intrinsic strength of the strong force and of the electroweak forces is given by three dimensionless coupling constants (designated $g_s$ for the strong force and $g_1$ and $g_2$ for the electroweak forces). Although the quantities are called constants, they vary slowly with energy. If the forces have a common origin, then at some energy the coupling constants are expected to have the same value; in other words, the three forces are expected to have the same strength. Under fairly general assumptions the unification energy is calculated to be about $10^{15}$ times the mass of the proton.**

A solution to this difficulty was proposed at Harvard University in 1974 by Howard Georgi, Helen R. Quinn and me. It had been known since the 1954 work of Murray Gell-Mann of the California Institute of Technology and Francis E. Low of the Massachusetts Institute of Technology that coupling constants depend somewhat on the energy of the physical processes in which they are measured. In 1973 independent calculations by H. David Politzer of Harvard and David Gross and Frank Wilczek of Princeton showed that the strong coupling constant $g_s$ decreases slowly with increasing energy. The larger of the two electroweak coupling constants, $g_2$, also decreases, but more slowly, whereas the smaller electroweak coupling constant, $g_1$, increases with increasing energy. What Georgi, Quinn and I proposed was that the scale of energies at which the strong interactions become unified with the electroweak interactions is enormously high, so high that the very slow decrease with increasing energy of the strong coupling constant and the even slower variation with energy of the two electroweak coupling constants bring them all to essentially the same value at this superhigh energy scale. Specifically, under rather general assumptions (in essence, that the strong and the electroweak interactions are unified by some set of symmetries of the kind known mathematically as a "simple" group, that there are no intermediate stages of unification and that elementary particles with one-half unit of spin form patterns more or less like the familiar patterns of the leptons and the quarks) we found that the energy scale at which the strong and the electroweak interactions are unified is on the order of $10^{15}$ to $10^{16}$ proton masses.

Any theory that unifies the strong and the electroweak interactions and puts leptons and quarks on the same footing would have to involve new kinds of particles to fill out the picture, and, as I have argued above, there is no good reason to think the interactions of those new particles would conserve baryon number. (The energy scale of $10^{15}$ to $10^{16}$ proton masses calculated by Georgi, Quinn and me is sufficiently high so that baryon-number nonconserving interactions that could be produced by exotic particles this heavy would not lead to a proton lifetime in contradiction with the present experimental lower limit of $10^{30}$ years. We estimated a lifetime very roughly on the order of $10^{32}$ years.)

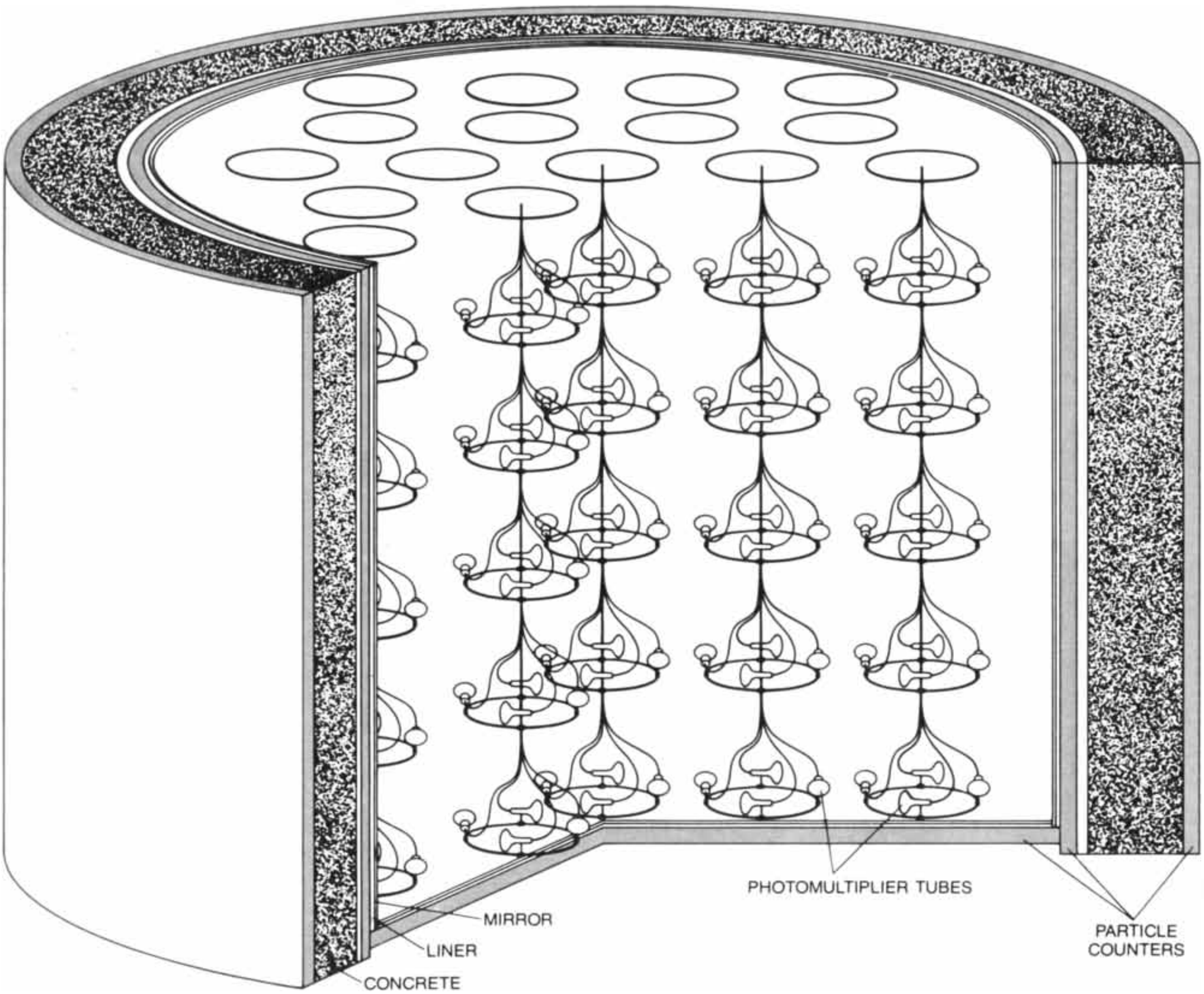
Starting in 1973 many theorists have worked to develop such theories, including Jogesh C. Pati of the University of Maryland and Abdus Salam of the International Centre for Theoretical Physics in Trieste, Georgi and Sheldon Lee Glashow of Harvard, Harold Fritsch and Peter Minkowski of Cal Tech and Feza Gürsey, Pierre Ramond and Pierre Sikivie of Yale University. The models are generally known by the name of the mathematical group of symmetries that connects the various forces, such as $SU(4)^4$, SU(5), SO(10), $E_6$, $E_7$ and SU(7). These models all include exotic particles that when emitted or absorbed convert a quark into an antiquark, a lepton or an antilepton; therefore, as already pointed out by Pati and Salam in the first paper on the unification of the strong and the electroweak interactions, they can violate the conservation of baryon number. Furthermore, all these models in at least some of their versions satisfy the general assumptions made by Georgi, Quinn and me, so that the mass scale of the exotic particles would be expected to be on the order of $10^{15}$ to

$10^{16}$ proton masses and the proton lifetime would be on the order of our rough estimate, $10^{32}$ years.

More recent refined calculations have been made by many theorists, including Andrzej Buras, John Ellis, Mary K. Gaillard and Demetres V. Nanopoulos of the European Organization for Nuclear Research (CERN), Terrence J. Goldman and Douglas A. Ross of Cal Tech, William J. Marciano and Alberto Sirlin of Rockefeller University and New York University, Cecilia Jarlskog and Francesco Yndurain of CERN and Lawrence Hall of Harvard. The newer calculations give an improved value of about $10^{15}$ proton masses for the superheavy-mass scale and a proton lifetime of about $10^{31}$ years. Unfortunately the calculation of the proton-decay rate is complicated by the presence of strong nuclear forces acting among the quarks and antiquarks in the proton and in the decay products, and so even if the properties of the superheavy exotic particles were known precisely, it would probably remain impossible to predict the proton lifetime more accurately than to within an order of magnitude or so.

Experimental studies of the weak interactions have already provided limited verification of the general analysis made by Georgi, Quinn and me. It would be no surprise to find that the graphs for any two of the three coupling constants intersect somewhere, but in order for the three curves of coupling constant v. energy to cross at the same point it is necessary to impose one condition on their starting points, that is, on the values of the coupling constants at low energies. We used this condition to calculate that under our general assumptions a certain parameter (related to the ratio of $g_1$ and $g_2$), which describes the unification of the weak and the electromagnetic interactions, has a value close to .2. Experiments in the physics of electron and neutrino interactions currently indicate a value for this parameter of from about .2 to .23. The theoretical and experimental values are close enough to encourage us to take this analysis seriously even though it involves an extrapolation of an unprecedented extent: by 13 orders of magnitude in energy.

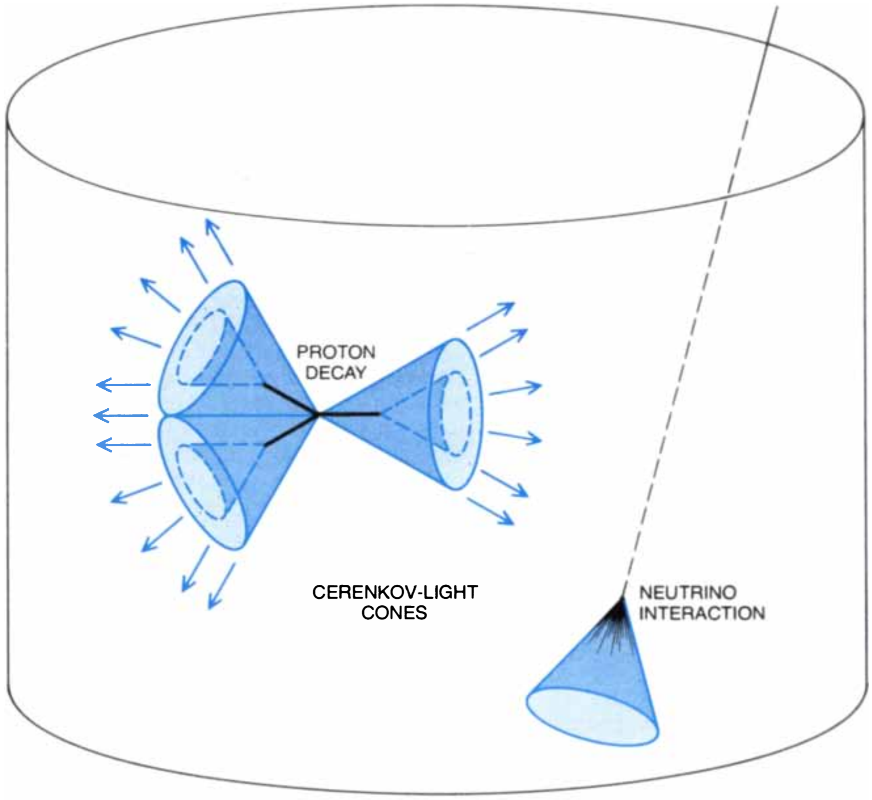
A mass on the order of $10^{15}$ proton masses is so large that the emission and absorption of particles this heavy is almost impossible at experimentally accessible energies and can therefore produce only tiny effects in any feasible experiment. The only hope for detecting these tiny effects lies in the possibility that they may violate otherwise exact conservation laws and thereby make possible processes that otherwise would be strictly forbidden. One of these conservation laws is baryon-number conservation, which is tested by looking for




**THOUSAND-TON DETECTOR is being constructed in the Silver King mine near Park City, Utah. The detector will have 800 photomultiplier tubes distributed throughout a volume of water. The photomultiplier tubes are supported by hoops of wire, and the chamber is surrounded by additional particle counters and a thick concrete shield. The monitored volume of water will have some $6 \times 10^{32}$ protons and neutrons. The detector is being built by workers from Harvard University, Purdue University and the University of Wisconsin.**

proton decay. The only other known conservation law that is not required for the consistency of theories of particle interactions and therefore could be violated by superhigh-energy effects is lepton-number conservation, the conservation of the total number of leptons (neutrinos, electrons, muons and so on) minus the number of antileptons. The nonconservation of lepton number could show up in such processes as neutrinoless double beta decay: the decay of two protons in a nucleus into two neutrons plus two positrons. Baryon number would remain constant in this reaction, but lepton number would decrease by 2. (The reaction does not violate the conservation of energy because the neutrons in the final state are not free but have a negative binding energy in the nucleus.) Another sign of the violation of lepton-number conservation would be a nonzero neutrino mass.

Several new attempts to determine the lifetime of the proton are now in progress. Experiments are already in operation in the Soudan mine in Minnesota, the Kolar gold field in southern India, the Mont Blanc tunnel between France and Italy and the Baksan valley in the Caucasus range of the U.S.S.R. By the end of the year three other experiments should be recording data; they are in the Morton salt mine in Ohio, the Silver King mine in Utah and at another site in the Mont Blanc tunnel. Other experiments are under consideration.




**CERENKOV RADIATION is a signal not only of possible proton decay but also of interactions of high-energy particles from space. Indeed, most flashes of light detected by the photomultiplier tubes will not be from the products of proton decay. The number of the cones of Cerenkov radiation and their orientation will help to distinguish proton decay from other events, such as the collision of a neutrino created in the earth's atmosphere by cosmic rays with a proton or neutron in the detector. The decay of a proton into a positron and a neutral pi meson, for example, could give rise to three cones of Cerenkov light at certain characteristic angles. The positron would generate one cone, and the decay of the neutral pi meson into two photons, each of which could give rise to a shower of electrically charged particles, would be responsible for the other two cones of light. The neutrino interaction generates only one cascade of electrically charged particles and so it is observed as a single cone of Cerenkov radiation.**

The basic technique of all the experiments is to compensate for the extreme slowness of proton decay by careful monitoring of a very large mass of material. The larger the mass, the larger the number of protons and bound neutrons and hence the greater the probability of observing a decay. In this way it is expected that it will be possible to detect the decay of the proton or the bound neutron with an average lifetime rather longer than the current bound of $10^{30}$ years. The experiments differ chiefly in the nature and amount of material monitored, the nature and arrangement of the devices used to detect a proton decaying in the material and the characteristics of the experiment that suppress spurious signals from cosmic rays, including the depth underground at which the experiments are carried out.

Since a very large mass must be monitored, the experiments must use some relatively inexpensive material, such as water, concrete or iron. In experiments such as those in the Soudan mine, the Kolar gold field and the Mont Blanc tunnel, which use iron or concrete, one must rely on detection devices such as proportional tubes or streamer tubes, which can directly detect the energetic charged particles that are expected to be emitted in proton decays. The charged particles have a short range in iron or concrete, and so the detector tubes must be closely spaced throughout the monitored material.

On the other hand, experiments such as those in the Morton salt mine, the Silver King mine and the Homestake gold mine, which monitor a transparent material such as water, can use a rather different strategy. The energy released in proton decay is great enough so that an electron, a positron, a muon or a pi meson emitted in the decay process is likely to have a very high speed, slower of course than the speed of light in vacuum but faster than the speed of light in water. When a charged particle travels through a transparent medium at a speed higher than the speed of light in that medium, there occurs what is known as the Cerenkov effect. It is like the sonic boom generated by an airplane traveling faster than the speed of sound in air, but the Cerenkov effect is an optical boom, in which the particle emits a cone of light rather than of sound. (The beautiful blue glow of Cerenkov light was noted in early experiments on radioactivity by Marie Curie, but its properties were first explored in detail in the 1930's by Pavel A. Cerenkov.) The angle between the Cerenkov light rays and the path of the charged particle depends on the ratio of the speed of the charged particle to the speed of light in the medium. Most of the light is emitted by particles traveling at nearly the speed of light in vacuum, and in this case the angle has a characteristic value in water of approximately 42 degrees.

Observation of a cone of Cerenkov light is a signal that something has happened in the medium to create a fast charged particle. Furthermore, for a given initial speed, the thickness of the cone and the amount of light emitted depend only on the distance the charged particle travels before its speed drops below the speed of light in the medium, which in turn depends on its initial energy. By registering the positions where light is received and the light's intensity, one can therefore deduce the initial energy as well as the direction of each charged particle. In certain circumstances the moving particle itself might decay, emitting other particles that could give rise to a second flash of Cerenkov light. A muon or an antimuon created in a proton decay could decay into an electron that emitted Cerenkov radiation. A charged pi meson might decay into a slow muon (or antimuon), which in turn could decay into a fast electron (or positron) that would give off a cone of light. A neutral pi meson could decay into two photons, each of which would give rise to a shower of charged particles and the accompanying Cerenkov light. The use of Cerenkov light to detect proton decays thus offers an alternative to the direct observation of charged particles as a means of reconstructing the decay process and verifying that one is really seeing a proton decay.

One advantage of using Cerenkov light over other means of detecting proton decay is that the light can travel greater distances in water than the charged particles themselves can. Hence with a given volume of monitored material one needs fewer detectors than one does in experiments that use opaque materials such as iron or concrete. Also, water is cheaper than iron or concrete. On the other hand, water Cerenkov detectors are sensitive only to charged particles traveling faster than the speed of light in water. Also, the relatively low density of water requires that larger cavities be excavated underground to hold a given mass of monitored material, and the water has to be kept very pure to maintain its transparency to Cerenkov light. As I have indicated, work is proceeding vigorously both on experiments using water as the monitored material and on those using denser materials.

What proton-decay rates could be detected in these experiments? To take one example, the planned experiment with the largest monitored mass is the one in the Morton salt mine. Of its monitored mass of 10,000 tons of water, an outer shell of perhaps 5,000 tons will be used to keep track of the background of cosmic rays. The remaining mass of about 5,000 tons of water includes $3 \times 10^{33}$ protons (and bound neutrons). If the proton has the average lifetime of about $10^{31}$ years that is indicated by refined versions of the analysis done by Georgi, Quinn and me there should be some 300 proton-decay events per year. A few years of observation would yield a few proton-decay events even if the lifetime were as long as $10^{33}$ years, but at such low decay rates the experiments would start to be jeopardized by an unremovable background of spurious events due to cosmic-ray neutrinos, and further improvement would be difficult.

What will be learned if proton decay is discovered? Of course one will imme-

| LOCATION | DATE OF OPERATION | DEPTH (EQUIVALENT METERS OF WATER) | MONITORED MATERIAL |
|---|---|---|---|
| HOMESTAKE GOLD MINE, SOUTH DAKOTA | NOW | 4,400 | 150 TONS OF WATER (TO BE INCREASED TO 900 TONS) |
| KOLAR GOLD FIELD, INDIA | NOW | 7,600 | 150 TONS OF IRON (½-INCH PLATES) |
| BAKSAN VALLEY, U.S.S.R. | NOW | 850 | 80 TONS OF LIQUID SCINTILLATOR |
| SOUDAN MINE, MINNESOTA | NOW | 1,800 | 30 TONS OF TACONITE CONCRETE AND IRON (TO BE INCREASED TO 1,000 TONS) |
| MONT BLANC TUNNEL BETWEEN FRANCE AND ITALY | NOW | 4,270 | 30 TONS OF IRON AND LIQUID SCINTILLATOR (TO BE INCREASED TO 200 TONS) |
| MORTON SALT MINE, OHIO | 1981 ESTIMATED | 1,670 | 10,000 TONS OF WATER |
| SILVER KING MINE, UTAH | 1981 ESTIMATED | 1,700 | 1,000 TONS OF WATER |
| MONT BLANC TUNNEL BETWEEN FRANCE AND ITALY | 1981 ESTIMATED | 5,000 | 150 TONS OF IRON INITIALLY (ONE-CENTIMETER-THICK PLATES) |
| GRAND SASSO TUNNEL, ITALY | UNDER CONSTRUCTION | 4,000 | 10,000 TONS OF IRON (THREE-MILLIMETER PLATES) |
| JAPAN | UNDER CONSIDERATION | 2,700 | 3,400 TONS OF WATER |
| FRÉJUS TUNNEL BETWEEN FRANCE AND ITALY | UNDER CONSIDERATION | 4,500 | 1,500 TONS OF IRON (PLATES THREE TO FOUR MILLIMETERS THICK) |
| JAPAN | UNDER CONSIDERATION | 2,700 | 600 TONS OF IRON |
| ARTYEMOVSK SALT MINE, U.S.S.R. | UNDER CONSIDERATION | 600 | 100 TONS OF LIQUID SCINTILLATOR |

**TABLE OF PROTON-DECAY EXPERIMENTS includes five detectors that are now operating, three detectors that are expected to begin recording data this year and five detectors that are under consideration. The experiments differ in the kind and the amount of materi-**

diately conclude that baryon number is not conserved, and this will support the growing belief that all conserved quantities have a dynamical significance similar to that of electric charge. Furthermore, if proton decay is discovered in the near future, the lifetime will have to be in the range of from $10^{30}$ to $10^{33}$ years, and this will lend some credence to the general assumptions about the unification of the strong and the electroweak forces that were used by Georgi, Quinn and me. There are a great many theories, however, that satisfy these general assumptions, including at least some versions of the models SU(4)[4], SU(5), SO(10) and so on mentioned above. It will be difficult to tell which (if any) of these theories describes physics at very high energy.

About one thing we can be sure. If proton decay is discovered, great new resources will be devoted to its study, and before long there will be a second generation of experiments, in which the effort will be to learn not whether protons decay but how they decay: what are the probabilities for the various modes of decay?

In preparation for this effort a number of theorists have been exploring the likely modes of proton decay. (The remarks below are based on independent work by Wilczek and by Anthony Zee and me.) Interestingly, one can go pretty far in this analysis without any assumptions about the unification of the strong and the electroweak interactions. All that one needs are the familiar conservation laws for charge, color and so on and the assumption that the exotic particles responsible for proton decay are very heavy, as they surely must be to explain the long proton lifetime. In this case, although the emission and reabsorption of these particles can produce a great many different modes of proton decay, the more complicated modes are more strongly suppressed by the large mass of the exotic particles than the simpler modes. Unless some special circumstance intervenes, the dominant modes of decay will in general be those in which the proton or the bound neutron decays into a positron, an antimuon or an antineutrino, plus some number of mesons, rather than into an electron, a muon or a neutrino, plus some number of mesons. One can go further and make predictions about ratios of decay rates. For instance, a neutron decays just twice as fast as a proton into a positron and a single pi meson or rho meson. The proton decays faster into a positron plus mesons than a neutron decays into an antineutrino plus mesons. The neutron decays faster into a positron and mesons than the proton decays into an antineutrino and mesons.

| DETECTORS | INSTITUTIONS |
|---|---|
| · PHOTOMULTIPLIER 3ES IN WATER | UNIVERSITY OF PENNSYLVANIA |
| S PROPORTIONAL COUNTERS TWEEN PLATES | TATA INSTITUTE OF FUNDAMENTAL RESEARCH, BOMBAY<br>UNIVERSITY OF OSAKA<br>UNIVERSITY OF TOKYO |
| 00 PHOTOTUBES | INSTITUTE FOR NUCLEAR RESEARCH, MOSCOW |
| 56 GAS PROPORTIONAL UNTERS IN CONCRETE | UNIVERSITY OF MINNESOTA<br>ARGONNE NATIONAL LABORATORY |
| OTOTUBES AND REAMER CHAMBERS | INSTITUTE FOR NUCLEAR RESEARCH, MOSCOW<br>UNIVERSITY OF TURIN |
| 00 PHOTOMULTIPLIER BES IN WATER | UNIVERSITY OF CALIFORNIA AT IRVINE<br>UNIVERSITY OF MICHIGAN<br>BROOKHAVEN NATIONAL LABORATORY |
| ) PHOTOMULTIPLIER BES IN WATER AND RRORED WALLS | HARVARD UNIVERSITY<br>PURDUE UNIVERSITY<br>UNIVERSITY OF WISCONSIN |
| REAMER CHAMBERS TWEEN PLATES | ITALIAN NATIONAL SYNCHROTRON LABORATORY AT FRASCATI<br>UNIVERSITY OF MILAN<br>UNIVERSITY OF TURIN |
| ASH CHAMBERS TRIGGERED STREAMER CHAMBERS | ITALIAN NATIONAL SYNCHROTRON LABORATORY AT FRASCATI<br>UNIVERSITY OF MILAN<br>UNIVERSITY OF TURIN<br>UNIVERSITY OF ROME |
| 56 20-INCH OTOMULTIPLIER TUBES | UNIVERSITY OF TOKYO<br>JAPANESE NATIONAL LABORATORY FOR HIGH-ENERGY PHYSICS<br>UNIVERSITY OF TSUKUBA |
| ASHTUBES AND REAMER CHAMBERS | ÉCOLE POLYTECHNIQUE<br>UNIVERSITY OF PARIS AT ORSAY<br>SCHOOL OF MINES AT SACLAY |
| ON FLASHTUBES | UNIVERSITY OF OSAKA<br>UNIVERSITY OF TOKYO<br>JAPANESE NATIONAL LABORATORY FOR HIGH-ENERGY PHYSICS |
| PHOTOMULTIPLIER TUBES | INSTITUTE FOR NUCLEAR RESEARCH, MOSCOW |

**al monitored. Moreover, they are done at different depths in the ground. The depth is given in terms of the equivalent depth of water that would provide the same shielding from cosmic rays.**

One cannot be sure these predictions will be borne out by experiment. If they are not, there must be exotic particles considerably lighter than $10^{14}$ proton masses to produce more complicated decay modes. For example, the decay of a proton or a bound neutron into an electron and mesons rather than into a positron and mesons could be produced at an observable rate if these were exotic particles no heavier than about $10^{10}$ proton masses. A decay into three neutrinos or three electrons (or some other combination of three leptons) could be observable if there were exotic particles no heavier than about $10^{4}$ proton masses. Such relatively light exotic particles, however, would have to have special properties to avoid producing the "ordinary" decay of the proton (into a meson and a positron or an antineutrino) at much too high a rate.

The verification of these predictions, that is, the finding that the proton decays into mesons and a positron or an antineutrino and with the above ratios of decay rates, would serve to confirm that proton decay is really due to exotic particles with masses greater than about $10^{10}$ proton masses, but it would not point toward any specific underlying theory. For that purpose it would be necessary to explore finer details of the decay process. (For example, a determination of the direction in which a positron or an antimuon spins when it is produced in a proton decay can be used to diagnose the spin of the superheavy exotic particles whose emission and reabsorption produced the decay.) If proton decay is discovered, it will rank as a triumph of experimental ingenuity and an unparalleled clue to the physics of very high energies, but it will also present experimental and theoretical physicists with many new tasks that will need to be done if the mechanism of proton decay is to be understood.

# The Beta-Lactam Antibiotics

*The penicillins and cephalosporins are chemically characterized by the beta-lactam ring. Their molecular structure has been much modified to increase their efficacy and combat resistant bacteria*

by E. P. Abraham

Nearly 40 years have passed since penicillin was introduced into medicine. Not long before that the response to sulfonamides of patients with puerperal fever and other dangerous illnesses caused by streptococci had shown for the first time that the effective use of small molecules to treat generalized bacterial infections was within the range of possibility. Many who practice medicine today have no memory of an earlier time when little could be done to help those with serious bacterial infections, when bacterial endocarditis was almost invariably fatal, when meningococcal meningitis left its few survivors with pitiable disabilities and when pneumococcal pneumonia was known as "the old man's friend," because it was a merciful cause of death. Penicillin, with its high activity against a number of pathogenic bacteria and its almost negligible toxicity to man, seemed to have properties that were almost miraculous, and it heralded a new era in the chemical treatment of disease.

Following the success of penicillin a vast international search uncovered a host of other naturally occurring antibacterial substances. Many were highly toxic and were soon forgotten. Others, such as bacitracin, were of marginal usefulness in human medicine. Those with enduring clinical value include streptomycin (and other amino glycosides such as gentamicin), the erythromycins and the family of the tetracyclines. The early penicillins were succeeded by a variety of modified penicillins that proved to be effective against bacterial strains that had developed penicillin resistance. The basis of antibiotic resistance and the mechanism of antibiotic activity has become a lively branch of cell biology and genetics. The penicillin family has also been enlarged to include the chemically related cephalosporins, discovered at the University of Oxford, which are said now to account for 15 to 20 percent of all drug expenditures in American hospitals.

The critical event that opened the antibiotic era in medicine was the decision by Howard W. Florey and Ernst B. Chain in 1938 to make a systematic survey, in the Sir William Dunn School of Pathology at Oxford, of antibacterial substances that had long been known to be manufactured by microorganisms. Florey and Chain were motivated mainly by scientific interest and not by the expectation that the project would lead to substances of clinical value. By good fortune penicillin was one of three substances chosen to be studied first.

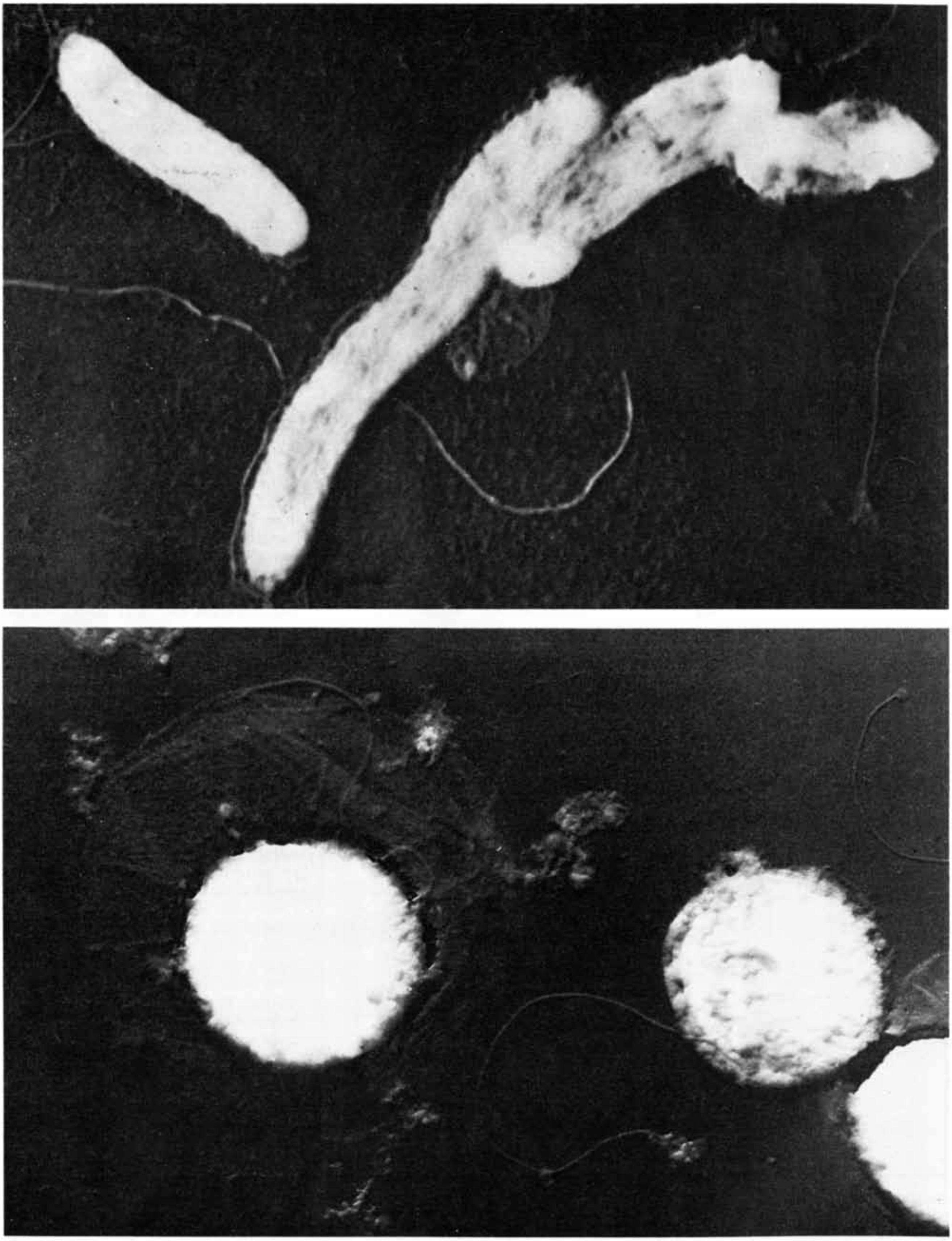
Penicillin itself was not new. It was the name given by Alexander Fleming in 1929 to an active "mold broth filtrate" he had obtained after the growth of a fungus, *Penicillium notatum*, in a nutrient broth at St. Mary's Hospital in London. Fleming had been studying variants of staphylococci, the bacteria that can produce boils, abscesses and sometimes generalized and fatal infections. The fungus alighted by chance on a plate of nutrient agar that Fleming had seeded with the bacteria and had put aside before a summer vacation. A period of cool weather allowed the fungus to grow and secrete penicillin before the growth of the bacteria, which required a higher temperature, was complete. On his return to the laboratory Fleming observed that the colonies of bacteria that had been growing in the vicinity of the *Penicillium* had partly lysed, or dissolved. The lysis would not have occurred if the weather had been warmer, enabling the staphylococci to mature first: penicillin brings about the lysis of growing bacteria but not of those in cultures in which growth has ended.

Fleming was not the first to observe that a fungus of the genus *Penicillium* was antagonistic to the growth of bacteria in its vicinity. This had been noted by John Burdon Sanderson in 1870, by Joseph Lister a year later, by William Roberts in 1874, by John Tyndall in 1876 and by a French medical student, Ernest Duchesne, in 1896. Whether any of those investigators were observing the effects of penicillin will never be known, because species of *Penicillium* secrete a number of other antibacterial substances.

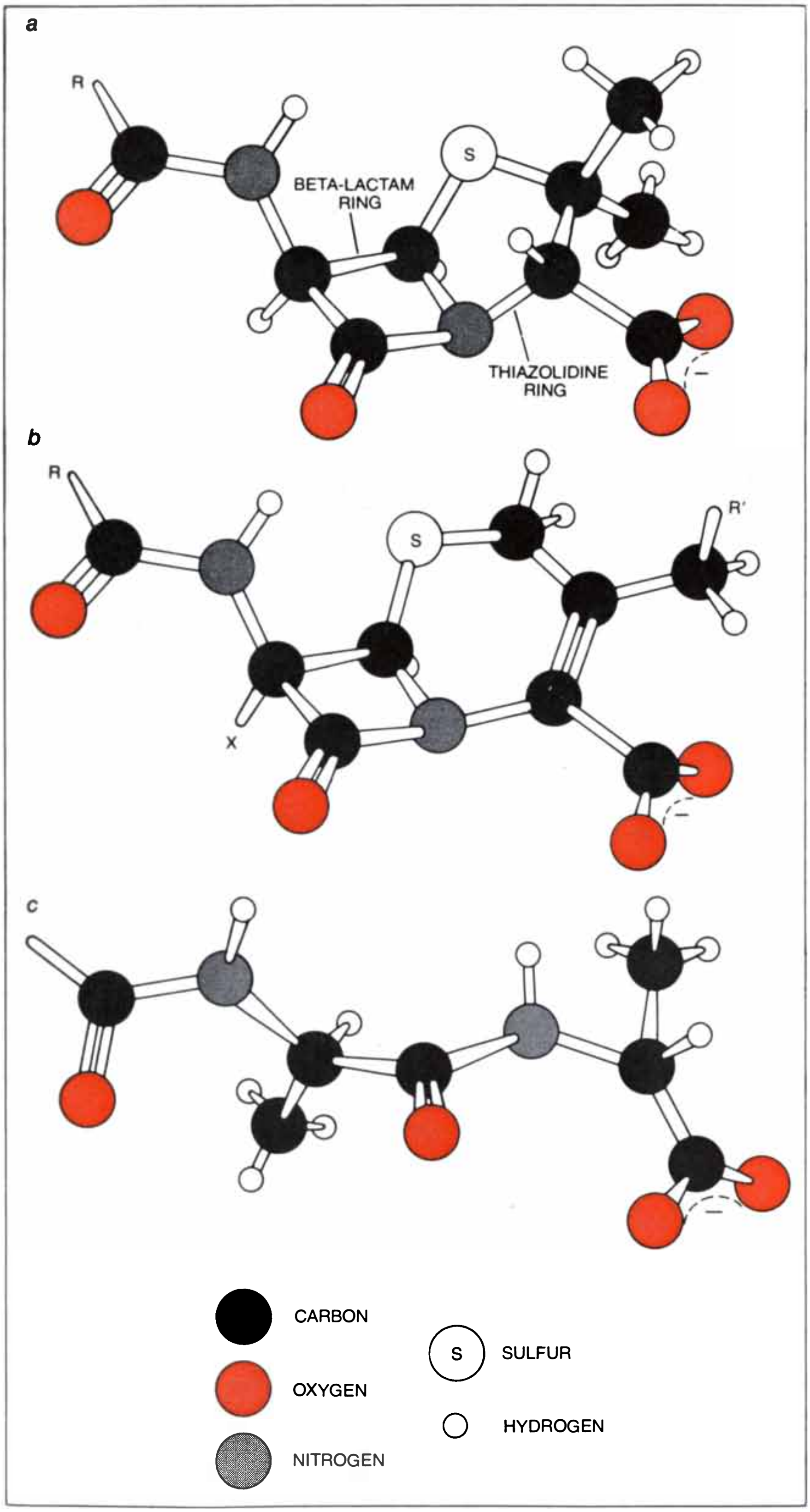
In any event Fleming was an acute observer of the unexpected who was intrigued by the phenomenon he had seen and studied it further. He showed that penicillin was highly active against several species of Gram-positive bacteria, such as the staphylococci, the streptococci and the pneumococci, but not against a variety of pathogenic Gram-negative bacilli such as *Salmonella typhi, Escherichia coli* and *Pseudomonas aeruginosa*. Gram-positive and Gram-negative bacteria are distinguished by their response to a staining procedure developed by Hans Christian Joachim Gram in 1884, which is now known to reflect important differences in the cell wall of the organisms.

Fleming also showed that his penicillin-containing broth was no more toxic than ordinary broth to the white cells of the blood and to a live rabbit. He had no conception, however, of the magnitude of his discovery. Although he applied penicillin on several occasions as a local antiseptic, he wrote in 1940 that it did not seem worth the trouble of making it. His main interest in the substance appeared to be in its selective antibacterial action, which enabled it to serve for the isolation of certain organisms from mixed cultures. It clearly never occurred to him that penicillin might be introduced into the bloodstream to cure infections. Hence for more than 10 years after his observation penicillin remained little more than a curiosity.

The situation changed dramatically when Florey, Chain and a small group of their colleagues at Oxford succeeded in partly purifying penicillin and demonstrating, first in mice and then in human beings, that it had remarkable therapeutic properties. In 1941 a small clinical trial, limited by the small supply of penicillin (which was husbanded to the point of recovering it from the urine of patients), provided the first of a series of results that Florey described in 1949 as "in truth so gratifying as to be at times almost unbelievable."



EFFECT OF PENICILLIN is shown in these electron micrographs made by Milton R. J. Salton of the New York University School of Medicine. In the micrograph at the top are normal bacteria of the species *Vibrio metchnikovi*. The bottom micrograph shows the "spheroplasts" that develop when the bacteria are grown in the presence of penicillin. Also visible around the spheroplast at the left is the weakened and detached cell wall of the bacterium. Penicillin exerts its effect by interfering with the synthesis of the cell wall as susceptible bacteria are growing. All that preserves the spheroplasts from disintegration is the fragile cytoplasmic membrane, faintly visible here.




**BASIC STRUCTURE of the penicillins and cephalosporins centers on the beta-lactam ring. In penicillin (*a*) the four-membered beta-lactam ring is fused to a five-membered thiazolidine ring, which consists of a sulfur atom, a nitrogen atom and three carbon atoms. The side chain *R* is variable; several side chains of penicillin are depicted in the illustration on the opposite page. In the chemically related cephalosporins (*b*) *R* and *R'* are variable and *X* can be a hydrogen or a methoxyl group (–$OCH_3$). An indication of how the beta-lactam antibiotics may exert their effect by acting on the developing cell wall of a bacterium can be seen in the structure (*c*) of the growing D-alanyl-D-alanine peptidoglycan chain of a bacterial cell wall. One conformation of the peptidoglycan chain is similar to that of part of the penicillin molecule.**

It was this demonstration of the clinical value of penicillin that stimulated the governments of Britain and of the U.S. to support strenuous efforts by pharmaceutical companies to make enough penicillin to treat battle casualties in World War II. The rapid progress that ensued was due in no small measure to the American contribution: the introduction of an improved growth medium, deep (instead of surface) fermentation and much higher-yielding mutant strains of *Penicillium chrysogenum*. These three advances solved what had seemed at first to be a formidable problem indeed. Today the yield of penicillin per liter of culture is many thousands of times the amount that could be obtained in the 1940's at Oxford.

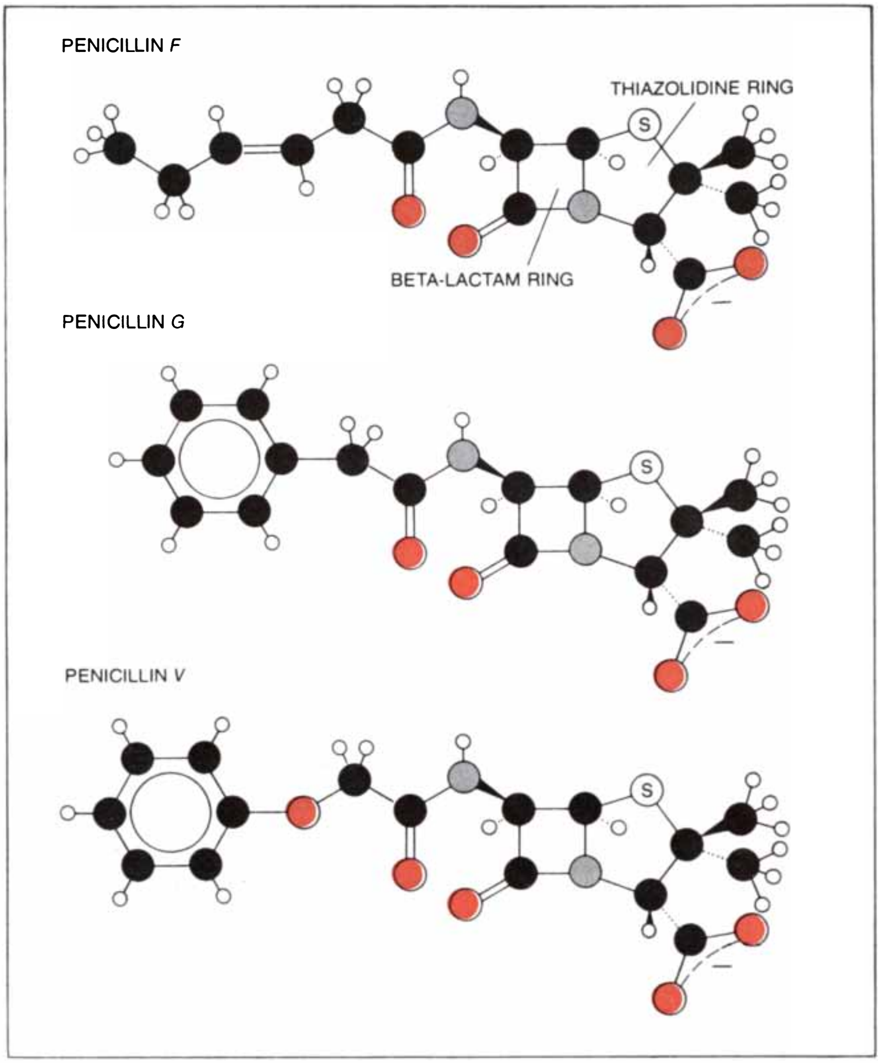
The war years also saw an intensive Anglo-American effort, fostered by both governments, to produce penicillin by chemical synthesis. A successful effort depended on first establishing the structure of the molecule. In 1943 Chain and I had proposed a then novel structure in which a four-membered beta-lactam ring was fused to a five-membered thiazolidine ring: a ring consisting of a sulfur atom, a nitrogen atom and three carbon atoms. The beta-lactam configuration was later strongly advocated by Robert B. Woodward of Harvard University, but it was not generally accepted until 1945, when an X-ray-crystallographic analysis by Dorothy C. Hodgkin and Barbara W. Low at Oxford left no doubt about the relative positions of the atoms in the molecule.

Meanwhile attempts to synthesize the rather unstable penicillin molecule had been disappointing. It was not until 1958 that John C. Sheehan and his colleagues at the Massachusetts Institute of Technology accomplished a successful synthesis with the help of a new reagent that enabled them to close the beta-lactam ring under mild chemical conditions. In spite of this achievement total chemical synthesis for the production of penicillin has not been economically competitive with fermentation. Semisynthesis starting from a fermentation product, however, has led to new compounds of great medical value.

By the early 1950's the advent of penicillin could be seen as a turning point in the history of medical science. The highly selective activity of the substance against bacterial cells gave it an outstanding importance, which it still has today. Nevertheless, it was clear that the penicillin then available had its limitations: some bacteria were resistant to it. As early as 1940 Chain and I had found an enzyme in *E. coli* that destroyed the Oxford penicillin. We suggested that the enzyme, which we named penicillinase, had a role in the resistance of bacteria that manufactured it.

Even before the structure of the characteristic penicillin ring system had been established it was known that the penicillin being made at Oxford differed from that being made in the U.S. These two penicillins, which were respectively named penicillin *F* and penicillin *G*, had the same molecular nucleus but different side chains. Penicillin *G*, which is also known as benzylpenicillin, became the first penicillin to be widely used in medicine and has retained its importance to the present day. It was obtained in the U.S. because a readily available corn-steep liquor, which had been added to the fermentation medium, contained phenylacetic acid ($C_6H_5CH_2COOH$), from which its side chain was derived.

Later many penicillins with chemically related side chains were obtained from fermentations of *Penicillium chrysogenum* by the addition of appropriate precursors to the growth medium. None of these penicillins was markedly superior to benzylpenicillin in its range of activity. In 1954, however, it was discovered that phenoxymethylpenicillin, unlike benzylpenicillin, was stable under the acidic conditions in the stomach and was clinically effective when it was given by mouth.




**SIDE CHAINS of three penicillins are portrayed, in each case on the left side of the basic beta-lactam structure. Penicillin *F* is 2-pentenylpenicillin; *G* is benzylpenicillin, and *V* is the first oral penicillin. The chains depicted here are the *R* side chains of the basic molecule.**

At this time there was little indication that any penicillin-like compound free from the major limitations of benzylpenicillin would be obtained. Indeed, there were signs that the medical value of penicillin would diminish because benzylpenicillin-resistant staphylococci, which manufactured a penicillinase, were becoming common in hospitals. Fortunately several unpredictable developments led to new penicillins, to the chemically related family of the cephalosporins and to other substances incorporating the beta-lactam ring that retain most of the desirable properties of benzylpenicillin and are effective against many benzylpenicillin-resistant bacteria. These compounds, together with the early penicillins, are now known collectively as the beta-lactam antibiotics. At the same time research has thrown light on the mode of action of beta-lactam antibiotics, on the reasons why different members of the family differ in activity, on the factors that can make bacteria resistant to them and on the ways resistance can spread.

Following early observations that penicillin killed only growing bacteria, Joshua Lederberg, who was then working at the University of Wisconsin, showed that rod-shaped cells of *E. coli* growing in a hypertonic medium (one abnormally high in osmotic pressure) changed into spherical cells when penicillin was added to the culture. These "spheroplasts" had a fragile cytoplasmic membrane but lacked most of the outer cell wall of their parents. As a result they burst when the osmotic pressure of the medium outside the cell was reduced by dilution. Lederberg thus deduced that penicillin interfered with the synthesis of the bacterial cell wall.

A partial understanding of the nature of this interference and its consequence has been accompanied by an increase in knowledge of the molecular structure of bacterial cell walls, for which we are indebted to Milton R. J. Salton of the New York University School of Medicine, to Jack L. Strominger of Harvard, to James Baddiley of the University of Newcastle upon Tyne and to many others. In both Gram-positive and Gram-negative bacteria the cytoplasmic membrane is covered with a layer of peptidoglycan, also known as mucopeptide or murein. The peptidoglycan forms a three-dimensional structure in which the backbones consist of alternate units of two amino sugars, N-acetylglucosamine (NAG) and N-acetylmuramic acid (NAM).

The NAM units, which are found only in the walls of bacteria and closely related microorganisms, are linked to peptide chains (short chains of amino acid units) in which both D (dextro, or right-handed) and L (levo, or left-handed) amino acids are represented. Many of these peptide chains in turn are cross-linked to one another. The resulting structure endows the wall with rigidity and the cell with its shape. The peptidoglycan layer of Gram-negative bacteria, however, is thinner than that of Gram-positive bacteria and has fewer cross-links. It is also surrounded by an outer membrane that incorporates phospholipid, lipopolysaccharide and protein and has a hydrophobic (water-repelling) interior.

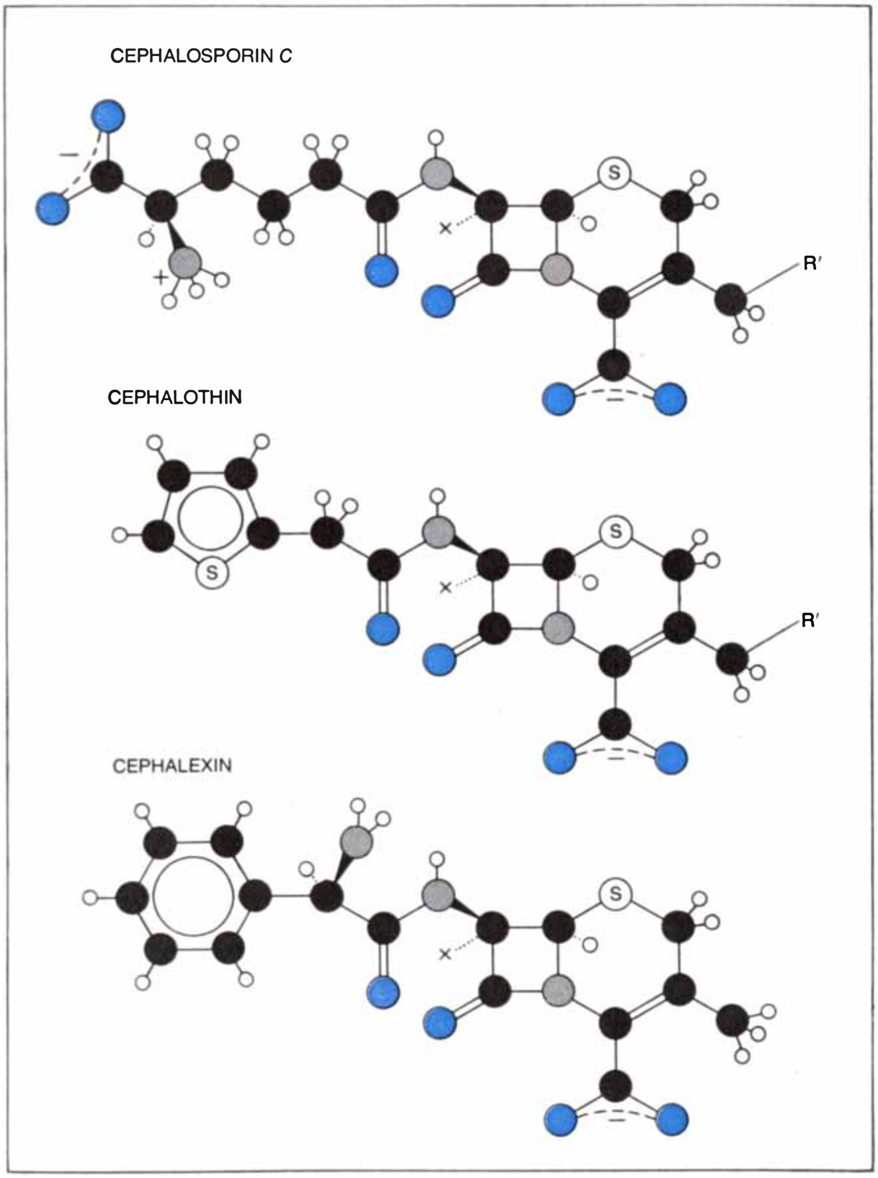
Studies by Strominger, by James T. Park of the Tufts University School of Medicine and by J.-M. Ghuysen of the University of Liège have shown that the cytoplasmic membranes of bacteria and the bacteria-like fungi of the *Streptomyces* family bear enzymes that are inactivated by penicillin and other beta-lactam antibiotics. The vulnerable enzymes are either transpeptidases (which cross-link one peptide chain of new peptidoglycan to another by displacing a terminal D-alanine) or D-carboxypeptidases (which remove a terminal D-alanine by hydrolysis without cross-linking taking place) or enzymes that can do both. The inhibition of cross-linking is believed to be the reaction by which beta-lactam antibiotics initiate events that lead to the death and often the lysis of the bacterial cell. And since peptidoglycan synthesis does not take place in the cells of man and other mammals, it is not surprising that substances selected in the course of evolution to inhibit it should be relatively harmless in the human body.

The events that are directly responsible for killing and lysis, however, are not yet understood. Bacterial cells contain both autolytic enzymes that break down peptidoglycan and enzymes that bring about its synthesis. Alexander Tomasz of Rockefeller University has found that pneumococci bearing a mutation that makes them defective in their autolytic system do not lyse in the presence of penicillin, although their growth is inhibited. Such organisms have been described as tolerant. The question therefore arises whether the autolytic system in normal strains is triggered by the primary change caused by penicillin and is then responsible for the lytic action of the antibiotic. To what degree that is so remains uncertain. It has nonetheless become clear that the mechanism of action of the beta-lactam antibiotics is one of considerable complexity.

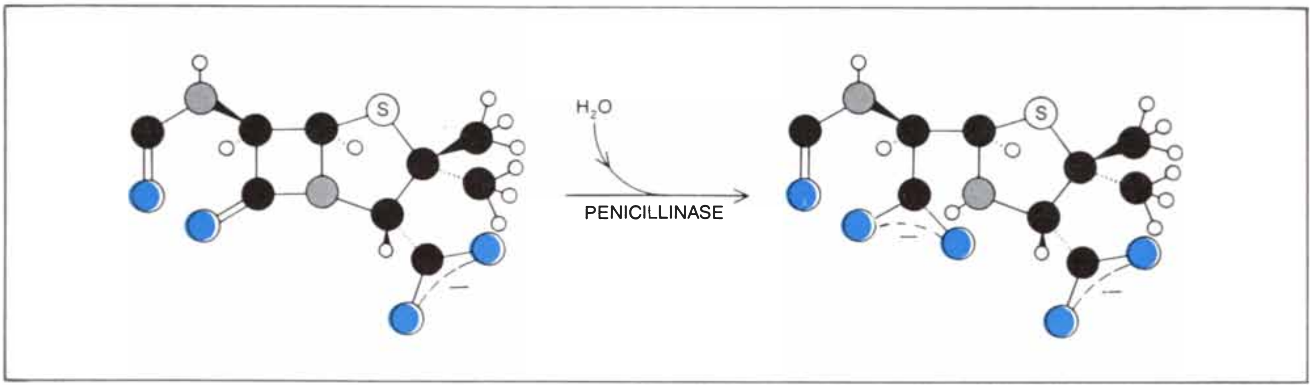
Recent work has revealed that in the cytoplasmic membrane of bacterial cells there are a multiplicity of proteins to which the beta-lactam antibiotics bind specifically, at least some of which appear to be enzymes with D-alanine carboxypeptidase activity or transpeptidase activity, or both. Such penicillin-binding proteins (PBP's) can be revealed by solubilizing the membrane proteins with a detergent after allowing them to react with penicillin that has been labeled with the radioactive isotope carbon 14 and observing which proteins are radioactive when they are separated from one another by electrophoresis in an agar gel. Different beta-lactam antibiotics have different affinities for one or more PBP's and cause different morphological changes in bacteria. When they act on the rod-shaped *E. coli* in sufficiently high concentrations, most of them give rise to the formation of protoplasts and to cell lysis, whereas in lower concentrations some of them induce the formation of osmotically stable spherical cells and others induce the formation of filaments of cells in which cell division has been inhibited.

The growth of heat-sensitive mutants of *E. coli* in which one or another of the PBP's is stable at 30 degrees Celsius but unstable at 42 degrees is sometimes accompanied by morphological changes at the higher temperature that mimic changes caused by beta-lactam antibiotics at the lower temperature. A study of such mutants by Brian G. Spratt of the University of Sussex has shown that the PBP designated 1B is concerned with cell elongation, PBP2 with cell shape and PBP3 with cell division. Inhibition of any of these PBP's appears to be lethal, whereas inhibition of others is not, even though the latter are present in much larger amounts.

It is evident that a decrease in the affinity of different beta-lactam antibiotics for essential PBP's, or a decrease in their ability to reach the site of these enzymes, will be reflected in their antibacterial activities. A temperature-sensitive mutant of *E. coli* has been isolated that shows increased resistance to two beta-lactam antibiotics (mecillinam and thienamycin) and a decrease in the affinity of its PBP2 for these substances. It is uncertain, however, whether resistance acquired in this way by mutation has so far been clinically significant. Mecillinam and thienamycin exert their lethal effects on *E. coli* by binding almost exclusively to PBP2. Mutations in the genes coding for several PBP's would be necessary to generate an increased resistance to other beta-lactam antibiotics. Moreover, the changes produced in the PBP's would need to be sufficiently specific to inhibit the binding of beta-lactam antibiotics but not the binding of the natural substrates required for the synthesis of the peptidoglycan essential to the survival of the cell.




**SIDE-CHAIN STRUCTURE of three cephalosporins is displayed as it is for the penicillins in the illustration on the preceding page. Cephalosporin *C* was the first of this family of new penicillins derived from the fungus *Cephalosporium* (rather than *Penicillium*) to be found.**



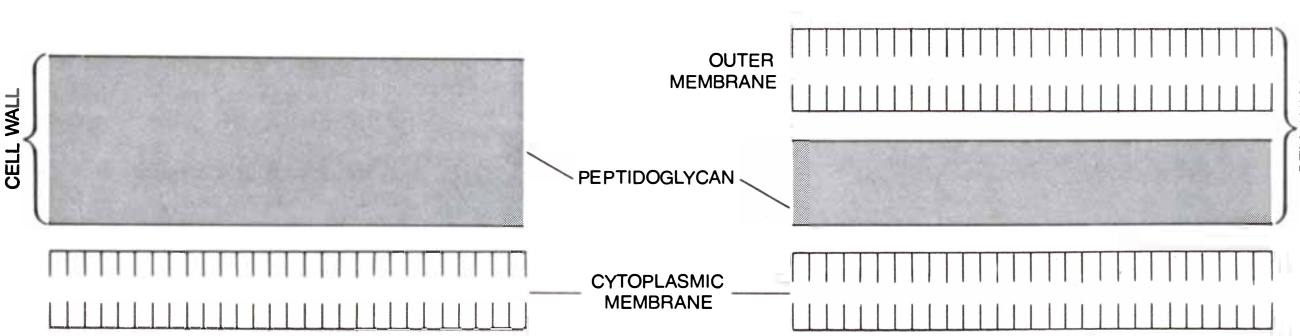

**RESISTANT BACTERIA presented a problem in the clinical use of penicillin. In 1940 Ernst B. Chain and the author at the University of Oxford discovered an enzyme in *Escherichia coli* bacteria that destroyed penicillin. They named it penicillinase. It and other penicillinases act by cleaving the amide bond in the beta-lactam ring of penicillin to form penicilloic acid, which is not active as an antibiotic.**

Gram-negative bacteria tend to be intrinsically more resistant than Gram-positive organisms to many beta-lactam antibiotics. The results of investigations of the properties of the outer membrane of Gram-negative bacteria, undertaken by Hiroshi Nikaido of the University of California at Berkeley and by others, are consistent with the hypothesis that this membrane contains proteins (called porins) that provide channels through which the small hydrophilic beta-lactam antibiotics can nonetheless gain access to the PBP's in the underlying cytoplasmic membrane. Whatever the mechanism of penetration, the concentrations of certain beta-lactam antibiotics required to inhibit the growth of Gram-negative bacteria are often higher than those required to inactivate the essential PBP's in cell-free systems, even when there is no significant enzymatic destruction of the antibiotic by a bacterial beta-lactamase. In contrast, a higher concentration of the antibiotic is not needed to inhibit the growth of a "permeability mutant" of *E. coli*, in which access to the cytoplasmic membrane is unimpeded.

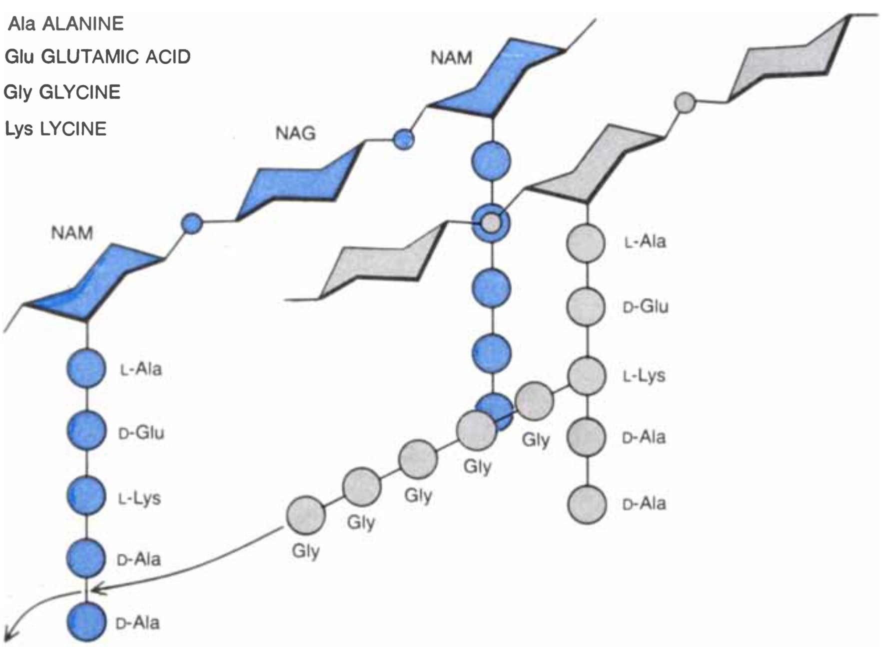
It seems that the barrier presented by the outer membrane of the cell is at least partly responsible for the relative insensitivity to many beta-lactam antibiotics of *Pseudomonas aeruginosa*, an organism that causes serious infections in damaged tissues, and for the insensitivity of many other Gram-negative bacteria to penicillins such as methicillin. The ability of both Gram-negative and Gram-positive bacteria to produce beta-lactamases that can open the beta-lactam ring, however, is also a major factor in resistance and one of undoubted clinical importance. This first became apparent with the emergence of strains of penicillinase-producing staphylococci, which multiplied while sensitive strains were suppressed. Further studies have shown that the penicillin-destroying enzyme may partly be excreted from the cells and partly remain bound in them, and that it is commonly induced by beta-lactam antibiotics. One consequence of this situation is that dense populations of the staphylococci, which manufacture large amounts of beta-lactamase, rapidly destroy many penicillins, although single cells, which manufacture only small amounts, remain intrinsically sensitive.

Many Gram-negative bacteria also produce a beta-lactamase, and in these organisms the enzyme is often constitutive, that is, it is synthesized in maximal amounts without the addition of an inducer. Moreover, it is normally bound to the cells, although it is often found in the periplasmic space between the cytoplasmic membrane and the outer cell wall and is easily released when the wall is damaged. As beta-lactamases from a variety of bacteria were tested against a series of penicillins and cephalosporins it became apparent there were many enzymes of this family that differed in their ability to hydrolyze various beta-lactam antibiotics.

The multiplicity of beta-lactamases has presented serious problems to the physician. So has the finding that the genes that specify the production of beta-lactamases can be transferred from one bacterial species to another. These genes are not essential to the life of the bacteria, and they are often carried on the self-replicating genetic units known as plasmids. Richard P. Novick of the Public Health Research Institute of the City of New York and Mark H. Richmond of the University of Bristol showed that the manufacture of penicillinase by staphylococci is mediated by plasmids that can be transferred from one staphylococcus to another by incorporation into the head of a virus that infects the cell. Transfer by this mecha-




**CELL-WALL STRUCTURE of Gram-positive (*left*) and Gram-negative bacteria (*right*) helps to account for the fact that penicillin G is active against Gram-positive bacteria but not against Gram-negative ones. In both types the cytoplasmic membrane is covered with a layer of peptidoglycan, but Gram-negative bacteria have an outer membrane and may also store beta-lactamase enzymes in the periplasmic space between the cytoplasmic membrane and the peptidoglycan layer. The cytoplasmic membrane incorporates penicillin-binding proteins. Some are enzymes involved in the synthesis of the cell wall and are therefore vulnerable to the action of penicillin.**




**PEPTIDOGLYCAN STRUCTURE of the bacterial cell wall consists of alternating units of two amino sugars, N-acetylglucosamine (NAG) and N-acetylmuramic acid (NAM). The NAM units are linked to peptide chains, many of which are also cross-linked to each other. In *Staphylococcus aureus* bacteria the cross-linking is achieved when the amino group of a terminal glycine unit is inserted into the bond that links the two D-alanine units of another chain, forming a new bond. The transpeptidase enzymes that make the insertion (when the bacteria are multiplying and thus building new cell wall) are attacked by penicillins and cephalosporins.**

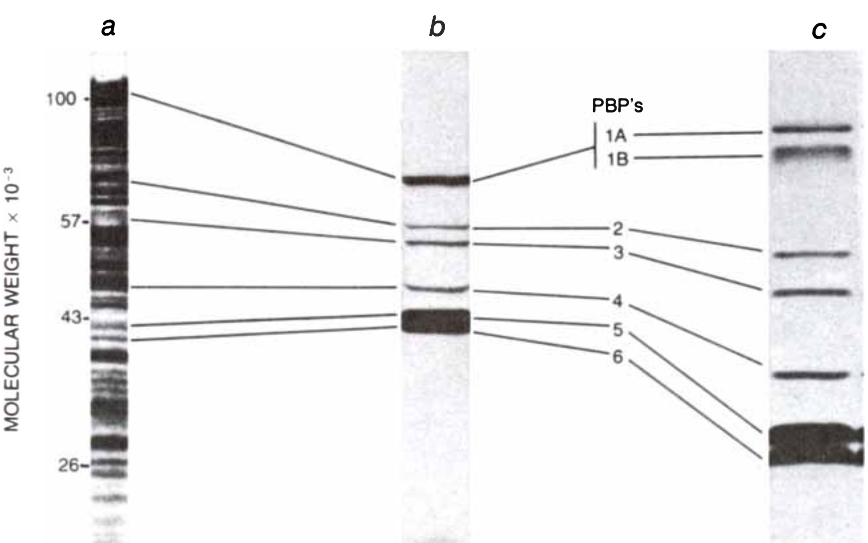
nism, known as transduction, may be of little clinical significance. In Gram-negative bacteria, however, plasmids can be transferred from one cell to another of the same species or of a different species by the mating process known as conjugation, in which a male donor cell adheres briefly to a female recipient cell. Furthermore, genes coding for beta-lactamase can be present in certain specialized fragments of DNA, known as transposons, that can be transposed from one plasmid to another.

The ability of one bacterial cell or one plasmid to infect another of its kind with genes that specify a beta-lactamase provides a potentially important mechanism for the spread of resistance. The resistance can be transferred not only from one pathogen to another but also from a nonpathogenic organism to a pathogenic one. In some instances, however, the resistance may be slow to appear. Virtually all gonococci remained highly sensitive to benzylpenicillin for many years, but a resistant strain carrying a plasmid-mediated beta-lactamase similar to one known in other Gram-negative bacteria was eventually identified in patients with gonorrhea.

Against this background, how successful has the search been for new beta-lactam antibiotics to overcome clinical problems? One significant advance stemmed from the finding at Oxford in 1953 that a fungus of the genus *Cephalosporium,* which had been isolated in 1945 by G. Brotzu in Sardinia, manufactured a new penicillin with a side chain derived from D-alpha-aminoadipic acid. G. G. F. Newton and I discovered that crude preparations of this penicillin contained small amounts of a second antibiotic having the same side chain but exhibiting a structural feature not seen heretofore: the beta-lactam ring was fused to a six-membered ring instead of to a five-membered one. The new antibiotic, named cephalosporin *C,* was of potential medical interest because it had the characteristic low toxicity of the penicillins when it was tested in animals but was resistant to hydrolysis by certain beta-lactamases.

When *Penicillium* and *Cephalosporium* are grown in the laboratory or in an industrial plant, they require distinctly different fermentation conditions. Both fungi, however, follow similar biosynthetic pathways in producing beta-lactam antibiotics. Studies done with cell-free preparations have shown that a chain of three amino acids is formed into a ring to give rise to isopenicillin *N,* a penicillin with an L-alpha-aminoadipyl side chain. In *Penicillium chrysogenum* an acyl transferase then catalyzes the replacement of this side chain by one derived from a utilizable precursor, such as phenylacetic acid. In the *Cephalosporium* species the transferase appears to be absent, because no compound with a new side chain is formed. Instead two sequential reactions occur. The asymmetric center of the L-alpha-aminoadipyl side chain is inverted to its three-dimensional mirror image. Some of the resulting penicillin *N,* with a D-alpha-aminoadipyl side chain, is then excreted, but recent experiments by Arnold L. Demain and his colleagues at M.I.T. have indicated that some of it is converted into deacetoxycephalosporin *C,* which then gives rise to cephalosporin *C* and other cephalosporins.

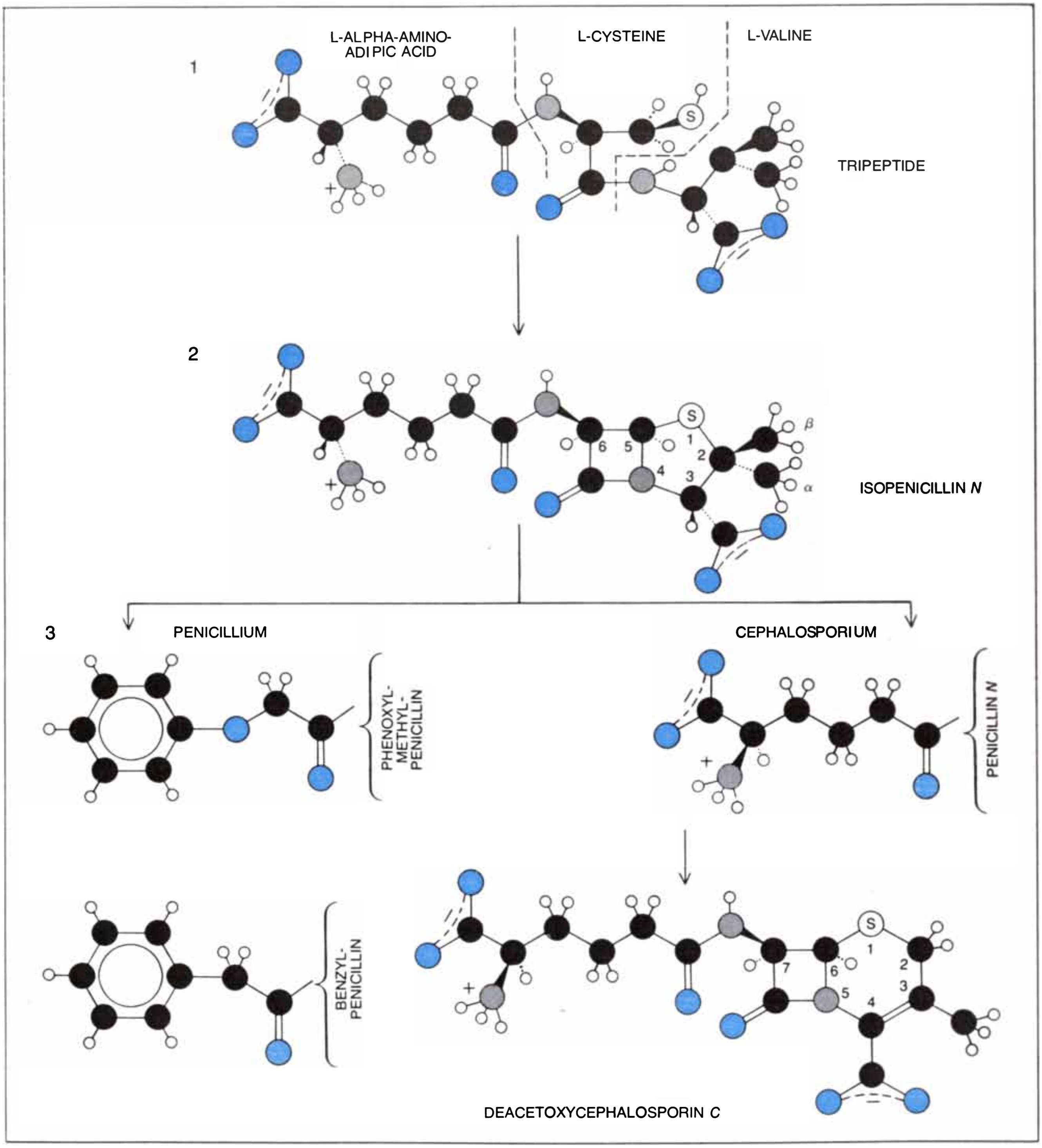
By fermentation alone it has been possible to obtain only certain types of penicillin and only cephalosporins with the D-alpha-aminoadipyl side chain. By semisynthetic methods, however, thousands of members of the beta-lactam




**PENICILLIN-BINDING PROTEINS in the cytoplasmic membranes of *E. coli* bacteria were detected by electrophoresis techniques applied by Brian G. Spratt of the University of Sussex. Penicillin labeled with the radioactive isotope carbon 14 was bound to the membranes. The proteins were then fractionated by electrophoresis on a gel, through which the proteins migrated at different rates depending on their molecular weight. The photographs show the gel stained for protein (*a*), the labeled proteins detected by autoradiography (*b*) and the separation of penicillin-binding protein 1A and 1B by electrophoresis procedures that gave more resolution (*c*).**

family of antibiotics have been prepared and studied. Semisynthesis became feasible with the discovery of methods for the production of the nucleus of the penicillin molecule, 6-aminopenicillanic acid (6-APA), and of the cephalosporin molecule, 7-alpha-aminocephalosporanic acid (7-ACA), in which an acyl side chain is absent. Following reports from Japan of the production of 6-APA by *Penicillium chrysogenum* this substance was synthesized by John Sheehan and was discovered independently at the Beecham Research Laboratories in Britain, where it was isolated from fermentations to which no side-chain precursor had been added. Later it was produced in quantity by the removal of the side chain of the readily available benzylpenicillin with acylases from bacteria.

No enzyme was found that would




**BIOSYNTHESIS of penicillins and cephalosporins from the fungi *Penicillium* and *Cephalosporium* begins (*1*) with the formation of a tripeptide; reading from left to right the three units, separated by broken lines, are L-alpha-aminoadipic acid, L-cysteine and L-valine. The tripeptide is converted into isopenicillin *N* (*2*). In the penicillins a transferase enzyme catalyzes the replacement of the L-alpha-aminoadipyl side chain of the isopenicillin *N* (*3*), whereas in the cephalosporins the transferase is apparently absent, since no compound with a new side chain is formed. Instead two sequential reactions take place. First, the center of the L-alpha-aminoadipyl side chain is inverted to its three-dimensional image to form penicillin *N*. Second, penicillin *N* is converted into deacetoxycephalosporin by a ring expansion involving the rupture of the bond between the sulfur atom and the second carbon and the linkage of a beta-methyl group to the sulfur through the carbon atom of the beta-methyl group. The mechanisms of the various enzymic reactions have yet to be elucidated.**

remove the D-alpha-aminoadipyl side chain from cephalosporin *C*. Nevertheless, 7-ACA was produced in small quantities at Oxford by mild acid hydrolysis of cephalosporin *C* and subsequently by a much more efficient chemical procedure that was devised at the Eli Lilly Research Laboratories. These achievements opened the door to the preparation of semisynthetic penicillins and cephalosporins by the chemical coupling of different side chains to 6-APA and 7-ACA. In addition the cephalosporin molecule could be further modified without altering its essential ring system by means of chemical procedures that exchange the acetoxy group ($-OCOCH_3$) attached to the carbon atom designated 3′ projecting from the six-membered ring for other groups or for a hydrogen atom.

Two other findings were of importance in the cephalosporin work. One was the discovery at the Lilly laboratories of an ingenious chemical route from the penicillin ring system to the cephalosporin one, which enabled certain cephalosporins to be obtained from a penicillin. The other was the discovery at the Lilly laboratories and the Merck Sharp & Dohme Laboratories that cephalosporins manufactured by *Streptomyces* bore a methoxyl group ($-OCH_3$) instead of a hydrogen on the carbon atom lying diagonally across from the nitrogen in the beta-lactam ring. Subsequently methods were developed for introducing the methoxy group into other cephalosporin molecules.

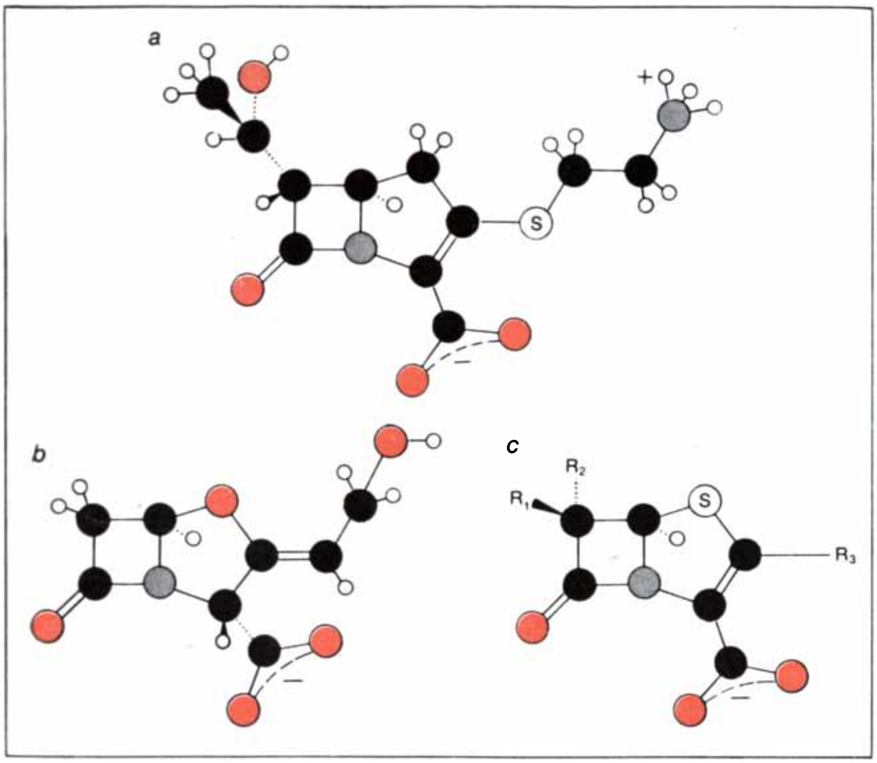
Many thousands of penicillins and cephalosporins have now been made by one or another of these methods, and some of them have found an important place in medicine. The first of the clinically valuable new penicillins, named methicillin, was prepared in the Beecham laboratories and proved to be effective against the benzylpenicillin-resistant staphylococcus, mainly because its bulky side chain hindered its interaction with the staphylococcal penicillinase. Later products were ampicillin, which showed higher activity against a number of Gram-negative bacteria, and carbenicillin, which showed useful activity against the bacterium *Pseudomonas aeruginosa*.

The first two semisynthetic cephalosporins, cephalothin from Lilly and cephaloridine from the Glaxo laboratories in Britain, were also active against the penicillin-resistant staphylococcus, but here it was the nature of the ring system, rather than the side chain, that protected the antibiotics from rapid hydrolysis by the staphylococcal enzyme. The two cephalosporins also exhibited useful activity against a number of Gram-negative bacteria and in addition provoked no reaction in most patients who were hypersensitive to penicillin. Although the cephalosporin ring system itself is reasonably stable under the acid conditions in the stomach, the first cephalosporins were not absorbed when they were given by mouth. Unpredictably, however, another semisynthetic cephalosporin, cephalexin, proved to be effective when it was taken orally.

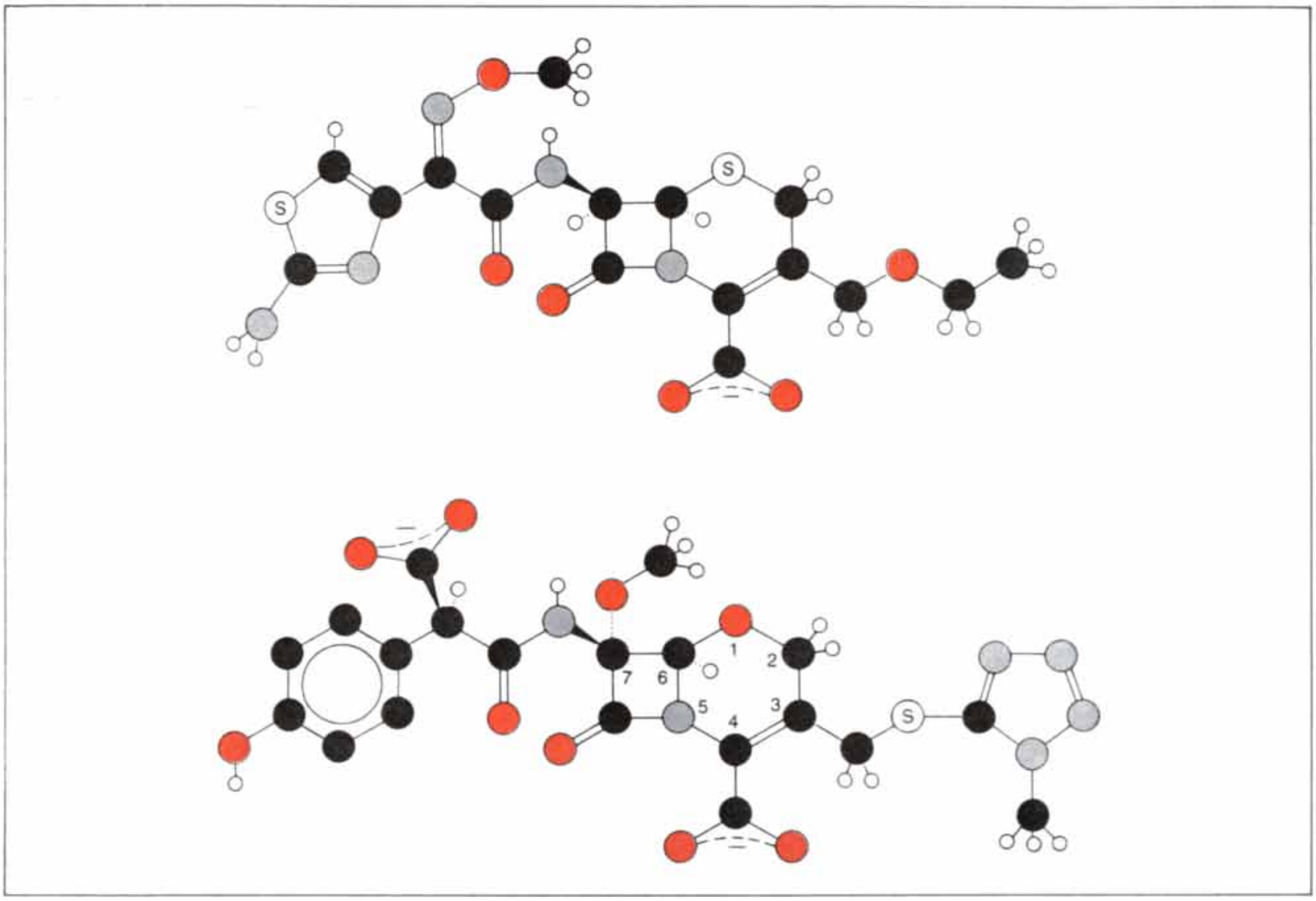
In spite of the valuable properties of these early semisynthetic penicillins and cephalosporins, some Gram-negative pathogens are insensitive to them. Other changes in the groups attached to the ring system have now yielded a second and a third generation of cephalosporins. The first three products were given the names cefamandole, cefuroxime and cefoxitin, and they were followed by cefotaxime and cefoperazone. In addition a highly active 1-oxacephalosporin, with an oxygen instead of a sulfur in the six-membered ring, has been obtained by semisynthesis from penicillin. Some of these compounds represent clear-cut advances in bacterial chemotherapy because they show a broader spectrum of activity than earlier beta-lactam antibiotics and resist hydrolysis by a wider variety of beta-lactamases.

A new chapter in the history of beta-lactam antibiotics opened in 1976 with the discovery in several pharmaceutical laboratories that certain *Streptomyces* manufacture substances that have a beta-lactam ring but neither the penicillin ring system nor the cephalosporin one. Some of the new substances were detected by screening for antibacterial activity; others were found by screening for substances with an ability to inhibit the action of beta-lactamases. One of them, thienamycin, has a broad antibacterial spectrum and a high resistance to beta-lactamases and may well find a place in medicine. Another, clavulanic acid, is a powerful inactivator of beta-lactamases. At the same time Robert Woodward described the synthesis and semisynthesis of reactive beta-lactams known as penems, in which the double chemical bond in the sulfur-containing ring of the cephalosporins was placed in the corresponding ring of the penicillins. Very recently still more novel beta-lactam antibiotics have been isolated from the culture fluids of several species of bacteria.

How far has it been possible to predict what kinds of beta-lactam antibiotics should be made in attempts to solve clinical problems? Some ideas readily presented themselves. It became known that a side chain capable of attracting an electron could render a penicillin sufficiently stable in the presence of acid for the drug to be given effectively by mouth. By analogy with the known dif-




**NEW BETA-LACTAM ANTIBIOTICS discovered since 1976 include thienamycin (*a*), clavulanic acid (*b*) and the penems, one of which is shown (*c*). Thienamycin and clavulanic acid are derived from *Streptomyces* bacteria. The antibiotics have a beta-lactam ring but neither the penicillin ring system nor the cephalosporin one. The penems, synthesized by the late Robert B. Woodward of Harvard University, are notable for the placing of the double bond of the sulfur-containing ring of the cephalosporins in the corresponding ring of the penicillins.**




**THIRD-GENERATION CEPHALOSPORINS include cefotaxime (*top*) and moxalactam (*bottom*), a highly active 1-oxacephalosporin derived from penicillin by semisynthesis. A feature of 1-oxacephalosporin is that oxygen replaces sulfur in the six-membered ring.**

ferences between penicillin *N* and benzylpenicillin it seemed likely that the exchange of the alpha-aminoadipyl side chain of the naturally occurring cephalosporins for less polar side chains (that is, side chains without positive and negative charges) would result in a large increase in intrinsic activity against some bacteria and the retention of stability to staphylococcal penicillinase.

This proved to be so. The finding that the O-acetyl group at the 3′ carbon in cephalosporins, to which I referred above, could be removed in the body by an acetyl esterase to yield less active compounds suggested that this group might profitably be replaced by stabler groups. The structural features required for a cephalosporin to be absorbed from the gut could not, however, have been predicted, nor could the features required for a beta-lactam antibiotic to be a powerful inactivator of essential enzymes in bacterial cell membranes but still be resistant to hydrolysis by many beta-lactamases. Thus although many thousands of compounds have been made, only a few have found a place in medicine. These few have come primarily from the exploitation of chance observations and the finding of new types of beta-lactam molecules in large and systematic screening programs.

One reason for this has been a lack of knowledge of the three-dimensional active sites in the various enzymes with which beta-lactam antibiotics react. Significant advances in our understanding of the nature and function of such sites are only now being made.

Penicillin-binding proteins with D-carboxypeptidase activity can be purified by affinity chromatography, in which an appropriate penicillin or cephalosporin is coupled to a resin. The PBP's become bound specifically to the column and can be eluted with hydroxylamine, which breaks the covalent bond formed by reaction of the PBP with the beta-lactam ring. Following this procedure the amino acid sequences in fragments of two major PBP's have been determined and the amino acid unit through which they bind penicillin has been identified. Purification of minor PBP's, which are presumably transpeptidases and are targets for inactivation by beta-lactam antibiotics, has proved difficult, but progress is being made by the insertion into *E. coli* of DNA that leads to the production of multiple copies of the gene specifying a PBP. This has been done by insertion of the genes into a multicopy plasmid or into a virus that produces many copies of the gene when it infects the bacterial cell and multiplies.

The purification of many beta-lactamases has already been accomplished. The amino acid sequences in enzymes from *Staphylococcus aureus, Bacillus licheniformis* and *B. cereus* have been determined by Richard P. A. Ambler of the University of Edinburgh, and the sequence of a plasmid-mediated beta-lactamase from *E. coli* has been deduced by J. G. Sutcliffe of Harvard from the sequence of the nucleotides in the corresponding gene. With one exception (a zinc-requiring enzyme from *B. cereus*) these sequences display considerable homology, that is, they show much closer similarities than can be attributed to chance. Significant homology has also been reported between beta-lactamases and two PBP's that are D-carboxypeptidases. It is therefore possible that a group of beta-lactamases and a group of penicillin-sensitive enzymes in bacterial cell membranes have had a common evolutionary origin.

These two enzyme families also demonstrate similarities in their mode of action. Through the use of benzylpenicillin compounds that remain bound at their active sites, the hydroxyl group of a specific amino acid unit, serine, has been shown to attack the beta-lactam ring. This attack results in the opening of the ring with the formation of an acyl enzyme that is a serine ester. With a beta-lactamase the acyl enzyme from many beta-lactam antibiotics is rapidly hydrolyzed to regenerate the enzyme and to yield a penicilloate or an analogous compound with no antibacterial activity. With a carboxypeptidase from the bacterial membrane the breakdown of the acyl enzyme from a beta-lactam antibiotic is very slow, so that inactive acyl enzyme accumulates.

In one D-carboxypeptidase the serine unit that reacts with penicillin has been found to be the same unit that reacts with a substrate of the enzyme containing D-alanyl-D-alanine. This is consistent with a hypothesis proposed in 1965 by Jack Strominger and Donald J. Tipper of the University of Wisconsin at Madison. To explain why penicillin inhibits cross-linking in cell-wall synthesis they postulated that the position in space of a sequence of atoms in the penicillin structure mimics the position of corresponding atoms in the D-alanyl-D-alanine terminus of nascent peptidoglycan side chains. How far this hypothesis can be extended to account for the relative abilities of different beta-lactam antibiotics to inhibit the functioning of the essential membrane PBP's is still uncertain.

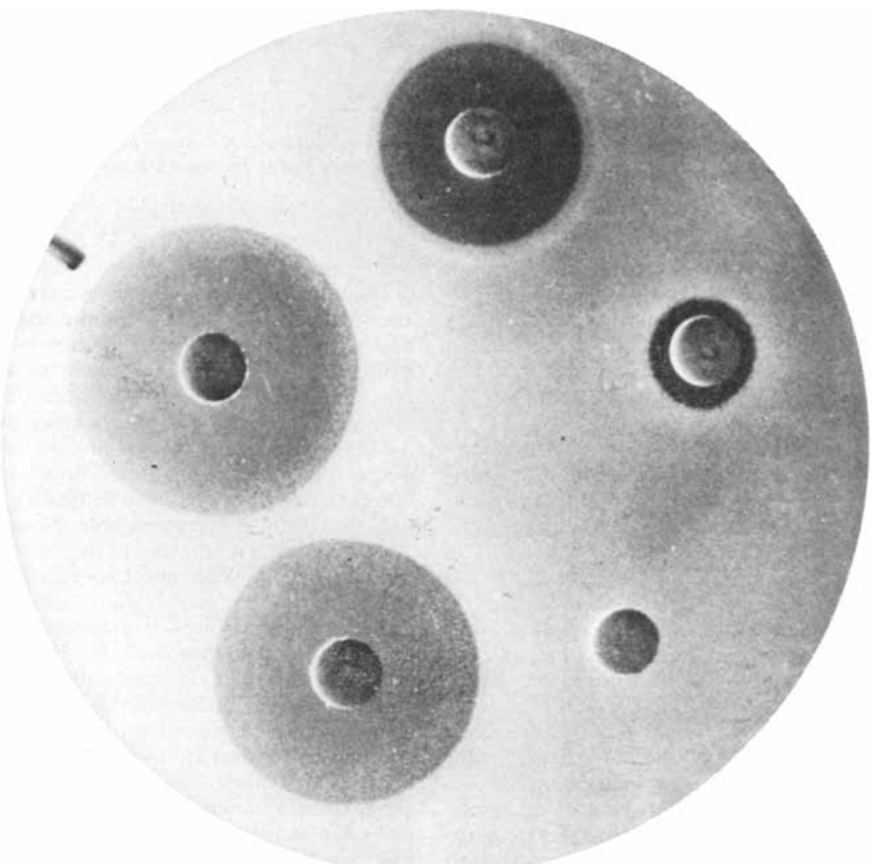
Among the new beta-lactam compounds with a much higher resistance than benzylpenicillin to hydrolysis by members of the beta-lactamase family are some whose interaction with a beta-lactamase brings about a temporary or permanent inactivation of the enzyme. This phenomenon is being studied by Jeremy Knowles of Harvard, Stephen G. Waley of Oxford and many others. If the rate of hydrolysis of an acyl-enzyme intermediate is slower than that of its formation, the intermediate will accumulate and the active site will not be available for a readily hydrolyzable substrate such as benzylpenicillin. After the slowly hydrolyzed substrate is completely utilized and none remains to form new acyl enzyme, however, active enzyme can be regenerated. In some instances, on the other hand, there is a structural change in the acyl component of acyl-enzyme molecules that entirely prevents the removal of the acyl group. When this happens, the enzyme may be said to have committed suicide.

The notion that one beta-lactam antibiotic could protect another from destruction by a beta-lactamase in the course of therapy dates from 1956, when the hydrolysis of benzylpenicillin by a beta-lactamase was found to be competitively inhibited by cephalosporin *C*. The administration of mixtures of two compounds that are synergistic for this reason has not been widely favored. Nevertheless, following the discovery of clavulanic acid and other substances that not only inhibit beta-lactamases competitively but also inactivate them irreversibly, the potentialities of such synergism may well be further explored.

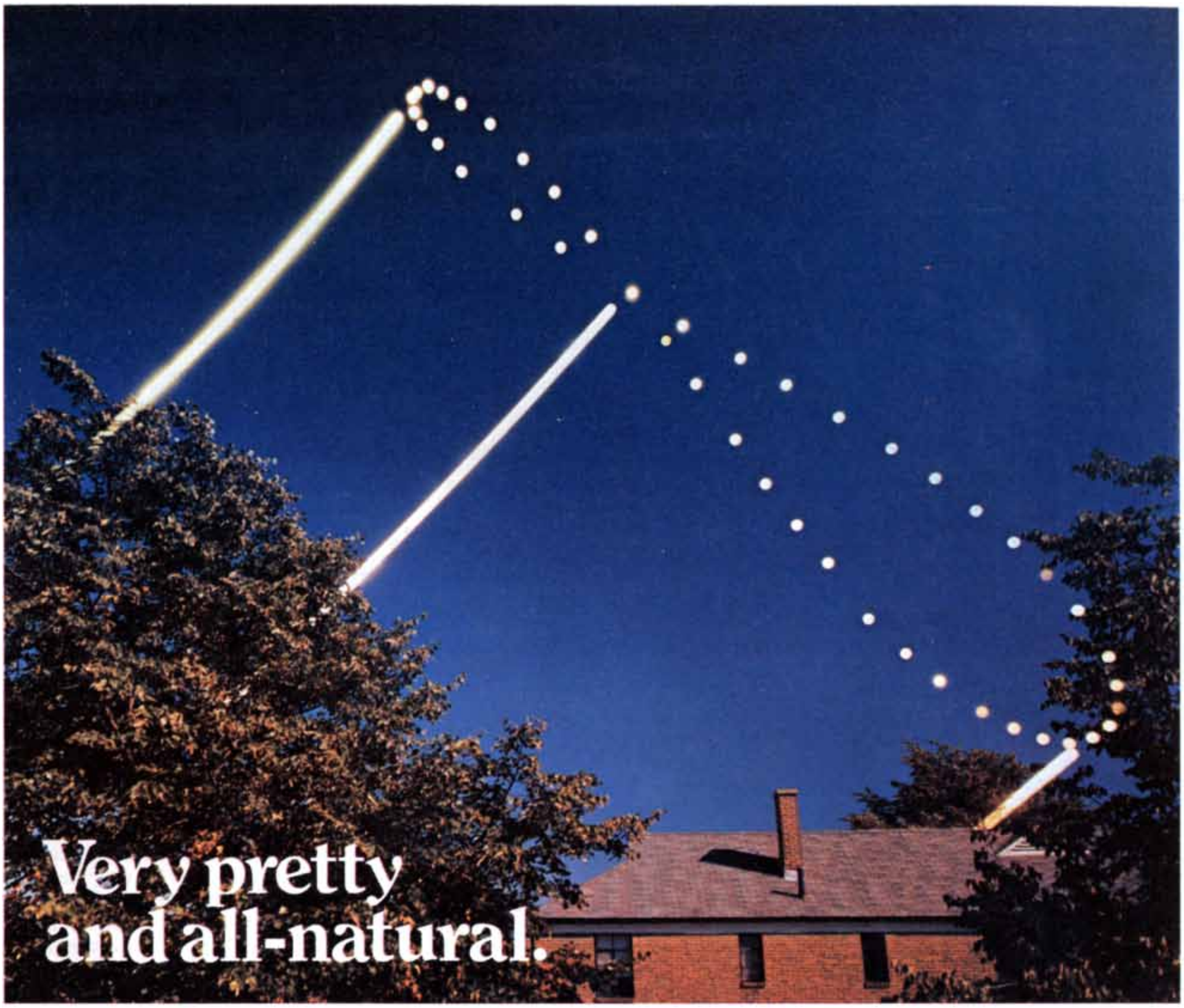
The striking developments in the field of beta-lactam antibiotics since the discovery of the therapeutic power of penicillin and the intensive research in this field at the present time have been stimulated in no small measure by the outstanding value of these nontoxic antibiotics in medicine. Major advances have been made toward the solution of problems presented by penicillin-resistant bacteria, although changing patterns of resistance and the great mutability of microorganisms make it unwise to assume that any solutions are final.

Thirty years ago the penicillin molecule seemed to embody the properties of a unique fused ring system. Now the only common feature in the structure of the known beta-lactam antibiotics is the beta-lactam ring. The nitrogen atom in the ring is commonly attached to a carbon atom bearing a carboxyl group (–COOH), a related ionizable group of similar shape or a lactone (a cyclic ester of a hydroxyl acid), but in the new beta-lactams from bacteria this nitrogen forms part of a sulfamate group ($-NSO_3^-$). For the molecule to show high intrinsic activity against bacteria by formation of an acyl enzyme it seems that the beta-lactam should have a high chemical reactivity. This is achieved in the penicillins and cephalosporins by fusion of the four-membered beta-lactam ring with a second ring to yield a system in which the normal configuration is strained.

If the beta-lactam is too reactive, however, the molecule will not be stable enough for clinical purposes, and its biological effect will no longer be highly selective. Moreover, unless the groups attached to the ring occupy appropriate positions in space, the molecule will not fit into the active sites of penicillin-binding proteins in the cell membrane in such a way that a relatively stable acyl enzyme can be formed. Over the next decade it may be possible to specify such features of a potentially useful beta-lactam with some precision.



**EFFECT OF PENEMS is indicated in this photograph of the inhibitory action of several antibiotics against penicillin-resistant *Staphylococcus aureus* bacteria growing on agar. Each of the five dark circles represents an area where bacteria have been killed by a particular antibiotic; the larger the circle, the more effective the antibiotic. Clockwise from the top the antibiotics were cephalexin; penicillin *V;* an inactive penem, in which the sulfur atom projects below the plane of the beta-lactam carbon atom instead of above as in penicillins and cephalosporins, resulting in an unnatural three-dimensional shape; a mixture of penems with natural and unnatural configurations, and an active penem with a natural three-dimensional configuration.**




**B**ut what does it prove?

It proves how clever were certain parties who, long before the availability of Kodak Vericolor II professional film, drew an 8-shaped analemma figure on globes and sundials to show the daily declination of the sun over the course of a year.

That, though, hardly needs proving. Nor is proof needed that a decent result is obtainable from a sheet of that film exposed 48 times over a full year.

Why, then, do the directions for use say "process promptly"?

**ANSWER:** This is one of the "professional" films we make for portrait and commercial photographers. These users refrigerate their film and process promptly anyway, out of business necessity. Their raw film inventory justifies dedicating a refrigerator to it. We take this into account in balancing the relative sensitivities of the three color layers for best color rendition of subjects like fashions and flowers by which such users earn their bread. We take enormous pride in being that fussy about nuances of color which most people can scarcely appreciate.

Not everyone blessed with such sensitivity to color depends on it for a living. Fortunately for us, they too buy film and expect the same high level of color rendition though their shooting schedules are not so tight, they may not process so promptly, and their refrigerators are more likely to be stocked with watermelon and hamburger than color film. This we must assume in balancing the relative layer sensitivities of "amateur" films.

All this crafty scheming for fine tuning of photographic properties to differences between use by professionals and non-professionals is just an example of the upgrading in product performance that goes on over the years.

The "process promptly, store at 55° F (13° C) or lower" advises which kind you are using. Whoever reads it as an indication that "professional" film is less stable than "amateur" film is much mistaken.

In *Sky and Telescope* for June, 1979, that magazine's Dennis di Cicco told all about how he took this one-year-long picture. Analemmas at once became a hot subject. The magazine's August '79 issue found it necessary to advise readers how to purchase a 16" x 20" print. The deal is still on. If you want to know all about analemmas, see page 20 of *Sky and Telescope* for July, 1972. If you want a copy of Kodak Pamphlet E-36, "Nine KODAK Color Films for Process C-41," write Department 412-L, Kodak, Rochester, NY 14650. If you need detailed information on astronomical photography, make that address "Scientific and Technical Photography" instead of "Dept. 412-L."

Kodak

# SCIENCE AND THE CITIZEN

## *Unmet Demand*

In spite of the well-publicized efforts of antiabortion groups in the U.S., the number of women resorting to abortion as a method of birth control continues to increase. Indeed, in free-market terms the demand for legal abortions currently exceeds the supply by a wide margin. In 1979, the latest year for which substantially complete national data are available, more than 1.5 million pregnant women (which is 30 percent of all pregnant women) chose to terminate their pregnancies by abortion. Another 641,000 women, it is estimated, were unable to get the abortion services they sought and went on to give birth although it had not been their intention to have a child.

The main obstacle to free choice in the latter cases appears to be the geographic and financial inaccessibility of abortion facilities. According to a recent survey conducted by workers at the Alan Guttmacher Institute in New York, approximately 80 percent of the counties in the U.S. have no facility where legal abortions are performed. Moreover, only 5 percent of all abortions are performed outside of metropolitan areas, whereas 26 percent of the women of childbearing age live outside such areas. As a consequence 8 percent of the women who obtained abortions in 1978 had to cross state lines to do so; a much larger percentage had to travel, often considerable distances, to the major cities in their states, where most abortion facilities are located. In the five states with the least adequate abortion services roughly half of the women who had abortions had to go outside the state to get them.

The number of hospitals offering abortion services actually decreased slightly from 1977 to 1978, although the number of nonhospital providers (clinics and physicians' offices) increased; overall there was a net increase of 2 percent in the number of facilities providing abortion services. In 1978 abortions were performed in only 18 percent of the public hospitals and 36 percent of the non-Catholic private and voluntary hospitals in the U.S.

The survey also noted a marked decline in the proportion of abortions done in hospitals: from 30 percent in 1977 to 25 percent in 1978. It is thought that this trend can be attributed in part to the decreased availability of Medicaid funding for abortion. Since late 1977 Federal and state funds to subsidize abortions for poor women have been drastically reduced, following the passage by Congress of the Hyde Amendment (named for its sponsor, Republican Representative Henry J. Hyde of Illinois). Last year's Supreme Court decision upholding the constitutionality of the Hyde Amendment dispelled hopes that Medicaid funding would be restored for poor pregnant women in need of abortion services.

Earlier studies had indicated that between 18,000 and 30,000 indigent women would have no alternative but to give birth in 1978 as a result of the new restrictions on Medicaid funding, reducing the total number of abortions by between 1.3 and 2.1 percent. Indeed, the survey takers found that the 1977–78 increase in the total number of abortions was smaller by from 1 to 3 percent than the rise expected from projecting the increases of previous years.

As predicted, nonwhite and unmarried women were affected disproportionately by the changes in Medicaid funding for abortions. The survey data show that the abortion rates for both nonwhite and unmarried women, after increasing consistently for the preceding few years, fell slightly for the first time in 1978, whereas the rates for white and married women continued to rise. These results, the investigators report in *Family Planning Perspectives,* a publication of the Alan Guttmacher Institute, tend to support the conclusion that "a significant proportion of Medicaid-eligible women who would choose to have an abortion instead carry their pregnancies to term when Medicaid funding is unavailable."

## *Brownian Black Body*

In 1905 Albert Einstein published three seminal papers: on Brownian motion, on the photoelectric effect and on the special theory of relativity. Twelve years later in another classic paper he analyzed the quantum-mechanical interaction of electromagnetic radiation with a gas of molecules. The 1917 paper established the concept of a quantum of electromagnetic energy (later named the photon) that has a definite energy and a definite momentum. In analyzing the effects of radiation on the gas molecules Einstein relied on his conclusions about Brownian motion and the photoelectric effect. Curiously, he ignored his work on special relativity; he considered only the case in which the gas molecules move with a nonrelativistic velocity, one that is low compared with the speed of light. This apparent oversight was not noticed until 1979, when Timothy H. Boyer of the City College of the City University of New York pointed it out and argued, moreover, that if Einstein had made the correct relativistic assumptions, his analysis would have led to a disturbing inconsistency between the theory of quantum mechanics and the theory of statistical mechanics.

Einstein's 1917 paper was the culmination of the early years of quantum mechanics, which began when Max Planck analyzed the exchange of energy between the wall of a cavity in a solid body and radiation trapped in the cavity. The wall of such a cavity is the best possible emitter of all wavelengths of electromagnetic radiation. A graph of the intensity of the radiation v. the wavelength is the same as the graph for a perfect black body and is called the black-body spectrum. The shape of the black-body spectrum depends on the temperature of the cavity wall, so that actually there is a family of graphs, one graph for each temperature. The determination of a general procedure for deriving the family of graphs on theoretical grounds was the outstanding unsolved problem in physics at the turn of the century. In October, 1900, Planck announced that he had developed a single formula for the family of graphs that was consistent with the observed data.

Planck went on to analyze the interaction of the radiation with the cavity wall. He assumed that the wall is made up of submicroscopic resonators, each of which oscillates at a characteristic frequency and emits and absorbs radiation at that frequency. The distribution of energy among the resonators in the wall is described by the formula of statistical mechanics called the Boltzmann distribution.

For the resonators to be in thermal equilibrium with the radiation, Planck reluctantly concluded, the resonators must emit and absorb energy only in discrete bundles: in units of $h\nu$, where $h$ is the universal constant now known as Planck's constant and $\nu$ is the frequency of the resonator. Planck carefully kept his comments on quantization to the exchange of energy between the material resonators and the radiation field; he drew no conclusions about the energy of the radiation itself. In 1905 Einstein proposed that the radiation acts as if it were made up of a finite number of localized energy quanta $h\nu$. Over the next 12 years Einstein extended his corpuscular theory and argued that light not only acts as if it were made up of particles but in fact consists of particles.

In his 1917 paper Einstein worked out the details of the quantum interaction of matter and radiation. With this paper quantum mechanics had come full circle. Planck had started with the black-body-radiation law and derived the Boltzmann energy distribution of material oscillators. Now Einstein did the opposite: he derived the Planck radiation law from the Boltzmann distribution. Einstein considered a gas of hypotheti-

# The Beefeater® $35,000 True-or-False Quiz.

## Important: These statements are all true but one. Tell us which one, and you may win the $25,000 Grand Prize, the $5,000 Second Prize, or one of five $1,000 Runner-up Prizes.

**1.** Every bottle of Beefeater you buy in the United States is distilled, bottled and sealed in the distillery in London. Beefeater is the only major imported gin that can make this statement.

**2.** Many gins which appear to be English are not made in England.

**3.** Four generations of the Burrough family have distilled Beefeater Gin.

**4.** Beefeater is the only gin to have received the Queen's Award... and four times at that!

**5.** The Beefeater goes back to the reign of William the Conqueror, some 900 years ago.

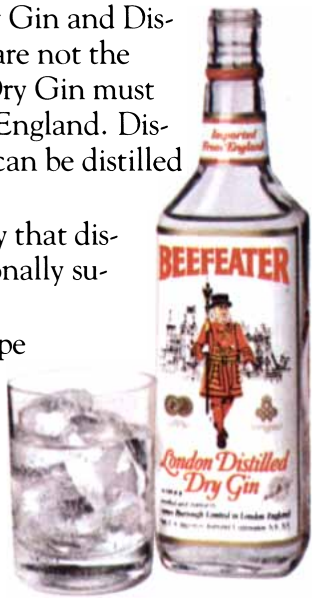
**6.** London Distilled Dry Gin and Distilled London Dry Gin are not the same. *London Distilled* Dry Gin must be distilled in London, England. Distilled London Dry Gin can be distilled anyplace.

**7.** To this day, the family that distills Beefeater Gin personally supervises each day's run.

**8.** The only known recipe for the perfect Beefeater Martini is in the secret archives of The Tower of London.

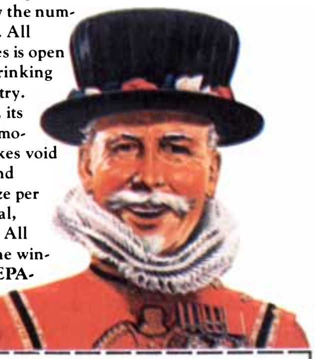
**9.** Each bottle of Beefeater is individually registered—one of the many steps in the quality control process.

**10.** Beefeater is famous as the "First name for the Martini."



Need help? You'll find the official answers in a free Beefeater pamphlet at leading spirits dealers and restaurants. Or, send a stamped, self-addressed #10 envelope to: Beefeater Pamphlet, Box 7187, Blair, NE 68009. Do not send your *entry* to this address.

**OFFICIAL RULES:** (No purchase required.) **1.** Read carefully the ten True or False statements appearing in this ad. **2.** On an official entry form or 3"x5" piece of plain paper, hand-print your name and address and the number representing the one statement that is false. **3.** Mail your entry in a hand-addressed envelope not larger than 4⅛"x9½" (#10 envelope) to: Beefeater Gin Sweepstakes, P.O. Box 9624, Blair, NE 68009. **4. IMPORTANT:** In order to be eligible for a prize, you must indicate on the lower left corner of your mailing envelope the same answer as on your entry form (the number of the statement you think is false). **5.** Entries must be received by August 31, 1981. Enter as often as you wish; each entry must be mailed separately. **6.** Winners will be determined via a random drawing from among all correct entries received. Drawings are under the supervision of D. L. Blair Corporation, an independent judging organization whose decisions are final. The odds of winning will be determined by the number of correctly answered entries received. All prizes will be awarded. **7.** This sweepstakes is open to residents of the United States of legal drinking age in their state of residence at time of entry. Not eligible: employees of Kobrand Corp., its distributors, retailers, advertising and promotion agencies and their families. Sweepstakes void in Ohio, Pennsylvania, Texas and Utah, and wherever prohibited by law. Limit one prize per family; no substitution of prizes; all Federal, State and local laws and regulations apply. All applicable taxes are the responsibility of the winner. **8.** For a list of prizewinners, send a SEPARATE stamped, self-addressed envelope to: Beefeater Winners List, P.O. Box 6541, Blair, NE 68009.



A

**Mail to: Beefeater Gin Sweepstakes**
**Box 9624, Blair, NE 68009**
(This address for entries only.)

**The one false statement is number** ______________

Name ______________

Address ______________

City ______ State ______ Zip ______

**BEEFEATER® GIN. The Crown Jewel of England.™**

BEEFEATER GIN. 94 PROOF DISTILLED FROM 100% GRAIN NEUTRAL SPIRITS. IMPORTED BY KOBRAND CORP., N.Y., N.Y.

cal molecules in thermal equilibrium with black-body radiation. Each molecule is assumed to have only two possible energy levels. A molecule in the lower energy level can absorb radiation whose frequency multiplied by Planck's constant is equal to the energy difference between the levels. Similarly, a molecule in the higher level can spontaneously fall to the lower level by emitting radiation of the same frequency.

Einstein assumed that the molecules undergo Brownian motion, like that of fine particles suspended in a liquid medium. Each molecule moves in a straight line until its trajectory is altered by a collision. The intervals between collisions are random and so the overall path of the particle is a zigzag. In his nonrelativistic calculations Einstein considered the molecules to have a low velocity and a large mass compared with the mass equivalent of the radiation energy. Therefore the velocity of a particle scarcely changes when it emits or absorbs a photon.

Boyer's analysis of these ideas, which appeared in *Physical Review,* suggested that Einstein's description could not be extended to relativistic molecular velocities. Under relativistic assumptions, Boyer argued, the Boltzmann energy distribution of the molecules cannot be reconciled with the energy distribution of the radiation as given by the Planck radiation law. Special relativity is an umbrella theory that should be satisfied by all other physical theories, and so Boyer's finding of an apparent inconsistency between a law of statistical mechanics and a law of quantum mechanics was a serious one. The Planck radiation law depends on the temperature of the oscillating resonators but not on any other property, such as their mass. Boyer concluded that the Boltzmann energy distribution of relativistic particles can be consistent only with a radiation law that depends on the particle mass.

Three further analyses with quite different results have now appeared in *Physical Review;* they are by Uri Ben-Ya'acov of the University of California at Santa Barbara, by Asher Peres of the Technion–Israel Institute of Technology and by Charles H. Braden, Ronald F. Fox and Harold A. Gersch of the Georgia Institute of Technology. These investigators point out that some care is needed in extending Einstein's analysis to fast particles. For example, Boyer adopted Einstein's assumption that a molecule can be considered to move inertially, or without acceleration, when it emits or absorbs a photon. The assumption works well at the low velocities Einstein investigated. At high velocities, however, the recoil of the molecule as a result of the emission or absorption of a photon cannot be neglected even when the energy exchanged in the frame of reference of the molecule is small compared with the mass of the molecule. To put it another way, Boyer's model violates the law of conservation of momentum. The model also violates another conservation law, that of energy. The model neglects the fact that the mass of the particle changes as a result of emitting or absorbing the energy of a photon.

When the appropriate relativistic assumptions are made, according to the three recent interpretations, no inconsistency is found between special relativity, quantum mechanics and statistical mechanics. Indeed, at high velocities as well as at low ones the Boltzmann energy distribution of the gas molecules implies the Planck radiation law. Although Einstein's conclusion holds for relativistic velocities, there remains the question of why he did not investigate such velocities himself.

## *Oncogenes*

Some 100 kinds of cancer can be distinguished, but they all begin with the transformation of a single normal cell into a tumor cell. The cell's shape changes and so does its biochemistry; it ceases to be governed by the rules that have made it part of a coherent tissue. Over the years a bewildering variety of conditions and events have been identified that seem to trigger oncogenesis: the formation of a tumor. The incidence of particular cancers has been correlated with particular environmental factors, diets and personal habits. Ionizing radiation, sunlight and a long list of chemical agents have been shown to induce mutations in DNA, the genetic material, that can lead to cancer. A number of animal cancers are caused by certain viruses; this may be the case for some human cancers too, although it has not yet been conclusively demonstrated. Some cancers run in families, suggesting a genetic component. Disorders of the immune system are thought to play a role.

How is one to make sense of a disease with such varied manifestations and such an array of apparent triggering events? Until recently there was no indication of how these disparate events can lead to a similar result: the transformation of a cell. Now investigators in a number of disciplines are beginning to perceive the outlines of what J. Michael Bishop of the University of California at San Francisco calls a possible "final common pathway of tumorigenesis." There seems to be a family of genes whose alteration or abnormal expression in the cell may be the proximate cause of malignant transformation. In addition there are indications of how at least one subset of such genes gives rise to transformation.

The discovery of the genes came from the study of the retroviruses, an unusual group of viruses that can cause cancer in animals. The genetic material of the retroviruses is not DNA but the related nucleic acid RNA. When a retrovirus infects a cell, the viral RNA is transcribed into DNA, which is then integrated into the cell's genetic material. The DNA derived from the virus is thereafter transcribed into messenger RNA and translated into protein along with the cell's DNA. The genome, or total genetic complement, of the most intensively studied group of retroviruses, the avian sarcoma viruses, has only four genes. One of them, designated *src,* was found to encode a protein that induces sarcoma, a cancer of the connective tissues, in infected birds.

At first it was assumed that the *src* gene is peculiar to the virus. Soon a number of observations suggested a remarkable alternative explanation: the *src* gene and other retrovirus "oncogenes" originated in animal cells and have been captured by the retroviruses from the host cells. In other words, various retroviruses have plucked out from the large genome of animal cells a few genes that have the potential of inducing cancers.

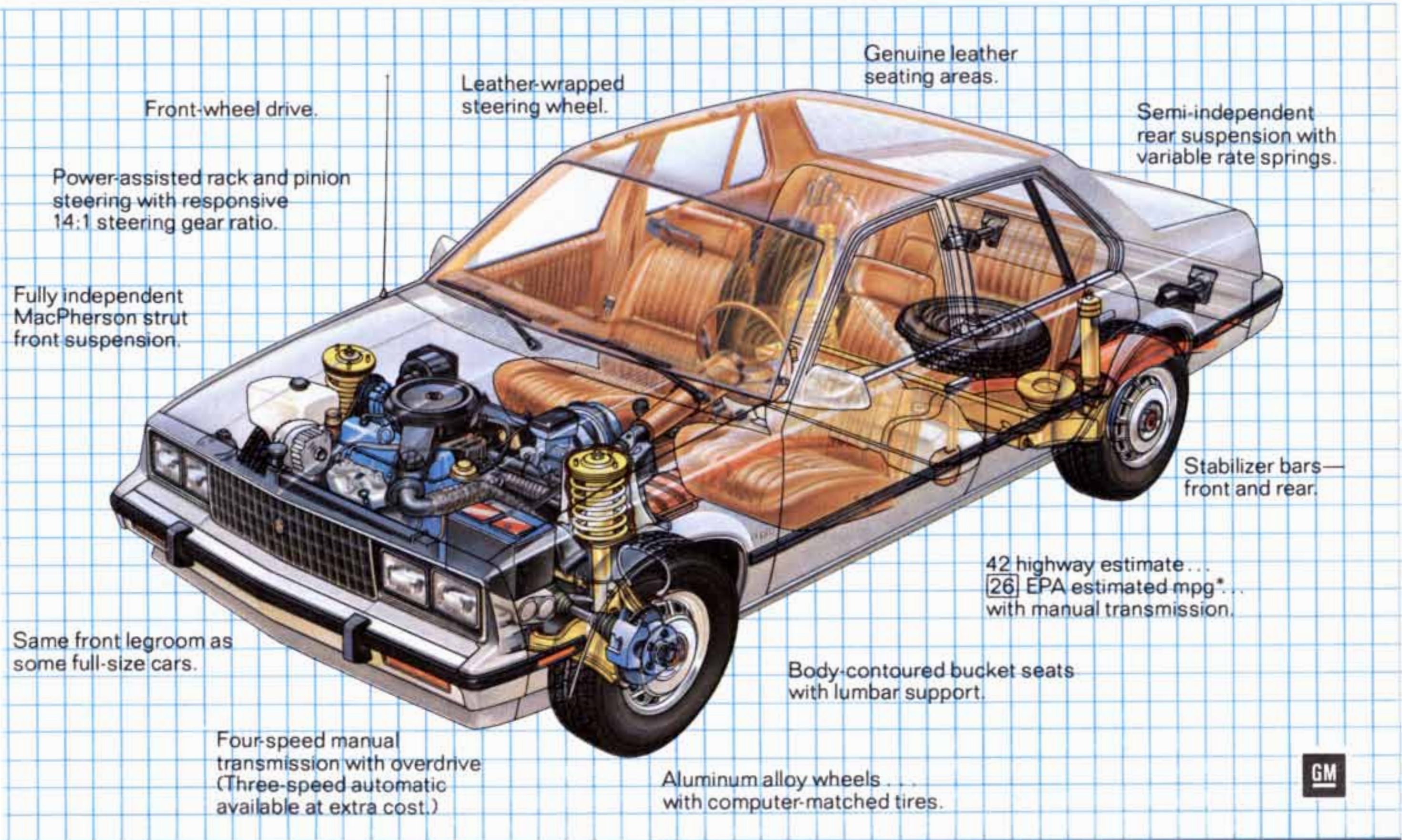
The original animal genes from which the oncogenes of the retroviruses are derived have been designated by Bishop "proto-oncogenes." Usually the proto-oncogenes do not lead to transformation; something happens to them that converts them into transforming genes. One possibility is that it is only in their viral form that they cause malignant transformation, perhaps because they are expressed (translated into protein) differently or more strongly. What is more likely is that viral infection somehow changes the expression of the cellular genes themselves, unmasking their oncogenic potential. Perhaps infection is only one of many events that can unmask a proto-oncogene.

Experiments conducted by Robert A. Weinberg of the Massachusetts Institute of Technology and by Geoffrey Cooper of Harvard University and their colleagues lend some support to the possibility that other stimuli can activate a proto-oncogene and thus also support the idea that there is indeed a common final pathway to transformation. Weinberg's group worked with a line of mouse fibroblasts (connective-tissue cells) that had been transformed by exposure to a chemical carcinogen, methylcholanthrene. They extracted the DNA from the cells and introduced it into normal mouse fibroblasts. Some of the normal cells were transformed; some of the DNA from the transformed cells was able in turn to transform other normal cells. The investigators established that no viral mediator was responsible for the transformation. This was the first direct demonstration that DNA damaged by a chemical agent is in itself carcinogenic.

In subsequent experiments Weinberg's group treated the transforming DNA with a battery of restriction enzymes, which can cut DNA at specific

## THIS IS THE NEW *Cimarron*

# WITH ALL THIS STANDARD...IT'S AN EXTRAORDINARY VALUE.




### EVERYTHING LISTED ABOVE IS STANDARD, AS WELL AS THE FOLLOWING:

- Air conditioning.
- Transverse-mounted engine to help provide interior roominess.
- Side and rear window defoggers.
- Front seat console with stowage compartment and coin holder.
- Tachometer and engine gauges.
- And more

Only Cimarron has all this... all standard. Before you buy—or lease—any car, test-drive Cimarron by Cadillac. Due to initial limited production, Cimarron is not available at all Cadillac dealers at this time.

***Use estimated mpg for comparison. Your mileage may differ depending on speed, distance, weather. Actual highway mileage lower. Cadillacs are equipped with GM-built engines produced by various divisions. See your Cadillac dealer for details.**

*Cimarron* BY CADILLAC

A NEW KIND OF CADILLAC
FOR A NEW KIND OF CADILLAC OWNER.

sites. Certain enzymes always abolished the ability to transform; others had no effect on it. A single gene was thereby implicated in all mouse-cell transformations. Recently Weinberg's group reported in *Nature* that DNA from several kinds of spontaneous or chemically induced tumors in mice, rabbits and human beings is similarly able to transform mouse fibroblasts. Weinberg and his colleagues may have uncovered a group of mammalian genes that, having been modified or deregulated by some carcinogenic trigger, can become transforming genes. At the genetic level, then, what happens in chemical carcinogenesis may be the same as what happens in carcinogenesis mediated by retroviruses. Just possibly some of the same genes may be at work.

If there are genes capable of initiating the growth of a tumor, what do they do? What proteins do they encode and how is their normal activity altered to bring about transformation? Hints of answers are emerging in the case of some sarcoma and leukemia retroviruses that infect birds and mice. The protein encoded by *src* and by a gene in some leukemia viruses is a kinase: an enzyme that transfers phosphate groups from adenosine triphosphate (ATP), the cell's major energy currency, to one of the amino acid subunits of a protein. Several retrovirus kinases (unlike most other kinases) phosphorylate the amino acid tyrosine in particular; several proteins in cells transformed by retroviruses show a significant increase in phosphorylated tyrosines. Phosphorylation is a biochemical step that can be expected to have broad effects on a cell's metabolism, structure and relations with other cells. It could account for many of the changes seen in transformation.

One of the most interesting effects of phosphorylation is on energy metabolism. In the 1920's the German biochemist Otto Warburg found that glycolysis, the fermentation of glucose to lactic acid, is enhanced in malignant cells. Efraim Racker of Cornell University showed some years ago that the glycolytic pathway is favored by an increased availability of adenosine diphosphate (ADP) and inorganic phosphate. These substances are generated by enzymes called ATPases that split ATP as they effect such cellular processes as the synthesis of large molecules and the transport of ions across membranes. Racker and his colleagues found that one ATPase in the cell membrane, which pumps sodium out of the cell and moves potassium into it, is inefficient in tumor cells: it breaks up ATP without pumping a corresponding amount of sodium, thereby making more ADP and inorganic phosphate available for glycolysis. Racker and his colleagues established that the tumor ATPase has an extra phosphate group on one amino acid unit and that a particular kinase in the tumor cell is responsible for the extra phosphorylation. That phosphorylation, it has turned out, is the result of a four-kinase "cascade" of enzymes. The cascade is initiated by a kinase very similar to the one encoded by the *src* gene. It phosphorylates a second kinase, which phosphorylates a third one, which phosphorylates a fourth; it is the fourth kinase that eventually phosphorylates the ATPase.

Perhaps the kinase cascade is the common element in at least some retrovirus-mediated transformations, and perhaps even in transformations effected by other agents. That remains to be demonstrated. So does the broader concept that a particular, limited family of transforming genes may be implicated in causing a broad range of cancers. As the matter stands now a variety of stimulating questions can be asked, and many of them appear to be amenable to experiment. Disparate lines of investigation have suggested fundamental generalizations. It is as if a number of blind men, each feeling a different part of the elephant, had begun to recognize the basic nature of elephantness.

## *Micromachines*

Since it is now routine to build something as complex as the central processing unit of a computer on a single chip of silicon by the techniques of photolithography and chemical etching, it would seem plausible that other miniature structures might be fabricated in a similar way. A substantial amount of activity along these lines is already under way, some of it in research laboratories and some at the manufacturing stage. The process by which the devices are made has come to be known as micromachining. It seems likely to play a significant role in instrumentation and control systems for health care, industrial processing and motor vehicles.

At present most of the devices being made by micromachining are components of larger apparatus. In general the devices function as sensors or actuators, often serving to establish a connection between an electronic apparatus and the nonelectronic world with which it deals. For example, a cantilever beam etched in silicon and supporting a weight on a silicon paddle operates as an accelerometer that is sensitive in the range from .01 to 50 times the acceleration of gravity at the earth's surface. The accelerometer is a transducer that converts a change in velocity into an electrical signal. The Integrated Circuits Laboratory of Stanford University has developed a miniature accelerometer of this kind for biomedical applications such as the measurement of the motion of the heart wall. Workers at the International Business Machines Research Laboratory in San Jose, Calif., have also made cantilever-beam devices. They function as electromechanical relays and as torsion bars that support mirrored surfaces in light modulators and laser scanners.

If a cavity etched in silicon is covered with a thin silicon membrane, one has the basis for a pump or a pressure sensor. Pumping is achieved by applying a voltage that deflects the membrane, thereby ejecting liquid from the cavity. The IBM workers have made a single-chip device consisting of a cavity, a membrane and a nozzle; it delivers ink in an ink-jet printer.

A more recent trend is toward self-contained microminiature apparatus. An example is a gas chromatograph devised by the Stanford group. On a silicon wafer five centimeters across they fabricate an input channel for the sample gas, an input channel for a carrier gas, a mixing valve, a spiral capillary column 1.5 meters long and a detector channel. With the addition of a thermal-conductivity detector that is fabricated separately and mounted on the wafer, the device separates a sample gas into its constituents in less than 10 seconds. The results of the analysis (made by a microcomputer on a separate chip) are reported by means of a liquid-crystal display. The system is expected to find application in portable monitors of air quality and in unmanned probes of the atmosphere of other planets.

## *Soft Suckers*

The two essentials of a mechanism for inhaling are that a chamber (such as a lung) must expand and a rigid framework (such as a rib cage) must keep the enlarged chamber from collapsing before it is filled with air or water. All vertebrate animals conform to this pattern, and so do the scorpions, spiders and crabs that are said to suck in their air. How then can an animal breathe if there are no hard parts in its body? The problem must have been solved by every soft-bodied animal. In particular it arises for the sea cucumbers, soft-bodied, almost cylindrical animals with a dark, warty skin, and for *Urechis*, a pink, sausage-shaped unsegmented marine worm that dwells in mud flats off the West Coast of the U.S. These animals breathe by drawing water into their body and extracting dissolved oxygen from it. Thomas G. Wolcott of North Carolina State University has now proposed an explanation of how this is done. His hypothesis appears in *Biological Bulletin*.

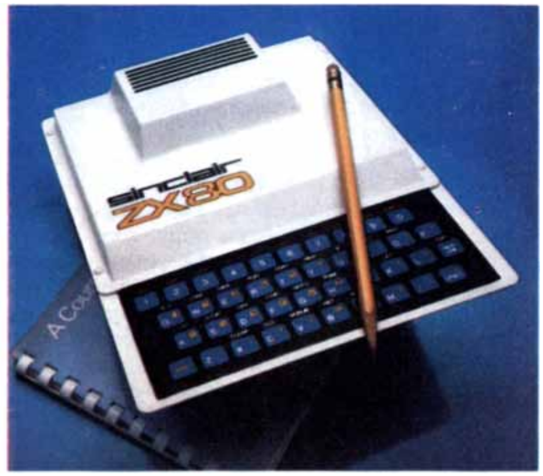
For the purpose of understanding how a soft-bodied animal breathes, its body can be considered a tube within a tube. The outer tube is the body wall, which includes laminations of muscle. Some of the muscles extend like hoops around the body; others extend longitudinally. The animal can move itself much like a worm. The inner tube is the cloaca (literally the sewer), a chamber that communicates at one end with the

# Send your kids on a 10-day trip into the computer age. Or send yourself.

Someday computers will be a common household item.

But you don't have to wait for the future. Your family can take a trip into the computer age now, at no risk and no obligation. Because for 10 days, you can have your very own Sinclair ZX80 computer. To play with, experiment with, learn with. It will give your children a valuable headstart in life.

And if you don't have kids? Then take the trip yourself, because it's your chance to keep up with a changing world.




**EVERYTHING YOU NEED.** For just $199.95, you get everything you need to operate your ZX80. You just hook it up to your TV (either black & white or color). Then to store your programs, you can use any ordinary cassette tape recorder. We'll also send you a 128-page guide to computing that's yours to keep as a free gift.

**$199.95**

## ANYONE CAN USE IT.

You'll find the Sinclair ZX80 is incredibly easy to use for anyone 10 years or older. You don't have to know anything about computers when you start out. We'll send you a complete 128-page guide that takes you step-by-step into the world of computing. In one day, you'll be writing your own programs!

The ZX80 will absorb you and your children for hours as you find new ways to use it. You can create games. Home budgets. Math lessons for your kids. Almost anything, in fact.

## THE ZX80 IS YOURS FOR 10 DAYS. THE COMPUTER GUIDE IS YOURS FOREVER.

For your 10-day home trial, just call our toll-free number and order with your Master Charge or VISA. Or send the coupon along with a check or money order for the remarkably low price of $199.95, plus shipping. (Other personal computers cost two or three times as much!)

For an extra $49.95, you can also have the Sinclair Computer Learning Lab—a set of 100 experiments and lessons on cassette tapes that make the ZX80 itself your teacher.

And we'll send you the Sinclair ZX80 personal computer. It's yours for 10 days. If you decide you don't want to keep it, then send it back. We'll give you a full refund. Either way, the 128-page guide to computing—a $5.95 value—is yours to keep as a free gift.

Why not take advantage of this unique offer today?

And send your family on a 10-day journey they'll never forget.

# 10-DAY MONEY-BACK HOME TRIAL
# CALL TOLL FREE 800-543-3000.

Ask for operator #508. In Ohio call 800-582-1364. Phones open 24 hours a day, 7 days a week.

Mail to: Sinclair Research Ltd.
One Sinclair Plaza, Nashua, NH 03061

Check one:

| | | | |
|---|---|---|---|
| ☐ ZX80 Personal Computer | $199.95 | ☐ ZX80 Personal Computer | $199.95 |
| | | Computer Learning Lab | $ 49.95 |
| Shipping | $ 4.00 | Shipping | $ 4.00 |
| Total | $203.95* | Total | $253.90* |

If I'm not fully satisfied with my purchase I can return it in 10 days for a full refund. The 128-page guide is mine to keep.

Name__________

Address__________

City, State, Zip__________

*U.S. dollars.

06SA

**sinclair**

SINCLAIR RESEARCH LTD., ONE SINCLAIR PLAZA, NASHUA NH 03061 anus and at the other end with the intestine and also with a pair of long branching hollows known as respiratory trees. The cloaca, like the body wall, is muscular and flexible. The annular space between the body wall and the cloaca is filled with water and with radial muscles, which span the distance between the tubes.

Three steps, Wolcott observes, make up each cycle of breathing. Water fills the cloaca; it is driven into the respiratory trees, where oxygen dissolved in the water is captured; finally the water is expelled through the anus. The last two steps are easy to understand; they require various combinations of muscle contractions and various openings and closings of the sphincters, or ring muscles, at the portals of the cloaca. The first step is more mysterious. One imagines that the radial muscles might contract and thereby enlarge the cloaca. Water would then be drawn in through the anus. The contraction of the radial muscles, however, would tend to collapse the body wall. And if the collapse of the body wall were resisted by increasing the pressure of the water in the space between the tubes, it seems as if the inner tube, the cloaca, would tend to collapse.

In an effort to resolve this apparent paradox Wolcott made models to determine "the maximum theoretical suction" in "a hydrostatically supported sac within a sac." In a mathematical model the body wall and the cloaca were described as concentric cylinders or as concentric half spheres. The morphology of the animal is intermediate between the two. In a mechanical model the cloaca was a toy balloon and the body wall was a rubber condom. The radial muscles were modeled by pieces of nylon fishing line.

From both of the models a plausible explanation emerged. Assume that a soft-bodied animal has made an effort to suck in water while its anal sphincter is closed, so that no water enters and the pressure of suction remains maximal. The radial muscles will then have contracted. The contraction indeed will tend to pull the body wall in and the cloacal wall out. Meanwhile the constriction of the annular space inside the body wall will have squeezed the water there. The result will be a hydrostatic pressure that resists the muscles' contractile force. The hydrostatic pressure is distributed throughout the water-filled volume, so that the pressure per unit area on the surfaces facing the water is equal everywhere. Since the body wall surrounds the cloaca, however, its surface area is larger, and so is the total pressure on it. The total pressure on the body wall and the smaller total pressure on the cloaca are mechanically coupled by the network of radial muscles. It follows that a net outward force acts through the muscles to expand the cloaca, and if the anal sphincter is opened, water will be drawn in.

The crucial point is the concentric geometry of the body wall and the cloaca. In the cylindrical model the pressure available for suction exceeds in magnitude the pressure inside the body wall (which is sustained throughout the hypothetical cycle) if the radius of the cloaca is less than half that of the body. The rubber model yielded even more striking results. "The suctions attainable at minimum cloacal diameters" were "up to 20 times as large" as the pressure inside the body wall.

Wolcott attempted to measure the pressures in living cloacal breathers by means of probes inserted through the skin or through the anus. The animals "spent much effort attempting to rid themselves of the tubing while producing meters of irrelevant pressure records." The animals also clogged the probes with mucus. Plugging the anus to maximize the sucking pressure "triggered writhing, a 'cough' reflex and cessation of normal pumping." The attempts revealed "that living animals are much less tractable research subjects than are mathematical models or rubber novelties." Ultimately the pressures were measured only in animals whose anal sphincter was open. The data confirmed Wolcott's hypothesis at least to this extent: the animals do maintain the pressure inside their body wall during what amounts to inhalation.

## *Alexander's Dial*

One of the many cities founded by Alexander the Great on his march to India (330–325 B.C.) was in ancient Bactria, just south of the Oxus River. The river is now called the Amu Darya, and the ruins of the city lie on the frontier between Afghanistan and Russian Turkistan. A French team of archaeologists, which has been working at the site for some years, has found among other evidence of the Hellenic presence in central Asia a curiously carved slab of limestone. It lay buried under limestone rubble in a room of what had been the city gymnasium; perhaps, like the rubble, the slab had been destined to feed a nearby limekiln and was spared when the city was destroyed by invaders from the steppe some two centuries after its founding.

The slab, 45 centimeters high, 35 centimeters wide and 15 centimeters thick, is pierced by a cylindrical hole 22 centimeters in diameter. The carvings consist of two sets of 13 lines cut into the lower half of the cylindrical surface. Twelve of the lines extend some five centimeters inward from one face of the slab and 12 extend the same distance inward from the other face; the remaining two lines, which are in the middle of each series, lie at the hole's lowest point and meet in the middle to form a single line that extends through the thickness of the slab. The investigators recognized the curious object as a sundial of the equatorial type, one that must be set with the faces of the slab parallel to the plane of the celestial equator. The finding of such an instrument was no small surprise; it had been thought that the equatorial sundial was not invented until late in the first millennium A.D.

René R.-J. Rohr, a French student of gnomonics, has recently summarized in English two French analyses of the instrument; writing in *Journal of the Royal Astronomical Society of Canada,* Rohr also speculates on the history of the artifact. As for analysis, the first point to be established was that the dial's hour divisions measure what are known as temporary hours. In this system of timekeeping the duration of an hour is not fixed; instead the period between sunrise and sunset is divided into 12 hours that grow longer in spring and summer and shorter in fall and winter. That is how hours were commonly reckoned from Classical Greek times until the Middle Ages. Second, knowing the declination of the sun at the solstice in Alexander's day, it was possible to calculate at what latitude the instrument was meant to be used. When the calculation was made, the "home" of the instrument was found to lie near the Tropic of Cancer, some 13 and a half degrees south of the city where it was discovered.

Rohr proposes the following reconstruction of the instrument's history. The closest Alexander came to the Tropic of Cancer was in India at ancient Pattala, near the mouth of the Indus, in 325 B.C. There part of his withdrawing army was to leave by sea for the Persian Gulf. Preparations for the voyage were protracted, and Rohr suggests that at this improvised port facility some anonymous but inspired member of Alexander's forces saw to the making of an equatorial dial with the offset angle of its base suited to Pattala's latitude. After the fleet sailed and Alexander, with the rest of his troops, began an overland return to Persia, the Pattala dial was transported to the Greek city in Bactria. There someone familiar with sundial practice but weak on theory cut a wedge off the bottom of the slab to adjust its tilt to the higher latitude. The increased inclination of the slab, however, threw the engraved hour lines out of their proper horizontal position. As a result the Pattala dial now showed the correct temporary hours only twice a year, at the vernal and autumnal equinox.

Did anyone in Greek Bactria complain? Rohr suggests that no one may have noticed, citing an episode in the First Punic War. The triumphant Romans carried a Greek sundial captured in Sicily back to the capital city. A century passed before the citizens discovered that the Sicilian dial was not telling Roman time.


MAN OF THE YEAR 1976
MICHAEL HILL
PRESENTED BY
DEERVILLE JAYCEES
Johnnie Walker
BOTTLED IN SCOTLAND
YEARS 12 OLD

# Archaebacteria

*These unusual bacteria are genealogically neither prokaryotes nor eukaryotes. This discovery means there are not two lines of descent but three: the archaebacteria, the true bacteria and the eukaryotes*

by Carl R. Woese

Early natural philosophers held that life on the earth is fundamentally dichotomous: all living things are either animals or plants. When microorganisms were discovered, they were divided in the same way. The large and motile ones were considered to be animals and the ones that appeared not to move, including the bacteria, were considered to be plants. As understanding of the microscopic world advanced it became apparent that a simple twofold classification would not suffice, and so additional categories were introduced: fungi, protozoa and bacteria. Ultimately, however, a new simplification took hold. It seemed that life might be dichotomous after all, but at a deeper level, namely in the structure of the living cell. All cells appeared to belong to one or the other of two groups: the eukaryotes, which are cells with a well-formed nucleus, and the prokaryotes, which do not have such a nucleus. Multicellular plants and animals are eukaryotic and so are many unicellular organisms. The only prokaryotes are the bacteria (including the cyanobacteria, which were formerly called blue-green algae).

In the past few years my colleagues and I have been led to propose a fundamental revision of this picture. Among the bacteria we have found a group of organisms that do not seem to belong to either of the basic categories. The organisms we have been studying are prokaryotic in the sense that they do not have a nucleus, and indeed outwardly they look much like ordinary bacteria. In their biochemistry, however, and in the structure of certain large molecules, they are as different from other prokaryotes as they are from eukaryotes. Phylogenetically they are neither prokaryotes nor eukaryotes. They make up a new "primary kingdom," with a completely different status in the history and the natural order of life.

We have named these organisms archaebacteria. The name reflects an untested conjecture about their evolutionary status. The phylogenetic evidence suggests that the archaebacteria are at least as old as the other major groups. Moreover, some of the archaebacteria have a form of metabolism that seems particularly well suited to the conditions believed to have prevailed in the early history of life on the earth. Hence it seems possible that the newest group of organisms is actually the oldest.

## The Evolutionary Record

The earth is four and a half billion years old, and on the basis of the macroscopic fossil record it would appear to have been inhabited for less than a seventh of that time: the entire evolutionary progression from the most ancient marine forms to man spans only 600 million years. The fossil imprints of unicellular organisms too small to be seen with the unaided eye tell a different story. Microfossils of bacteria in particular are plentiful in sediments of all ages; they have been found in the oldest intact sedimentary rocks known, 3.5-billion-year-old deposits in Australia. Over an enormous expanse of time, during which no higher forms existed, the bacteria arose and radiated to form a wide variety of types inhabiting a great many ecological niches. This age of microorganisms is the most important period in evolutionary history not only because of its duration but also because of the nature of the evolutionary events that took place over those billions of years.

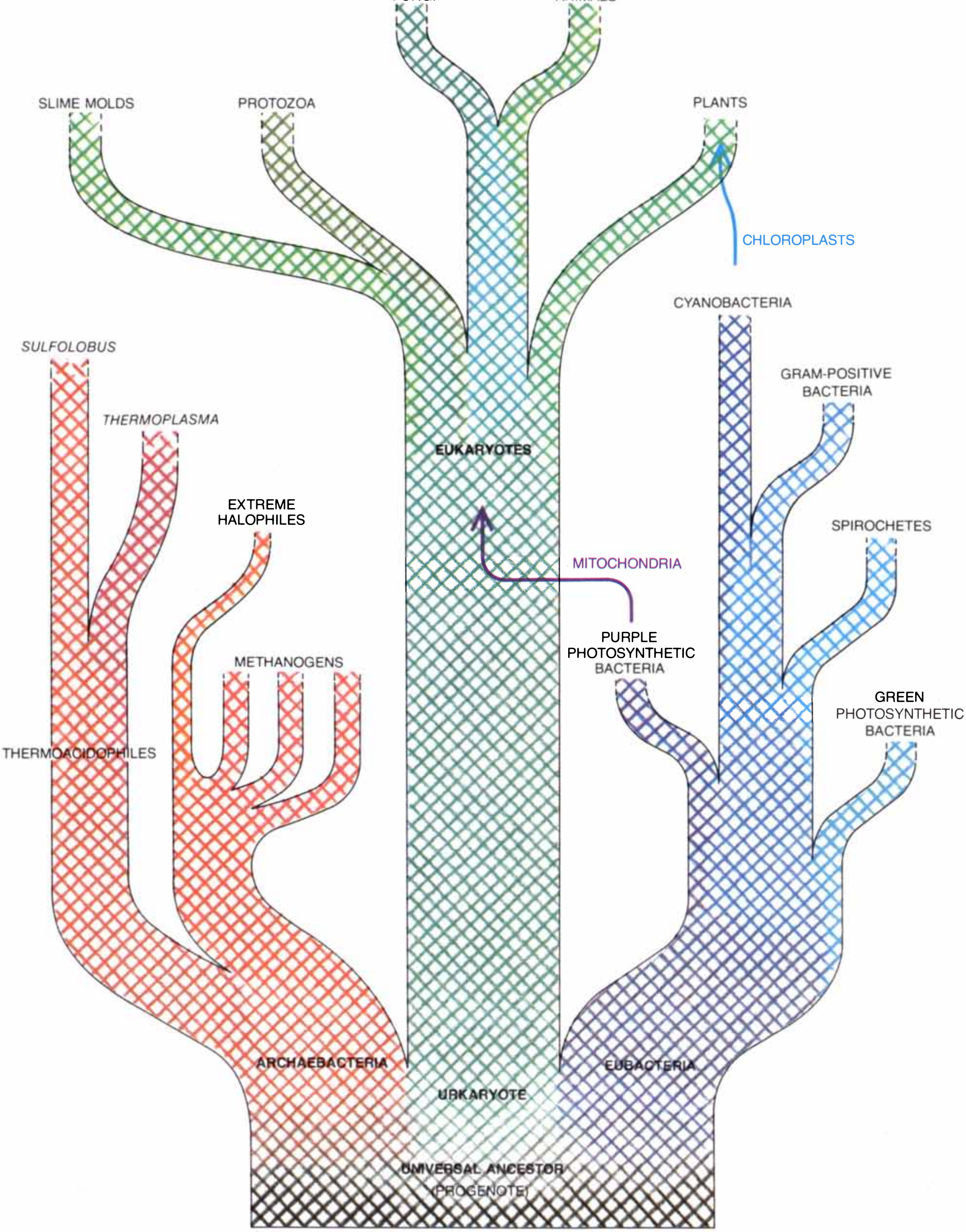
Until recently, however, almost nothing was known about the age of microorganisms. Bacterial microfossils are not very informative structures; little can be inferred from the imprint of a small sphere or rod. The main paleontological indications of the nature of the early bacteria have come not from the individual microfossils but from the macroscopic structures called stromatolites, which are thought to be fossilized bacterial mats: colonies of bacteria embedded with minerals. Today such structures are formed primarily by several kinds of photosynthetic bacteria, usually the cyanobacteria. Stromatolites fossilized recently resemble the ancient ones to such an extent that it seems entirely reasonable to think the ancient structures were also made by photosynthetic bacteria. Therefore at least some of the ancient bacteria must have been photosynthesizers. Apart from that one fact virtually nothing could be established with certainty about the earliest microorganisms. The entire evolutionary tree of the bacteria remained obscure, as did the base of the tree for the higher forms of life.

In reconstructing early evolutionary events, however, biologists are not limited to the geologic fossil record. The cell itself retains evidence of its past in the amino acid sequences of its proteins and in the nucleotide sequences of its nucleic acids: DNA and RNA. This living record is potentially far richer and more extensive than the fossil record, and it reaches back in time beyond the oldest fossils, to the period when the common ancestor of all life existed.

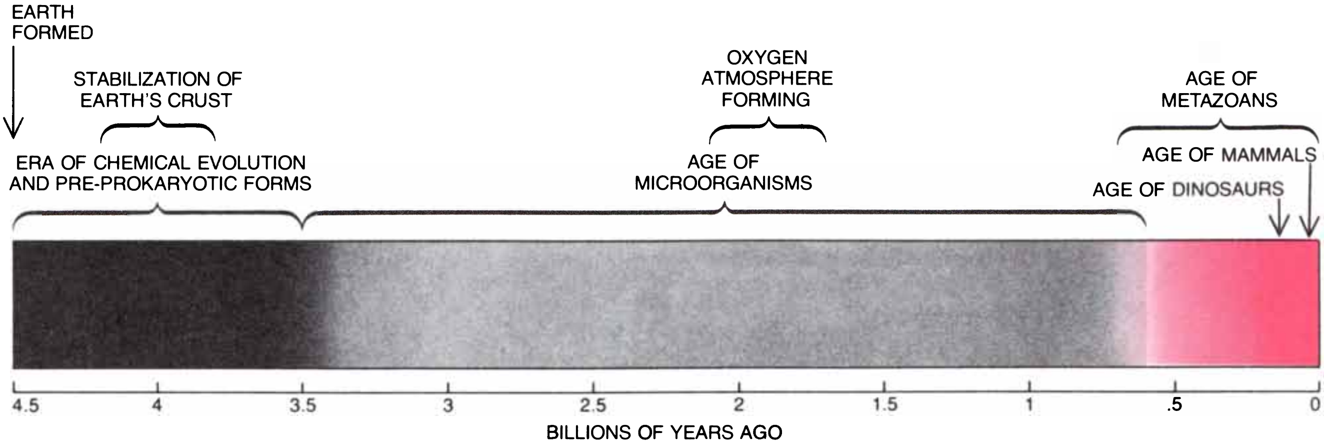
In order to read the biochemical record it was necessary to develop a technology for determining (at least in part) the sequence of a gene or of the RNA or protein product encoded by a gene. For proteins this has been possible for about 25 years, but the direct sequencing of DNA and RNA has been feasible for only the past five years or so. The new technology for sequencing nucleic acids should enable biologists to uncover in relatively short order far more about the history of life than had been thought possible. It was by applying techniques of sequencing to the century-old problem of the natural relations among bacteria that my colleagues and I recognized the archaebacteria as a third form of life.

## Eukaryotes and Prokaryotes

In order to appreciate the special status of the archaebacteria it is helpful to consider some of the defining characteristics of eukaryotes and prokaryotes. The eukaryotic cell is comparatively large: roughly 10 micrometers on a side. It is surrounded by a double membrane,



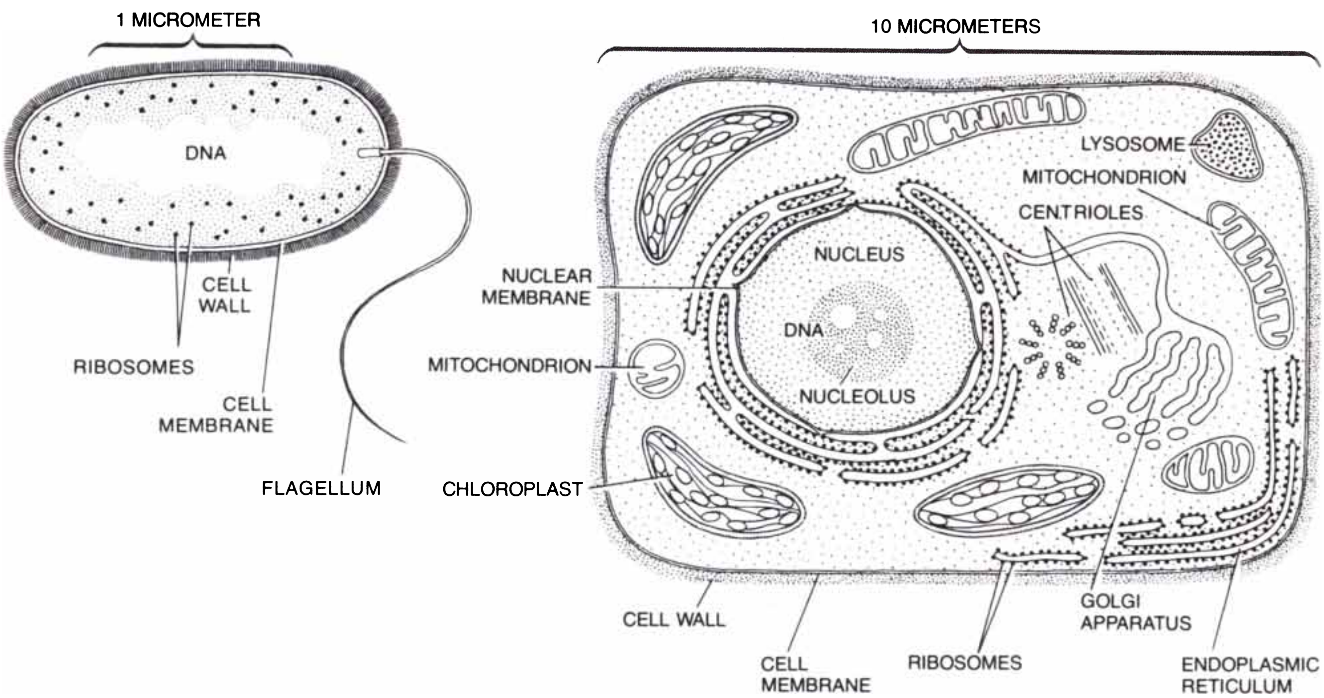

**THREE PRIMARY KINGDOMS** are proposed by the author to comport with the discovery that the archaebacteria are fundamentally different from all other bacteria, which are designated the eubacteria, or true bacteria. Both eubacteria and archaebacteria are alike in being prokaryotic cells: simple cells that lack a nucleus and are very different in their structural properties from eukaryotic cells, which have a nucleus and several other subcellular organelles. Genealogically, however, archaebacteria and eubacteria are no more closely related to each other than either group is to eukaryotes. It is proposed that the archaebacteria, the eubacteria and an urkaryote—the original eukaryotic cell—stemmed from a common ancestor (the progenote) much simpler than the simplest present-day cells (prokaryotes). Eukaryotes evolved after the urkaryote became a "host" for bacterial endosymbionts that developed into mitochondrion and chloroplast.




**AGE OF MICROORGANISMS, which lasted some three billion years, dominates the time span of biological evolution. Microfossils of prokaryotic cells have been found in deposits 3.5 billion years old, and those cells must have been preceded by simpler ones. The earliest eukaryotic fossils are only about 1.3 billion years old. Almost nothing is known about evolution during the age of microorganisms. The macroscopic fossil record goes back only about 600 million years to the time of the earliest metazoans, or multicellular organisms.**

within which a number of structures can be discerned that are themselves defined by membranes. The nucleus contains the bulk of the cell's genetic material. The rod-shaped mitochondria are the site of cellular respiration, which generates the cell's main energy currency, adenosine triphosphate (ATP). In plant cells the chloroplast, another rod-shaped body, converts the energy of light into the chemical energy of ATP. Other specialized structures such as the Golgi apparatus (for secretion) and cilia (for motility) are often present. Many eukaryotic cells are laced with a membrane system, the endoplasmic reticulum, that provides a surface on which important reactions such as the synthesis of proteins take place.

The prokaryotic cell is vastly different. It is typically far smaller than the eukaryotic cell: by a factor of 10 in linear measure and hence by a factor of 1,000 in volume. The prokaryotic cell too is circumscribed by a double membrane, and in addition it almost always has a rigid cell wall. On the other hand, none of the internal structures characteristic of the eukaryotic cell are present; there are no mitochondria or chloroplasts and of course there is no membrane-bounded nucleus. The genome—the total complement of genetic




**PROKARYOTES AND EUKARYOTES are fundamentally different at the structural level, as is shown by these schematized drawings of a typical prokaryotic cell (*left*) and eukaryotic cell (*right*). The prokaryote is by far the smaller cell. Little subcellular structure is seen even at the scale revealed by the electron microscope; a single circular strand of the genetic material DNA lies loose in the cytoplasm. Both the archaebacteria and the eubacteria are prokaryotes and share prokaryotic structural properties. The eukaryotic cell is much larger and has a number of discrete subcellular structures. Its DNA, complexed with proteins, is organized into chromosomes within a membrane-bounded nucleus. Mitochondria carry out cellular respiration; in a plant cell there are chloroplasts, which conduct photosynthesis. The Golgi apparatus is a secretory organelle; the endoplasmic reticulum is a membrane system along parts of which some of the cell's ribosomes (on which genetic information is translated into protein) are arrayed. All cells more complex than the bacteria are eukaryotes.**

# Citicorp Travelers Checks are accepted in places that've never heard of a split-level.

That's right, even in places where there's no need to worry about aluminum siding and bicycles left in the driveway.

The fact is, you never know where you might end up. That's why it's nice to know Citicorp Travelers Checks are accepted in literally millions of hotels, restaurants and stores throughout the world.

So don't worry if you're visiting a part of the world where the people throwing a block party own the entire block. Because you can use Citicorp Travelers Checks even there.

Travel the world with Citicorp—America's leading financial institution, worldwide.


CITICORP®
TRAVELERS CHECKS

# The Polaroid SX-70 Auto

**With New Time-Zero Supercolor Film.**

The photographs shown here not only illustrate the sophistication of one of the world's finest cameras, they also offer a Polaroid-eye-view of one of the world's most exclusive auction houses, Sotheby Parke Bernet.

**1** This detail of an Aubusson tapestry testifies both to the visual splendor which abounds at Sotheby's and to the advantage of single lens viewing. Unlike all other instant cameras, the SX-70 AutoFocus is an SLR. So the composition you see through the lens is the picture you get.

**2** As you can see, not all of the lovely things at Sotheby's are objets d'art. Note how new Polaroid Time-Zero Supercolor film faithfully reproduces not only the rich jewel tones of this young woman's costume but the subtler hues of her complexion as well.

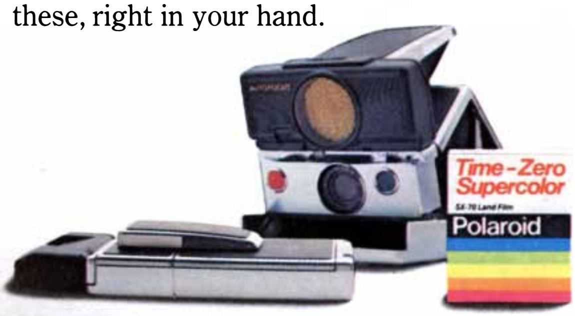
**3** The hands of Sotheby's master craftsman Gian Asaro, captured here restoring a 16th century figure, demonstrate Polaroid's amazing sonar autofocus system. While all you do is aim and shoot, the SX-70 AutoFocus sets the aperture and shutter and whips the lens into precise focus. Automatically.

**4** Each time you open your SX-70 AutoFocus, a unique sharing experience unfolds. Both you and your subject can react to the world's fastest emerging color image in just 10 seconds. And as the picture develops, so does the chemistry.

**5** Pictured here is an extreme close-up of a rare, Early American coverlet. No other camera lets you move in this close—as close as 10.4 inches—without expensive extra lenses.

**6** You can almost feel the cool countenance of this magnificent marble bust. Because while other instant cameras have plastic lenses, the SX-70 AutoFocus has a fine, four-element glass lens that can give you the crisp, clear images you demand.

Perhaps now you can see why, no matter where your imagination takes you, no other camera in the world lets you leave with remarkable photographs like these, right in your hand.



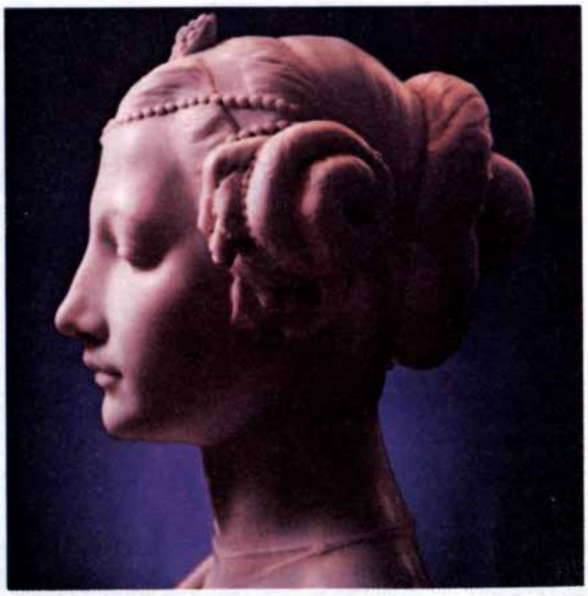

**No other instant camera comes close.**

1

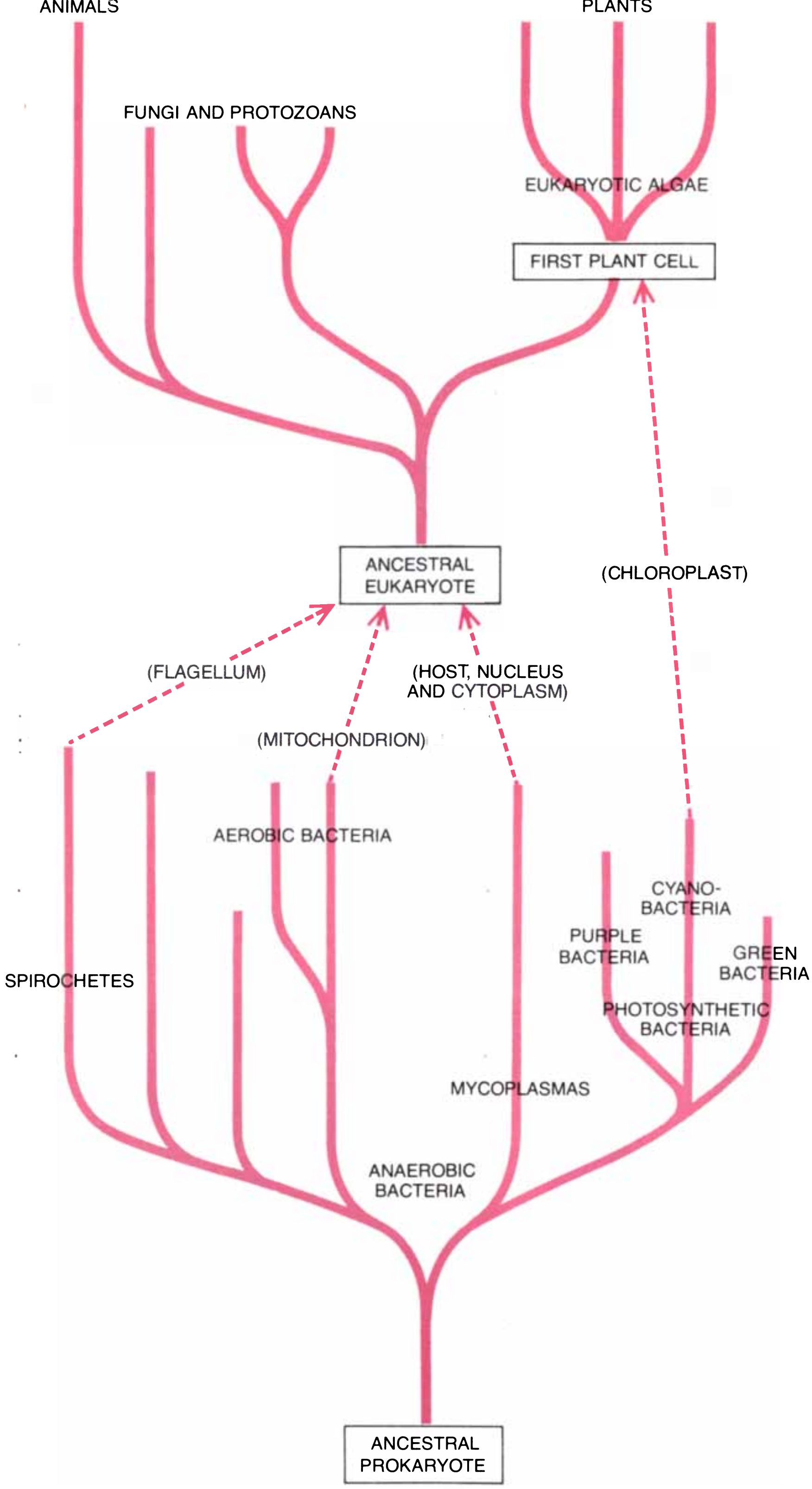
4

## ocus goes to Sotheby's.

3



6




**CONVENTIONAL TREE OF LIFE prior to the discovery of archaebacteria had two primary lines of descent: prokaryotic and eukaryotic, the latter derived from the former. The first cells were assumed to have been anaerobic bacteria (prokaryotes) that derived energy from fermentation. They gave rise to a variety of sublines. After the atmosphere became enriched with oxygen, certain anaerobic cells that had lost their cell wall (mycoplasmas) established an endosymbiotic relation with smaller bacteria they had ingested. An endosymbiotic aerobic (oxygen-respiring) bacterium evolved into the mitochondrion, a photosynthesizing cyanobacterium into the chloroplast and (perhaps) a spirochete into the flagellum (an organ of motility). In this way an ancestral eukaryotic cell evolved, and it in turn gave rise to the protozoa, fungi, animals and plants. The drawing is based on a scheme devised by Lynn Margulis of Boston University.**

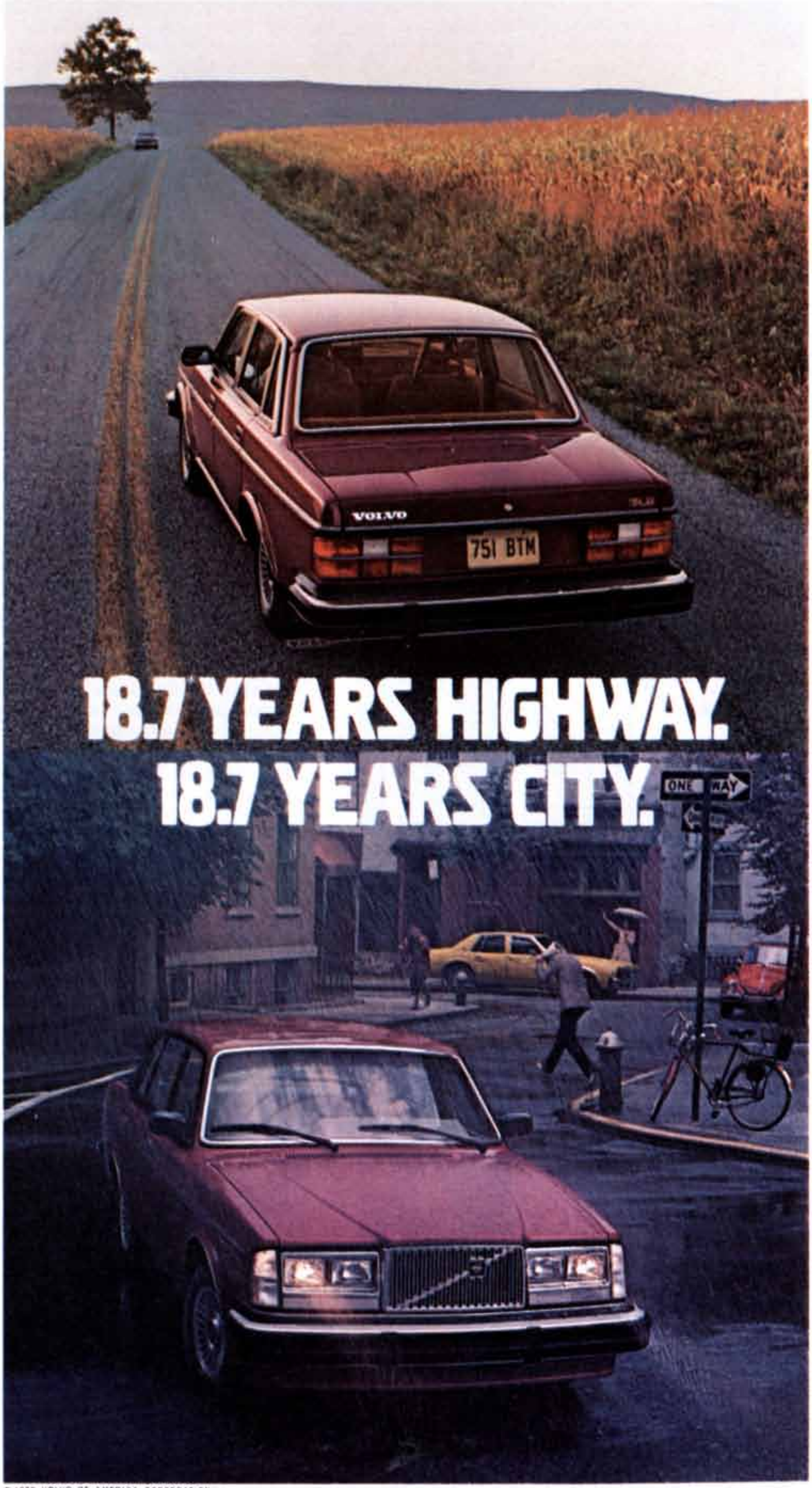
material—is limited to between 2,000 and 3,000 genes in a prokaryotic cell; the typical eukaryotic genome is larger by several orders of magnitude.

The distinction between eukaryotes and prokaryotes was initially defined in terms of subcellular structures visible with a microscope. At that level all cells appeared to be either large and complex, and so eukaryotic, or small and simple, and so prokaryotic. The distinction between the two cell types was ultimately carried to the most basic biological level, the level of molecules. Here eukaryotic and prokaryotic cells have many features in common. For instance, they translate genetic information into proteins according to the same genetic code. Even where the molecular processes are the same, however, the details in the two forms are different; they are either characteristically eukaryotic or characteristically prokaryotic. For example, the amino acid sequences of various enzymes tend to be typically prokaryotic or eukaryotic. All these differences between groups and similarities within each group made it seem certain to most biologists that the tree of life had two main stems, one stem prokaryotic and the other eukaryotic.

That conclusion was drawn too hastily; the aesthetic appeal of a dichotomy was too great. Simply because there are two types of cell at the microscopic level it does not follow that there must be only two types at the molecular level.

The evolutionary relation of prokaryotes and eukaryotes is actually more complicated than the evidence cited above would indicate. Two eukaryotic organelles, the mitochondrion and the chloroplast, each have their own DNA. Moreover, the pigments in the chloroplast (the chlorophylls and the carotenoids) are similar to those found in the cyanobacteria. Both mitochondria and chloroplasts are the size of bacteria; their apparatus for translating genetic information into proteins differs from the eukaryotic cell's own apparatus and has a number of properties in common with that of prokaryotes.

These facts and others have led to the idea that mitochondria and chloroplasts are descended from prokaryotes that became trapped in a larger cell and established an endosymbiotic relation with it. The mitochondrion is thought to have been a respiring bacterium and the chloroplast to have been a photosynthesizing relative of the cyanobacteria. This conjecture, which in its simplest form is more than a century old, was essentially proved in the case of the chloroplast by the demonstration that the nucleotide sequence of one of the kinds of RNA in the organelle, ribosomal RNA, is specifically related to ribosomal-RNA sequences in cyanobacteria. Similarly, the ribosomal RNA of the mitochondrion in plants appears to be of the bacterial






A lot of car makers today are trying to sell you economy with EPA figures. But at Volvo, we believe true economy isn't more miles per gallon. It's more years per car.

So if you just want to buy less gas and save a little money, look at EPA figures. But if you like the idea of buying fewer cars and saving a lot, consider Volvo's figures.

***Average life expectancy of a Volvo in Sweden. Driving conditions in the United States may differ. So your Volvo may not last as long. Then again, it may last longer.**

**VOLVO**

A car you can believe in.

type. Thus it seems that at least two lines of prokaryotic descent are represented in the eukaryotic cell.

## The Urkaryote

Logically the next question is: Where does the rest of the eukaryotic cell come from? What was the original host cell: the urkaryote? It is generally agreed that the bulk of the eukaryotic cell (the nucleus and the cytoplasmic structures) represents a separate line of descent. The exact nature of the ancestral cell is not clear. Most investigators have tended to believe the main eukaryotic line also arose from among the ordinary bacteria. The idea is that some anaerobic bacterium deriving its energy from the fermentation of nutrients (rather than from their oxidation) at some point happened to lose its tough cell wall. Organisms of this kind are known; they are the mycoplasmas. A strain of mycoplasma then evolved the capacity to engulf other organisms, an ability retained by many eukaryotes today. Among the many kinds of organisms such a mycoplasma might have ingested two appear to have established a stable endosymbiotic relation with their host and to have become the mitochondrion and the chloroplast. In this way the eukaryotic cell was born. (The origin of its defining characteristic, the membrane-bounded nucleus, is still far from clear.)

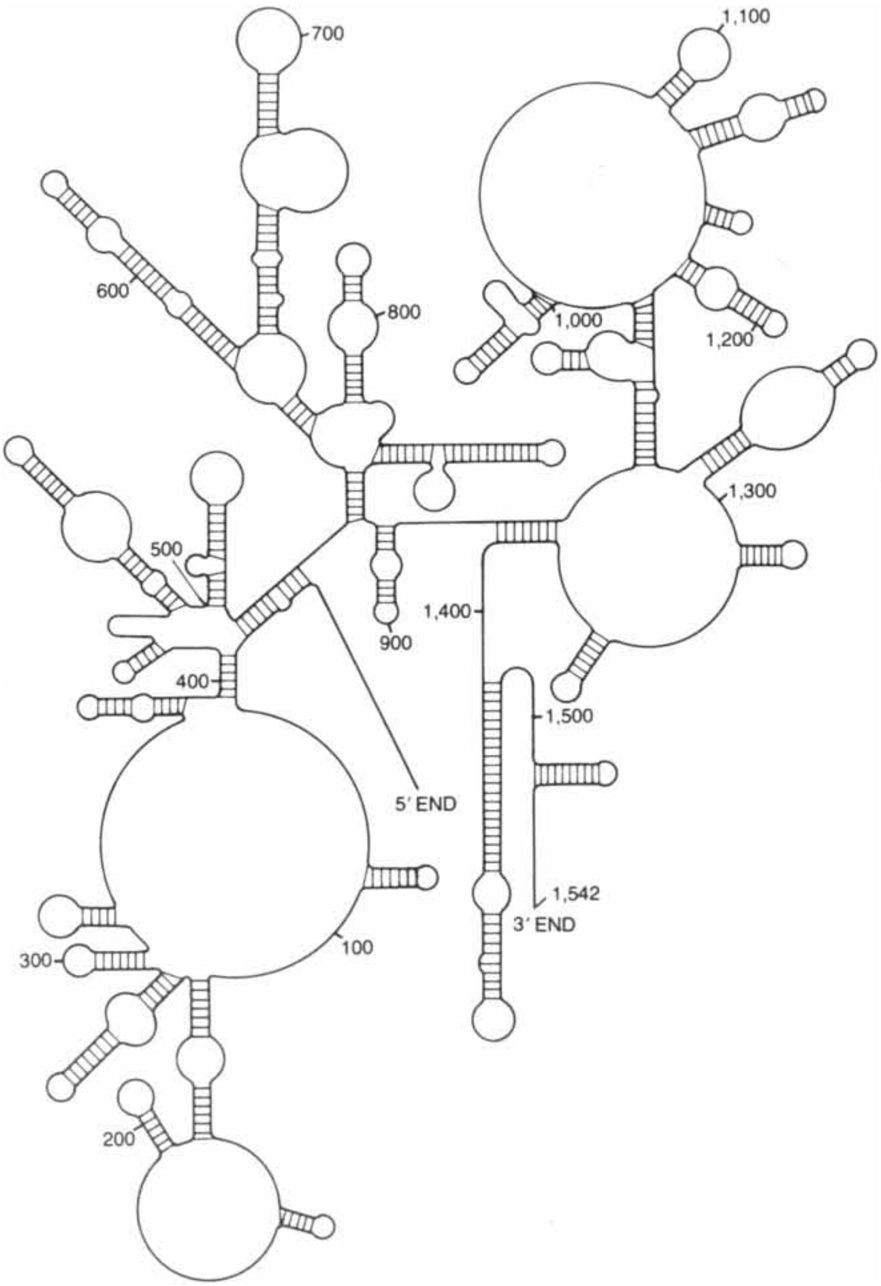
This view is satisfying in some respects, but it fails to explain the many differences between eukaryotes and prokaryotes. In particular it does not account for the different details of molecular processes or for the large differences in the amino acid sequences of functionally analogous proteins in the two kinds of cells. It is often taken for granted that the differences are merely a consequence of the many small changes in cellular design that would be necessary in passing from the simple prokaryotic condition to the more complex eukaryotic one. It is questionable that so many changes (changes in the composition of almost all enzymes, for example) can reasonably be accounted for in this way.

Essentially for this reason some biologists think the line of descent that gave rise to the putative urkaryotic species may have diverged from the prokaryotic line at some earlier point, before the ancestor of the bacteria had itself arisen. The urkaryote could then have evolved independently to a form comparable in complexity to that of the bacteria. Such an assumption would at least provide more time for differences to emerge between prokaryotes and the urkaryote. The urkaryote, then, would represent a line of descent distinct from that of the prokaryotes, in accordance with the basic phylogenetic dichotomy.

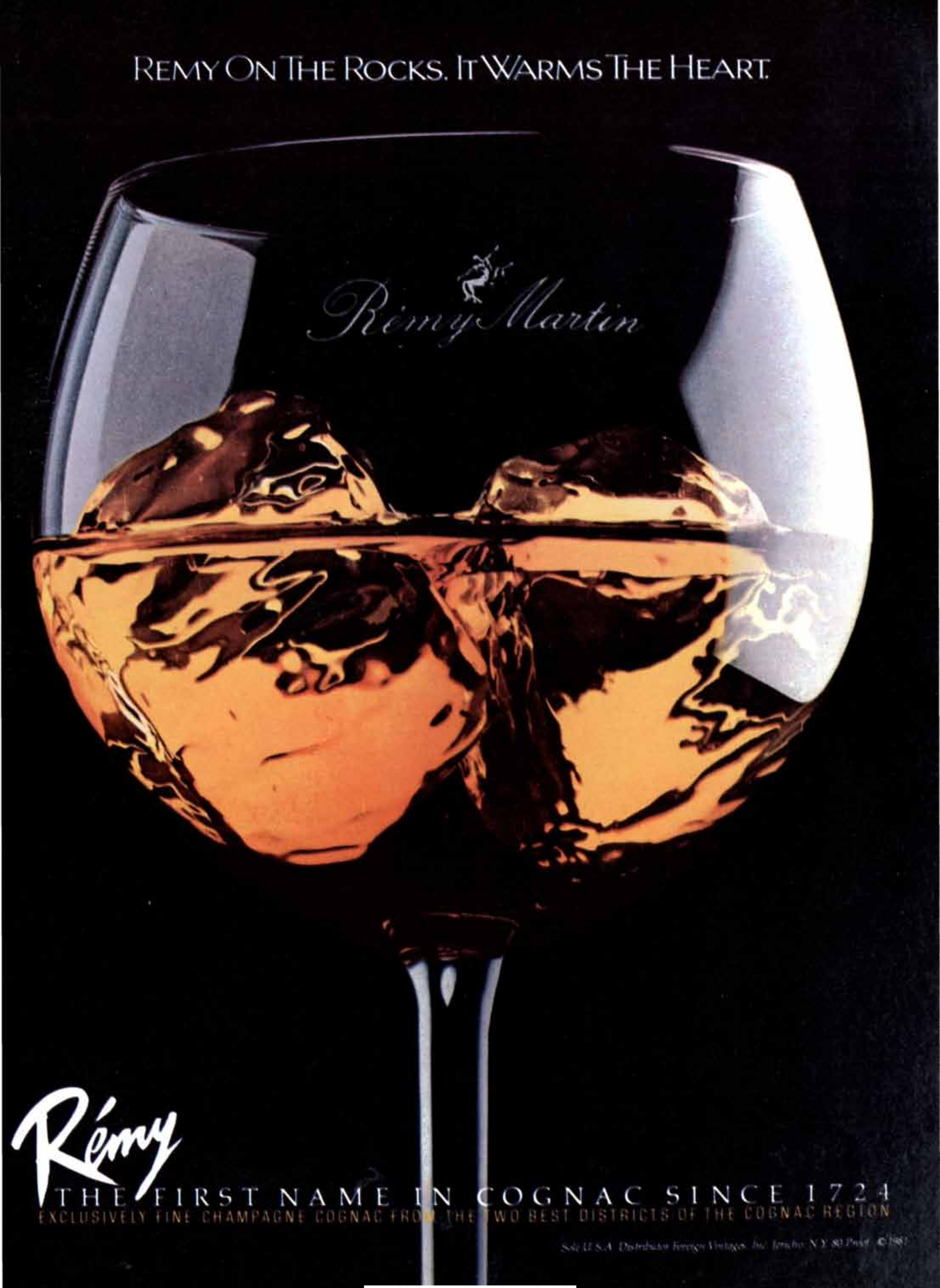
So it stood at the beginning of the 1970's. The phylogeny of the higher eukaryotes, spanning some 500 million years, was reasonably well understood except for the all-important joining of the main eukaryotic branches. There was a definite, widely accepted hypothesis concerning the way in which the eukaryotic cell had evolved. Tests of the hypothesis and answers to the remaining questions, however, lay in the unexplored recesses of bacterial phylogeny, in the age of microorganisms.

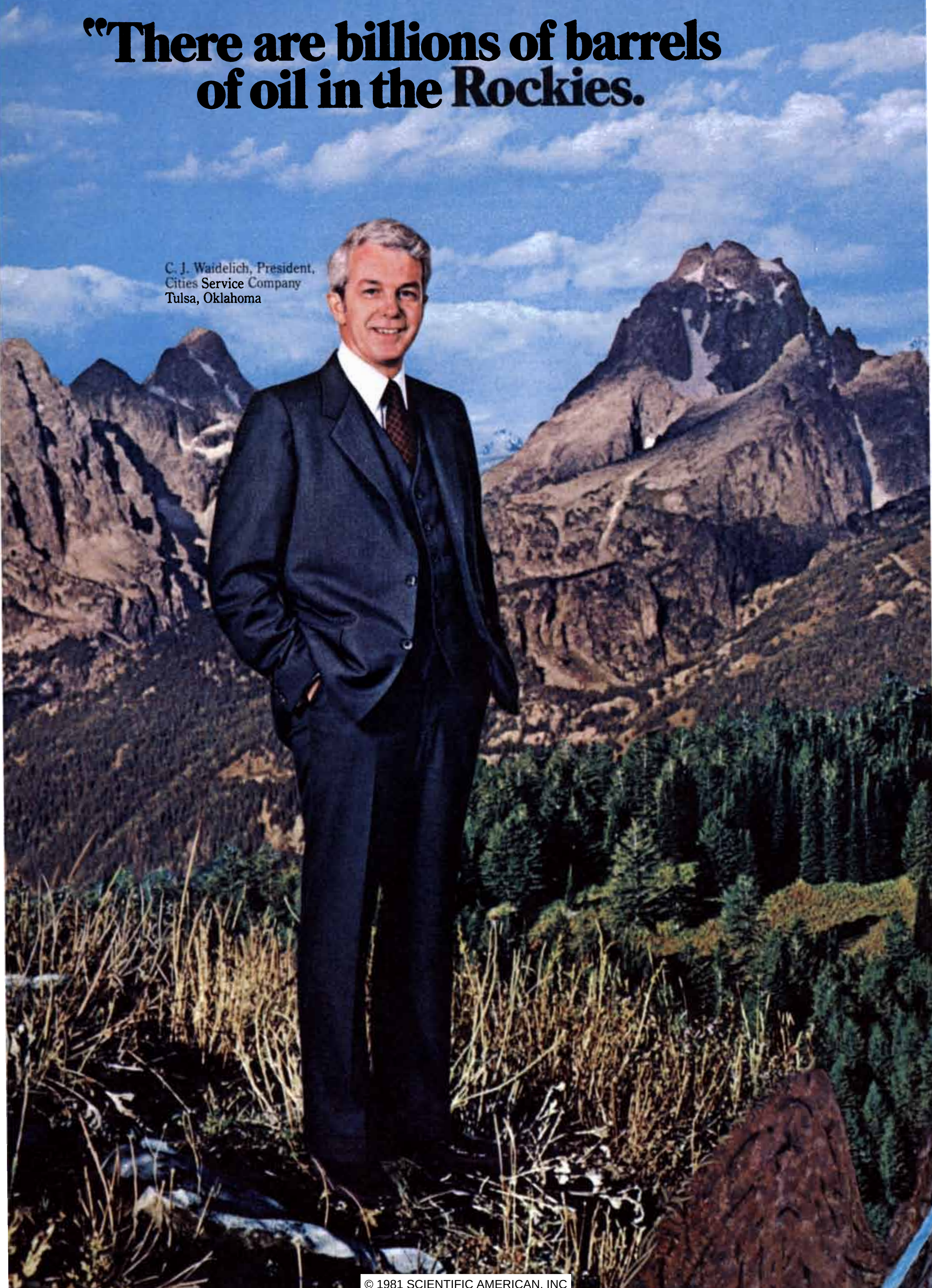
## Genetic Sequencing

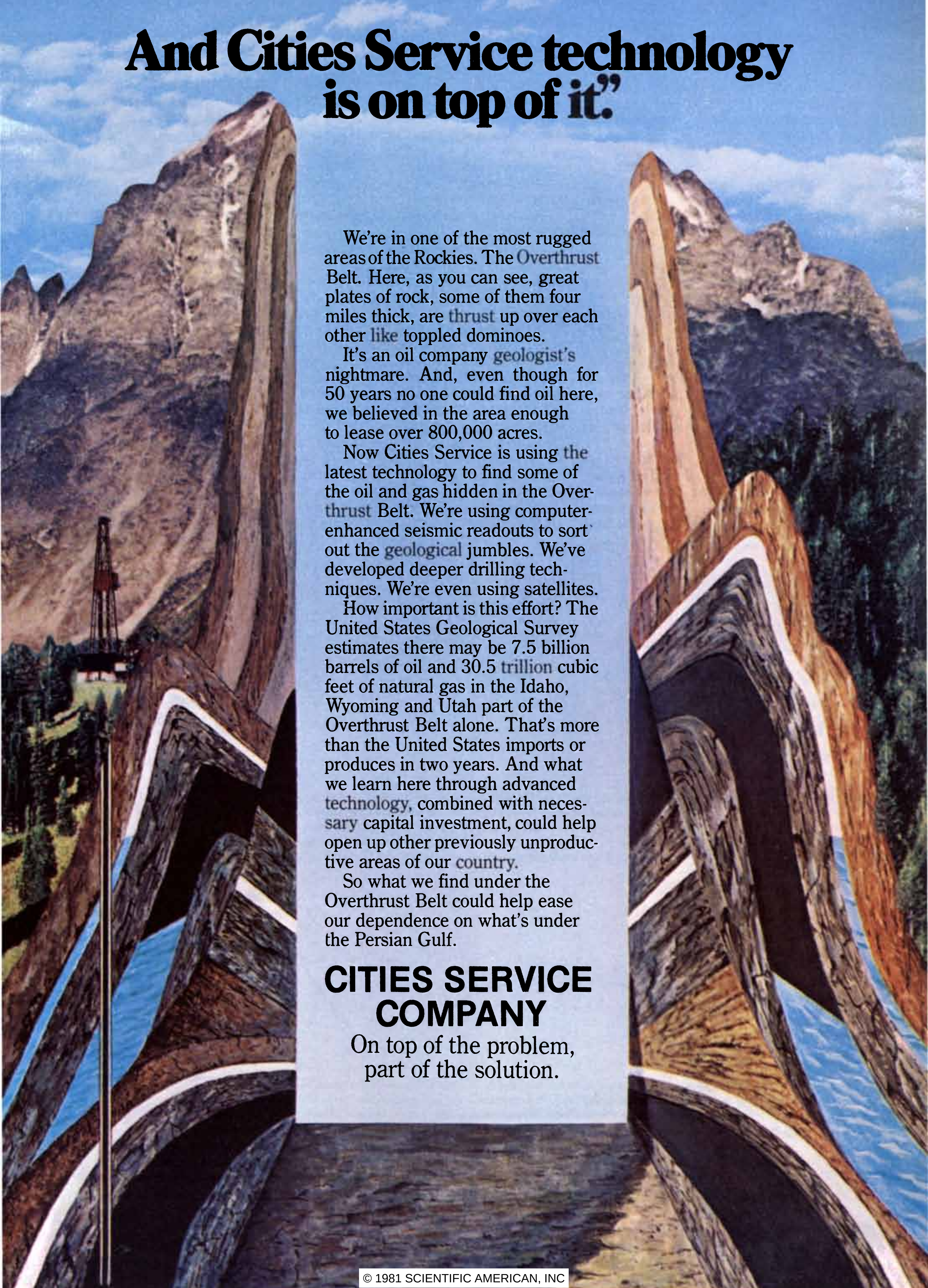
Bacteria constitute a world of extraordinary variety, far more than the microscope reveals. The ecological niches in




**16*S* RIBOSOMAL RNA is the molecule whose nucleotide sequences in a number of organisms have been compared in order to establish phylogenetic relations. The molecule is a component of the ribosome, the molecular machine that synthesizes proteins; the designation 16*S* refers to the speed with which the molecule sediments in a centrifuge, measured in Svedberg units. The RNA molecule is a long chain of the subunits called nucleotides, each of which is characterized by one of four bases: adenine (*A*), uracil (*U*), guanine (*G*) or cytosine (*C*). The first two bases and the last two are complementary: they can be linked by hydrogen bonds to form pairs, *A* pairing with *U* and *G* pairing with *C*. Base pairing determines what is called the secondary structure of the molecule, or the way in which it initially folds, by forming some 50 short double-strand structures in which the bases are paired (*barred regions*). The drawing shows secondary structure of the 16*S* RNA of the eubacterium *Escherichia coli*, full sequence of which was determined by Harry F. Noller, Jr., of the University of California at Santa Cruz.**


REMY ON THE ROCKS. IT WARMS THE HEART.
Rémy Martin
Rémy
THE FIRST NAME IN COGNAC SINCE 1724
EXCLUSIVELY FINE CHAMPAGNE COGNAC FROM THE TWO BEST DISTRICTS OF THE COGNAC REGION


"There are billions of barrels of oil in the Rockies.
C. J. Waidelich, President, Cities Service Company Tulsa, Oklahoma


And Cities Service technology is on top of it."
We're in one of the most rugged areas of the Rockies. The Overthrust Belt. Here, as you can see, great plates of rock, some of them four miles thick, are thrust up over each other like toppled dominoes.
It's an oil company geologist's nightmare. And, even though for 50 years no one could find oil here, we believed in the area enough to lease over 800,000 acres.
Now Cities Service is using the latest technology to find some of the oil and gas hidden in the Overthrust Belt. We're using computer-enhanced seismic readouts to sort out the geological jumbles. We've developed deeper drilling techniques. We're even using satellites.
How important is this effort? The United States Geological Survey estimates there may be 7.5 billion barrels of oil and 30.5 trillion cubic feet of natural gas in the Idaho, Wyoming and Utah part of the Overthrust Belt alone. That's more than the United States imports or produces in two years. And what we learn here through advanced technology, combined with necessary capital investment, could help open up other previously unproductive areas of our country.
So what we find under the Overthrust Belt could help ease our dependence on what's under the Persian Gulf.
CITIES SERVICE COMPANY
On top of the problem, part of the solution.

which they are found far exceed in variety those occupied by the higher forms of life. For a century microbiologists have tried in vain to understand the natural relations among bacteria and to impose some order on the bewildering array of forms, physiologies and ecologies. Variety among bacteria is mostly variety within simplicity, and so it provides little information about phylogenetic relations. In higher organisms the eye, for example, has evolved a number of times, but the eye is complicated enough for the independently evolved examples to be readily distinguishable from one another. Such is generally not the case for the form and structure of bacteria; rods, spheres and spirals, which are the typical bacterial shapes, are easily arrived at and have evolved many times. The same principle applies to bacterial biochemistry. Although some bacterial characteristics are valid phylogenetic indicators, it is impossible to tell in advance which ones are and which are not.

The simplest way in which the cell is a record of its past is in terms of genetic sequences. Every gene that exists in a cell today is a copy of a gene that existed eons ago. It is not an exact copy because mutations have altered the original genetic sequence, but vestiges of the original state often persist. What makes a gene a superb record of the past is its simplicity (it is a linear array) and the fact that genetic-sequence "space" is enormous, so that over the entire span of evolution only a small fraction of the possible genetic sequences can ever be realized. Hence if two genes are similar over a stretch comprising a significant number of nucleotides, this can only mean they have an ancestor in common; such genetically related molecules are said to be homologous.

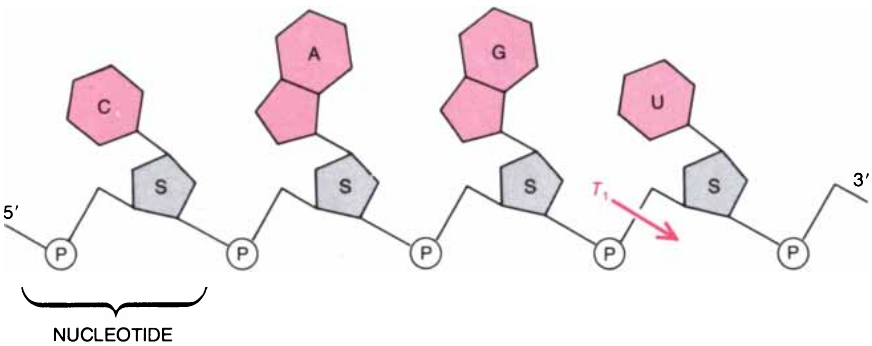
A genetic sequence yields three kinds of evolutionary information. The sequence can reveal genealogical relations, it can measure evolutionary time and it is a record of ancestral characteristics. To the extent that two genes for the same function in different organisms are related, the organisms are related. The extent to which two such sequences differ measures the time since the organisms diverged from a common ancestor. From an extensive set of related sequences one can construct a phylogenetic tree in which the branch points measure (approximately) the relative times of the bifurcations. Finally, comparisons among an extensive set of homologous sequences make it possible to reconstruct with some accuracy various ancestral versions of a gene.

Since the relation between a gene and its product (either a protein or one of several kinds of RNA molecule) is generally a colinear one, the sequence of the product is ordinarily as useful for phylogenetic studies as the sequence of the gene itself. Because until recently only proteins could be sequenced it was through comparisons of proteins that the first phylogenies based on molecular evolution were constructed. Comparisons of the respiratory protein cytochrome *c* proved to be particularly valuable for confirming and extending the phylogenetic tree of the higher organisms. On the other hand, molecules such as cytochrome *c* are not as effective in establishing relations among bacteria. Such proteins are not universally distributed; they are not strictly constant in function and so are not entirely comparable, and because of the greater antiquity of bacterial lineages differences in sequence can be far greater among bacteria than they are among eukaryotes. These factors make bacterial phylogenies deduced from protein evolution incomplete and uncertain.

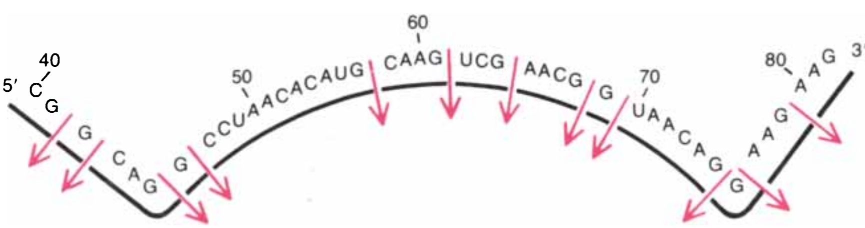
## Ribosomal RNA

There are other gene products that can serve as indicators of bacterial relations. All self-replicating entities necessarily have systems for maintaining and propagating genetic information and for translating it into the chains of amino acids that constitute proteins. Most of the large molecules engaged in these processes must trace their origin back to the very early stages in the evolution of the cell; they certainly emerged before cells became complex enough to be called prokaryotes. Therefore one would expect these molecules to have the requisite properties of a phylogenetic marker.

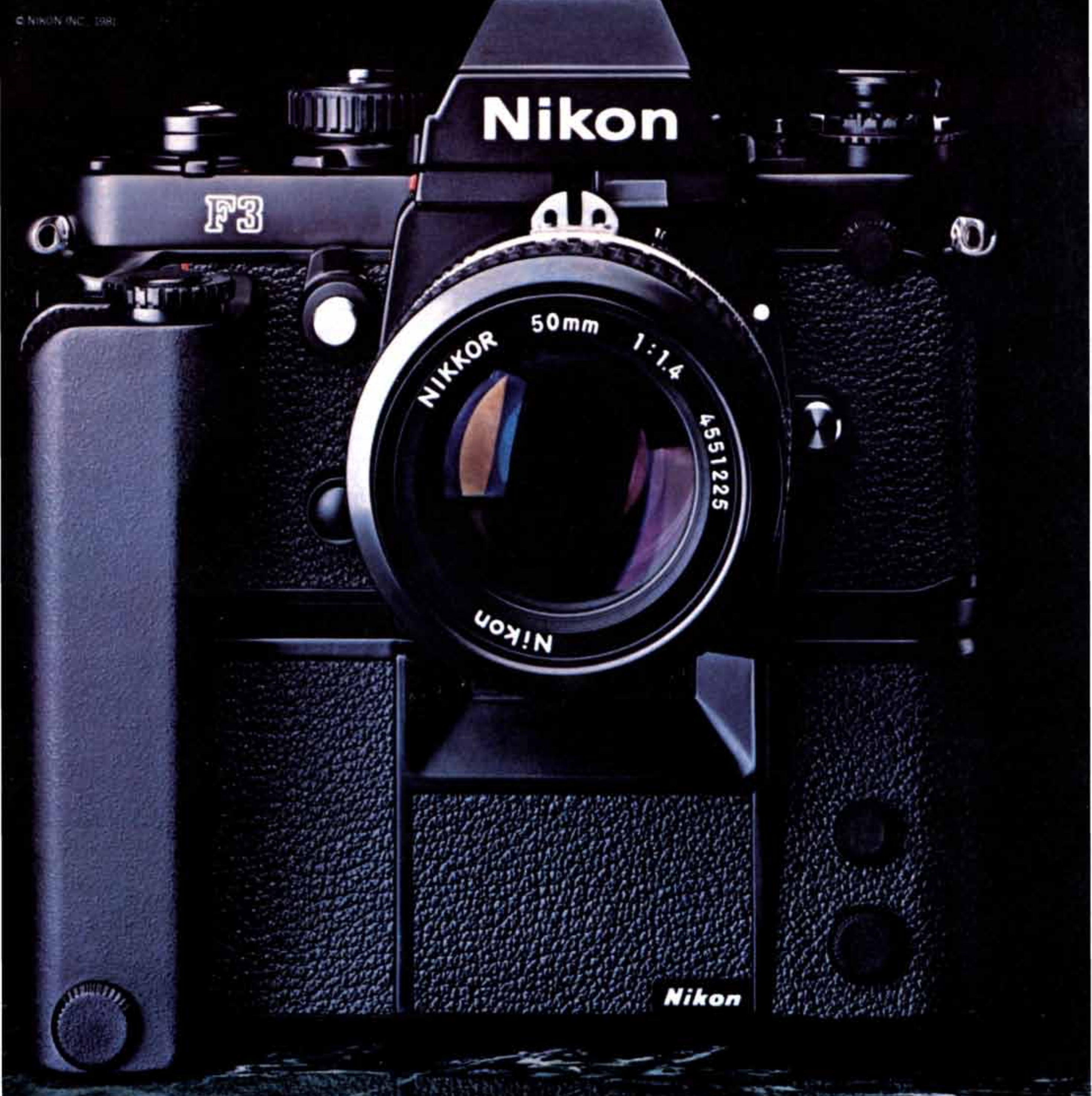
The most reasonable first choices among such molecules are the RNA molecules that are complexed with proteins to form the ribosomes. It is on the ribosomes that genetic information is translated into proteins. The ribosomal RNA is easy to isolate in workable quantities because a typical bacterial cell has from 10,000 to 20,000 ribosomes. Moreover, ribosomal-RNA molecules seem to have remained constant in function over great evolutionary distances. This is important because functional changes in a molecule bring with them additional changes in sequence that make it difficult or even impossible to compare one molecular sequence with another and thereby deduce phylogenetic relations. Still another advantage of the ribosomal RNA's is that at




**EACH NUCLEOTIDE in an RNA molecule is composed of a base, a ribose sugar (*S*) and a phosphate group (*P*). The enzymes called ribonucleases cleave the chain at specific sites. $T_1$ ribonuclease cuts it by hydrolysis (insertion of a water molecule) on the 3′ side of the phosphate that follows any guanine nucleotide. It therefore cuts a long RNA molecule into a number of short fragments, each consisting of one nucleotide or more and ending with guanine (*G*).**




**EFFECT OF $T_1$ RIBONUCLEASE on a short stretch of the *E. coli* 16*S* molecule is demonstrated. When a typical 16*S* RNA is cleaved in this way, its sequence is cut into fragments ("words") ranging in length from one nucleotide ("letter") to 20 nucleotides. The base sequence of each such word is determined. Words of six letters or more are compiled into a dictionary. Dictionaries of two organisms can be compared in terms of an association coefficient $S_{AB}$. The coefficient is a fraction equal to twice the number of letters in words (at least six letters long) common to organisms *A* and *B*, divided by the total number of letters in all such words in *A* and *B*.**




# How the man who has everything takes pictures of it all.

One of the great joys of owning beautiful things is being able to take beautiful pictures of them.

The Nikon F3 offers you limitless possibilities for doing just that. It's the latest in the line of cameras that made Nikon the choice of more professional photographers than all other 35mm cameras combined. (As you might expect with a heritage like this, the F3 is in great demand. So some dealers may be out of stock. Rest assured more are on the way.)

You'll find the Nikon F3 is the ultimate means of expression for the man of means.

**Nikon**
We take the world's greatest pictures.

least some portions of their sequences change slowly enough for the common ancestral sequence not to be totally obliterated. In other words, the sequences make it possible to detect the deepest phylogenetic relations.

There are three kinds of ribosomal-RNA molecules. In bacteria the "large" ribosomal RNA is the 23*S* RNA (*S* stands for Svedberg unit, a measure of the rate of sedimentation in an ultracentrifuge and hence an indirect measure of molecular size); it is approximately 2,900 nucleotides long. The "small" one, designated 16*S* ribosomal RNA, is about 1,540 nucleotides long. A very small one (5*S*) has only 120 nucleotides. The sizes are similar in eukaryotic cells: 18*S*, 25–28*S* and 5*S*. One might think that ease of characterization would make the small 5*S* RNA the most suitable one for phylogenetic studies. Actually it is not as accurate an indicator of phylogenetic relations as the larger ribosomal RNA's, chiefly for statistical reasons. (The 5*S* RNA sometimes exhibits anomalous large differences in sequence from one species to another.) The 16*S* ribosomal RNA is the molecule of choice, because the 23*S* molecule is almost twice as large and more than twice as difficult to characterize.

## RNA Dictionaries

At the University of Illinois in 1969 I decided to explore bacterial relations by comparing the sequences of the 16*S* ribosomal RNA's in different species. It was not yet feasible (as it is now) to determine the nucleotide sequence of the entire molecule. The technology did exist, however, for sequencing short segments of the molecule. The enzymes called ribonucleases yield short fragments of RNA by cutting an RNA strand at specific sites. Each nucleotide of RNA is composed of a sugar called ribose, a phosphate group and one of four nitrogenous bases: adenine (*A*), uracil (*U*), guanine (*G*) or cytosine (*C*). The enzyme $T_1$ ribonuclease cuts an RNA strand at a particular bond on one side of each nucleotide that incorporates a guanine base. The $T_1$ enzyme therefore digests an RNA "text" into short "words," called oligonucleotides. Each oligonucleotide includes, and ends with, a single *G*, as in *AACUCG* or *UCCUAUCG*.

The oligonucleotides made in this way were short enough to be sequenced by the available techniques. The smallest words are of little value because they recur many times in each molecule. By the time the word length reaches six nucleotides, however, a particular sequence is unlikely to appear more than once in a 16*S* RNA molecule. (Given the constant terminal *G*, there are $3^5$, or 243, possible six-letter sequences of this kind, and a typical 16*S* RNA molecule has roughly 25 such words.) When 16*S* RNA's from different organisms include the same six-letter sequence, it almost always reflects a true homology. By confining attention to words six letters long or longer one can generate a "dictionary" characteristic of a given organism, which can readily be compared with other such dictionaries to determine genealogical relations.

A simple way to analyze the data is in terms of an association coefficient $S_{AB}$, which is defined as twice the number of nucleotides in the words common to both of the dictionaries *A* and *B* divided by the number of nucleotides in all words in the two dictionaries. $S_{AB}$ ranges from 1 when dictionaries *A* and *B* are identical to less than .1 when they are unrelated. (The coefficient is usually greater than zero even for unrelated sequences because of chance correspondences.) By compiling the $S_{AB}$ values for a number of organisms in a matrix one can discern a pattern of relatedness or unrelatedness among organisms. Moreover, it is possible by straightforward statistical methods to construct from a set of $S_{AB}$ values for a group of organisms a dendrogram, or tree, showing the relations among members of the group.

To date the ribosomal RNA's of almost 200 species of bacteria and eukaryotes have been characterized. Most of the bacteria form a coherent but very large (which is to say ancient) group. They are the eubacteria, or true bacteria, and as would be expected they are quite distinct from the eukaryotes. The relations among the various genera (represented by the branchings of the tree) determined through ribosomal-RNA analysis are at variance with many of the established prejudices about bacterial relations. What is important at this point is that the eubacteria are divided into a number of major branches and

| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | *Saccharomyces cerevisiae* | — | .29 | .33 | .05 | .06 | .08 | .09 | .11 | .08 | .11 | .11 | .08 | .08 | .10 | .07 | .08 |
| 2 | *Lemna minor* | .29 | — | .36 | .10 | .05 | .06 | .10 | .09 | .11 | .10 | .10 | .13 | .07 | .09 | .07 | .09 |
| 3 | *L* cell | .33 | .36 | — | .06 | .06 | .07 | .07 | .09 | .06 | .10 | .10 | .09 | .07 | .11 | .06 | .07 |
| 4 | *Escherichia coli* | .05 | .10 | .06 | — | .24 | .25 | .28 | .26 | .21 | .11 | .12 | .07 | .12 | .07 | .07 | .09 |
| 5 | *Chlorobium vibrioforme* | .06 | .05 | .06 | .24 | — | .22 | .22 | .20 | .19 | .06 | .07 | .06 | .09 | .07 | .05 | .07 |
| 6 | *Bacillus firmus* | .08 | .06 | .07 | .25 | .22 | — | .34 | .26 | .20 | .11 | .13 | .06 | .12 | .10 | .07 | .09 |
| 7 | *Corynebacterium diphtheriae* | .09 | .10 | .07 | .28 | .22 | .34 | — | .23 | .21 | .12 | .12 | .09 | .10 | .10 | .06 | .09 |
| 8 | *Aphanocapsa* | .11 | .09 | .09 | .26 | .20 | .26 | .23 | — | .31 | .11 | .11 | .10 | .10 | .13 | .10 | .10 |
| 9 | Chloroplast (*Lemna*) | .08 | .11 | .06 | .21 | .19 | .20 | .21 | .31 | — | .14 | .12 | .10 | .12 | .12 | .06 | .07 |
| 10 | *Methanobacterium thermoautotrophicum* | .11 | .10 | .10 | .11 | .06 | .11 | .12 | .11 | .14 | — | .51 | .25 | .30 | .34 | .17 | .19 |
| 11 | *Methanobrevibacter ruminantium* | .11 | .10 | .10 | .12 | .07 | .13 | .12 | .11 | .12 | .51 | — | .25 | .24 | .31 | .15 | .20 |
| 12 | *Methanogenium cariaci* | .08 | .13 | .09 | .07 | .06 | .06 | .09 | .10 | .10 | .25 | .25 | — | .32 | .29 | .13 | .21 |
| 13 | *Methanosarcina barkeri* | .08 | .07 | .07 | .12 | .09 | .12 | .10 | .10 | .12 | .30 | .24 | .32 | — | .28 | .16 | .23 |
| 14 | *Halobacterium halobium* | .10 | .09 | .11 | .07 | .07 | .10 | .10 | .13 | .12 | .34 | .31 | .29 | .28 | — | .19 | .23 |
| 15 | *Sulfolobus acidocaldarius* | .07 | .07 | .06 | .07 | .05 | .07 | .06 | .10 | .06 | .17 | .15 | .13 | .16 | .19 | — | .13 |
| 16 | *Thermoplasma acidophilum* | .08 | .09 | .07 | .09 | .07 | .09 | .09 | .10 | .07 | .19 | .20 | .21 | .23 | .23 | .13 | — |

**MATRIX OF ASSOCIATION COEFFICIENTS reveals the degree of relatedness of any two organisms; the higher the $S_{AB}$ fraction, the closer the relation. The pattern is significant. The eukaryotes (*1–3*), the eubacteria (*4–9*) and the archaebacteria (*10–16*) each form a distinct group (*color*). The archaebacteria are no more closely related to the eubacteria than to the eukaryotes. *Saccharomyces* is yeast; *Lemna* is duckweed; *L* cells are a line of mouse cells. Chloroplast is descended from endosymbiotic cyanobacterium and is therefore eubacterial.**

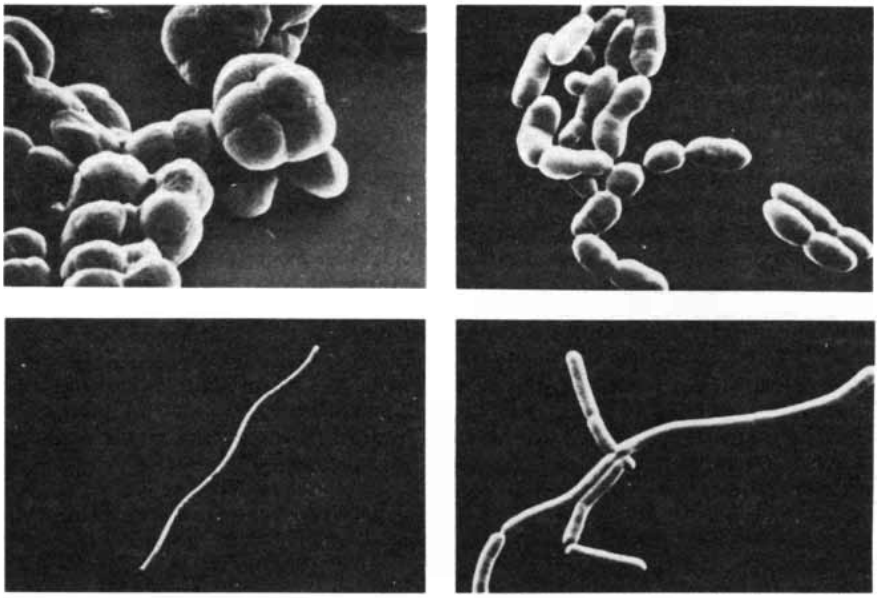
# The diamond solitaire.

# A rare gift.

One single diamond.
Set simply and elegantly, to sparkle on its own.
The diamond solitaire.
A jewel that becomes more precious with
every passing year.
The gift that makes a rare and beautiful
moment last a lifetime.

A diamond is forever.

The 1¼ carat diamond shown is enlarged for detail. DeBeers.



**METHANOGENS, anaerobic bacteria that generate methane ($CH_4$) from hydrogen and carbon dioxide, make up the largest group of archaebacteria identified so far. Four genera of methanogens that differ widely in size and morphology are seen here in scanning electron micrographs made by Alexander J. B. Zehnder of the Swiss Federal Institute of Technology. They are *Methanosarcina* (*top left*), *Methanobrevibacter* (*top right*), *Methanospirillum* (*bottom left*) and *Methanobacterium* (*bottom right*). The cells are shown enlarged respectively 2,500, 5,000, 1,000 and 5,000 diameters. The methanogens are found only in oxygen-free environments.**

that several of the branches include photosynthetic bacteria. This finding suggests all eubacteria stem from a common photosynthetic ancestor.

## The Discovery of Archaebacteria

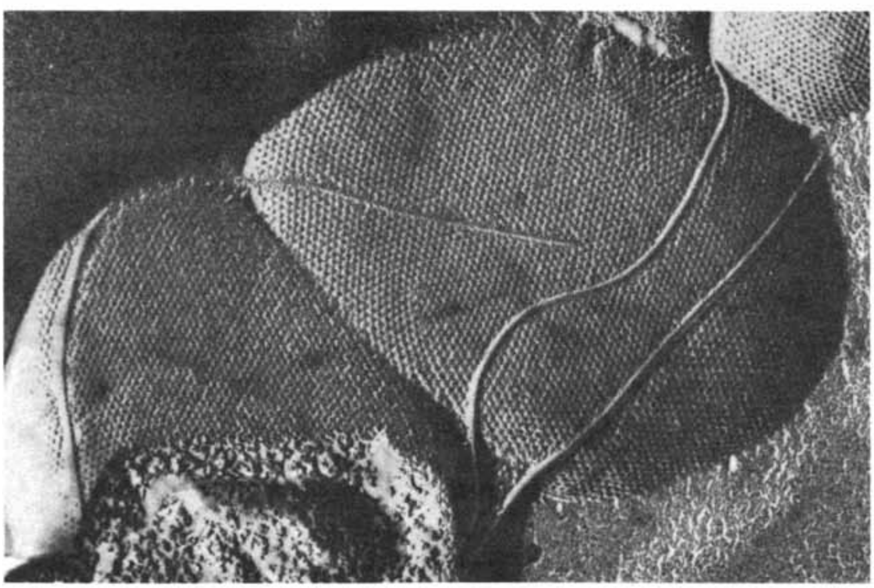
As the screening of bacteria continued a surprise emerged. In collaboration with Ralph S. Wolfe I looked at the ribosomal RNA of the methanogenic bacteria. These unusual organisms live only in oxygen-free environments and generate methane ($CH_4$) by the reduction of carbon dioxide ($CO_2$). We discovered that methanogens do not fall within the phylogenetic group defined by the other bacteria. Indeed, they appear to represent an evolutionary branching that far antedated the common ancestor of all true bacteria. Not only were the methanogens separate but also the group they formed seemed to be about as deep phylogenetically—as ancient—as the group defined by the eubacteria.

There can be no doubt that the methanogens and their relatives are bacteria. They are the size of bacteria, they have no nuclear membrane, they have a low DNA content and so on. Surely, then, one would have expected them to be related more closely to other bacteria than to eukaryotes. Our analysis showed they are not. Methanogens are related as closely to eukaryotes as to eubacteria.

How could this be? There were supposed to be only two primary lines of descent, the eukaryotic and the prokaryotic. Here was a new group of organisms: the methanogens and their relatives, which together have come to be called archaebacteria. They were obviously like other bacteria in their superficial characteristics, and so they had been assumed to be in the prokaryotic line of descent. It is not striking differences in morphological characteristics, however, that distinguish the prokaryote phylogenetically from the eukaryotic cell; it is the subtler and more ancient differences in molecular sequences and in details of function at the molecular level that distinguish them. Hence there is no reason two prokaryotic lines of descent cannot be just as distinct from each other as either one is from the eukaryotic line.

This idea was too novel to be easily accepted, and initially some biologists rejected out of hand the notion of a "third form of life." How could something that looked like a bacterium not be a bacterium and indeed not be related to bacteria? In time the simplicity of our argument and the accumulation of evidence prevailed. Although a few biologists still dispute our interpretation, the idea that archaebacteria represent a separate grouping at the highest level is becoming generally accepted.

The supposed great antiquity of the archaebacteria remains an unproved prejudice, but it is a plausible one. The methanogenic phenotype seems to cover a phylogenetic span as great as or greater than the span covered by any other comparable bacterial phenotype. This implies that the methanogens are as old as or older than any other bacterial group. Moreover, methanogenic metabolism (the reduction of carbon dioxide to methane) is ideally suited to the kind of atmosphere thought to have existed on the primitive earth: one that was rich in carbon dioxide and included some hydrogen but virtually no oxygen. The name archaebacteria implies that these organisms were the dominant ones in the primeval biosphere. When conditions changed, the methanogens' need



**UNUSUAL CELL WALL of *Methanogenium marisnigri*, a methanogen found on the floor of the Black Sea, is enlarged 70,000 diameters in a scanning electron micrograph made by Frank Mayer of the University of Göttingen. The mosaic pattern of proteinaceous subunits is characteristic of several archaebacteria. It is different from the typical eubacterial cell wall made up of peptidoglycan subunits, which are not components of cell wall of archaebacteria.**

Once every generation it seems, something comes along that changes the entire nature of home entertainment. That time has come again.

Pioneer introduces LaserDisc,™ a whole new technology that makes your television and stereo do things they've never done before. Now you can watch a movie or hear and *see* a concert as easily as playing a record.

And when you hook up the LaserDisc player to your TV and hi-fi, for the first time, along with a television picture you get true stereo.

And a television picture that's actually a 40% sharper picture than home videotape.

And because the disc is read by a light beam rather than a needle or video head, nothing touches the disc. So, with normal use, it doesn't wear out from play. You can enjoy the disc forever.

But Pioneer LaserDisc offers you more than superb fidelity, it's truly the first form of personal entertainment. Your television viewing is no longer limited to what's on television. Because you can see what you want to see when you want to see it. A whole library of Pioneer Artists and MCA/Discovision discs with movies like Jaws, Animal House, The Blues Brothers; concerts like Paul Simon, Liza Minelli, Loretta Lynn, and Abba; sports like NFL football, and title fights.

You can study the standard-play discs in fast motion, slow motion, stop motion. Even create your own instant replay.

But there's something even more remarkable. A micro-computer built into the Pioneer LaserDisc player lets you access any individual frame on the disc at the touch of a button. You can go right to your favorite scene in a movie, or a song in concert in seconds.

And because you can view the up to 108,000 frames on the disc one frame at a time, it just may be the future of education.

The cost of Pioneer LaserDisc is surprising as well. Just $749* (with $50* more for optional remote control). And a full-length movie on LaserDisc can cost less than taking your family out to the movies.

The only way to even begin to experience the magic of this remarkable LaserDisc system is to see it. And we've arranged for that. Just call at **800-621-5199**** for the store nearest you. **(In Illinois 800-972-5855).

*Suggested retail price. Actual price set by dealers.

# Finally, high-fidelity television with stereo sound.

*Simulated TV picture from* Paul Simon.

for an anaerobic environment confined them to a limited range of relatively inaccessible niches.

The measurements that revealed the existence of the archaebacteria (differences in RNA sequences) were genetic ones and were purely quantitative. They revealed nothing about the quality of the differences—the phenotypic differences—between the archaebacteria and the true bacteria. If our interpretation of the archaebacteria as a primary kingdom separate from that of the true bacteria is correct, then on detailed inspection the archaebacteria should prove to be as different from true bacteria in their molecular phenotype as either group is from eukaryotic cells.

## Archaebacterial Forms

The archaebacteria are indeed unusual organisms. The group is now known to include three very different kinds of bacteria: methanogens, extreme halophiles and thermoacidophiles.

The dominant form (in the sense that it constitutes a deep phylogenetic grouping) is the methanogen. Bacteria that give off methane have been known for some time. Alessandro Volta discovered in 1776 that "combustible air" is generated in bogs, streams and lakes whose sediments are rich in decaying vegetation, but the fact that a microorganism is responsible for generating "marsh gas" became known only much later. Methanogens are widely distributed in nature, but they are not commonly encountered because they are killed by oxygen and do not exist in the open.

In ancient times methanogens could have existed almost anywhere. Today they live only where oxygen has been excluded and where hydrogen and carbon dioxide are available. This generally means living in close association with other bacteria, such as the clostridia, that metabolize decaying organic matter and give off hydrogen as a waste product. Methanogens are found in stagnant water and in sewage-treatment plants (in amounts that have made it commercially feasible to manufacture methane). They are also found in the rumen of cattle and other ruminants and in the intestinal tract of animals in general. Methanogens can be isolated from the ocean bottom and from hot springs. In spite of their intolerance of oxygen they are obviously globally distributed.

The extreme halophiles are bacteria that require high concentrations of salt in order to survive; some of them grow readily in saturated brine. They can give a red color to salt evaporation ponds and can discolor and spoil salted fish. The extreme halophiles grow in salty habitats along the ocean borders and in inland waters such as the Great Salt Lake and the Dead Sea. Although the extreme halophiles have been studied by microbiologists for a long time, they have recently become particularly interesting for two reasons. They maintain large gradients in the concentration of certain ions across their cell membrane and exploit the gradients to move a variety of substances into and out of the cell. In addition the extreme halophiles have a comparatively simple photosynthetic mechanism based not on chlorophyll but on a membrane-bound pigment, bacterial rhodopsin, that is remarkably like one of the visual pigments.

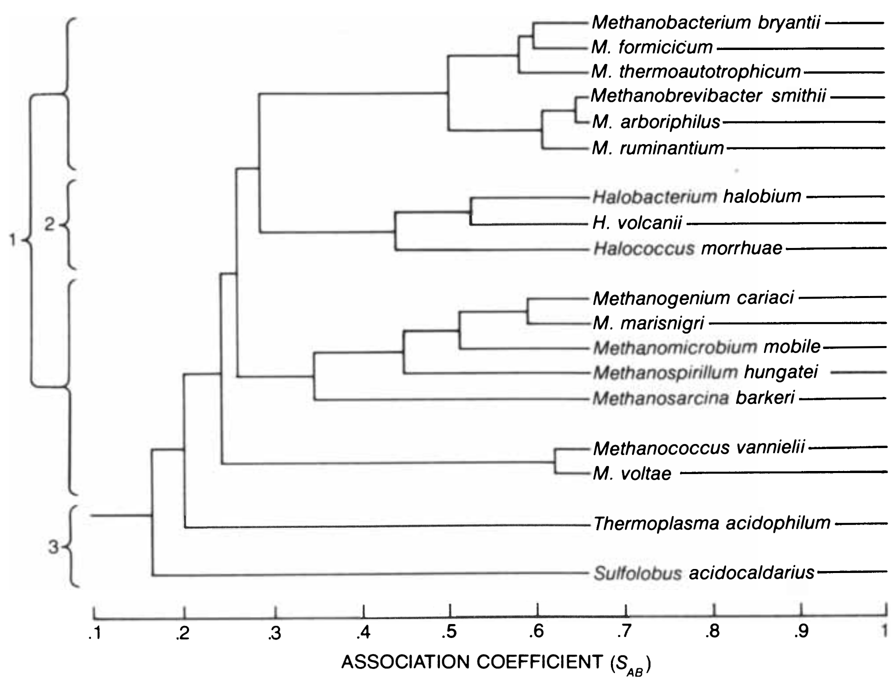
## The Thermoacidophiles

The third known type of archaebacterium is the thermoacidophile, and the members of this group too are notable for their habitat. *Sulfolobus,* one of the two genera of thermoacidophiles, is found in hot sulfur springs. Its various species generally grow at temperatures near 80 degrees Celsius (176 degrees Fahrenheit); growth at temperatures above 90 degrees has been observed for some varieties. Moreover, the springs in which *Sulfolobus* flourishes are extremely acidic; *p*H values lower than 2 are common (a *p*H of 7 is neutral). *Thermoplasma,* the other genus of thermoacidophile, has so far been found only in smoldering piles of coal tailings. It is a mycoplasma: it has no cell wall but merely the limiting cell membrane.

Although archaebacterial thermoacidophiles can grow only in an acidic environment, the internal milieu of the cell has a quite moderate *p*H, near neutrality; this requires that a sizable *p*H gradient be maintained across the cell membrane. As in the extreme halophiles the gradient may play a role in pumping other molecules into and out of the cell. It is interesting that when the temperature is reduced and as a consequence *Sulfolobus* stops metabolizing, the cell's internal *p*H can no longer be maintained near neutrality and the cell dies.

For some time it had been recognized that various organisms now classified as archaebacteria are individually somewhat peculiar. In each instance the idiosyncrasy had been seen as just that: an adaptation to some peculiar niche or a biochemical quirk. The ribosomal-RNA phylogenetic measurement, however, showed at least some of the idiosyncrasies might instead be general characteristics of a new group of organisms. Thus informed, investigators in many countries have undertaken to find the general properties that link archaebacteria to one another and to see how those properties either distinguish the archaebacteria from the other two major forms or relate them specifically to one or the other of those forms.

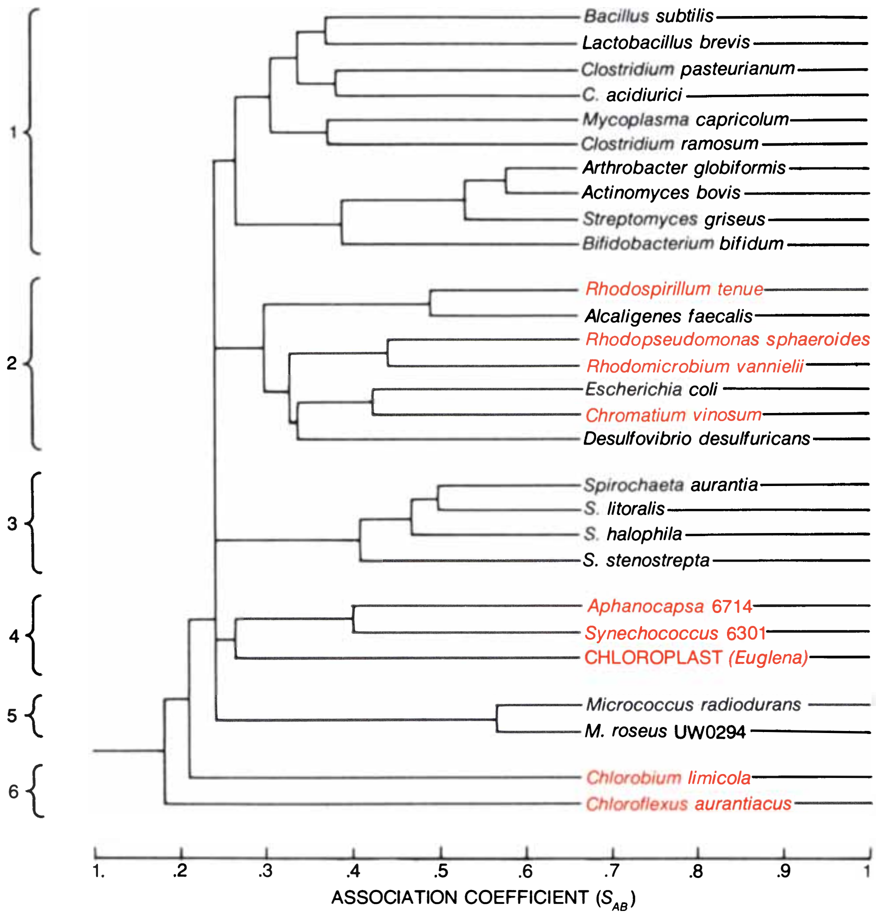
One generalization about bacteria has




**ARCHAEBACTERIAL DENDROGRAM, or tree, is derived from $S_{AB}$ values and shows the phylogenetic relations among members of this primary kingdom. Most of them are methanogens (*1*), which are anaerobes (organisms that survive only in the absence of oxygen) and generate methane by the reduction of carbon dioxide or certain other very simple sources of carbon. The extreme halophiles (*2*) are aerobic (oxygen-respiring) bacteria that exist only in environments with a high salt concentration. The thermoacidophiles (*3*) are aerobic bacteria that live only in very hot, highly acidic environments. The methanogens appear to be an ancient group within which the halophiles arose; the thermoacidophiles may have arisen separately.**


The Ultimate Wagon
Imagine having gas mileage about the same as a full size 2-wheel drive wagon . . . and 4-wheel drive for added security when you need it most. 21 HWY EST. (16) EPA EST MPG.*
That's what makes Jeep Wagoneer the ultimate family wagon. No conventional 2-wheel drive wagon can perform so confidently under pressure when the rains are torrential, the streets unplowed . . . or when the beauty of the off-road beckons.
Wagoneer may take a tough-minded approach to conditions outside . . . but inside it's a soft touch, with room, luxuries and driving comfort you'd marvel at in any station wagon. And with a full range of Wagoneer models to choose from, you'll find the price of owning one surprisingly affordable.
The mileage of 2-wheel drive full size wagons . . . the security of 4-wheel drive. Doesn't your family deserve the ultimate?
Jeep Wagoneer
AT AMERICAN MOTORS
*Figures are for comparison. Your results may differ due to driving speed, weather conditions and trip length. Actual highway mileage will be less. California mileage will be different.
Jeep Corporation, a subsidiary of American Motors Corporation.


Computers for people.™
Gordon Cooper thinks NASA computers are out of this world. But his personal computer is the best on earth.
I've been around computers most of my life, and I'm impressed with the Atari®800™. Nothing so complex has ever been so simple. It's as easy to set up as a stereo. It's so easy to operate, your kids can use it.
My life didn't end when I left NASA. Today, I'm a businessman and writer. I've even gone on the lecture circuit. I trust my Atari computer to help me.
So, if you're looking for a way to stay on top of your affairs, look into the Atari 800.
Ask your local computer dealer for a demonstration.
The ATARI® 800™ Computer
For further information write:
Atari Inc., Computer Division,
1196 Borregas Avenue, Dept. AC-11
Sunnyvale, CA 94086
ATARI®
Computers for people.™
Gordon Cooper
Former Astronaut




**EUBACTERIAL DENDROGRAM shows six major subgroups, three of which include photosynthesizing bacteria (*color*); other groups remain to be defined. The Gram-positive bacteria (*1*) have a thick cell wall with a unique composition that absorbs and retains the Gram stain. The purple photosynthetic bacteria are grouped (*2*) with a number of close relatives that are not photosynthesizers, presumably having lost an ancestral ability. The spirochetes (*3*) are long, spiral bacteria. The cyanobacteria (*4*) are photosynthetic, oxygen-producing organisms; the chloroplast is descended from them. Some spherical bacteria with an atypical cell wall (*5*) are notable for their resistance to radiation. The green photosynthetic bacteria (*6*) are anaerobic.**

been that they have a cell wall incorporating the sugar derivative muramic acid, which is the basis of a complex polymer called a peptidoglycan. One extreme halophile and the thermoacidophile *Sulfolobus* were known to be exceptions to this generalization; they were considered to have an idiosyncratic wall structure. Otto Kandler of the University of Munich, collaborating with Wolfe, made a systematic study of cell-wall structure in other known archaebacteria. All of them turned out to be atypical. The archaebacteria have a variety of wall types, but none of them is of the muramic-acid-based peptidoglycan type.

## Lipids and RNA's

It was also known that the cell membrane of the extreme halophiles and of the thermoacidophiles is composed of unusual lipids. The lipids of both eukaryotes and eubacteria consist in the main of two straight-chain fatty acids bound at one end to a glycerol molecule through an ester linkage (–CO–O–). The lipids of the extreme halophiles and the thermoacidophiles are also composed of a glycerol group linked to two long hydrocarbon chains, but the connection between the glycerol and the chains is an ether (–O–) link rather than an ester link. Moreover, the hydrocarbon chains are not straight but branched: every fourth carbon atom in the chain carries a methyl group ($CH_3$). The basic archaebacterial lipid, in other words, is a diether composed of glycerol and two molecules of an alcohol, phytanol. When a number of methanogens were examined for lipid composition, our expectation was confirmed: their lipids turned out to be typically archaebacterial branched-chain glycerol ethers.

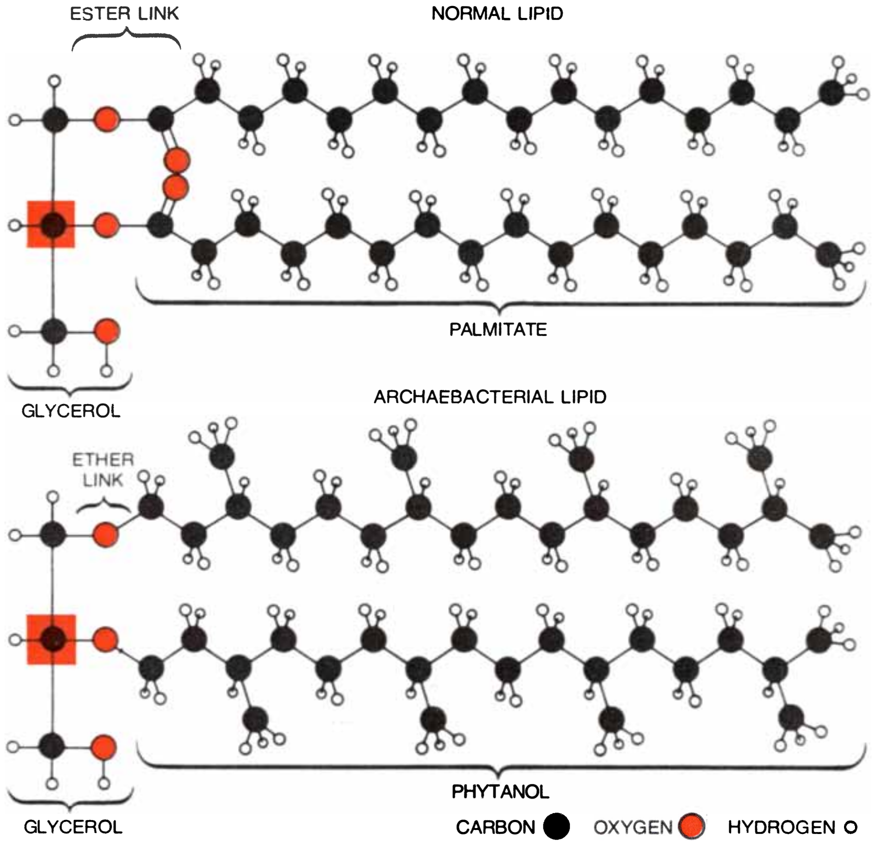
In the course of the ribosomal-RNA studies another unexpected archaebacterial property emerged, one that was to provide the first clue to the significance of the differences between archaebacteria and true bacteria. Central to the process of translation is the transfer-RNA molecule. It recognizes a three-base "codon" in messenger RNA specifying a particular amino acid, and it delivers that amino acid to be incorporated into the protein chain. A number of the nucleotides in a transfer-RNA molecule are modified, that is, their structure is altered chemically after they have been incorporated into the molecule; most often a methyl group is added to the nucleotide at some position on either the base or the sugar. Biologists had come to believe one particular modification was characteristic of a certain position in almost all transfer-RNA molecules in almost all organisms: at that position the base uracil has been methylated to form thymine (which is normally present only in DNA, not in RNA). It turns out that all transfer RNA's of all archaebacteria lack this thymine unit; instead the uracil has been modified in one of two other ways to yield a pseudouridine or an as yet unidentified nucleotide.

If one compares both ribosomal RNA and transfer RNA in eukaryotes, eubacteria and archaebacteria, one finds a general pattern, of which the replacement of thymine in archaebacterial transfer RNA's is only one example. The same regions in the RNA's tend to be modified in all three primary lines of descent, but the nature of the modification tends to vary from one kingdom to another. The differences are of two kinds. Either the modification of a given base is different in each of the kingdoms, or a given base is modified in one kingdom and in another kingdom the modification is made to an adjacent base. These modes of variation suggest that the modifications have evolved separately in each major line of descent.

Several other molecular distinctions between the archaebacteria and the other groups are known (for example, in the subunit structure of the enzyme RNA polymerase), but the list is not long. The reason is not that additional differences do not exist; it is rather that the world of archaebacteria remains virtually unexplored. The study of archaebacterial genetics is in a primitive state; few mutants have even been isolated for genetic study. Nothing whatever is known about the control of gene expression in archaebacteria. The basic molecular biology of archaebacteria is not understood. And yet to the extent that the archaebacteria have been characterized they have been found to differ significantly from both of the other major groups.

## A New Perspective

The discovery of a new primary kingdom of organisms is a major finding in its own right (comparable to going into




**MEMBRANE LIPIDS of archaebacteria are different from the lipids found in other organisms. The lipids of both eubacteria and eukaryotes are glycerol esters of straight-chain fatty acids, that is, they are composed of a three-carbon alcohol, glycerol, attached to fatty-acid chains such as palmitate by an ester link. Archaebacterial lipids, on the other hand, are diethers in which a glycerol unit is connected by an ether link to phytanols: branched chains in which carbon atoms at regular intervals carry a methyl ($CH_3$) group. Moreover, glycerol has two optical isomers, distinguished by the configuration of the molecule about the central carbon atom (*colored box*); the optical isomers rotate polarized light in opposite directions. The configuration about the central carbon atom of glycerol that is found in archaebacterial lipids is the mirror image of the configuration that is found in both eubacterial and eukaryotic lipids.**

the backyard and seeing an organism that is neither a plant nor an animal), but the real importance of the discovery lies in what it may reveal about the early history of life. When there were only two known primary lines of descent, one could not readily interpret the differences between the two. The recognition of three lines of descent equidistant from one another gives a much better perspective for judging which properties are ancestral and which have evolved recently. With the discovery of the archaebacteria two central evolutionary problems therefore become approachable: the nature of the common ancestor of all life and the evolution of the eukaryotic cell.

At what stage in the evolution of the cell did the fundamental division into the primary kingdoms take place? What was the nature of the universal ancestor? The assumption has been that the universal ancestor was a prokaryote, the simplest of today's living forms. Long ago, however, there must have been still simpler forms of the cell. Although nothing is known about such forms, one can make an educated guess as to certain of their general properties.

Consider the following argument. The translation mechanism is complex, comprising on the order of 100 large molecular components. It is also highly accurate in its functioning. In making proteins of the size common today (chains of from 100 to 500 amino acids), obtaining a flawless product 90 percent of the time or more requires an error rate of no more than a few parts in 10,000. Moreover, this accuracy must be attained by a mechanism of molecular dimensions. For such a mechanism to have evolved in a single step is clearly impossible. The primitive version of the mechanism must have been far simpler, smaller and less accurate.

Imprecision in translation would have required the synthesis of proteins that were smaller, and therefore less specific in their action, than proteins are today. (Otherwise the probability of error in making a protein strand would have been too great.) Among the smaller and less specific proteins would have been the enzymes required to process genetic information. If those enzymes were less precise than today's versions are, the cell's mutation rate must necessarily have been higher and the size of its genome correspondingly smaller. The translation process is the link between genotype and phenotype, between information and its expression; as the process evolved to become more precise, the cell itself necessarily passed through a corresponding series of evolutionary refinements. It evolved from an entity having simple properties, imprecise and general functions and a rather small complement of genes to an entity that functioned with many highly specific enzymes and a complex, precise genetic apparatus. To emphasize the primitive genetic and translational mechanisms of the earlier, simpler cells, I call them progenotes.

## The Progenote as Ancestor

The discovery of the archaebacteria provides the perspective needed to approach the question of whether the universal ancestor was a prokaryote or a progenote. Although the question is far from settled, the initial indications are that the universal ancestor was indeed a progenote. First of all consider that true bacteria and archaebacteria have probably existed for at least 3.5 billion years. The time needed for the evolution of the first true bacteria or archaebacteria, then, had to be less than a billion years, and perhaps much less. Still, the kinds of evolutionary changes that have arisen within each of the bacterial kingdoms over the later interval of three billion years or more are minor compared with the differences that separate archaebacteria from true bacteria, such as the differences in lipids, in transfer-RNA and ribosomal-RNA sequences and modification patterns and in enzyme-subunit structure.

It would seem that the nature of evolution some four billion years ago was very different from what it was later. This implies that the organisms undergoing the evolution were also very different. The possibility that the universal ancestor was in the process of developing the cell wall is suggested by the fact that archaebacterial cell walls are as unlike true bacterial walls as eukaryotic walls are. Perhaps the universal ancestor was still developing or refining biochemical pathways as well; lipids are synthesized differently in the two bacterial kingdoms, and many coenzymes are different. If archaebacteria should be found to differ from true bacteria in their mechanisms for controlling gene expression (a possibility that has yet to be investigated), the implication would be that their common ancestor may

| | ARCHAEBACTERIA | EUBACTERIA | EUKARYOTES |
|---|---|---|---|
| CELL SIZE (LINEAR DIMENSION) | ABOUT 1 MICROMETER | ABOUT 1 MICROMETER | ABOUT 10 MICROMETERS |
| CELLULAR ORGANELLES | ABSENT | ABSENT | PRESENT |
| NUCLEAR MEMBRANE | ABSENT | ABSENT | PRESENT |
| CELL WALL | VARIETY OF TYPES; NONE INCORPORATES MURAMIC ACID | VARIETY WITHIN ONE TYPE; ALL INCORPORATE MURAMIC ACID | NO CELL WALL IN ANIMAL CELLS; VARIETY OF TYPES IN OTHER PHYLA |
| MEMBRANE LIPIDS | ETHER-LINKED BRANCHED ALIPHATIC CHAINS | ESTER-LINKED STRAIGHT ALIPHATIC CHAINS | ESTER-LINKED STRAIGHT ALIPHATIC CHAINS |
| TRANSFER RNA'S: THYMINE IN "COMMON" ARM | ABSENT | PRESENT IN MOST TRANSFER RNA'S OF MOST SPECIES | PRESENT IN MOST TRANSFER RNA'S OF ALL SPECIES |
| DIHYDROURACIL | ABSENT IN ALL BUT ONE GENUS | PRESENT IN MOST TRANSFER RNA'S OF ALL SPECIES | PRESENT IN MOST TRANSFER RNA'S OF ALL SPECIES |
| AMINO ACID CARRIED BY INITIATOR TRANSFER RNA | METHIONINE | FORMYLMETHIONINE | METHIONINE |
| RIBOSOMES: SUBUNIT SIZES | 30*S*, 50*S* | 30*S*, 50*S* | 40*S*, 60*S* |
| APPROXIMATE LENGTH OF 16*S* (18*S*) RNA | 1,500 NUCLEOTIDES | 1,500 NUCLEOTIDES | 1,800 NUCLEOTIDES |
| APPROXIMATE LENGTH OF 23*S* (25–28*S*) RNA | 2,900 NUCLEOTIDES | 2,900 NUCLEOTIDES | 3,500 NUCLEOTIDES OR MORE |
| TRANSLATION-ELONGATION FACTOR | REACTS WITH DIPHTHERIA TOXIN | DOES NOT REACT WITH DIPHTHERIA TOXIN | REACTS WITH DIPHTHERIA TOXIN |
| SENSITIVITY TO CHLORAMPHENICOL | INSENSITIVE | SENSITIVE | INSENSITIVE |
| SENSITIVITY TO ANISOMYCIN | SENSITIVE | INSENSITIVE | SENSITIVE |
| SENSITIVITY TO KANAMYCIN | INSENSITIVE | SENSITIVE | INSENSITIVE |
| MESSENGER-RNA BINDING SITE *AUCACCUCC* AT 3′ END OF 16*S* (18*S*) RNA | PRESENT | PRESENT | ABSENT |

**MOLECULAR TRAITS of archaebacteria, eubacteria and eukaryotes clearly distinguish the three primary kingdoms. In some instances the archaebacteria are unique; in others they are similar either to eubacteria or to eukaryotes, as is indicated by color pattern. Thymine and dihydrouracil are modified bases that replace uracil in transfer RNA's. Elongation factor is a component of translation machinery.**

have had only rudimentary mechanisms of genetic control.

The key question is whether the universal ancestor was still developing the genotype-phenotype link when it gave rise to its descendant lines. Two observations suggest it may have been. RNA polymerase is the enzyme that transcribes the gene into its messenger-RNA complement (which is then translated into protein). The subunit structure of the RNA polymerase is quite constant among the true bacteria, whereas the archaebacterial polymerase structure is different. Could this mean the RNA-polymerase function was still being refined at the time the two bacterial lines separated?

The second observation concerns the modified nucleotides in transfer and ribosomal RNA's. As I mentioned above, the patterns of modification are almost invariant within any primary kingdom, but they tend to differ between kingdoms. Although the function of modified nucleotides in transfer and ribosomal RNA's is not understood, it is reasonable to assume most of them serve to "fine tune" translation: to make it more precise. If that is so, it would appear that many of the modifications have evolved independently in each of the primary kingdoms. The independence of these modifications implies in turn that the universal ancestor did not have today's highly specialized transfer-RNA and ribosomal-RNA molecules but made do with more rudimentary translation machinery. At this stage one can say only that the facts are consistent with, and indeed suggestive of, the universal ancestor's having had rudimentary transcription and translation mechanisms, and so having been a progenote.

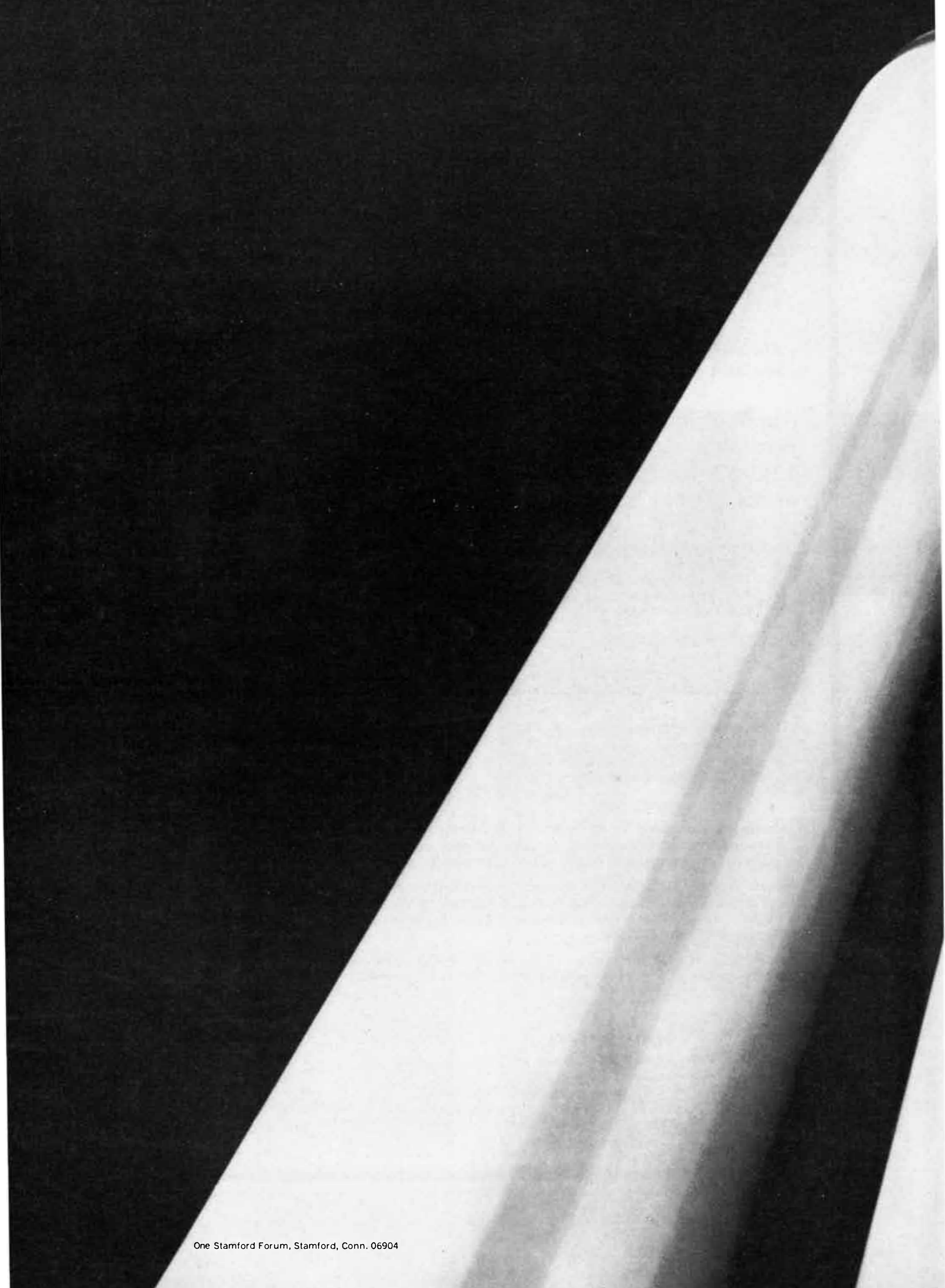
## The Origin of the Urkaryote

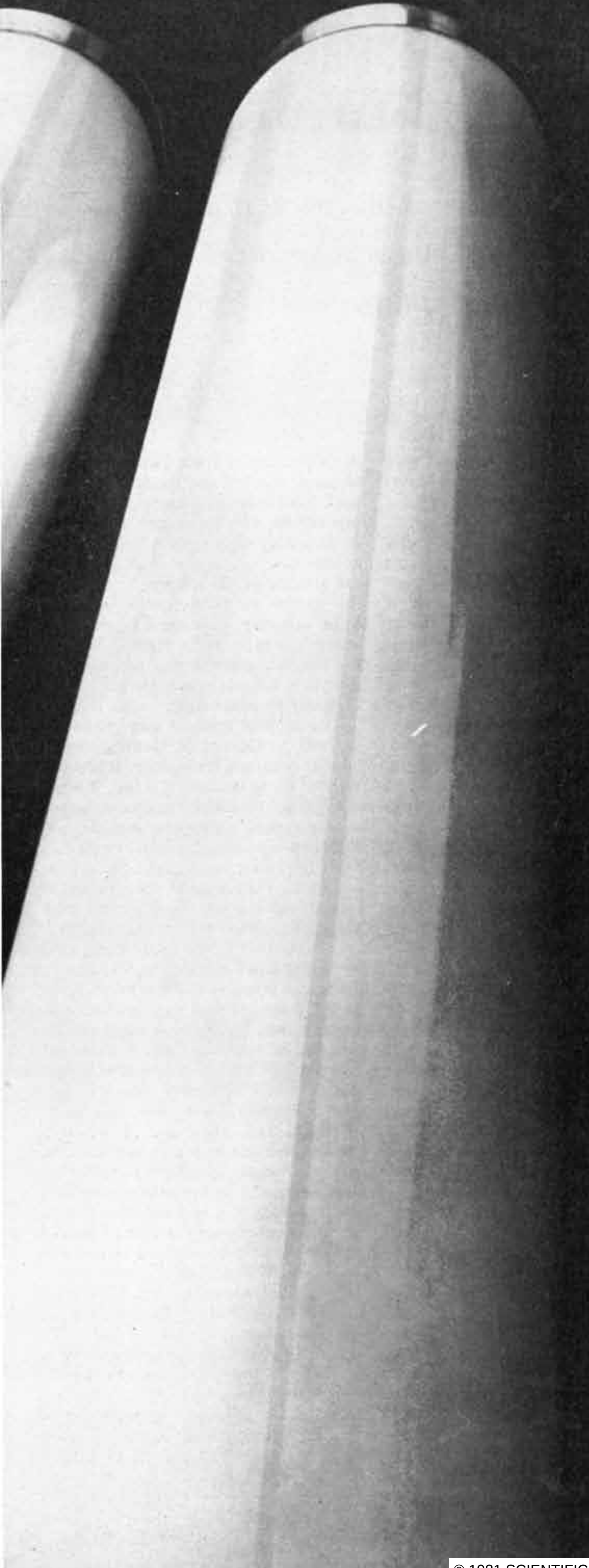
In terms of their ribosomal-RNA catalogues the archaebacteria, eubacteria and eukaryotes appear to be equidistant from one another genealogically; no specific relation between any two of the three has been detected. Nevertheless, in terms of the amino acid sequence of one protein, the ribosomal *A* protein, the archaebacteria clearly seem to be relatives of the eukaryotes. Therefore it may be that archaebacteria as well as true bacteria participated in forming the eukaryotic cell. Perhaps it is to the archaebacteria one should look for the origin of the unexplained stage of the eukaryotic cell: the urkaryote that played host to the endosymbiont ancestors of the mitochondrion and the chloroplast. (Rather than searching for the hypothetical host, however, one should instead question whether there was such an entity. This is not a time to shape new discoveries in accordance with old prejudices.)

As I have indicated, the differences between the eukaryotic cell and the other major cell types at the molecular level are more extensive and pervasive than any of the differences visible with a microscope. The eukaryotic nucleus appears to contain at least three kinds of genes: those of eubacterial origin (presumed to have been appropriated from the genomes of the eukaryote's organelles), those of archaebacterial origin (for example the gene for the ribosomal *A* protein) and those of an unidentified third origin (exemplified by the cytoplasmic ribosomal RNA). To what extent is the eukaryotic nucleus genetically a chimera: an entity composed of parts assembled from disparate sources? At what stage (or stages) of evolution did the presumed assembly take place? And what was the nature of the organisms that supplied the various genes and structures?

Biologists have tended to look at the eukaryotic cell as having been formed by the association of fully evolved prokaryotic cells; their association is assumed to have created a "higher" type of cell, the eukaryotic. (The term prokaryote—"before the nucleus"—carries just this implication.) The question that can profitably be asked now is whether the evolutionary events that gave the eukaryotic cell its basic molecular character really were of this nature. The eukaryotic cell appears to be a chimera at a very basic level. Even the eukaryotic ribosome seems to be chimeric, its component RNA's having come from a source other than that of at least one of its proteins. If this is a correct interpretation of the data (and future investigations must settle the point), the eukaryotic cell may be a different kind of entity than it is now taken to be. It may have been chimeric even before it reached a stage of complexity comparable to that of today's prokaryotes; it may have been chimeric as it emerged from the progenote condition. Rather than being an advanced, "higher" form, the eukaryotic cell may represent a throwback to the evolutionary dynamics of its long-gone ancestor, the progenote.

Perhaps the most exciting thing about the recent discoveries in molecular phylogeny is that they show how much information about the very early stages of evolution is locked into the cell itself. It is no longer necessary to rely solely on speculation to account for the origins of life. It has become customary to think of the last decades of this century as a time in biology when "genetic engineering" will make possible exciting developments in medicine and industry. It must also be recognized that biology is now on the threshold of a quieter revolution, one in which man will come to understand the roots of all life and thereby gain a deeper understanding of the evolutionary process.


One Stamford Forum, Stamford, Conn. 06904



# The Allocation of Resources by Linear Programming

*Abstract, crystal-like structures in many geometrical dimensions can help to solve problems in planning and management. A new algorithm has set upper limits on the complexity of such problems*

by Robert G. Bland

Consider the situation of a small brewery whose ale and beer are always in demand but whose production is limited by certain raw materials that are in short supply. Suppose the scarce ingredients are corn, hops and barley malt. The recipe for a barrel of ale calls for the ingredients in proportions different from those in the recipe for a barrel of beer. For instance, ale requires more malt per barrel than beer does. Furthermore, the brewer sells ale at a profit of $13 per barrel and beer at a profit of $23 per barrel. Subject to these conditions, how can the brewer maximize his profit?

It may seem that the brewer's best plan would be to devote all his resources to the production of beer, his more profitable product. This choice may not be well advised, however, because making beer may consume some of the available resources much faster than making ale does. If five pounds of corn are required for brewing a barrel of ale and 15 pounds are needed for brewing a barrel of beer, it may be possible to make three times as much ale as beer. Moreover, in brewing only beer the brewer may find that all his corn is used up long before his supplies of hops and malt are exhausted. It may turn out that by producing some beer and some ale he can take better advantage of his resources and thereby increase his profit. Determining such an optimum production program is not a trivial problem. It is the kind of problem that can be solved by the technique of linear programming.

Linear programming is a mathematical field of study concerned with the explicit formulation and analysis of such questions. It is a part of the broader field of inquiry called operations research, in which various methods of mathematical modeling and quantitative analysis are applied to large organizations and undertakings. Linear programming was developed shortly after World War II in response to logistical problems that arose during the war and immediately after it. One of the earliest publications on the uses of linear programming discussed a model of the 1948 Berlin airlift.

Although the computer is an indispensable tool for solving problems in linear programming, the term "programming" is employed in the sense of planning, not computer programming. "Linear" refers to a mathematical property of certain problems that simplifies their analysis. In the brewery problem the amount of any one resource needed to make either ale or beer is assumed to be proportional to the amount of the beverage produced. Doubling the amount of beer doubles the amount of each ingredient required for the brewing of beer, and it also doubles the profit attributable to the sale of beer. If the amount of corn consumed in making beer is plotted as a function of the amount of beer produced, the graph is a straight line. In order to apply the techniques of linear programming one must also assume that products and resources are divisible, or at least approximately so. For example, half a barrel of beer can be produced, and it has half the value of a full barrel.

Problems in linear programming are generally concerned with the allocation of scarce resources among a number of products or activities, under the proportionality and divisibility conditions I have described. The scarce resources may be raw materials, partly finished products, labor, investment capital or processing time on large, expensive machines. An optimum allocation may be one that maximizes some measure of benefit or utility, such as profit, or minimizes some measure of cost. In this era of declining productivity and dwindling resources a technique that can aid in allocating resources with the greatest possible efficiency may be well worth examining.

The properties of linear-programming problems derive from elementary principles of algebra and geometry. Solving such problems efficiently depends on algorithms, or step-by-step procedures, that cleverly exploit these algebraic and geometrical principles. The algorithms are also simple in conception, although the details of their operation can be quite intricate. It is the efficiency and versatility of a single algorithm called the simplex method that is largely responsible for the economic importance of linear programming.

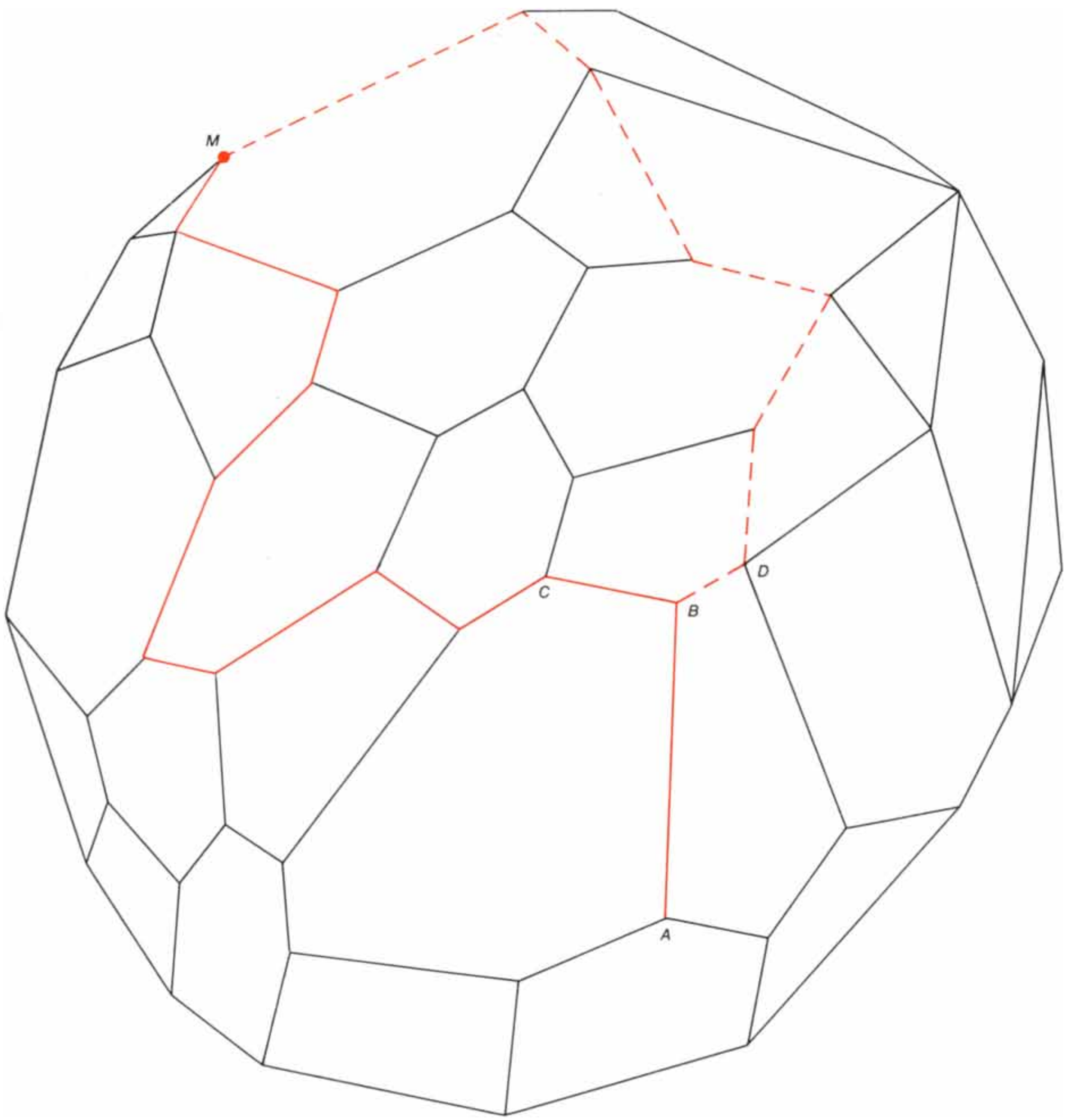
The simplex method was introduced in 1947 by George B. Dantzig, who is now at Stanford University. It has substantial value because it is fast, it is rich in applications and it can answer important questions about the sensitivity of solutions to variations in the input data. Such questions might take the following form: How should the production plan of the brewer change if the availability of hops or the profitability of beer is altered? How much should he be willing to pay for additional supplies of the scarce resources? What price should he ask from another entrepreneur who wants to buy some of his supplies? The simplex method can help in determining whether to buy a machine or sell one, whether to borrow money or lend it and whether to pay overtime wages or discontinue overtime labor. With the simplex method one can also impose additional constraints and solve the problem again to examine their effect. For example, the method can quickly tell an entrepreneur the cost of providing an unprofitable service in order to maintain the goodwill of a customer.

The simplex method has proved extremely efficient in solving complex linear-programming problems with thousands of constraints. For theoreticians, however, its speed in solving such problems has been somewhat puzzling. There is no definitive explanation of why the method does so well. Indeed, there are problems devised by mathematicians for which the simplex method is intolerably slow For reasons that are not entirely clear, such problems do not seem to arise in practice.

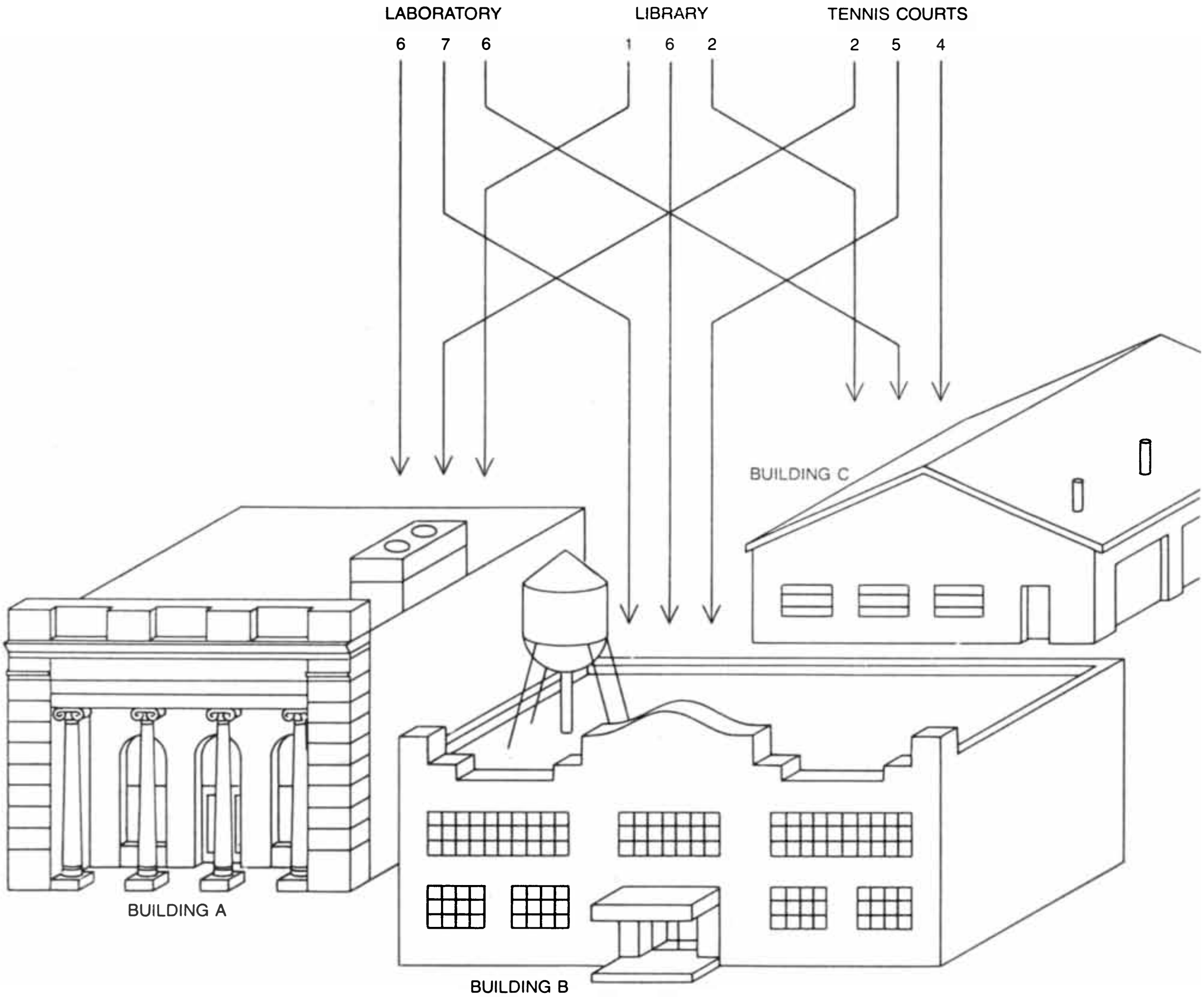
Mathematicians in the U.S.S.R. have recently developed a new algorithm for linear programming that in a certain sense avoids some of the theoretical difficulties that have been attributed to the simplex method. The development was reported in front-page articles in newspapers throughout the world, which suggests the economic significance of linear programming today. Unfortunately the new algorithm, which is called the ellipsoid method, has so far shown little prospect of outperforming the simplex method in practice. For now there is a curious divergence of practical and theoretical measures of computational performance.

Even when an efficient algorithm is employed, the setup costs of solving a large linear-programming problem can be considerable. Expressing a practical set of circumstances in terms of linear programming is not a trivial enterprise, and neither is the collection and organization of the data describing the circumstances. Moreover, solving the problem is feasible only with the aid of a high-speed computer.

Nevertheless, the economic benefits




**SIMPLEX METHOD of solving problems in linear programming finds an optimum allocation of resources by moving from one vertex to another along the edges of a polytope, a three-dimensional solid whose faces are polygons. Each point in the region where the polytope is constructed corresponds to a particular program, or plan, for allocating labor, capital or other resources. Associated with each such program is a net benefit or a net cost. The object of linear programming is to find the program with the maximum benefit or the minimum cost. The polytope defines a region of feasibility: all programs of allocation represented by a point within the polytope or on its surface are feasible, whereas those represented by a point outside the polytope are infeasible because of a scarcity of resources. When the relation between resources and benefits or costs is linear, the maximum and minimum values must lie at one of the vertexes of the polytope. The simplex algorithm examines the vertexes selectively, tracing a path (*ABC...M*) along the edges of the polytope. At each step along the path the measure of benefits or costs is improved until at the point *M* the maximum or minimum is reached. Often there are many paths from *A*, the starting vertex, to *M*. The polytope need not be three-dimensional and it commonly has thousands of dimensions.**




| | LABORATORY | LIBRARY | TENNIS COURTS |
|---|---|---|---|
| BUILDING A | **6** | 1 | 2 |
| BUILDING B | 7 | **6** | 5 |
| BUILDING C | 6 | 2 | **4** |

6 + 6 + 4 = 16

| | LABORATORY | LIBRARY | TENNIS COURTS |
|---|---|---|---|
| BUILDING A | 6 | **1** | 2 |
| BUILDING B | **7** | 6 | 5 |
| BUILDING C | 6 | 2 | **4** |

1 + 7 + 4 = 12

| | LABORATORY | LIBRARY | TENNIS COURTS |
|---|---|---|---|
| BUILDING A | 6 | 1 | **2** |
| BUILDING B | **7** | 6 | 5 |
| BUILDING C | 6 | **2** | 4 |

2 + 7 + 2 = 11

| | LABORATORY | LIBRARY | TENNIS COURTS |
|---|---|---|---|
| BUILDING A | **6** | 1 | 2 |
| BUILDING B | 7 | 6 | **5** |
| BUILDING C | 6 | **2** | 4 |

6 + 5 + 2 = 13

| | LABORATORY | LIBRARY | TENNIS COURTS |
|---|---|---|---|
| BUILDING A | 6 | **1** | 2 |
| BUILDING B | 7 | 6 | **5** |
| BUILDING C | **6** | 2 | 4 |

1 + 5 + 6 = 12

| | LABORATORY | LIBRARY | TENNIS COURTS |
|---|---|---|---|
| BUILDING A | 6 | 1 | **2** |
| BUILDING B | 7 | **6** | 5 |
| BUILDING C | **6** | 2 | 4 |

2 + 6 + 6 = 14

**ASSIGNMENT PROBLEM seeks to minimize the cost of matching buildings available for renovation with functions that must each be confined to one building. With three buildings and three functions there are $3^2$, or nine, renovation costs that must be considered. The costs (which are given here in millions of dollars) are conveniently represented in a matrix of numbers; a feasible assignment picks one number from each row and from each column of the matrix. There are $3 \times 2 \times 1$, or six, ways of accomplishing this. In general, for assignments of size $n$-by-$n$, the number of assignments is $n$ factorial (written $n!$), which is equal to $n$ multiplied by all positive integers less than $n$. Because the value of $n!$ increases rapidly, determining the minimum cost by enumeration of all possible assignments is impractical.**

of linear programming are often substantial. In the mid-1950's, when linear-programming methods were developed to guide the blending of gasoline, the Exxon Corporation began saving from 2 to 3 percent of the cost of its blending operations. The application soon spread within the petroleum industry to the control of additional refinery operations, including catalytic cracking, distillation and polymerization. At about the same time other industries, notably paper products, food distribution, agriculture, steel and metalworking, began to adopt linear programming. Charles Boudrye of Linear Programming, Inc., of Silver Spring, Md., estimates that a paper manufacturer increased its profits by $15 million in a single year by employing linear programming to determine its assortment of products.

Today "packages" of computer programs based on the simplex algorithm are offered commercially by some 10 companies. Roughly 1,000 customers make use of the packages under licenses from the developers. Each customer pays a sizable monthly fee, and so it is likely that each makes use of the method regularly. Many more organizations have access to the packages through consultants. In addition special-purpose programs for solving problems of flows in networks have been developed, and they may be in even wider service than the general-purpose linear-programming algorithms.

The broad applicability of linear programming can be illustrated even within a single organization. Exxon currently applies linear programming to the scheduling of drilling operations, to the allocation of crude oil among refineries, to the setting of refinery operating conditions, to the distribution of products and to the planning of business strategy. David Smith of the Communication and Computer Sciences Department of Exxon reports that linear programming and its extensions account for from 5 to 10 percent of the company's total computing load. This share has kept pace for the past 20 years with rapidly expanding general applications of information processing.

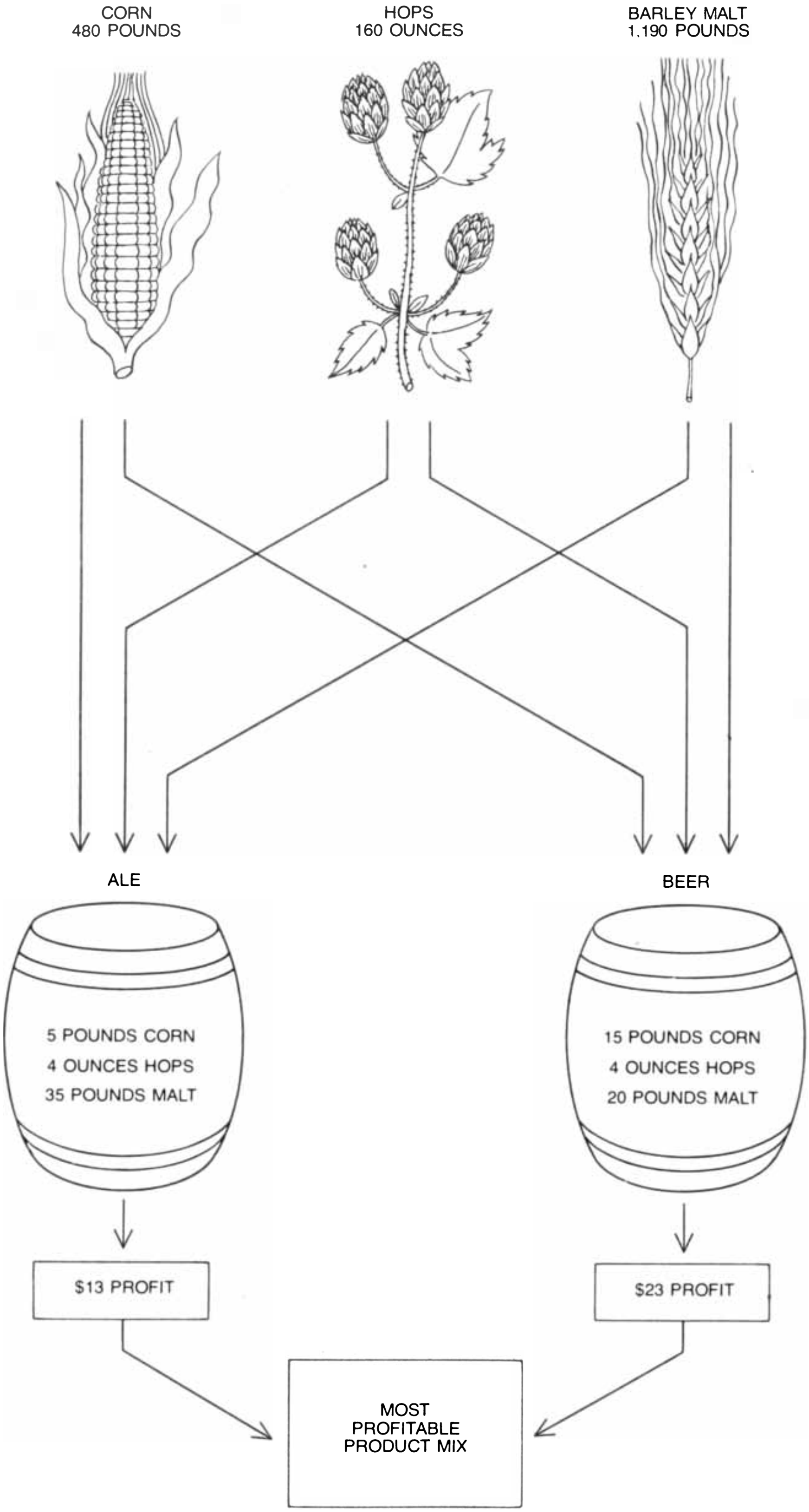
Although the simplex method is a powerful tool, it is founded on elementary ideas. In order to understand some of these ideas it is useful to examine a specially structured allocation problem called the assignment problem. Consider the situation of a university planning committee with three buildings available for renovation and three functions the buildings are to serve. Suppose the functions are those of a laboratory, a library and indoor tennis courts, so that there can be only one function to a building. The tables in the illustration on the opposite page indicate the cost of renovation for each of the nine possible matches of a building with a function. How can the committee minimize the renovation cost?

The problem can be solved by marking three of the squares in the 3-by-3 table. Precisely one square must be marked in each row and one must be marked in each column, so that each building has a function and all the functions are accommodated. The solution being sought is the one in which the sum of the costs in the marked squares is as small as possible. Finding an optimum solution is not difficult because there are only a few possible ways of marking the squares. Once one of the three squares in the first row is marked, only two squares in the second row remain available for marking. In the third row the choice is forced, because only one square appears in a heretofore unmarked column. Hence there are 3 × 2 × 1, or six, ways of making the assignment. It is an easy matter to enumerate all six possibilities,

| | SUBTRACT 3 | SUBTRACT 4 | SUBTRACT 8 | SUBTRACT 2 | SUBTRACT 9 | SUBTRACT 7 | SUBTRACT 0 | SUBTRACT 5 | SUBTRACT 1 | SUBTRACT 6 |
|---|---|---|---|---|---|---|---|---|---|---|
| SUBTRACT 2 | 6 | 6 | 12 | 4 | 14 | 11 | 3 | 9 | 4 | 10 |
| SUBTRACT 5 | 8 | 12 | 17 | 8 | 16 | 16 | 6 | 13 | 7 | 13 |
| SUBTRACT 6 | 11 | 13 | 18 | 9 | 15 | 16 | 8 | 14 | 8 | 14 |
| SUBTRACT 4 | 8 | 10 | 15 | 10 | 16 | 13 | 7 | 9 | 7 | 11 |
| SUBTRACT 8 | 13 | 15 | 16 | 11 | 19 | 19 | 10 | 16 | 12 | 18 |
| SUBTRACT 5 | 9 | 12 | 17 | 8 | 16 | 15 | 7 | 13 | 7 | 11 |
| SUBTRACT 3 | 8 | 8 | 13 | 8 | 13 | 12 | 6 | 8 | 4 | 11 |
| SUBTRACT 1 | 6 | 8 | 11 | 3 | 14 | 10 | 2 | 8 | 3 | 9 |
| SUBTRACT 0 | 5 | 5 | 12 | 2 | 12 | 7 | 2 | 9 | 6 | 12 |
| SUBTRACT 7 | 15 | 14 | 17 | 13 | 17 | 18 | 7 | 18 | 10 | 16 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 2 | 0 | 3 | 2 | 1 | 2 | 1 | 2 |
| 0 | 3 | 4 | 1 | 2 | 4 | 1 | 3 | 1 | 2 |
| 2 | 3 | 4 | 1 | 0 | 3 | 2 | 3 | 1 | 2 |
| 1 | 2 | 3 | 4 | 3 | 2 | 3 | 0 | 2 | 1 |
| 2 | 3 | 0 | 1 | 2 | 4 | 2 | 3 | 3 | 4 |
| 1 | 3 | 4 | 1 | 2 | 3 | 2 | 3 | 1 | 0 |
| 2 | 1 | 2 | 3 | 1 | 2 | 3 | 0 | 0 | 2 |
| 2 | 3 | 2 | 0 | 4 | 2 | 1 | 2 | 1 | 2 |
| 2 | 1 | 4 | 0 | 3 | 0 | 2 | 4 | 5 | 6 |
| 5 | 3 | 2 | 4 | 1 | 4 | 0 | 6 | 2 | 3 |

**CONVINCING A SKEPTIC that an assignment is optimum (*colored areas in upper illustration*) does not entail evaluating the cost of all the feasible assignments. Here the cost matrix is a 10-by-10 one, and so there are 10!, or 3,628,800, possible assignments. If a number is subtracted from each entry in a row or column, the number is also subtracted from the total cost of any possible assignment and the relative costs remain unchanged. The reason is that every assignment picks exactly one number from the transformed row or column of the matrix. The set of numbers to be subtracted can be chosen so that the original matrix is transformed into a matrix that has no negative entries and has at least one entry with a cost of zero in each row and column (*lower illustration*). Because no assignment can have a total cost less than zero the optimum assignment for the transformed matrix must be one that has a total cost of zero. It follows that the entries in the corresponding positions in the original matrix are also optimum. Efficient algorithms have been devised for generating the set of numbers to be subtracted.**




**BREWER'S DILEMMA illustrates an application of linear programming to problems of optimizing the apportionment of resources among various products. The brewer's production of beer and ale is limited by the scarcity of three essential ingredients: corn, hops and malt. Feasible production levels are determined not only by the total amount of each ingredient on hand but also by the proportions of the ingredients required to make the two products. The objective function, or the quantity to be optimized, is the brewer's profit. In linear programming all the resources, all the products and all the benefits are assumed to be divisible quantities: the brewer can use half a pound of corn, sell a fourth of a barrel of ale and realize proportional profits.**

evaluate the total cost of each one and select the least expensive assignment.

This enumerative approach solves the problem of a 3-by-3 matrix, but it becomes impractical for larger problems. Suppose there are four buildings and four functions; the number of possible assignments is then $4 \times 3 \times 2 \times 1$, or 24. In the general statement of the problem there are $n$ buildings and $n$ functions, and the number of assignments is $n$ factorial (written $n!$), which signifies $n$ multiplied by all the integers from 1 to $n - 1$. For $n = 10$ there are 10!, or more than 3.6 million, distinct assignments.

The rapid growth of $n!$ dispels any enthusiasm one might have for the enumerative method. Suppose one had to solve a 35-by-35 assignment problem by enumeration and one had a computer that could sort through the possible assignments, evaluate the cost of each one and compare it with the lowest-cost assignment encountered up to that point at a rate of a billion assignments per second. (A computer capable of this speed would be much faster than any available now.) Even if the task of enumerating the 35! assignments were to be shared by a billion such computers, only an insignificant fraction of the required computations would be completed after a billion years.

The 35-by-35 assignment problem is not large. If the problem were one of assigning personnel to jobs in order to minimize the total cost of job training, $n$ might well be equal to 35. There are numerous other assignment problems in which $n$ is equal to 1,000 or more. Clearly such problems require a procedure cleverer than enumeration.

The burden of enumeration might be greatly reduced if one could avoid examining assignments that are costlier than the ones already checked. This effect would be achieved if there were a stopping rule or optimality criterion that would allow an optimum assignment to be recognized quickly once it was encountered. Any algorithm that incorporates such a criterion offers important collateral benefits. The benefits are summarized by what Jack Edmonds of the University of Waterloo in Ontario calls "the principle of the absolute supervisor," or what might also be called "the problem of the skeptical boss."

Suppose after tedious enumeration you have solved the 10-by-10 assignment problem indicated in the upper illustration on the preceding page by examining all 3,628,800 possible assignments. The optimum assignment, you maintain, is the one shaded in color in the illustration. You present the solution to your boss, who looks you in the eye, puffs on his cigar and demands, "How do I know there is no less costly solution?" You might swallow hard at


STOP
STOP
INTRODUCING THE SUBARU HILL-HOLDER.™*
WHAT YOU ALWAYS NEEDED IN THE CLUTCH.
You know that sinking feeling that can come over you when you stop on a hill with a standard shift automobile?
Well, now, in a Subaru, you don't have to hold your breath. With an exclusive Subaru feature called Hill-Holder, your car won't roll backwards.
Just depress the clutch and brake, and the Hill-Holder automatically engages.
When your foot is removed from the brake, the car will not roll backward.
In fact, it will remain motionless until the clutch is released and the accelerator depressed to proceed up the hill.
No drifting back. No white knuckles. No sweat. You no longer go downhill on your way uphill.
SUBARU®
INEXPENSIVE. AND BUILT TO STAY THAT WAY.
*AVAILABLE ON CERTAIN MODELS.
© SUBARU OF AMERICA, INC. 1981

this question, because it seems the only way to demonstrate the merits of your solution is to repeat the examination of all 3,628,800 possibilities under the scrutiny of your boss.

A stopping rule offers a concise proof of optimality. Suppose you go to your boss not only with the optimum assignment but also with a set of numbers to be subtracted from the entries in each row and column. Setting forth how these numbers are obtained would require a detailed discussion of the assignment problem; it will suffice to point out that the set of numbers can be specified by an efficient computer algorithm. The utility of the numbers, once they have been found, is readily appreciated. Note that subtracting the same number from every entry in a given row or column is equivalent to subtracting this amount from the total cost of every possible assignment. The reason is that every feasible assignment must choose one and only one entry from each row and each column. For example, if 5 is subtracted from every entry in the sixth row, every possible assignment will include exactly one entry that is 5 less than the corresponding assignment made with the original array of costs. The relative costs of all the assignments will therefore remain unchanged. Such subtraction can be done repeatedly, provided it is always applied uniformly to every entry in a given row or column.

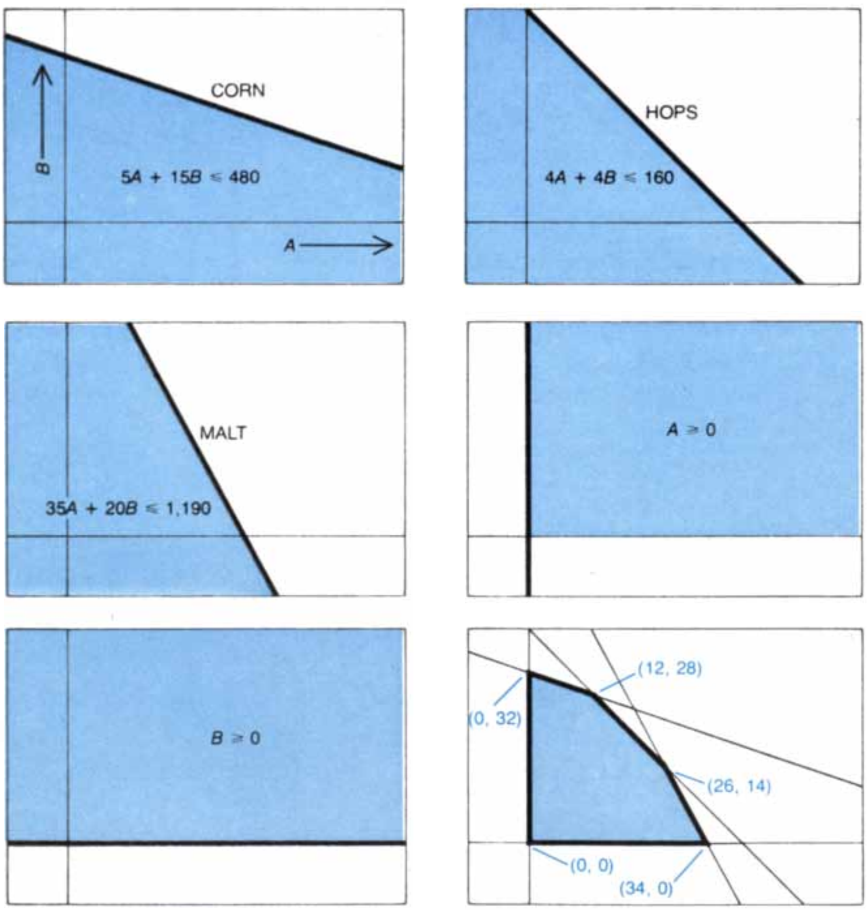
By means of repeated subtraction you can transform the original cost matrix into the matrix shown in the lower illustration on page 129. The latter array of costs has a remarkable property. You can now point out to your boss that the costs corresponding to the squares selected by your assignment are all zero and that no entries in the matrix are less than zero. Since the sum of the costs marked by your assignment is zero and there are no negative costs, no other possible assignment can have a lower cost. In short, you have shown your boss, with a few hundred calculations rather than tens of millions, that no assignment can be less expensive than the one you have chosen.

Although I have not demonstrated how to solve an assignment problem or how to find the row and column numbers to subtract, the assignment problem does illustrate the necessity of avoiding enumeration and the possibility of doing so by means of a stopping rule that recognizes an optimum solution. These are design characteristics of algorithms that can be applied not merely to the assignment problem but to linear-programming problems in general.

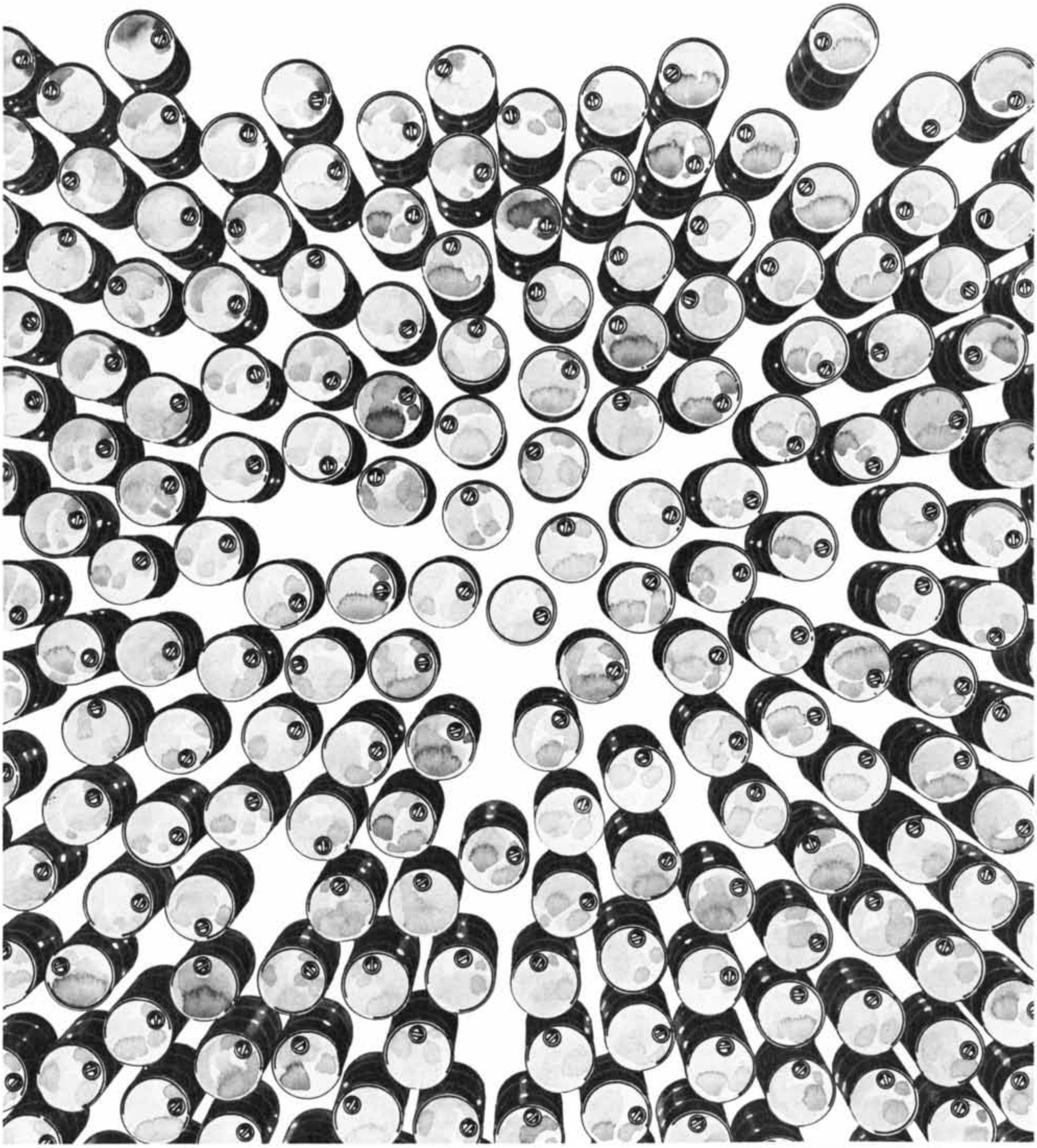
Consider again the situation of a brewer whose two products, ale and beer, are made from different proportions of corn, hops and malt. Suppose 480 pounds of corn, 160 ounces of hops and 1,190 pounds of malt are immediately available and the output is limited by the scarcity of these raw materials. Other resources, such as water, yeast, labor and energy, may be consumed in the manufacturing process, but they are considered to be plentiful. Although they may influence the brewer's willingness to produce beer and ale because of their costs, they do not directly limit the ability to produce. Assume that the brewing of each barrel of ale consumes five pounds of corn, four ounces of hops and 35 pounds of malt, whereas each barrel of beer requires 15 pounds of corn, four ounces of hops and 20 pounds of malt. Assume further that all the ale and beer that can be produced can be sold at current prices, which yield a profit of $13 per barrel of ale and $23 per barrel of beer.

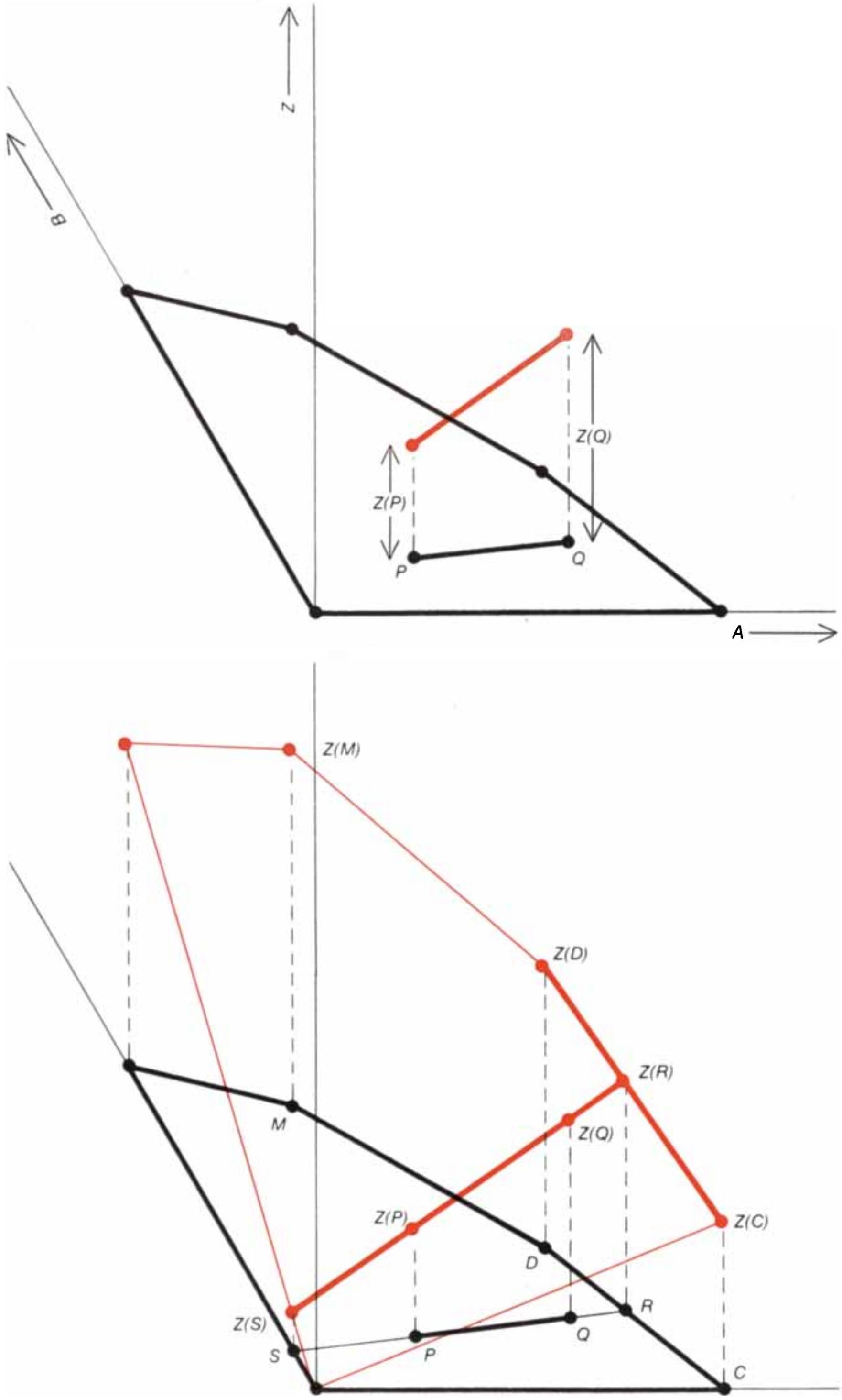
The scarcity of corn, hops and malt limits the feasible levels of production. For example, although there are enough hops and malt to brew more than 32 barrels of beer, production of this much beer would exhaust the supply of corn, allowing no greater output of beer and no output of ale at all. Another feasible production program calls for no beer and 34 barrels of ale, depleting all 1,190 pounds of malt. The first alternative seems preferable to the second. The profit realized by the first program is 32 × $23, or $736, whereas the second program yields only 34 × $13, or $442.

There are other production programs that are better than either of these. Six barrels of ale and 30 barrels of beer use all 480 pounds of corn, 154 of the 160 ounces of hops and 810 of the 1,190 pounds of malt, yielding a profit of (6 × $13) + (30 × $23), or $768. Many programs earn even greater profit. In this case it is not merely impractical to




**FEASIBILITY REGION in the brewer's problem is made up of the intersection of five half planes. Any point $(A,B)$ in the plane corresponds to a production program that calls for making $A$ barrels of ale and $B$ barrels of beer. The first three half planes graphically represent all the production programs that are achievable, given the available quantities of each ingredient. For example, the amount of corn required is $5A + 15B$ (the weight in pounds of the corn needed to make a barrel of ale times $A$ plus the weight of the corn needed to make a barrel of beer times $B$). This quantity must not exceed the 480 pounds of corn available to the brewer. Hence any point in the half plane to the lower left of the line $5A + 15B = 480$ represents a feasible production program that requires no more than the 480 pounds of corn available. The half planes associated with hops and malt can be constructed in a similar way. The remaining two half planes express the fact that only those programs having nonnegative production are feasible.**






**VALUE OF THE OBJECTIVE FUNCTION** for any point *P* in the feasible region can be graphed as the distance *Z*(*P*) above or below *P* measured on the *z* axis. (Distances along the *z* axis have not been drawn to the same scale as distances in the plane of the feasible region.) One can think of the distance *Z*(*P*) as a point in three-dimensional space. If the function is linear, the graph of the values the function takes along any straight line in the feasible region is also a straight line (*upper graph*). For any point *P* in the interior of the region, there is some line through *P* that intersects the boundary of the region at two points, say the points *R* and *S* (*lower graph*). If the line segment in space connecting *Z*(*R*) and *Z*(*S*) is not parallel to the plane of the feasible region, the objective function must assume its maximum value along the line segment at one of its endpoints, say *Z*(*S*), which corresponds to the point *S* on the edge of the feasible region. The graph of the objective function along an edge is also a straight line and it too assumes its maximum at one of its endpoints, say *Z*(*D*), which corresponds to a vertex of the feasible region. Thus there is always an edge point that dominates a given interior point, and there is always a vertex that dominates any point along an edge. To find the maximum value *Z*(*M*) of the objective function one need examine only the vertexes. When the feasible region is two-dimensional, the objective function forms a plane whose maximum height is attained at a vertex.

enumerate all the possibilities, as it was in the assignment problem; here it is not even possible. There are infinitely many production programs that meet the conditions of the brewer's problem. Each such program is called a feasible solution. Fortunately there is a small set of feasible solutions called extremal solutions to which attention can be confined.

The importance of the extremal solutions becomes apparent when the set of all feasible solutions is represented graphically as a set of points in a plane; the set of points constitutes the feasible region. Let $A$ designate the number of barrels of ale brewed according to any possible production strategy and let $B$ designate the number of barrels of beer. $A$ and $B$ are known in linear programming as decision variables. They can be associated with the coordinate axes of the plane. Any point on the plane can be specified by a pair of coordinates $(A,B)$, which also correspond to a particular set of production levels.

Because negative levels of production are not possible, the feasible region can immediately be confined to the upper right-hand quadrant of the plane, where $A$ and $B$ are both nonnegative. How does the scarcity of malt affect production? Since 35 pounds of malt are needed for each barrel of ale and 20 pounds are needed for each barrel of beer, the total amount of malt needed to make $A$ barrels of ale and $B$ barrels of beer is $35A + 20B$. If all 1,190 pounds of malt are used, $35A + 20B = 1{,}190$. The set of points $(A,B)$ that satisfy this equation form a straight line. All points $(A,B)$ corresponding to production plans that call for more than 1,190 pounds of malt lie on one side of the line and the points that require less malt lie on the other side. Only the latter set of points and the points actually on the line are feasible because of the limited supply of malt.

In a similar way the scarcity of hops confines the feasible region to one side of the line $4A + 4B = 160$ and the scarcity of corn confines the region to one side of the line $5A + 15B = 480$. The points that satisfy all these requirements make up the feasible region [*see illustration on page 132*]. Note that the feasible region is convex: any line segment that connects two points in the region (including the points on the perimeter lines) lies entirely within the region.

Since the brewer's profit is \$13 per barrel of ale and \$23 per barrel of beer, his problem is to maximize his total profit: $13A + 23B$. In order to do so he must find a point $(A,B)$ in the convex feasible region where $13A + 23B$ has its maximum value. In linear programming such a measure of benefits to be maximized (or sometimes of costs to be minimized) is called an objective function.

The objective function can be incorporated into the graph of the feasible

ENTRY®
computer services news
MCDONNELL DOUGLAS AUTOMATION COMPANY

# MCAUTO UNIGRAPHICS–the CAD/CAM workhorse for Division Engineering at 3M.

Supporting a variety of engineering needs in a multi-billion dollar company can tax even the most professional organizations, the most advanced techniques and equipment.

But for Miland Meek, Manager of Computer Aided Design Service for Division Engineering at 3M, the work has been simplified with UNIGRAPHICS,™ the minicomputer-based graphics system from MCAUTO.

"We can use UNIGRAPHICS on a variety of capital equipment design projects—from a piece of equipment a block long and three stories high, to a very small die assembly," says Mr. Meek. "We need that versatility, and we get it with UNIGRAPHICS."

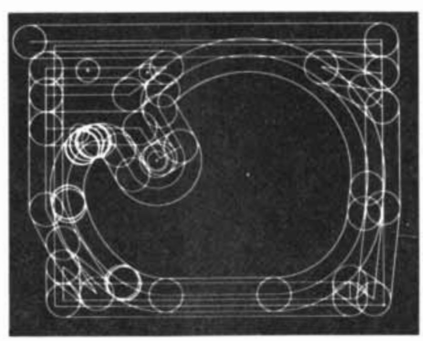
Besides capital equipment design, 3M uses UNIGRAPHICS to produce engineering drawings and to perform graphics machining for numerical control equipment. "It's a truly integrated tool here at 3M. We're now using most of the basic modules and we're exploring the use of both the Finite Element and Schematics modules," says Mr. Meek.

He also reports that UNIGRAPHICS has proven easy to learn and use. With 15 hours of instruction and 80 hours of practice, a trainee is well prepared to use it.

"Productivity is the goal. We've trained 85 people on the system and we're pleased to see that a trained user can be away from the system for some months, come back, and with the system's tutorial features be productive very quickly. Also, with remote stations—some are five miles from the host CPU—engineers turn travel time into productive time," says Mr. Meek.

With a single data base and full Automatic Programmed Tool capabilities, MCAUTO UNIGRAPHICS is the only totally-integrated CAD/CAM turnkey system available. It's the direct link between design and manufacturing that delivers the productivity and competitive edge you need. For more facts, call MCAUTO at (314) 232-8021, or mail the coupon.

*"UNIGRAPHICS gives us the versatility and the productivity we need to keep our competitive edge," says Miland Meek, manager of CAD Service for Division Engineering at 3M.*



*The ubiquitous 3M tape dispenser, as seen on a UNIGRAPHICS terminal.*

**W. R. Vickroy, MCAUTO, Dept. KI-500**
**Box 516, St. Louis, MO 63166**

☐ Please send me information on MCAUTO UNIGRAPHICS.
☐ Have your representative phone me for an appointment.

Name ____ Title ____
Company Name ____ Phone ____
Address ____
City ____ State ____ Zip ____

MCAUTO
MCDONNELL DOUGLAS
Setting the standard in data processing

region by adding a third dimension. For any point (*A,B*) representing a production plan the expected profit is given by the height of the function $13A + 23B$ above the plane at that point. Clearly the task confronting the brewer is to find a point in the feasible region where the objective function is at its greatest height. If the function had to be evaluated at all the points, the task would be impossible, but two distinctive properties of the problem act to narrow the search. The properties are the convexity of the feasible region and the linearity of the objective function.

Since the region is convex, any point in the interior of the region can be included in a line segment whose endpoints lie on the boundary of the region. (Indeed, an infinite number of such line segments can be drawn through any given point; which line segment is chosen is immaterial.) In the space above each such line segment it is possible to construct a graph of the objective function giving the profit for each point on the line segment. Since the objective function is linear, the graph is a straight line [*see illustration on page 134*]. The straight-line graph may be parallel to the plane, in which case all the brewing strategies along the line segment have the same profit. If the graph of the objective function over the line segment is not parallel to the plane, it must assume its maximum value at one of the two endpoints, which lie on the boundary of the feasible region. Hence the maximum of the objective function over such a line segment must always be attained at one of the points where the line segment intersects the boundary. Because the same analysis can be applied to any line segment in the feasible region, it follows that the overall maximum of the objective function is invariably found somewhere on the boundary of the region. The brewer in search of maximum profit can ignore the entire interior of the feasible region and consider only those brewing strategies that correspond to points on the boundary.

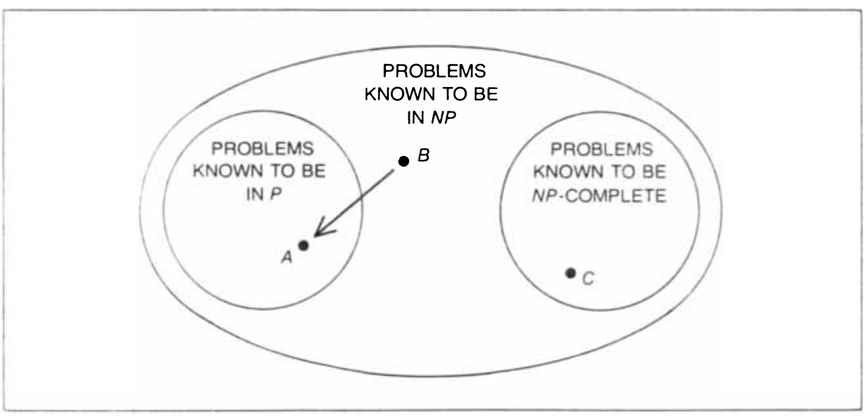
The analysis can be taken a step further by the same argument. If the feasible region is a polygon, every point on the boundary lies on a line segment whose endpoints are two of the vertexes of the polygon. A graph of the objective function for such a boundary segment can be constructed in the same way it is for a line segment that crosses the interior. Again the maximum must be found at one of the endpoints (or at both endpoints if the objective function is constant and parallel to the plane). Thus a maximum value of the objective function throughout the feasible region can be found among the vertexes. The brewer needs only to check, at most, the value of the function at all the vertexes of the feasible region and select the one yielding the best profit. He can then be certain that no other brewing strategy would bring a higher profit. In the example considered here there are five vertexes. The one at point (12,28), which represents the production of 12 barrels of ale and 28 barrels of beer, yields a profit of (12 × \$13) + (28 × \$23), or \$800. That is the maximum profit the brewer can realize.

The inclusion of an additional constraint (such as a shortage of yeast) would not significantly alter the geometrical interpretation of the brewer's problem. The polygon might then have six sides instead of five. The introduction of a third product, on the other hand, would have a more profound effect: the geometrical model would then be three-dimensional. Inequalities in three variables correspond to half spaces defined by planes in three-dimensional space instead of half planes defined by lines in two-dimensional space. The feasible region is no longer a polygon but might instead look like a cut gemstone, a three-dimensional polytope whose faces are all polygons. As the number $n$ of decision variables increases further, the geometrical interpretation remains valid but it becomes more difficult to visualize the $n$-dimensional polytope formed by the intersections of $(n - 1)$-dimensional hyperplanes. Vertexes retain their special status, however, and their positions can be determined by algebraic methods that replace geometrical intuition.

It may appear that by confining the evaluation of the objective function to the vertexes of the feasible region, the solution of linear-programming problems by enumeration becomes practical. As in the assignment problem, however, the number of possibilities to be enumerated grows explosively. A problem with 35 decision variables and 35 constraints would be impossible.

Dantzig's simplex method examines vertexes, but it does so selectively. In the brewery example the method might begin with the vertex at the origin, (0,0). Here nothing is produced and the profit is zero. Following either of the incident edges away from the origin leads to points with larger objective-function values. The simplex method selects such an edge, say the *B* axis, and follows it to its other end, the vertex (0,32). The program here calls for making 32 barrels of beer but no ale, for a profit of \$736. From this vertex an incident edge leads to still greater objective-function values. The simplex algorithm therefore proceeds to the vertex (12,28) at the other end of the edge. Here the profit from brewing 12 barrels of ale and 28 barrels of beer is \$800. At (12,28) all incident edges lead in unfavorable directions; the algorithm therefore halts with a declaration that the vertex (12,28) is optimal.

In general the simplex method moves along edges of a polytope from vertex to adjacent vertex, always improving the value of the objective function. The procedure can begin at any vertex, and it halts when no adjacent vertex has a better objective-function value than the current one. This stopping rule is val-

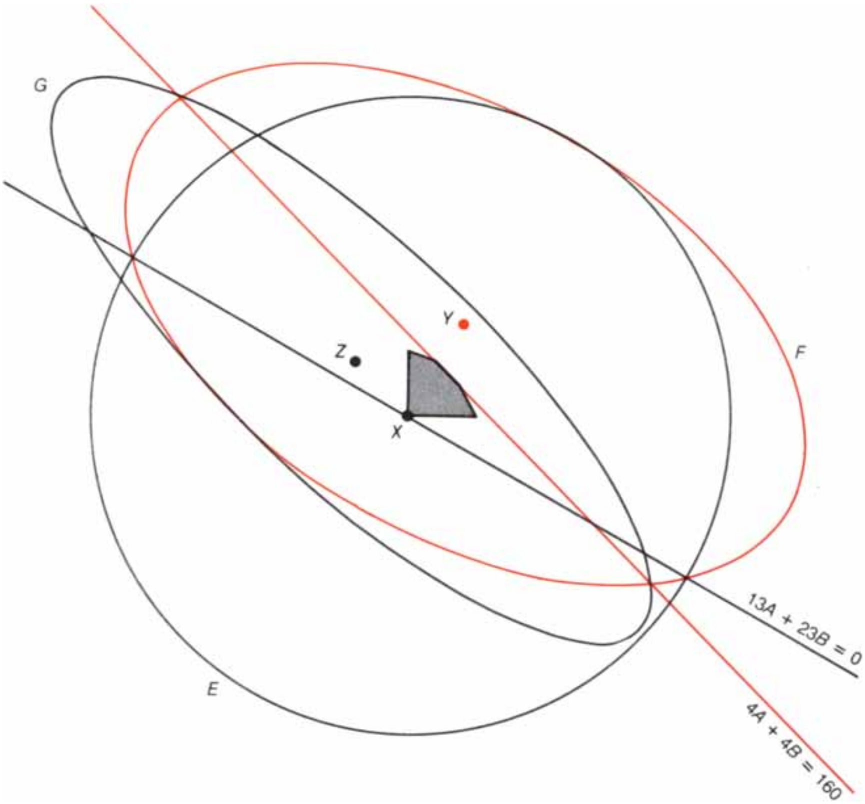



**COMPUTATIONAL-COMPLEXITY THEORY has recently been able to place linear programming in the set *P* of polynomially bounded classes of problems. This assignment, represented by point *A*, was made possible because of a proof by the Russian mathematician L. G. Khachian that a newly discovered algorithm for linear programming, called the ellipsoid method, is polynomially bounded. A polynomial function of a number *n* is a finite sum of powers of *n*, each power being multiplied by a constant. A class of problems is polynomially bounded if the number of elementary arithmetical operations needed to solve any problem in the class is bounded by a polynomial function of some measure *s* of the size of the problem. Until Khachian's result it was not known whether such an algorithm existed for linear programming (although there was no proof that it did not exist). Linear programming was previously known to be in the larger set *NP* of nondeterministic polynomial functions (point *B*). Roughly speaking, *NP* is the set of problem classes for which the feasibility of a proposed solution can quickly be checked. A second subset of *NP* is called the set of *NP*-complete problems; proving that any *NP*-complete class of problems is polynomially bounded would show that all classes in *NP* are also in *P*. The *NP*-complete problems are, in a sense, the hardest problems in *NP* (point *C*).**

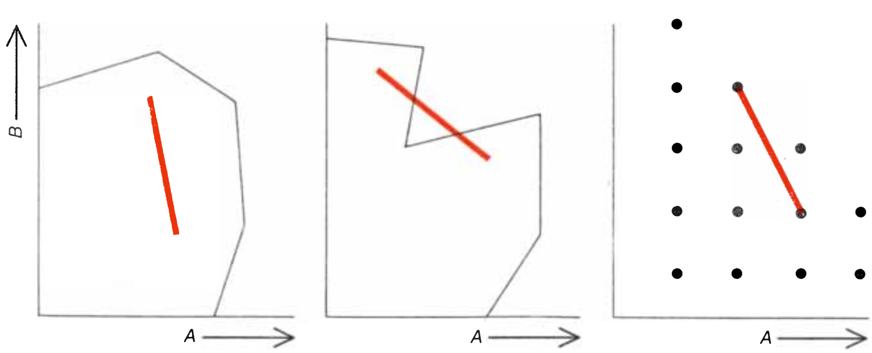
id only because the feasible region is convex: convexity guarantees that any locally maximum vertex is a globally maximum vertex. One need look only in the vicinity of each point to tell whether improvement is possible.

How does one tell that an edge leading away from a given vertex will improve the value of the objective function? The key is the concept of a "marginal value" attributed to each resource at each vertex of the feasible polytope. For many-dimensional problems in linear programming the algebraic method that is commonly employed to choose a path from vertex to vertex can be understood when one understands the significance of the marginal values.

At the maximum vertex in the brewery problem there are 210 pounds of excess malt that are not utilized in the production program represented by that vertex. The addition or subtraction of a pound of malt from the initial supply of 1,190 pounds would not change the obtainable profit. On the other hand, another ounce of hops would make possible an increase of $2 in total profit. This increase is the marginal value of hops at the maximum vertex. It can be interpreted as the effect of "pushing" the constraint line for hops farther from the origin to reflect the extra available ounce of hops [*see illustration on opposite page*]. Similarly, the marginal value of corn at this vertex is $1.




**ELLIPSOID METHOD, like the simplex method, can be interpreted geometrically. For the brewer's problem it begins with a large circle *E* centered at the origin *X* and enclosing the feasible region. The size of the circle is determined by the problem data. The method proceeds by constructing a sequence of smaller ellipses, each of which contains the optimum solution. Because the center of the circle is feasible the next ellipse in the construction must contain all the points in the circle whose objective function is at least as great as it was at *X*. The smallest ellipse that accomplishes this is *F*. The center of *F*, at point *Y*, is not feasible. Hence the algorithm requires the next ellipse *G*, centered at *Z*, to contain every point in the ellipse *F* that lies on the feasible side of the constraint $4A + 4B = 160$, which was violated by the point *Y*. The process continues by cutting the current ellipse with a contour of the objective function if the center of the current ellipse is feasible. If the center is not feasible, the successor ellipse circumscribes the part of the current ellipse that satisfies a constraint violated at its center. The areas of the ellipses shrink rapidly enough to ensure that their centers converge to an optimum solution.**




**CONVEXITY PROPERTY of linear programming states that for any two points in the feasible region a line segment connecting the points must lie entirely within the region. Only the feasible region at the left is convex. The region in the diagram at the right is a set of discrete points. Convexity ensures that any local maximum of a linear objective function is a global maximum.**

Marginal prices have a natural economic interpretation. If the brewer could buy an additional ounce of hops, he could increase his profit by $2. If hops were available for less than $2 per ounce, it would be worthwhile to buy hops. On the other hand, if a price higher than $2 per ounce were offered for hops, it would be worthwhile for the brewer to divert hops from the production of beer and ale and sell them on the market. This does not mean that the buying or selling of hops should continue indefinitely at $2 per ounce, but for increases of about 19 ounces or decreases of up to 32 ounces, $2 remains the break-even price in this example.

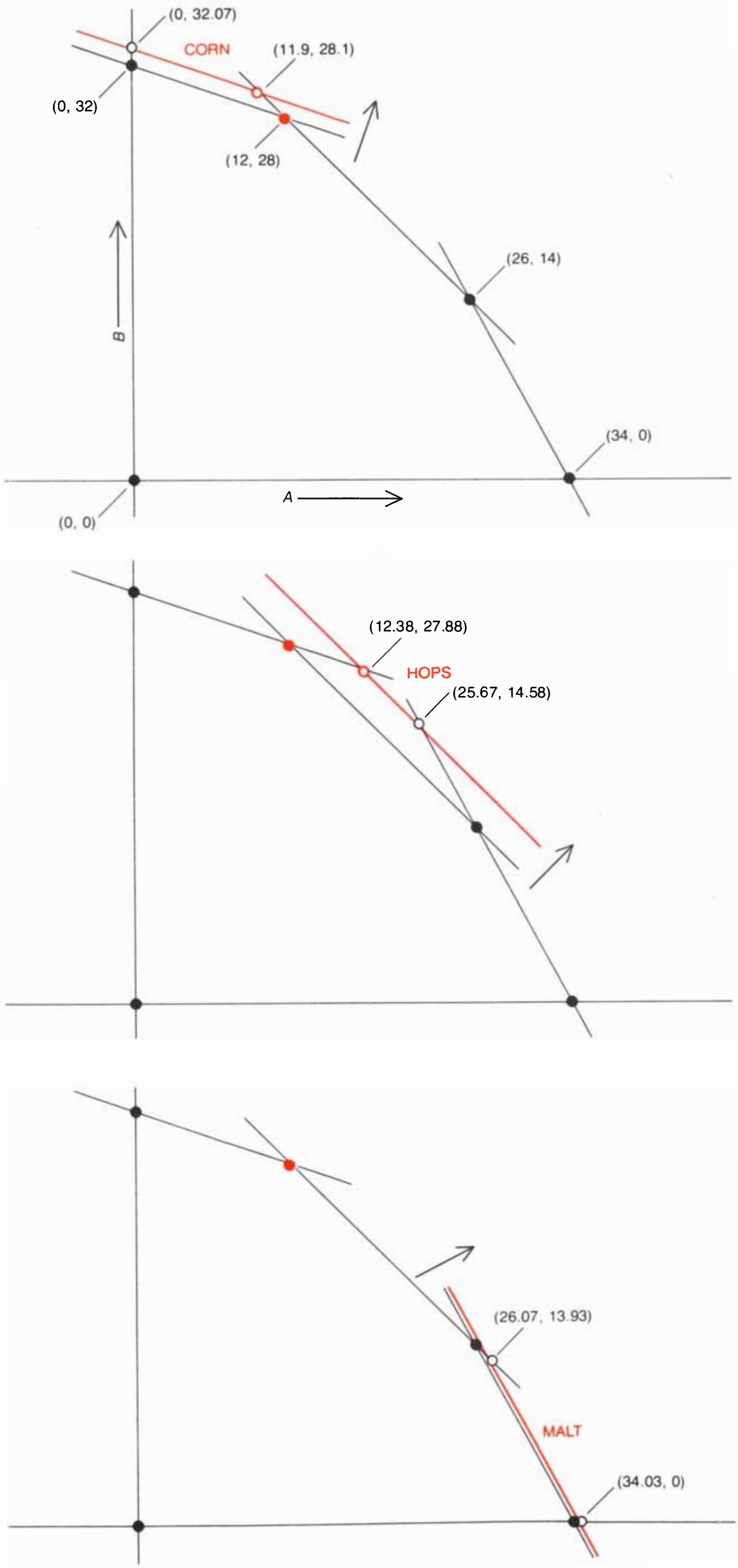
Marginal values are sometimes called shadow prices or imputed prices. They indicate the relative amount that each scarce resource contributes to the profitability of each item in production. For example, a barrel of ale requires five pounds of corn with an imputed price of $1 per pound, four ounces of hops with an imputed price of $2 per ounce and 35 pounds of malt with an imputed price of zero. The total imputed price of ale is equal to the profit of $13 per barrel.

Suppose a new product, light beer, is proposed. Making a barrel of light beer requires two pounds of corn, five ounces of hops and 24 pounds of malt. How much profit per barrel must be obtained from light beer to justify diverting resources from the production of beer and ale? The imputed prices of the resources answer the question. The total imputed price of the ingredients in light beer is $(2 \times \$1) + (5 \times \$2) + (24 \times \$0)$, which is equal to $12. This measures the profit that would be lost by diverting resources from the brewing of ale and beer to the production of one barrel of light beer. Thus for the brewing of light beer to be worthwhile it must yield a profit of at least $12 a barrel.

The role of marginal prices in judging

the wisdom of introducing a new product can help to explain how the simplex method determines which edges of a polytope lead from a given vertex to a vertex with an improved value of the objective function. Suppose the vertex currently under examination is the one at (0,32) in the brewery problem. First one can determine the marginal price of each resource in this production program. Malt and hops are in excess supply at this point and so their marginal values are zero. Corn, however, is in short supply and so its marginal value is positive. Since only beer is being made at (0,32), the marginal value of the ingredients in a barrel of beer should be equal to the profit associated with beer: $23. Since only corn has nonzero marginal value and since 15 pounds of corn are needed for each barrel of beer, the marginal value of corn for this production program is $23 divided by 15 pounds, or about $1.53 per pound.

What is the marginal value of the ingredients in ale measured at the (0,32) vertex? If this value is less than the profit that can be realized from selling a barrel of ale, the diversion of some resources from beer production to ale would increase the total profit of the brewer. Making a barrel of ale requires five pounds of corn. Therefore the marginal value of the ingredients in a barrel of ale is ($23/15) × 5, or about $7.67. As in the example of light beer, the marginal value represents the loss in profit that would result from diverting resources from beermaking to the brewing of ale. Because this value is less than $13 (the profit expected from selling each barrel of ale) it would be profitable to brew more ale than the production program at the point (0,32) specifies. Indeed, by keeping the total consumption of corn constant and diverting corn from beer to ale, the brewer can increase his profit by $13 − $7.67, or $5.33, for each barrel of ale. Since the amount of corn is held constant, producing more ale corresponds to moving along the edge of the feasible region that represents the con-




**MARGINAL INCREASES in the availability of scarce resources alter the potential profits of a hypothetical brewer in a predictable way. If one additional pound of corn were available, the maximum profit would increase by $1, a change in the objective function that is reflected in the movement of the maximum vertex in the feasible region from the colored dot to the colored circle (*upper graph*). If there were an extra ounce of hops, the maximum profit would increase by $2 (*middle graph*). A change in the available malt would not alter the attainable profit: malt is already a surplus resource (*lower graph*). The changes in profit associated with changes in the availability of each resource are known as shadow prices or imputed prices. They are used to direct the algorithm from vertex to vertex. The marginal changes are exaggerated for clarity.**

# THE LEADING EDGE

## #1 in a series of reports on new technology from Xerox

*About a year ago, Xerox introduced the Ethernet network—a pioneering new development that makes it possible to link different office machines into a single network that's reliable, flexible and easily expandable.*

*The following are some notes explaining the technological underpinnings of this development. They are contributed by Xerox research scientist David Boggs.*

The Ethernet system was designed to meet several rather ambitious objectives.

First, it had to allow many users within a given organization to access the same data. Next, it had to allow the organization the economies that come from resource sharing; that is, if several people could share the same information processing equipment, it would cut down on the amount and expense of hardware needed. In addition, the resulting network had to be flexible; users had to be able to change components easily so the network could grow smoothly as new capability was needed. Finally, it had to have maximum reliability—a system based on the notion of shared information would look pretty silly if users couldn't get at the information because the network was broken.

**Collision Detection**

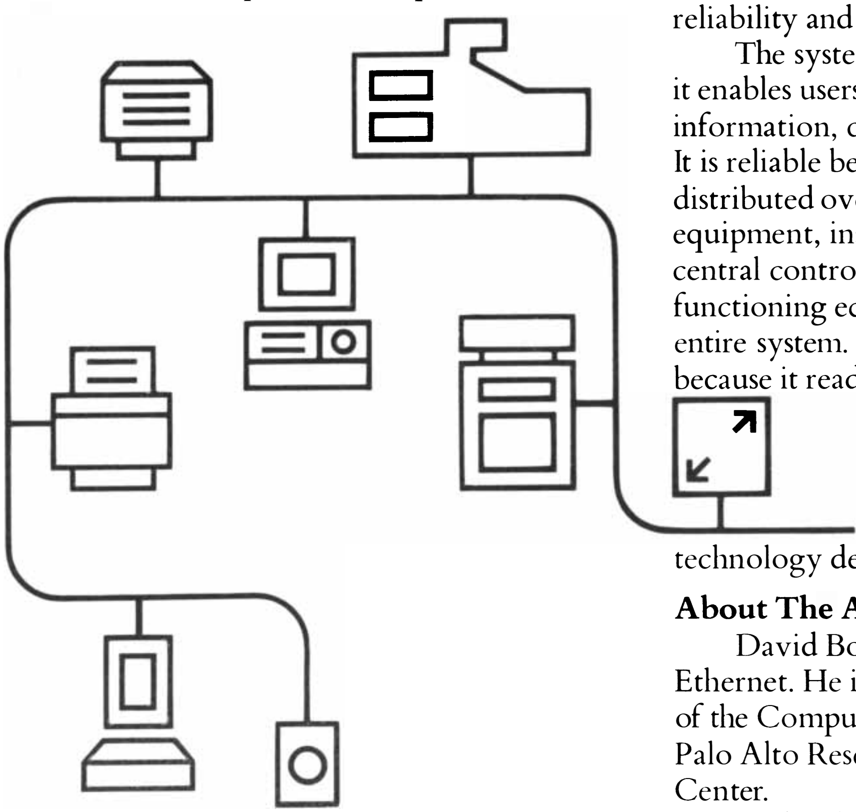
The Ethernet network uses a coaxial cable to connect various pieces of information equipment. Information travels over the cable in packets which are sent from one machine to another.

A key problem in any system of this type is how to control access to the cable: what are the rules determining when a piece of equipment can talk? Ethernet's method resembles the unwritten rules used by people at a party to decide who gets to tell the next story.

While someone is speaking, everyone else waits. When the current speaker stops, those who want to say something pause, and then launch into their speeches. If they *collide* with each other (hear someone else talking, too), they all stop and wait to start up again. Eventually one pauses the shortest time and starts talking so soon that everyone else hears him and waits.

When a piece of equipment wants to use the Ethernet cable, it listens first to hear if any other station is talking. When it hears silence on the cable, the station starts talking, but it also listens. If it hears other stations sending too, it stops, as do the other stations. Then it waits a random amount of time, on the order of microseconds, and tries again. The more times a station collides, the longer, on the average, it waits before trying again.

In the technical literature, this technique is called carrier-sense multiple-access with collision detection. It is a modification of a method developed by researchers at the University of Hawaii and further refined by my colleague Dr. Robert Metcalfe. As long as the interval during which stations elbow each other for control of the cable is short relative to the interval during which the winner uses the cable, it is very efficient. Just as important, it requires no central control—there is no distinguished station to break or become overloaded.



**The System**

With the foregoing problems solved, Ethernet was ready for introduction. It consists of a few relatively simple components:

Ether. This is the cable referred to earlier. Since it consists of just copper and plastic, its reliability is high and its cost is low.

Transceivers. These are small boxes that insert and extract bits of information as they pass by on the cable.

Controllers. These are large scale integrated circuit chips which enable all sorts of equipment, from communicating typewriters to mainframe computers, regardless of the manufacturer, to connect to the Ethernet.

The resulting system is not only fast (transmitting millions of bits of information per second), it's essentially modular in design. It's largely because of this modularity that Ethernet succeeds in meeting its objectives of economy, reliability and expandability.

The system is economical simply because it enables users to share both equipment and information, cutting down on hardware costs. It is reliable because control of the system is distributed over many pieces of communicating equipment, instead of being vested in a single central controller where a single piece of malfunctioning equipment can immobilize an entire system. And Ethernet is expandable because it readily accepts new pieces of information processing equipment. This enables an organization to plug in new machines gradually, as its needs dictate, or as technology develops new and better ones.

**About The Author**

David Boggs is one of the inventors of Ethernet. He is a member of the research staff of the Computer Science Laboratory at Xerox's Palo Alto Research Center.

He holds a Bachelor's degree in Electrical Engineering from Princeton University and a Master's degree from Stanford University, where he is currently pursuing a Ph.D.

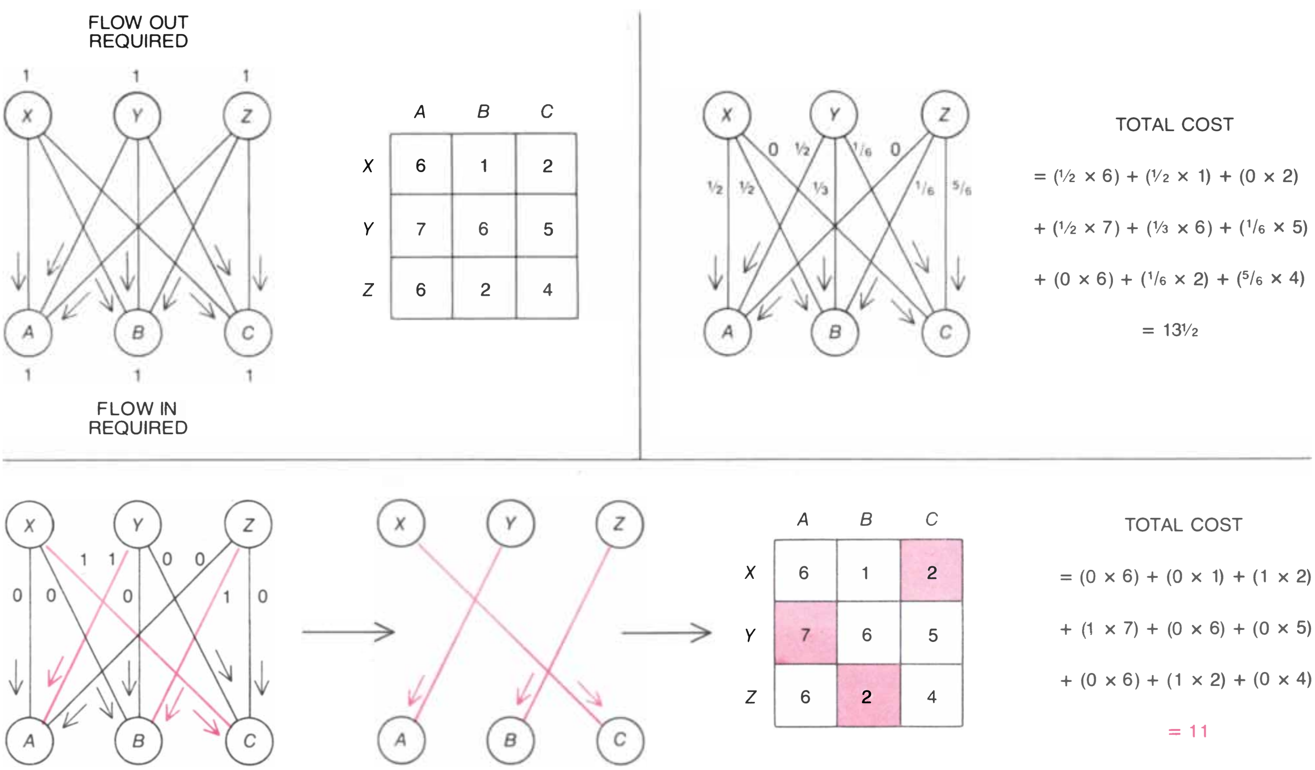
straint on the supply of corn. This is precisely how the simplex method recognizes that the objective function can be improved along this edge. In this case the edge connects the vertex (0,32) with the vertex (12,28) of the polygon.

Even in small problems of linear programming the feasible region can have an enormous number of vertexes, and there is the possibility that the simplex method will require an enormous number of iterations, or vertex-to-vertex moves, in order to solve the problem. Usually, however, the selective search encounters only a vanishingly small fraction of all the vertexes. People who regularly solve problems having from 2,000 to 3,000 resource constraints and from 10,000 to 15,000 decision variables find that such problems generally take only from a few minutes to a few hours to run on a large computer. One would like to be able to make a more precise statement concerning the efficiency of the simplex method.

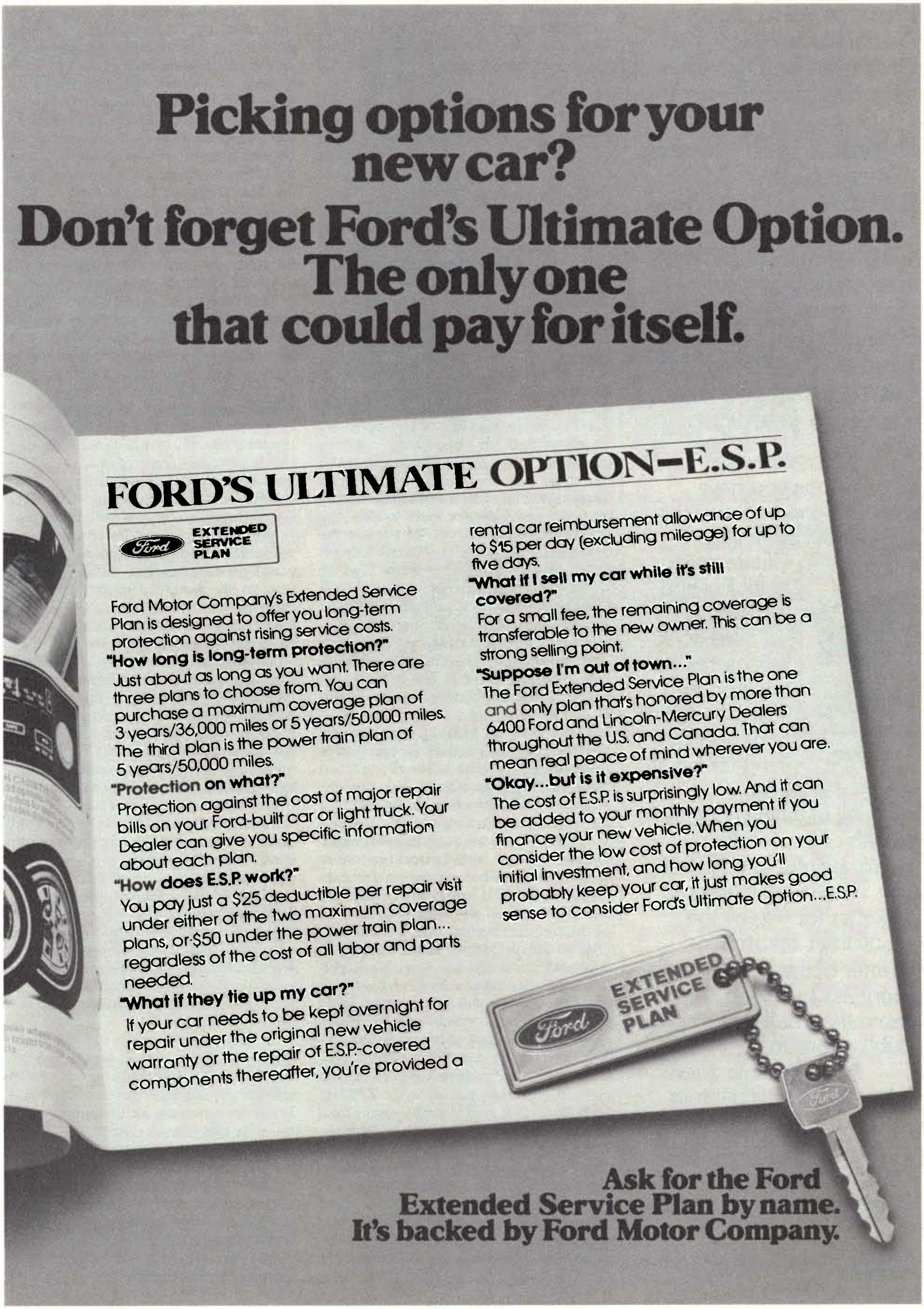
One way of convincing a skeptic that an algorithm is efficient is to provide some kind of guarantee. The guarantee takes the form of a function $G$ and a statement that on a problem of size $s$ the running time of the algorithm will not be greater than $G(s)$. The size $s$ of a problem is a measure of the amount of information (often expressed as a number of binary digits) needed to specify the input data. This is called a worst-case analysis, since the value of the guarantee is determined by the most troublesome problems of size $s$ the algorithm might be asked to solve. The running time actually required to solve a problem of size $s$ depends on the characteristics of the particular computer. To make the measure $G(s)$ machine-independent the function is usually expressed not in minutes or hours but as the maximum number of elementary arithmetical operations (such as addition, multiplication and comparison) that must be done by the algorithm.

If $G(s)$ is appropriately small for the values of $s$ one expects to encounter, one would be inclined to acknowledge that the proposed algorithm is efficient. For example, Harold W. Kuhn of Princeton University has devised an algorithm for the assignment problem that requires at most $cn^3$ elementary operations to solve any $n$-by-$n$ problem, where $c$ is a small constant. When $n$ is equal to 2, 3, 4 or 5, the algorithm can actually take longer than enumeration (although with a computer neither method would take longer than the blink of an eye). Once $n$ increases beyond this range, $n!$, which measures the size of the enumeration, quickly surpasses $cn^3$. In fact, $n!$ grows so explosively that, as I have shown, no existing computer could carry out the computation. On the other hand, $n^3$ grows more slowly as $n$ increases. For $n$ equal to 35 Kuhn's algorithm can still solve the problem in the blink of an eye. Furthermore, as $n$ increases from 35 to 36, the computational burden on Kuhn's algorithm increases by a factor of about $(36/35)^3$, or approximately 1.09. The work required for enumeration increases by a factor of 36.

Computer scientists have a particular interest in algorithms that are said to be polynomially bounded, where the function $G$ is a polynomial function of $s$. A polynomial function of $s$ is a sum of various terms; in each term $s$ is raised to some constant power and multiplied by a constant coefficient. If $s$ is sufficiently large, any algorithm whose worst-case guarantee increases as $s!$ or as any exponential function of $s$, say $2^s$, will run more slowly than any polynomi-




**NETWORK-FLOW PROBLEMS are a special class of problems called integer-programming problems that yield to linear-programming techniques. In an integer-programming problem the decision variables are not divisible: they must assume integer values only. In general such problems are in the set *NP*-complete and so are regarded as hard. The cost of each link of the network (*diagram at upper left*) is given by the matrix of numbers. Note that the matrix is identical with a possible matrix for a 3-by-3 assignment problem. A solution of the assignment problem must be integer-valued, however, and it appears that the constraints on the network flow are not sufficient to guarantee this result. Many fractional flows meet the constraints (*diagram at upper right*). Nevertheless, extremal solutions of network-flow problems found by linear-programming methods are always integer-valued, provided that the flows both into and away from each node are integer-valued. Hence the assignment problem can be interpreted as a special case of the network-flow problem (*lower diagram*).**


Picking options for your new car?
Don't forget Ford's Ultimate Option.
The only one that could pay for itself.
FORD'S ULTIMATE OPTION–E.S.P.
Ford EXTENDED SERVICE PLAN
Ford Motor Company's Extended Service Plan is designed to offer you long-term protection against rising service costs.
"How long is long-term protection?"
Just about as long as you want. There are three plans to choose from. You can purchase a maximum coverage plan of 3 years/36,000 miles or 5 years/50,000 miles. The third plan is the power train plan of 5 years/50,000 miles.
"Protection on what?"
Protection against the cost of major repair bills on your Ford-built car or light truck. Your Dealer can give you specific information about each plan.
"How does E.S.P. work?"
You pay just a $25 deductible per repair visit under either of the two maximum coverage plans, or $50 under the power train plan... regardless of the cost of all labor and parts needed.
"What if they tie up my car?"
If your car needs to be kept overnight for repair under the original new vehicle warranty or the repair of E.S.P.-covered components thereafter, you're provided a rental car reimbursement allowance of up to $15 per day (excluding mileage) for up to five days.
"What if I sell my car while it's still covered?"
For a small fee, the remaining coverage is transferable to the new owner. This can be a strong selling point.
"Suppose I'm out of town..."
The Ford Extended Service Plan is the one and only plan that's honored by more than 6400 Ford and Lincoln-Mercury Dealers throughout the U.S. and Canada. That can mean real peace of mind wherever you are.
"Okay...but is it expensive?"
The cost of E.S.P. is surprisingly low. And it can be added to your monthly payment if you finance your new vehicle. When you consider the low cost of protection on your initial investment, and how long you'll probably keep your car, it just makes good sense to consider Ford's Ultimate Option...E.S.P.
Ford EXTENDED SERVICE PLAN
Ask for the Ford Extended Service Plan by name.
It's backed by Ford Motor Company.

ally bounded algorithm. Consequently computer scientists consider polynomial boundedness a theoretical criterion of algorithmic efficiency.

Because the simplex method performs well in practice one might guess it is a polynomially bounded algorithm. Indeed, practitioners have found that a good rule of thumb is that the number of simplex iterations increases as a linear function of the number of constraints. In standard implementations, however, the simplex method is not polynomially bounded. Victor La Rue Klee, Jr., of the University of Washington and George J. Minty, Jr., of Indiana University have constructed an infinite family of linear-programming problems in which the simplex method is essentially as bad as enumeration.

That an algorithm so successful in practice could be regarded as other than efficient in the theoretical sense may seem perplexing. Keep in mind, however, that the theoretical criterion examines worst-case results. Examples like those first constructed by Klee and Minty that cause the simplex method to perform badly do not seem to occur in practice, at least not yet. To explain the divergence between practical experience with the simplex method and its worst-case performance, investigators have attempted to demonstrate that in some probabilistic sense the bad examples really are "pathological." The demonstration faces considerable difficulties, although Dantzig has recently reported some interesting results in this direction.

There is some flexibility in how the simplex method chooses at each iteration among several adjacent vertexes when each vertex would improve the value of the objective function. The Klee-Minty construction shows that one popular strategy leads to nonpolynomial results. Other investigators have since constructed similar examples for other strategies. It is not known, however, whether there is some implementation of the simplex method that is polynomially bounded. Indeed, until recently it was not known whether there is any polynomially bounded algorithm for linear programming; it was an unsolved problem of great theoretical interest. Finally L. G. Khachian of the Academy of Sciences of the U.S.S.R. showed how an algorithm developed by three other Russian mathematicians, N. Z. Shor, D. B. Iudin and A. S. Nemirovskii, could be implemented in polynomial time.

Because the work of Shor, Iudin and Nemirovskii does not assume linearity the new algorithm differs noticeably from the simplex method. Their algorithm is the one called the ellipsoid method. It begins with a very large ellipsoid centered at the origin. In general the initial ellipsoid is a sphere with a radius large enough to ensure that it contains the optimum solution. The algorithm proceeds to construct progressively smaller ellipsoids, each of which is also known to contain the optimum solution. The volume of the successive ellipsoids shrinks quickly enough to guarantee that their centers converge on the optimum solution. One need not draw the ellipsoids in order to employ the algorithm; there are simple recursive formulas that describe them. When one accounts for the work needed to construct each ellipsoid from its predecessor, the bound on the total number of operations is on the order of $(m + n)n^6l$, where $m$ is the number of constraints, $n$ is the number of decision variables and $l$ is the logarithm to the base 2 of the largest coefficient in the problem data.

This expression is a polynomial bound, but, like any such bound, it can still grow rather large for large values of $m$, $n$ and $l$. Could it be that the ellipsoid method, like the simplex method, does much better in practice than its worst-case bound would predict? Preliminary results are not encouraging. Computational testing so far has shown extremely slow convergence for the ellipsoid method. Remarkably simple families of problems can cause the method to run as long as the bound predicts. This is in contrast with the complicated form of the problems that have been constructed to defeat the simplex method.

Another major difficulty of the ellipsoid method concerns memory requirements. Very large linear-programming problems tend in practice to be sparse: the proportion of nonzero coefficients in the data is usually small. The simplex method can be implemented so that sparseness is maintained as the data are manipulated during the execution. This is a crucial property, since one need only store in the computer the nonzero entries. There is no apparent way to exploit sparseness in the ellipsoid method.

Although the prospects for a practical implementation of the ellipsoid method seem dim at the moment, one must see what develops. In the short time since Khachian's note considerable work has been done, with some substantial improvements. The difficulties noted here do appear formidable.

Linear programming is a practical tool of great theoretical interest. To the mathematician it constitutes a theory of linear inequalities, an elegant counterpoint to the classical theory of linear equations. To the government and corporate planner it is a valuable aid in decision making and long-range planning. Perhaps the most important consequence of Khachian's theoretical breakthrough will be increased interest in both practical and theoretical aspects of linear programming.

# FROM HERE TO SERENDIPITY.

In the land of Serendip, so an ancient fairy tale says, people who travelled unknown roads made wonderful discoveries when they least expected to.

To some, the fable of Serendip is just another bedtime story, but at the IBM research laboratories it's a story that hits close to home.

Recently, for instance, IBM scientists were experimenting with X-rays to explore new ways to make microscopic computer circuits.

A colleague wondered if this new technique might also help medical researchers look at living cells.

One thing led to another.

Working with biologists from the National Institutes of Health, the IBM team tested their idea by photographing an ordinary human blood platelet at a resolution of 1/5,000,000th of-an-inch.

What they saw stopped them cold.

For they found themselves staring at internal cell structures no one had ever seen before. Structures that may tell us more about how cells work.

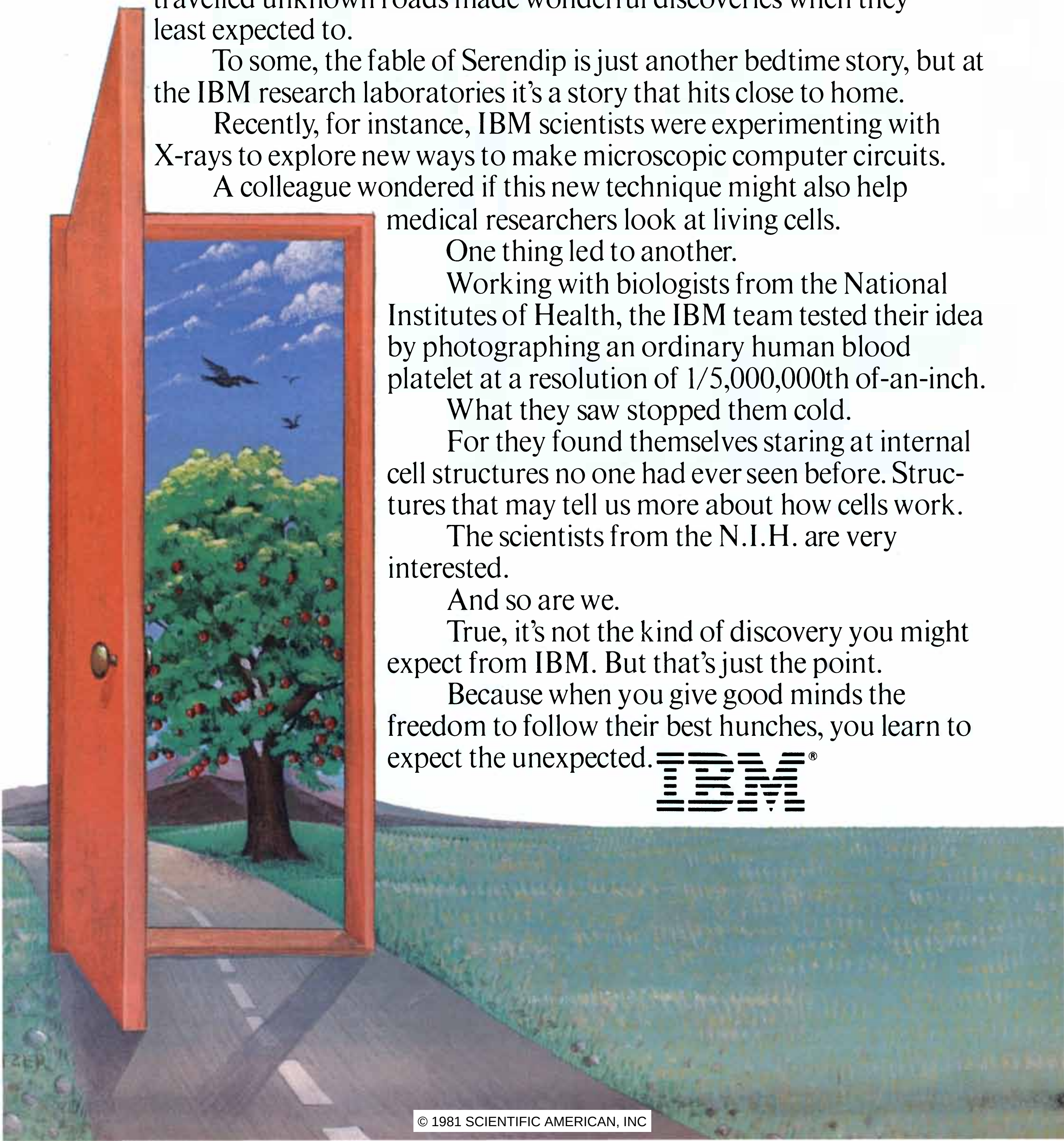
The scientists from the N.I.H. are very interested.

And so are we.

True, it's not the kind of discovery you might expect from IBM. But that's just the point.

Because when you give good minds the freedom to follow their best hunches, you learn to expect the unexpected.

IBM®



**HONEYBEE SWARM** forms on the branch of a tree. A typical swarm consists of a queen and several thousand other bees, representing about half of the population of a hive they have left abruptly in the late spring in order to establish a new colony. The conelike shape and close packing of bees in this swarm are characteristic and will be maintained for hours or days until a new nest site has been found.

# The Regulation of Temperature in the Honeybee Swarm

*Honeybees must be warm in order to fly. Therefore the temperature of an immobile swarm must be closely controlled so that when scout bees have found a new nest site the swarm is ready for a fast takeoff*

by Bernd Heinrich

The swarming of honeybees is a late-spring phenomenon in which a queen and several thousand other bees depart abruptly from a hive, leaving another queen and the remainder of the residents to stay on. In a matter of hours or a few days the swarm establishes a domicile of its own in a new place. In the interlude, however, the bees behave strangely. The swarm coalesces into a beardlike mass on a tree branch or some other firm object. It remains there virtually immobile except for a few scouts that go off to search for a suitable living place, such as a hollow tree. If the swarm is shaken, most of the bees drop to the ground. They are too cool to fly: their flight muscles cannot function until they are warmed up. Yet when the time comes to move to the new home, the swarm is airborne in less than a minute. Apparently the swarm regulates its temperature in some way. I have investigated the matter in both free and captive swarms and have found that the bees have a remarkably fine-tuned system of thermoregulation. On reflection one can see that in a Temperate Zone climate such a system is essential to the swarm's ultimate success in establishing a viable hive.

The primary function of swarming in the "European" honeybee *Apis mellifera* (it was imported to North America from Europe) is the reproduction of colonies. Usually the queen and about half of the workers of a colony leave when the hive becomes overcrowded. A daughter queen (perhaps more than one) is left behind.

A related phenomenon is "absconding," in which the entire colony leaves if its domicile becomes unsuitable. Some species of honeybee abscond more readily than others. The difference may lie in the adaptation of the bees to their climate. In the Asian Tropics the giant honeybee (*Apis dorsata*) and the dwarf honeybee (*A. florea*), which normally build their hanging combs in the open, abscond frequently, probably in order to move to areas where food is more abundant. The third tropical honeybee, the Indian bee (*A. indica* or *A. cerana*), builds its combs in cavities and also absconds readily, as does the African race of *A. mellifera*, which is the "killer bee" that has spread northward at a rate of about 50 kilometers per year since its introduction into South America.

Warm climate explains why the three tropical species and the southern races of *A. mellifera* abscond so readily. The colonies have much to gain by taking advantage of shifting food resources, and they do not need to lay up large stores of food to survive a Temperate Zone winter. To be this mobile means to travel light and not to spend much time choosing a site for a nest.

The honeybee in the north Temperate Zone, however, has to be quite discriminating about the kind of cavity it occupies. The opening must be small, and the size of the cavity must be such that the swarm can control its temperature during the winter. Moreover, once a cavity has been occupied the bees have to lay up much larger stores of food than they can carry with them, since the food is in effect the fuel that heats the hive. The bees must therefore have the social responses necessary for managing their food resources with maximum efficiency. They will not be able to leave the nest or the area when conditions are the worst—in the winter. Such bees are adapted for staying rather than for moving. Their swarming is focused more on reproduction than on finding better housing.

No amount of stored food and no social response in managing the internal environment of the hive can prevail long against the wind and cold of the northern winter if the bees are not in a suitable shelter. Finding, evaluating and competing to occupy suitable domiciles is therefore central to the biology of north Temperate Zone honeybees. It calls on a number of behavioral adaptations, in all of which thermoregulation is of primary importance. It is instructive to consider the adaptations as a backdrop to the thermoregulation.

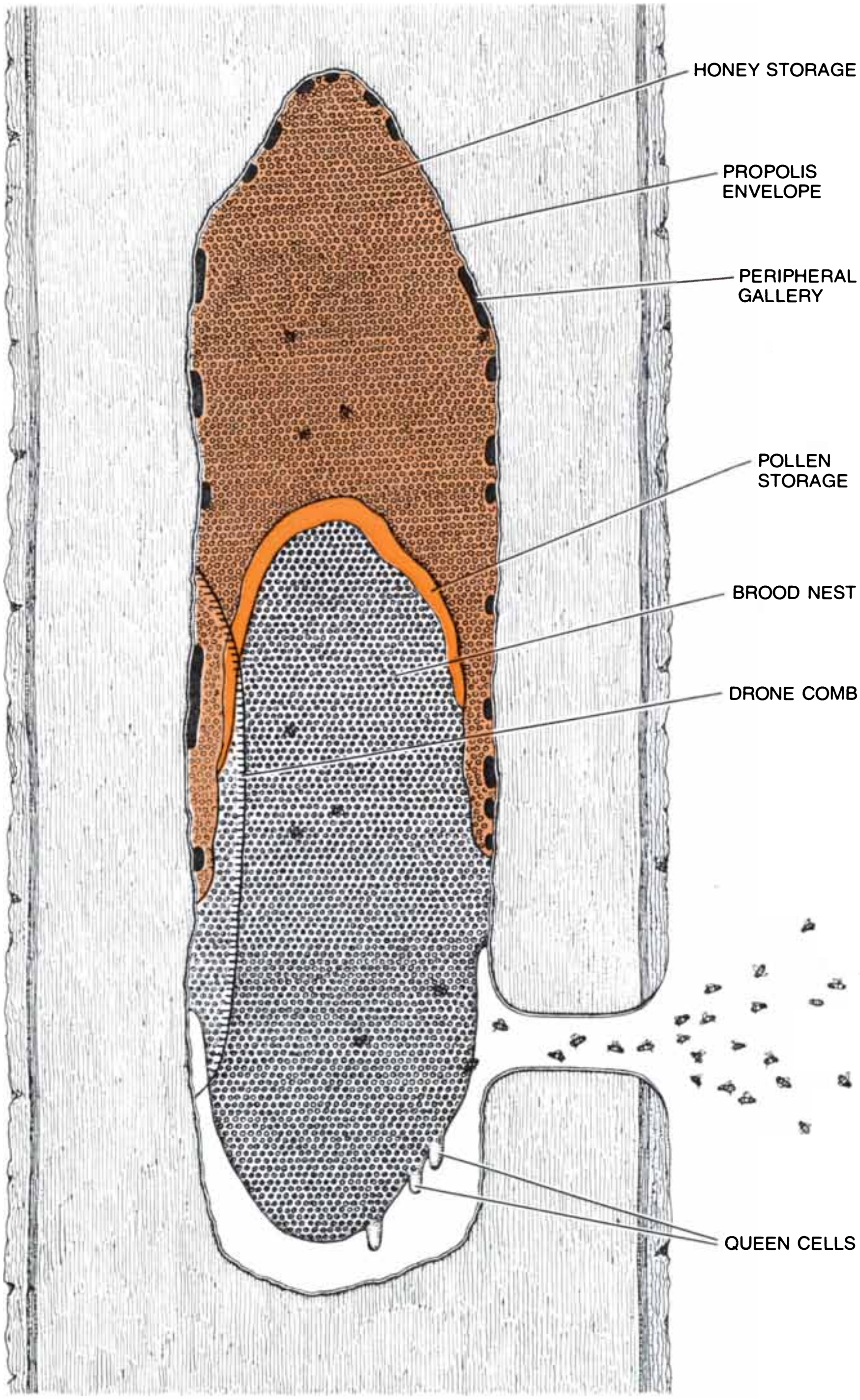
An important prerequisite to success in swarming is the timing. The reason swarms usually issue from hives in the spring is that the bees then have time to find a suitable place to live, to occupy it, to fill it with honeycomb and to stock the comb with stores of honey for the coming winter.

While the queen and most of the other bees in a newly formed swarm hang almost motionless from some object that is usually near the hive they have left, several scout bees begin the search for new quarters. They make up the small part (less than 5 percent) of the total population that seems to be primarily concerned with exploring rather than with the work of the hive. In an established hive the scouts are usually the bees that look for new food sources, such as flowers coming into bloom and weak hives that can be robbed. In the swarm they are the bees that look for potential hive sites.

The importance of the kind of site they find is suggested by studies that Thomas D. Seeley of Yale University has made of swarms near Ithaca, N.Y. He has found that the highest rate of mortality in the hive is during the winter and that the death rate is inversely correlated with the quality of the nest cavity. He has also found that scout bees give close attention to the size of the opening to a potential nest site, to the internal temperature and to internal convection currents. The scouts pace off the interior dimensions of the cavity. According to Seeley's measurements, the volume of the cavity chosen as a domicile is usually between 20 and 100 liters (roughly between a cubic foot and four cubic feet).

Martin Lindauer of the University of Würzburg studied honeybee swarms in Munich after World War II when he was working at the Munich Zoological Institute. He found that most of the swarms chose cavities in the ruins that then abounded in the city. Eventually he was able to predict which potential hive a given swarm would occupy.

Lindauer observed that scout bees, some of which he had marked, would sooner or later begin to "dance" on the surface of a clustered swarm, in the same way that dancers in an established hive indicate the distance and direction to potential sources of food. Sometimes several scouts simultaneously advertised different cavities. Before dancing, a scout visited a cavity several times; it would stop advertising the site if the cavity became too wet or too hot or changed unfavorably in some other way. The scouts also were attentive to one another's dance, visiting the other potential home sites indicated by the dances. Eventually all the scouts repeatedly visited and advertised the same site.




**TYPICAL NEST SITE in a partly hollow tree is depicted on the basis of a number of natural nests inspected by Thomas D. Seeley of Yale University and Roger A. Morse of Cornell University. It is such a site that scout bees from a swarm examine carefully and repeatedly while the swarm waits in its characteristic clustered form. Once the scouts have reached a consensus on a site they alert the swarm, which is able to take off quickly because the bees have maintained its core temperature near 35 degrees Celsius, the optimum level for flight. The swarm flies to the site, moves in rapidly and begins to build the various structures portrayed here.**

At this stage the question becomes how the swarm is aroused and guided to the site. Seeley and his colleagues observed marked scouts that had been investigating an empty hive and dancing on the swarm. About half an hour before the swarm lifted off the scouts made rapid buzzing runs into the hive. The runs appear to be an alerting signal and are not always followed by a takeoff of the swarm, which suggests that the almost instantaneous takeoff may be prompted by another signal.

Once a swarm has taken flight it moves slowly at first in a cylindrical formation some 10 meters in diameter. Presumably the flight is slow because the bees are checking for the odor of the queen substance: 9-oxodec-2-enoic acid. (A swarm may take off without a queen, but it will not travel far.) The swarm soon accelerates to a speed of about 11 kilometers per hour, but some bees, the "streakers," dart much faster through the group in the direction of the new cavity. Lindauer and Seeley think the streaker bees may be the scouts pointing the way.

The scouts also point the way chemically, particularly when the goal is near. They arrive at the entrance of the cavity ahead of the mass of bees and release scent from the Nassanoff glands at the tip of their abdomen. The scent appears to serve as a final marker for the swarm. In a few minutes the thousands of bees stream into the cavity. Within hours they are cleaning out debris, building comb and foraging for nectar and pollen. A new colony has been established.

If the scouts from a swarm are unsuccessful in finding a suitable cavity in a few days or do not reach a consensus on which of several alternatives is the best one, the workers start to build comb on the object from which they are hanging. The more they build, the less likely they are to leave, since the comb represents a substantial investment of effort. Open hives may survive in places where the winter is mild, but in the Temperate Zone the bees in an open hive invariably die during the winter because they are unable to keep warm without a suitably insulated shelter.

The measurement of temperature in beehives dates back at least a century. It has long been known that the hive serves as an incubator of the larvae and pupae. The bees maintain the temperature of the brood at 35 or 36 degrees Celsius (close to 95 degrees Fahrenheit). When the air temperature is low, the bees crowd on the brood and shiver; the rapid contraction of their muscles generates body heat. When the air temperature is high, the bees reduce the internal

Summer Vacation '81

Your **Q**uestions fully **A**nswered.

## How much should a vacation cost?

Ideally, any vacation you choose should be flexible enough to allow you to put your own package together, so you can really control costs. Start for instance, with a choice of 5 hotel categories, beginning as low as $10 per person per night. (You should also have the flexibility to change your itinerary once you're there.) Then add a choice of ground transportation, either a car from 8 categories, or rail (insist on unlimited mileage), that can be as low as $80 per week. Shop around for the best airfares. You can find peak season mid-week air fares that will get you into the heart of Europe for $681 round-trip from New York. Finally, make sure the prices are guaranteed. If all of this sounds too good to be true, you haven't seen The Swiss Travel Invention or The Austrian Travel Invention.

## Where should you go?

Go someplace where you can make the most of your time. A place where there's a high concentration of culture, activities and sights, and a place that's small enough, so it's all within easy reach. A place like the Alps, where the people genuinely like visitors and services are reliable. Everything is worked out in The Swiss Travel Invention, The Austrian Travel Invention and The Complete Summer Vacation.

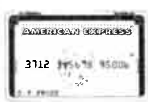
## What do you want to do?

If you like to travel on your own and see just the things you like, you need a reliable place, and an organized system of travel. This way, making accommodations will be trouble-free. Maybe you'd really like to rough it a bit and take to the outdoors, or take an escorted tour, where you don't have to bother with any planning details. In the Alps, the options are unlimited. With The Swiss Travel Invention, The Austrian Travel Invention and The Complete Summer Vacation, you can do it all.

## How much extra cash will you need?

With some packages, the sky's the limit, because there are so many hidden costs. Make sure the package you want clearly defines what is included, like meals and transfers. If it doesn't, scrap it and find one that does. Also, try to pre-pay as much as you can, at guaranteed prices, so there are no surprises when you arrive. This will limit the cash you need to personal expenses and gas, if you're traveling by car. If you're looking for a trip that spells it all out and is reasonably-priced, check The Swiss Travel Invention, The Austrian Travel Invention and The Complete Summer Vacation.

**The American Express® Card** Don't leave home without it.[SM] To make your vacation in Switzerland or Austria even easier, carry the American Express Card. With it, you can charge almost all of your expenses, and with the American Express "Sign & Fly®" Plan, you can extend your air fare payments.




For an application, call American Express toll-free (800) 528-8000.

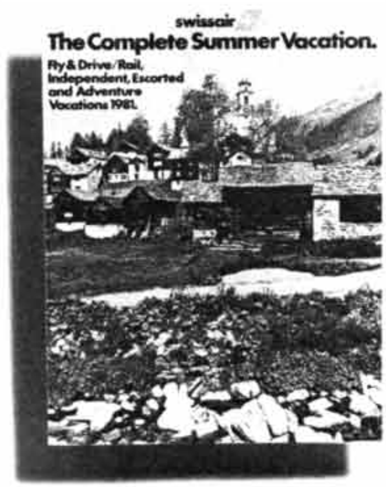

swissair

## What do you do about it?

Once you've decided what you want to do and where, go to someone who has a reputation for handling consistent quality arrangements, particularly someone who specializes in your area of interest. If our concept of a vacation sounds like yours, go to a professional travel agent or send in the coupon below to Swissair. If we don't have what you're looking for in the Alps, let us know.

- - - - - - - - -

☐ I'm interested in all of the different packages available in the Alps—independent travel, escorted, unescorted and adventure tours. Send me The Complete Summer Vacation.




☐ I'd like to travel on my own in Switzerland, plan my itinerary, and pre-pay my land costs and pre-select air fare, all at a price I choose. Send me The Swiss Travel Invention.

☐ I'd like to do all of the above in Austria. Send me The Austrian Travel Invention.

Name __________

Address __________

City __________

State ______ Zip ______

Tel. __________ SA-6

Mail to: Swissair
Box 845, Radio City Station
New York, N.Y. 10101

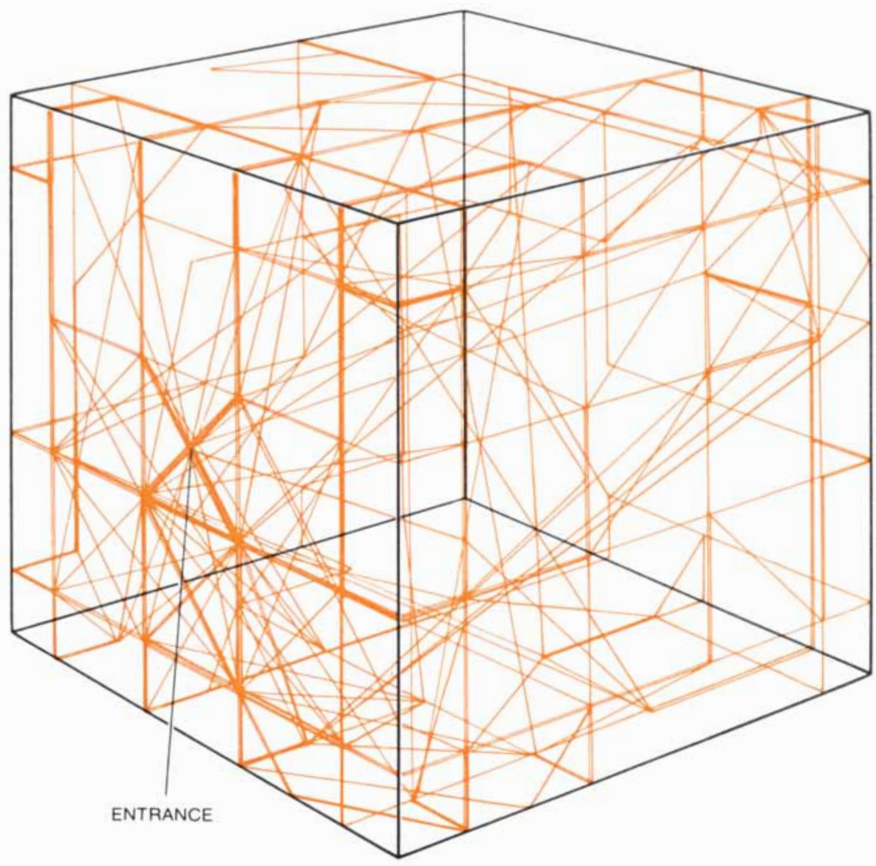
temperature by fanning and by evaporative cooling. If the brood is allowed to cool below 30 degrees C., the larvae or pupae die or emerge with developmental defects such as deformed wings. The temperature is not rigidly controlled in sections of the hive that do not contain brood, and so thermoregulation has been thought to be needed primarily for the immature bees.

A swarm consists entirely of adult bees without eggs, larvae, pupae or comb, suggesting that thermoregulation would be absent. My preliminary measurements showed, however, that the core of the swarm is generally maintained at almost the same temperature as the temperature of the brood in a hive and either remains constant or increases when the outside temperature decreases. These observations raised two questions I sought to answer. First, how do the bees in the core of the cluster regulate their body temperature with respect to changing external temperature when they are individually experiencing the relatively constant or increasing temperature of the core? Second, what is the functional significance of thermoregulation in the swarm?

I captured many swarms by shaking them from their branches and brought them into the laboratory to assess their temperature and metabolic rate in large respirometry vessels. The lid of such a vessel has many small holes that serve for ventilation, as a foothold for the bees (enabling them to form their swarm cluster under the lid) and as openings for the insertion of measuring devices. When I measured respiration, I sealed a second lid of solid Plexiglas over the perforated lid; air from the sealed system was circulated through an oxygen analyzer to determine the rate of expenditure of energy, which is to say the production of heat.

I also observed free swarms that were maintained on a window ledge outside my laboratory at the University of California at Berkeley. In some of them I had implanted a dozen or more thermocouples, the leads from which came into the laboratory so that I could record the reading of each thermocouple continuously for hours or days on a strip-chart recorder. The free swarms responded to ambient conditions, but with the captive swarms I could vary the ambient temperature from zero to 30 degrees C.

All the swarms showed steep temperature gradients from the core to the mantle, or perimeter. The highest temperatures were near the core. Both small swarms of 2,000 bees and swarms 15 times larger maintained a core tempera-




**PATTERN OF MOVEMENT of a single scout bee on several inspections of a cubical box is represented in a somewhat simplified form on the basis of a reconstruction made by Seeley. The colored lines show the paths of the bee; they have been straightened for simplification. Seeley employed a grid system that enabled him to keep track of the position of the exploring bee.**

# INTRODUCING DODGE MINI RAM. MORE MILEAGE, RANGE AND SEATS THAN VW VANAGON FOR THOUSANDS LESS.

Who moves more people than VW for less money? It's Ram Tough Dodge's new Mini Ram Wagon.

Compare Dodge Mini Ram 250 to VW Vanagon—passenger seating, mileage, fuel capacity and range, horsepower and rear loading area. Mini Ram gives you more.

Then compare factory-installed options like power steering, and your choice of 6 sound systems: all available in Mini Ram. Not in Vanagon.

Now compare price. Mini Ram costs $2,553 less than Vanagon. For panel-side van buyers, there's a Dodge Van with the same performance, efficiency and interior dimensions as

**Here's how Mini Ram beats Vanagon**

| | Dodge Mini Ram Wagon 250 | VW Vanagon |
|---|---|---|
| Sticker Price (MSRP at base) | **$8,137.00*** | **$10,690.00*** |
| EPA EST. MPG | 18 MPG** | 17 MPG** |
| Horsepower | 95 | 67 |
| Passenger Seating | 8 | 7 |
| Overall Length | 178.9" | 179.9" |
| Wheelbase | 109.6" | 95.8" |
| Fuel Capacity | 36 gal. | 15.9 gal. |
| EPA Est. Range | 648 miles** | 270 miles** |
| Side Door Width | 49" | 42" |
| Removable Rear Seat | Yes | No |
| Factory-Installed Air Conditioning | Option | Not Available |
| Power Steering | Option | Not Available |
| Color-Keyed Carpeting | Standard | $410 with int. pkg. |

Mini Ram Wagon at a price VW hasn't seen in years—just $6,418.*

Underneath it all, Mini Ram is a Ram Tough Dodge Wagon. With the best rust protection of all wagons. There's 100% electrocoating, plus 370 sq. ft. of galvanized steel in critical areas... compared to VW's 4 sq. ft.

New Dodge Mini Ram Wagon. It does more than VW Vanagon and it does it for less. As a pure people mover, it just might be the best buy on the road today.

Dodge Trucks

**DODGE TRUCKS ARE RAM TOUGH**

*Sticker price comparison, excluding title, taxes and destination charges, unadjusted for different levels of standard equipment. Two-tone paint and whitewall tires $340 extra.

**Use these numbers for comparison. Your mileage and range may differ depending on speed, distance and weather. Calif. est. lower.

ture of about 35 degrees C. Although the core temperature varied considerably, it was independent of the ambient temperature. Since the rate of passive cooling in a swarm is a function of the surface area and the ambient temperature, the results showed clearly that the core temperature is actively regulated. This was a surprising result because the core includes only a small part of the total population of the swarm.

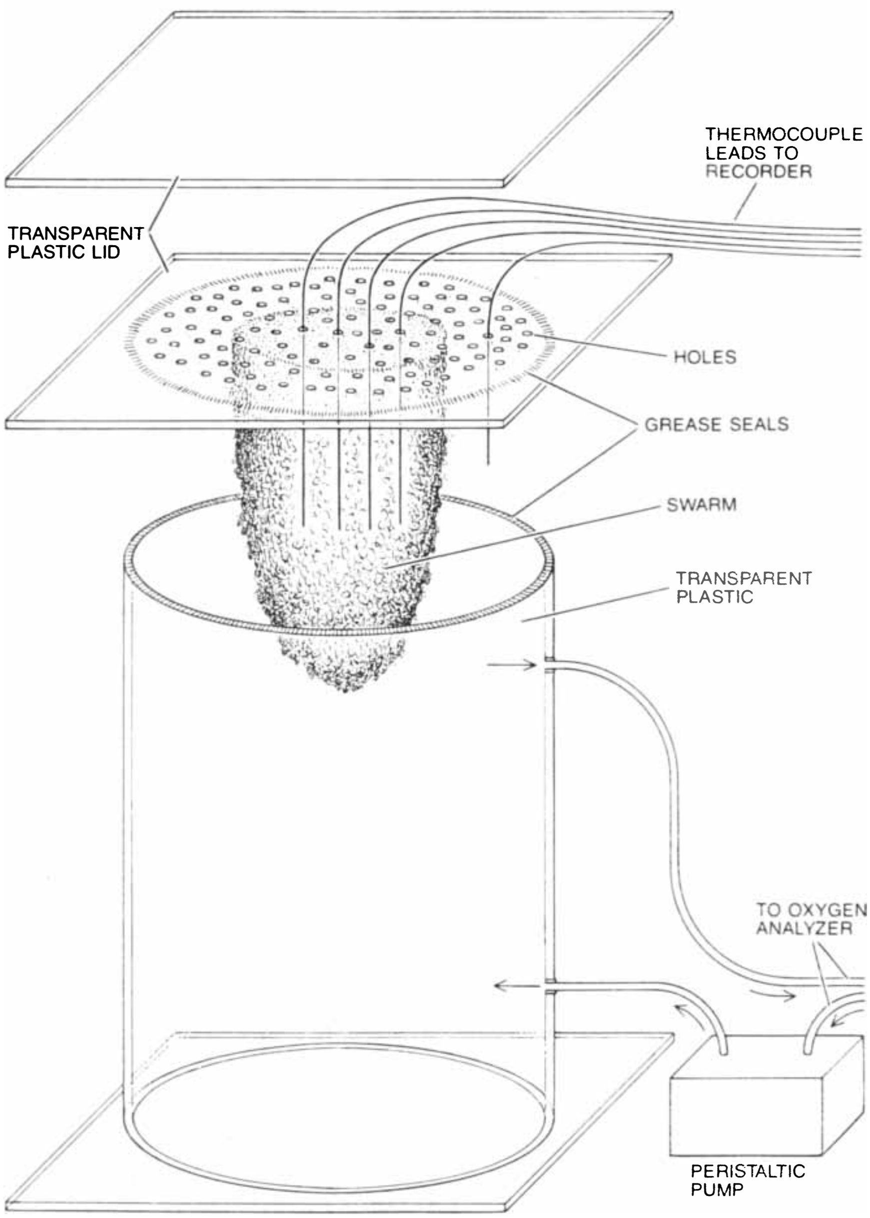
The mantle of the swarms was usually maintained at a much lower temperature. If the ambient temperature went down, the temperature of the mantle followed, remaining two or three degrees higher. This pattern appeared, however, only when the ambient temperature was relatively high. If it went below 20 degrees C., the temperature of the mantle stabilized. These results show that the mantle temperature is regulated in such a way that it remains close to the ambient temperature but does not go below about 17 degrees. The temperature of the rest of the swarm usually varied in a fairly smooth gradient between the 36-degree maximum of the core and the 17-degree minimum of the mantle.

Such were the general patterns revealed by many measurements, but I also found instructive variations and exceptions. For example, the captive swarms often had lower temperatures than the free ones, presumably because there is less activity in a captive swarm. When a small swarm was maintained at five degrees C. or less for several days, bees hung in seemingly lifeless chains from the tightly contracted cluster. The body temperature of these bees was close to the ambient temperature and the bees eventually dropped off. Most of them had an empty honey crop, possibly because they had exhausted their food reserves by shivering. The bees that remained on the mantle of the swarm, maintaining a body temperature of about 17 degrees C., were able to crawl but were not able to fly.

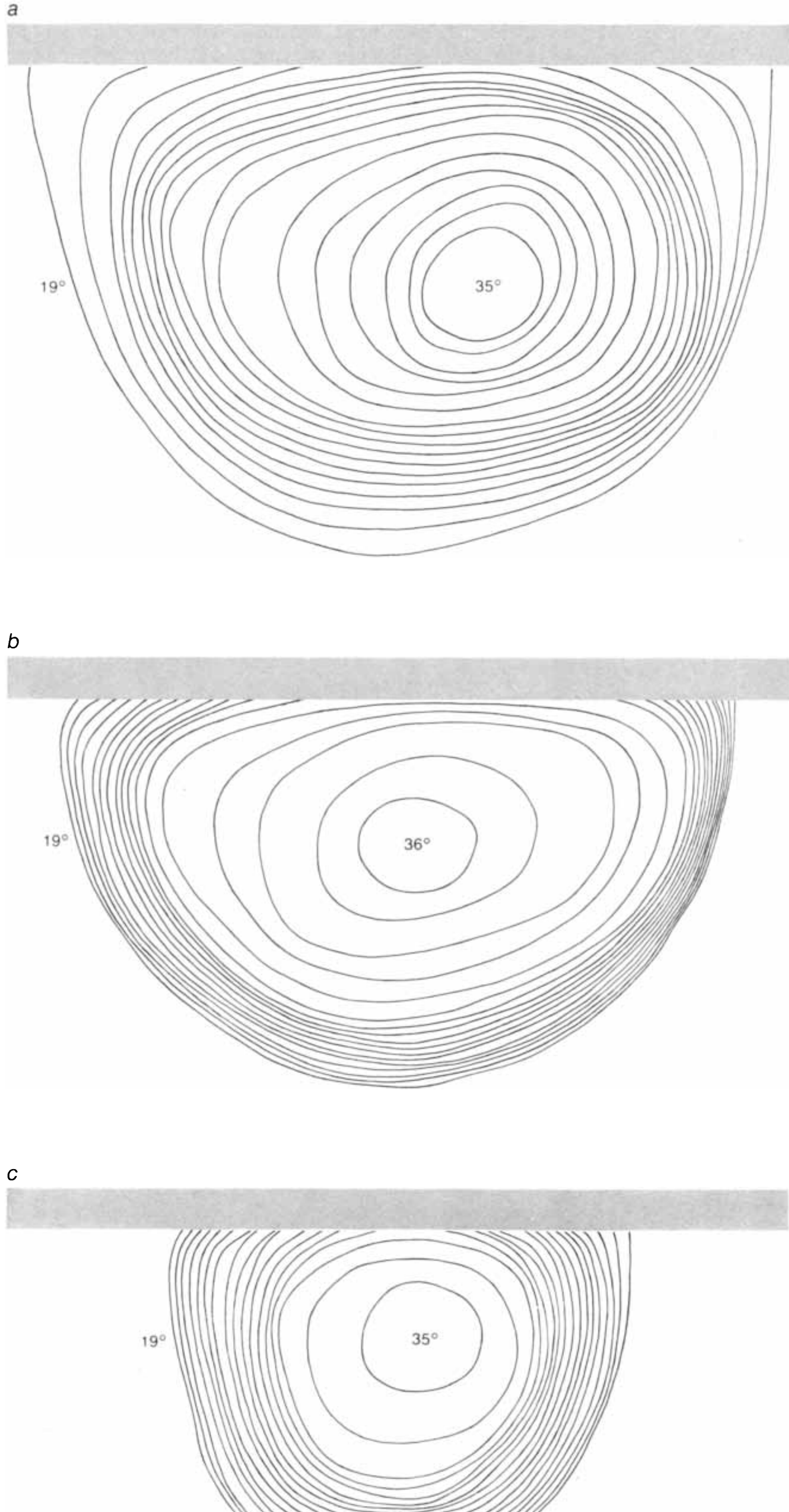
Bees usually came and went from a free swarm, and during the daytime the mantle of such a swarm was significantly warmer than the mantle of a captive one. The mantle temperature of the free swarms fluctuated above the minimum during the day. In a free swarm that was ready to take off, however, the temperature gradient disappeared as the temperature of the mantle reached the core temperature of 36 degrees C.

How is the temperature of the core regulated? It seemed reasonable to suppose that when the external temperature is dropping or the entire swarm is warming up for flight, the swarm as a whole expends more energy to counteract the increasing loss of heat. If that is the case, a small swarm should have a higher metabolic rate than a large one, and all swarms should show an elevated metabolism at relatively low external temperatures. Measurements did not support this supposition. The metabolic rate in all swarms was surprisingly low, averaging close to the resting metabolism of individual bees, with only a small increase at lower air temperatures. Moreover, in swarms of from 1,800 to 16,000 bees where the passive cooling rate varied by a factor of eight there was little or no tendency for small groups to have a higher metabolic rate than large ones. In fact, some of the highest metabolic rates were recorded in large swarms. Since for geometric reasons a large swarm has a greater proportion of its bees at the core than a small one has, and since the resting metabolic rate is a direct function of body temperature, the observed metabolic rates of large swarms could be mainly understood in terms of the passive production of heat by heated bees rather than in terms of heat production to keep warm.

The gross measurements did not reveal how much effort the bees have to put into generating heat actively, but calculations based on the passive cooling rate of bees made an estimate possible. A bee in the core cooling only passively would need to consume oxygen at a rate of .39 milliliter per gram of body weight per hour in order to maintain its temperature. This calculated rate is some 10 times lower than the consumption that would inevitably be required




**EXPERIMENTAL APPARATUS was devised by the author to measure the temperature and metabolic rate of captive swarms. It is shown here in an exploded view. The holes in the lower lid serve as a foothold for the bees and as a source of ventilation. During a measuring period the solid lid is sealed over the ventilating lid and air is circulated from the closed system to the oxygen analyzer to determine the rate of expenditure of energy by the bees. At the same time the apparatus provides a means of measuring the temperature in parts of the swarm.**




**INTERNAL TEMPERATURES** of a swarm at a low external temperature of 2.5 degrees C. (36.5 degrees Fahrenheit) were recorded in a captive swarm. Each isotherm line shows a temperature difference of one degree C. One reading (*a*) is for a swarm of some 5,300 bees, including a queen; the swarm was divided into two swarms, one with a queen (*b*) and one without (*c*).

by the idling metabolism of a resting bee with a body temperature of 35 degrees C. Hence even when the temperature outside the swarm is a chilly five degrees, bees in the core of a large swarm should have a problem getting rid of excess heat. Similar calculations indicate that in a small swarm (about 1,000 bees) the bees in the core would need approximately equal rates of passive heat production and heat loss to maintain their core temperature near 35 degrees. In a still smaller swarm even the bees in the core would have to shiver to maintain a temperature of 36 degrees.

The data suggested that most swarms, consisting of from 5,000 to 30,000 bees, would have no problem heating up internally. Since in such swarms the core temperature is generally held near 35 degrees C., this temperature must represent an upper set point the bees actively avoid exceeding. A small swarm of fewer than 1,000 bees should therefore be found to have a core temperature somewhat lower, and that is what my measurements showed.

How do the bees in the core avoid overheating? Again the problem can be analyzed by first examining the response of the entire swarm and then observing the behavior of individual bees. As the ambient temperature about a swarm rises, the swarm greatly expands in length and circumference. By thus increasing its surface area it achieves a larger passive loss of heat. In addition the bees on the mantle are more widely spaced and their heads are pointed outward, whereas at a lower air temperature they are bunched tightly together and their heads are pushed like shingles under the abdomens of the bees above them. The high-temperature configuration should facilitate the flow of air and heat through the mantle.

The excess heat, however, is mainly in the interior of the swarm. There must be a mechanism to move the heat from the core to the mantle. Watching swarms maintained in a transparent container, I observed that at low air temperature a swarm is an almost solid mass of bees. As the air temperature rises and the swarm expands one can see hanging curtains of immobile bees. The curtains form passageways along which bees travel rapidly from the core to the mantle and back. The passageways also serve as ventilating ducts.

The overall pattern of temperature regulation in the swarm does not reveal how the responses are coordinated. Is there some kind of central direction? Do the bees in the core communicate a need for heating or cooling by sound or chemical signals to the mantle bees that control the swarm's rate of heat loss? Or does each bee act independently?

Several experiments failed to demonstrate direct communication or cooperation. First, swarms with and swarms without a queen maintained the same

# SCIENCE/SCOPE

A new atomic clock being developed for navigation satellites will perform better than previous devices. The clock, which incorporates a hydrogen maser, will use a new microwave cavity design to provide a compact and lightweight package, and new electronic techniques to maintain long-term stability. The clock can provide precise navigation information to military forces because it is stable to one second in 3 million years. The differences in the time when signals from four satellites arrive at one location can be used to calculate that position to within a few yards. Hughes is developing the clock under a U.S. Navy contract.

By producing a coldness almost to where molecular motion freezes, a refrigeration unit cools the detector "eyes" of a U.S. Navy infrared sensor so they are sensitive enough to see infrared radiation emitted by warm objects. The cryogenic dewar base is built by Hughes for the A-6E TRAM/DRS (Target Recognition and Attack Multisensor Detecting and Ranging Set). The TRAM/DRS, also built by Hughes, is a combination laser and infrared device that lets crewmen of the A-6E Intruder aircraft locate and attack ground targets day or night.

A structural and thermal test model of NASA's Galileo Probe is undergoing a series of tests simulating every environment that the Hughes-built probe will experience from launch through descent into the Jovian atmosphere in the late 1980s. The model recently passed a descent simulation with temperatures ranging from -260°F through 240°F and pressures ranging from a vacuum through 235 pounds per square inch -- all in a span of 48 minutes. The Probe contains six instruments that will measure atmosphere characteristics to a depth corresponding to at least 10 times the pressure of air at sea level of Earth. This will be the first direct sampling of the Jovian atmosphere in an attempt to learn its composition and what causes the stunningly colorful weather patterns.

Hughes Missile Systems Group, located in Canoga Park, California, an attractive suburb of Los Angeles, is seeking engineers and scientists for a growing list of development and engineering programs. The list includes AMRAAM, Wasp, multimode guidance, TOW 2, Phoenix, and IR-Maverick. Typical openings are in LSI design, radar and millimeter-wave seekers, IR seekers, signal and data processing, pattern recognition, computer software, and advanced electronic packaging. Please send your resume to Gary Jong, Hughes Engineering Employment, Dept. SSA, Fallbrook at Roscoe, Canoga Park, CA 91304. Equal opportunity employer.

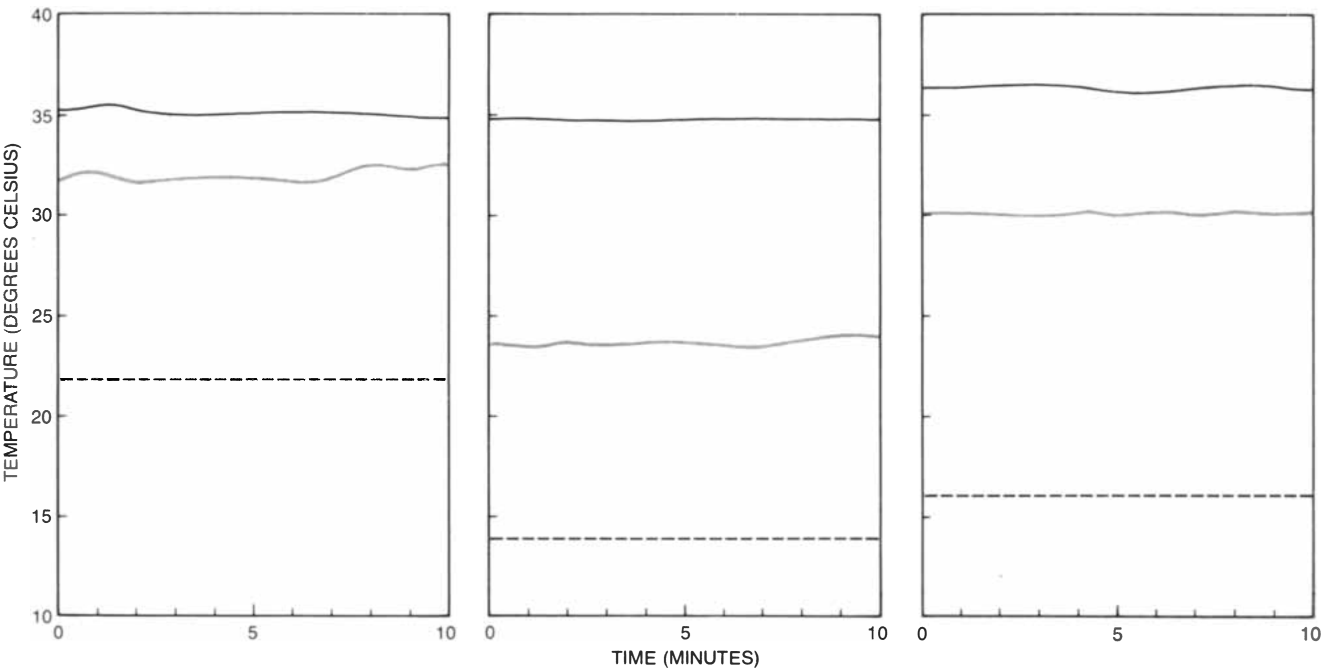
Heat pictures shown in 16 different colors are possible now with an infrared viewer. A new thermal image processing system extends the capabilities of the Hughes Probeye® infrared viewer to display scenes in distinct hues, each color representing a particular temperature range. The system consists of the viewer, an image conversion unit, and a video display. Its features include the ability to display only one temperature level and cancel all others, a zoom capability up to four times, freezing one frame, and simultaneous display of alphanumerics. The system is designed for use in industrial processing and energy auditing. Examples are temperature monitoring in the manufacture of glass, metal, and ceramic products, and helping to spot insulation leaks.

Creating a new world with electronics

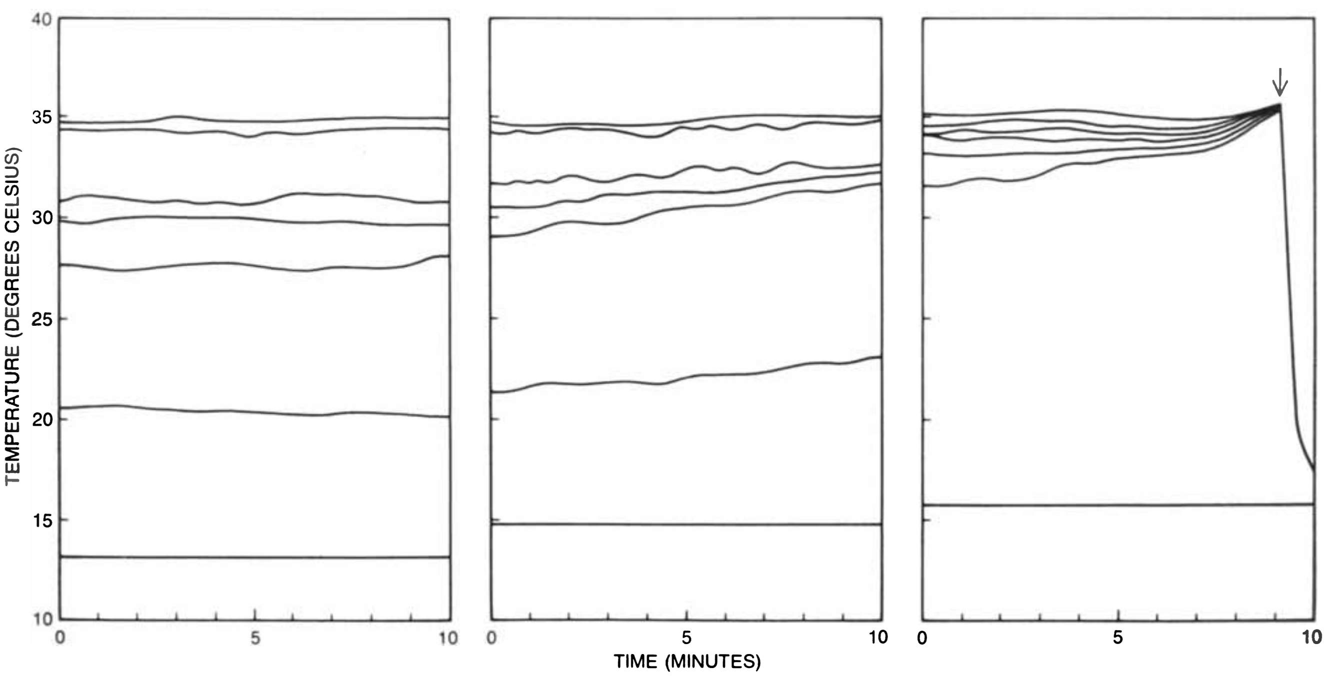
HUGHES

HUGHES AIRCRAFT COMPANY
CULVER CITY, CALIFORNIA 90230

core temperature. This finding ruled out the possibility that chemical signals from the queen in the interior of the swarm served as a thermoregulatory directive. Next I created a swarm in which the bees of the core were separated from the bees of the mantle by thin gauze. The swarm had the same temperature profile as other swarms. This finding showed that even when the bees of the core could not individually sample the temperature surrounding the swarm, the core could nonetheless be maintained at a constant temperature in spite of fluctuating exterior conditions. Third, when I made tape recordings of the sounds generated by the bees in the core or the mantle at a low exterior temperature and played the recordings back into the core or the mantle at a high exterior temperature (and vice versa), no changes in the core temperature were evoked. Finally, pumping air from the core of a swarm held at five degrees C. into one at 20 degrees (and vice versa) had no effect on either core temperature. The findings indicate that neither sounds nor chemi-




**TEMPERATURE RECORDINGS of a free swarm of 5,000 bees were made continuously over a period of several days and are reflected here for three times: 6:30 P.M. on June 7, 3:00 A.M. on June 8 and 10:00 A.M. the same day. The broken black lines show the ambient temperature. The gray line records the temperature of the mantle (perimeter) of the swarm, the solid black line the core temperature.**




**TEMPERATURE OF SWARM before it took off to move to a new nest was recorded at 7:00 A.M., 8:00 A.M. and 10:00 A.M. on June 4. The arrow indicates when the swarm took off. The bottom curve records the ambient temperature, the next curve the mantle temperature, the next three curves record the temperature at points about halfway into the swarm and the top two curves the core temperature. The temperatures converged toward the core temperature in the two hours preceding takeoff. This swarm contained some 28,000 bees.**

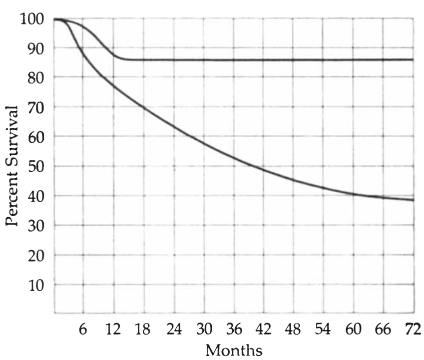
# Medicine is always changing, now there's a text that changes right along with it.

SCIENTIFIC AMERICAN *Medicine.*

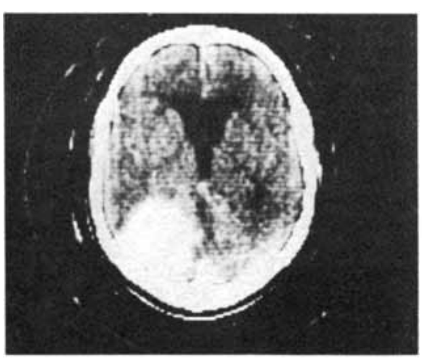
The answers that you get from SCIENTIFIC AMERICAN *Medicine* are current, because each month new chapters replace the old. You will know where to look for the answers, because this advanced text follows the familiar structure of medical knowledge, changing as medicine changes.

**Comprehensive.** Leading authorities from Harvard and Stanford cover the full range of medicine for you the seasoned clinician.

**Complimentary CME.** At no extra cost, we offer a CME program through which you can earn 32 top credits* per year.




**Top curve represents survival in chronic hepatitis patients treated with steroids.**



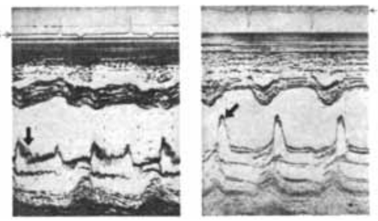
**CT scan reveals large meningioma and mild hydrocephalus.**

**Trial Offer.** We invite you to try SCIENTIFIC AMERICAN *Medicine* for two months with our compliments. Send us the coupon and you will receive the two-volume text and two monthly updates. You may also take a CME test for credit. At the end of 60 days, if you decide to continue the subscription, you will owe us $185 for the first 12 months (renewal is $160); otherwise, return the volumes; there will be no charge.

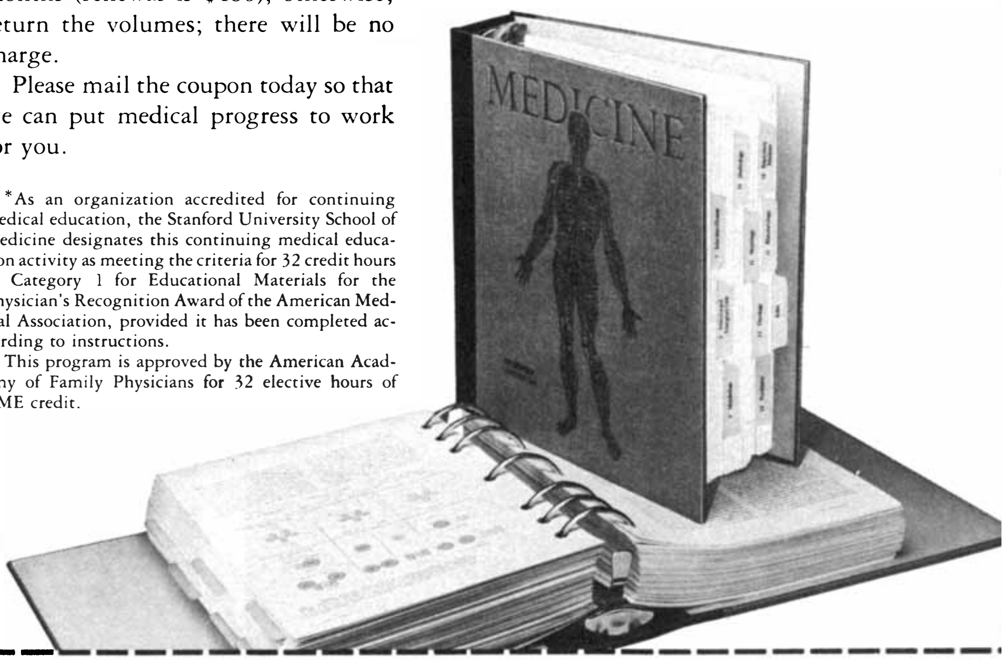
Please mail the coupon today so that we can put medical progress to work for you.

*As an organization accredited for continuing medical education, the Stanford University School of Medicine designates this continuing medical education activity as meeting the criteria for 32 credit hours in Category 1 for Educational Materials for the Physician's Recognition Award of the American Medical Association, provided it has been completed according to instructions.

This program is approved by the American Academy of Family Physicians for 32 elective hours of CME credit.



**Echocardiograms from patients with aortic regurgitation.**




## SCIENTIFIC AMERICAN MEDICINE

415 Madison Avenue, New York, New York 10017

Please enroll me as a subscriber to SCIENTIFIC AMERICAN *Medicine.* On receipt of this coupon you will send me the advanced two-volume text described in your announcement and update it regularly by sending me new monthly subsections. I understand that the annual subscription of $185 for SCIENTIFIC AMERICAN *Medicine* is tax deductible. If I am not entirely satisfied, I may cancel at any time during the first 60 days, returning all materials for a *complete refund.*

- ☐ Please enter my subscription for SCIENTIFIC AMERICAN *Medicine*
- ☐ I shall also enroll in the CME Program
- ☐ I enclose a check made out to SCIENTIFIC AMERICAN *Medicine* for $185*
- ☐ Please bill me

*Please add sales tax for California, Illinois, Massachusetts, Michigan, Ohio, and New York

Name ______________________ MD Specialty ______________

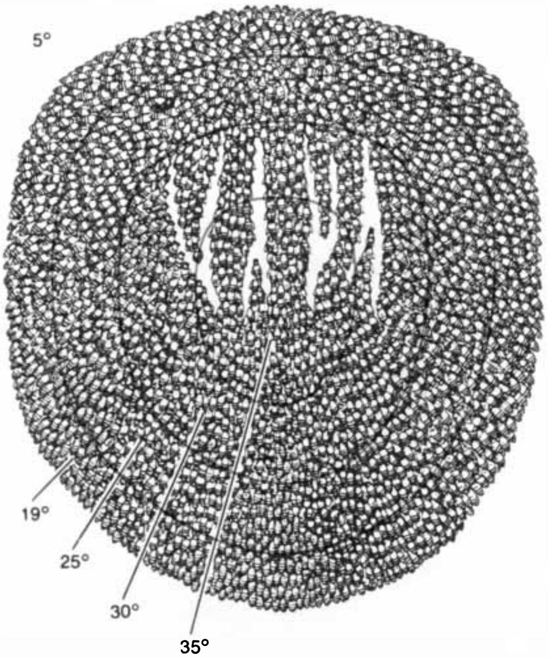
Address ______________________

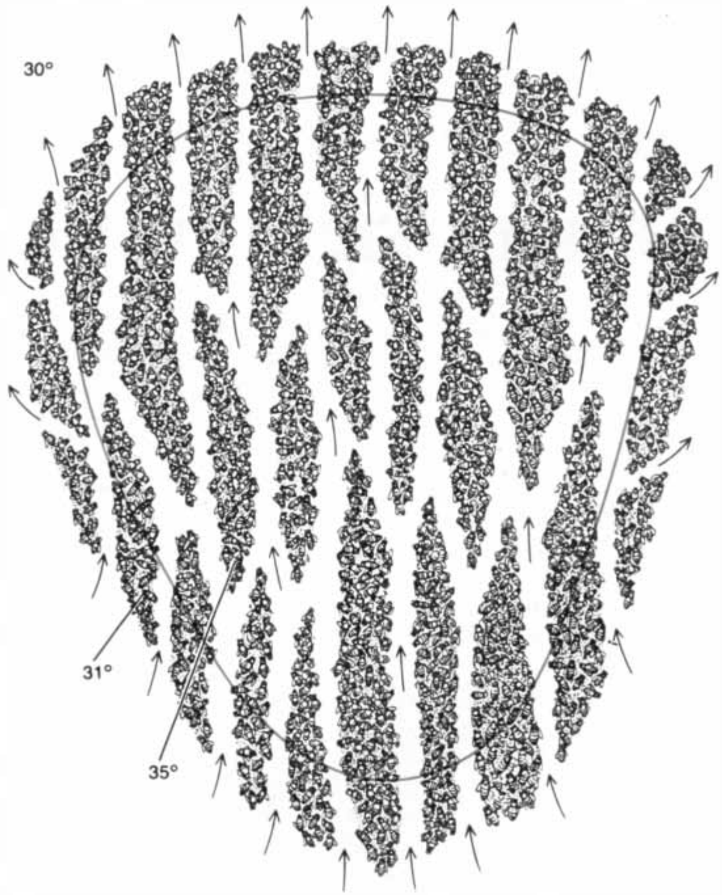
City ______________ State ______________ Zip ______________

Signature ______________________

Subscribers outside the U.S. and possessions will be charged extra for shipping and handling updates will be sent by surface routes unless airmail delivery is requested. Please allow 6-8 weeks for delivery

1Y




**SCHEME OF TEMPERATURE REGULATION** by a typical bee swarm is indicated for a low ambient temperature (*left*) and a high one (*right*). In the low-temperature condition the bees cluster close together, with the bees on the perimeter adopting a shinglelike formation, the adjacent layer of bees shivering to generate heat and the bees near the core giving off heat at their resting-metabolism level and forming corridors to dissipate heat. When the external temperature is high, the bees of the mantle assume a more open stance, most of the bees are at resting metabolism and the swarm forms a number of corridors that contribute toward the dissipation of heat (*arrows*).

**BEES OF THE MANTLE** assume different positions at low and high external temperatures. When the outside temperature is three degrees C. (*left*), the bees are packed close together and nearly every bee has its head under the abdomen of a bee above it. At an external temperature of 25 degrees C. (*right*) the bees of the mantle are more widely spaced and most of them have their head oriented outward.

cal signals are involved in the thermoregulatory response of the swarm. Apparently the bees in the core and the mantle do not "tell" one another about the local temperature.

It was more informative to examine the behavior of individual bees. The German biologist Anton Büdel had shown that the bees in the mantle tend to be predominantly the older workers, the foragers, and that the young "house" bees tend to be inside the swarm. Moreover, M. Delia Allen of the North of Scotland College of Agriculture had shown that the older bees are better at regulating their body temperature, particularly by shivering. The young bees have little capacity for shivering and generating their own heat. In experiments that give bees a choice of temperature the younger ones seek out higher temperatures than the older ones. If bees in a swarm behave similarly, these facts fit my model of temperature regulation, in which the core bees are heated passively and the mantle bees shiver to generate heat in order to keep from being chilled below 13 degrees C., the minimum body temperature below which they lose motor coordination and fall from the cluster.

I fitted bees with a hairlike thermocouple embedded in the thorax and returned them to the mantle. The temperature measurements showed that they regulated their thoracic temperature, keeping it above 13 degrees. If they were disturbed, they crawled deeper into the cluster, and their thoracic temperature rose within minutes to more than 30 degrees. Thus I learned that the mantle bees function at temperatures well below the core temperature and also derive considerable heat from their fellows in both the mantle and the core.

It is evident that the mantle bees affect and are affected by the thermoregulatory responses of bees in the interior without direct communication. When the core temperature is high, the bees in the mantle receive heat from the core, replacing heat they lose to the environment and thereby retarding their cooling rate and reducing their burden of heat production for thermoregulation. Indeed, since the excess heat that must be dissipated is generated passively in the interior of the swarm, it should help the thermoregulation of the core if the mantle bees do not generate heat, so that the temperature gradient in the swarm is outward and the outward flow of heat is facilitated.

Do the bees in the mantle shiver to maintain an elevated temperature in the interior or do they merely attempt to regulate their own temperature at about 19 degrees C.? The data are consistent with the hypothesis that they regulate only their own temperature but that in doing so they also help the core bees. When mantle bees are cooled by a low ambient temperature, they crawl into the swarm, plugging the passageways for convective heat loss from the core and also reducing the mantle's surface area and porosity. I tested the hypothesis by transferring a swarm from an environment of about one degree C. to room temperature (about 20 degrees). As the hypothesis predicted, the immediate response was a drop in the core temperature and a rise in the mantle temperature as the mantle bees loosened their tight formation.

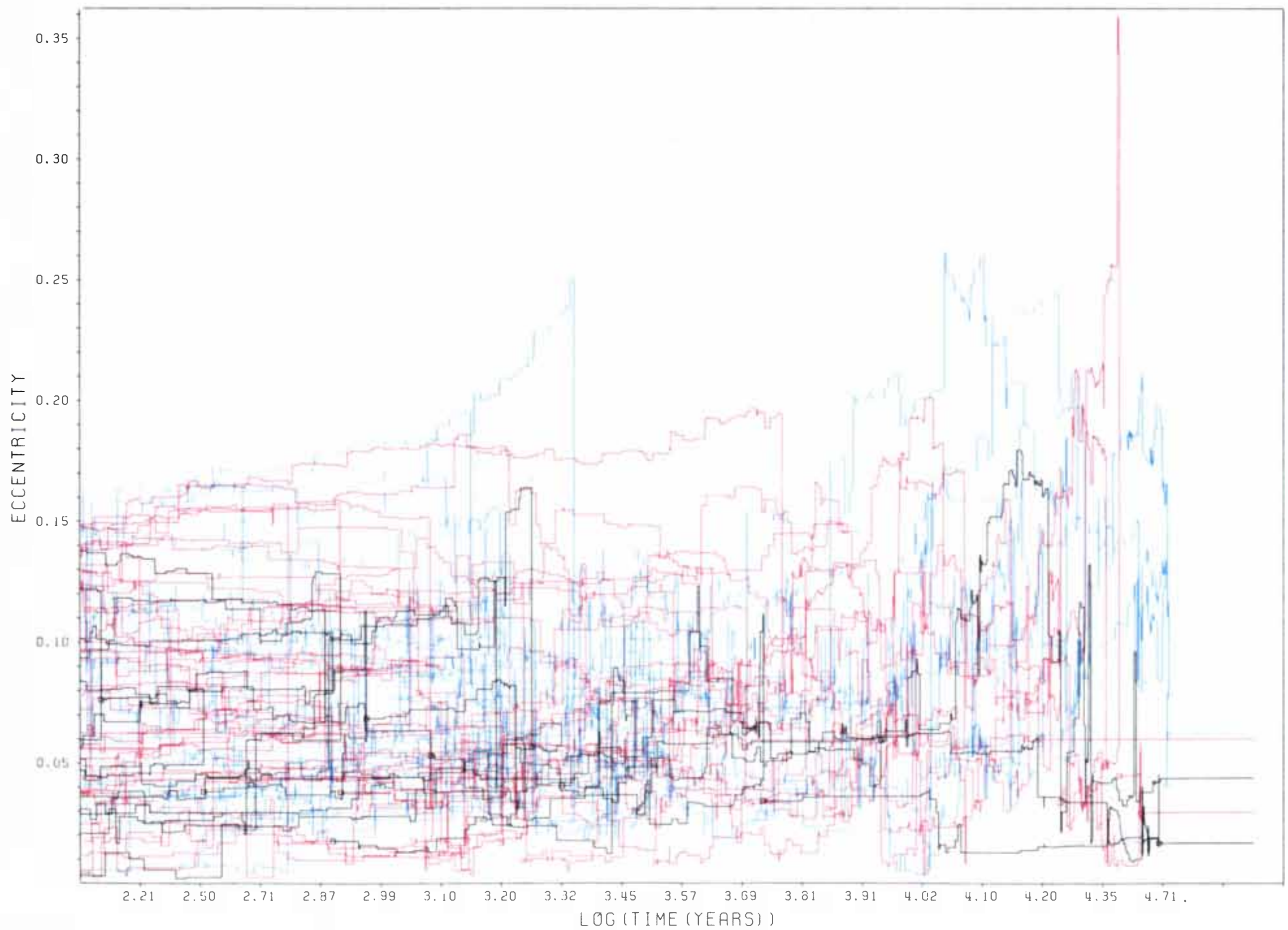
Alternatively, by crowding and shivering the mantle bees retard the loss of heat from the core. If they are too effective in retaining heat within the swarm, the core bees can partly overcome the effect by moving to the exterior, thereby creating convection currents in the swarm. There is no evidence that bees in one part of the swarm know what the temperature is in another part and modify their behavior accordingly. The bees act independently, but the result is a coordinated response that is beneficial to the entire swarm.

It is important for all the bees of a swarm to be able to leave together to occupy a suitable nest site soon after it has been found and evaluated. Thermoregulation of the swarm is critical in all aspects of finding a new nest and occupying it. The mantle must be maintained above 15 degrees C. at all times, otherwise the bees in it could not be aroused for the takeoff.

It is not immediately obvious why the core should be held at about 35 degrees C., particularly at night. That temperature is, however, the one required for taking flight quickly. A honeybee needs from four to six minutes to warm up to 35 degrees from about 20 degrees by shivering. Hence there is no great disadvantage in letting the mantle temperature fall below the flight temperature, provided that the core is kept warm and bees can enter it (or shiver) to warm up. By skimping on active metabolism (shivering) for generating heat the bees of the mantle prolong the swarm's limited reserves of food.

Honeybees that are about to swarm gorge themselves on honey before they leave their hive. This honey not only must fuel thermoregulation while the swarm waits to occupy a new hive but also must serve as a fuel for the scouts. It is also the substrate from which the initial combs of the new hive are built. The bees cannot skimp too much on the expenditure of energy for regulating the temperature of the swarm, however, or the swarm's response to the discovery of a new domicile would be too slow. Thermoregulation makes possible a quick response in a range of weather conditions at the most critical stage of the colony cycle: the competition for a suitable nest site that will thereafter serve the new colony indefinitely.


SEMIMAJOR AXIS (A.U.)
1.80
1.70
1.60
1.50
1.40
1.30
1.20
1.10
1.00
0.90
0.80
0.70
0.60
0.50
2.21
2.50
2.71
2.87
2.99
3.10
3.20
3.32
3.45
3.57
3.69
3.81
3.91
4.02
4.10
4.20
4.35
4.71
LOG (TIME (YEARS))


ECCENTRICITY
0.35
0.30
0.25
0.20
0.15
0.10
0.05
2.21
2.50
2.71
2.87
2.99
3.10
3.20
3.32
3.45
3.57
3.69
3.81
3.91
4.02
4.10
4.20
4.35
4.71
LOG (TIME (YEARS))

# The Formation of the Earth from Planetesimals

*The rocky inner planets of the solar system may have formed by coalescence in a swarm of small, colliding bodies in orbit around the sun. The process can be simulated with a computer*

by George W. Wetherill

Gaining a complete understanding of how the earth formed will require learning how the sun condensed from a cloud of gas and dust and how a small proportion of the matter in the cloud escaped falling into the sun and instead came to make up the planets. Such knowledge will not be easily acquired. For one thing, simulations done with the aid of a computer suggest that if such a cloud, or solar nebula, is dense enough to collapse gravitationally, it will form a large number of multiple-star systems instead of a single star. Even if the condensation of the sun were understood, it might not reveal much about the origin of the terrestrial planets (the rocky bodies Mercury, Venus, the earth and Mars). The mass of the terrestrial planets is only .0005 percent of the mass of the sun. Plainly it is unreasonable to expect detailed knowledge of a process as minute as the formation of the earth to emerge from a consideration of no more than the solar nebula's general properties.

Another approach seems more promising. In spite of the complexity attending the origin of the solar system from the solar nebula, there are only two basic ways a terrestrial planet could have formed. It could have arisen from a local concentration of the nebula's gas and dust that collapsed into a planet by self-gravitation, or it could have been built up by the accumulation of smaller bodies (planetesimals) that had earlier condensed from the gas and dust. Through simulations of these alternative processes theorists are improving their understanding of how the solar system arose. For example, it is becoming possible to specify conditions in the solar nebula that lead to the formation of the planets, and to distinguish them from conditions that do not. It is also becoming possible to determine what each hypothesis predicts about such properties as the initial temperature of the earth. One can be optimistic that a comparison of such predictions with the geologic record will show which alternative represents the most probable way the terrestrial planets formed.

In this article I shall describe the current understanding of the formation of planets by the accumulation of planetesimals. It is too early to decide whether the earth did form in this way, rather than by the alternative of gravitational collapse. It appears quite likely, however, that the terrestrial planets could have formed from planetesimals. If they did, the simulations show that the formation was virtually complete about 100 million years after it began and that the newly formed earth was so hot that much of its interior was molten.

The first question to be faced in the building of planets from planetesimals is whether when two bodies collide they coalesce and thereby form a larger body. It is obvious that a collision does not always lead to coalescence. For example, if two rocks a meter in diameter collide at a relative velocity of 1,000 kilometers per hour (.3 kilometer per second), they break into pieces and fly apart. That is not the way to form the terrestrial planets.

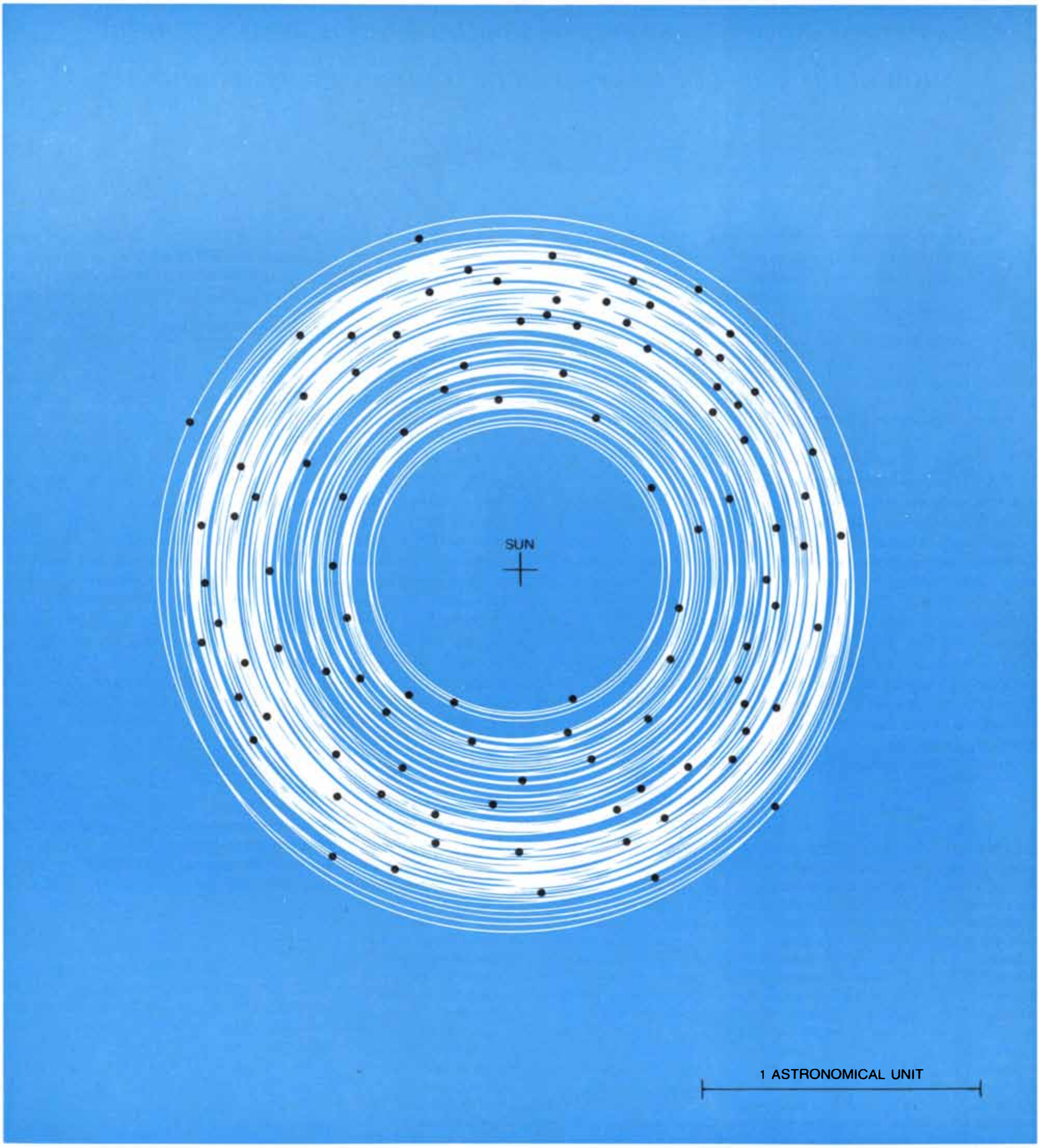
A collision at a low relative velocity does allow the bodies to coalesce. How low must the velocity be? In answering this question a quantity called the gravitational escape velocity turns out to be important. It is the minimum relative velocity required for two bodies to fly free of their mutual gravitational attraction. To escape from the earth, for example, a projectile shot from the surface must have a velocity of about 11 kilometers per second. To escape from the moon a velocity of 2.4 kilometers per second is sufficient. If two planetesimals were to collide at precisely their escape velocity, and if no energy were wasted, they would have just enough energy to rebound and regain their independent trajectories.

In a real collision the planetesimals would lose energy because colliding bodies deform and break apart to some extent. The lost energy would be dissipated as heat. The two planetesimals

**ACCRETION OF PLANETS** of the inner solar system from collisions of smaller bodies called planetesimals was modeled with a computer by Larry P. Cox of the Massachusetts Institute of Technology. The simulation began with 100 bodies of equal mass. In obedience to Kepler's laws each body followed an elliptical orbit around the sun. Each orbit was specified by two quantities; its semimajor axis (*upper chart*) and its eccentricity (*lower chart*). The semimajor axis is a measure of the orbit's distance from the sun; it is given in astronomical units, or A.U. (One A.U. is the mean distance between the earth and the sun, equal to approximately $1.5 \times 10^8$ kilometers.) The eccentricity measures the departure of an ellipse from circularity. An eccentricity of 0 signifies a circle; an eccentricity of 1 signifies a parabola, in which the semimajor axis is infinitely long. Although the initial parameters of the orbits were generated at random, the maximum initial eccentricity was chosen to be .15 and the semimajor axes were chosen to range from .5 to 1.5 A.U. The simulation is two-dimensional: it places all the orbits in a plane. Collisions among the bodies in the plane are more likely than they would be in three-dimensional space; hence the accretion in this "flatland" simulation is completed in only some 61,000 years. Small boxes in the charts mark collisions, which make the colliding bodies coalesce. Discontinuities at some of the boxes in the lower chart mark large decreases in eccentricity. In the end six planets remain. Separate colors distinguish the sequences by which they accreted. The surviving planet nearest the sun is a planetesimal that never collided, although the vagaries of its career show that its orbit was repeatedly perturbed by near-misses with another body. The other five planets are larger. They might whimsically be called Mercury, Venus, the earth, the moon and Mars. The success of the simulation is that bodies resembling the earth in their mass and in their orbit emerge from the process. The time axis has been distorted so that equal numbers of perturbing events (collisions and near-misses) occupy equal intervals of the printout.

colliding at their exact escape velocity would therefore be unable to escape from each other, and their fragments could combine to form a single larger body. Indeed, experiments done by William K. Hartmann of the Planetary Science Institute in Tucson, Ariz., and numerical calculations made by Thomas J. Ahrens and John D. O'Keefe of the California Institute of Technology have shown that when the impact velocity is twice and sometimes even three times the escape velocity, the planetesimals can coalesce.

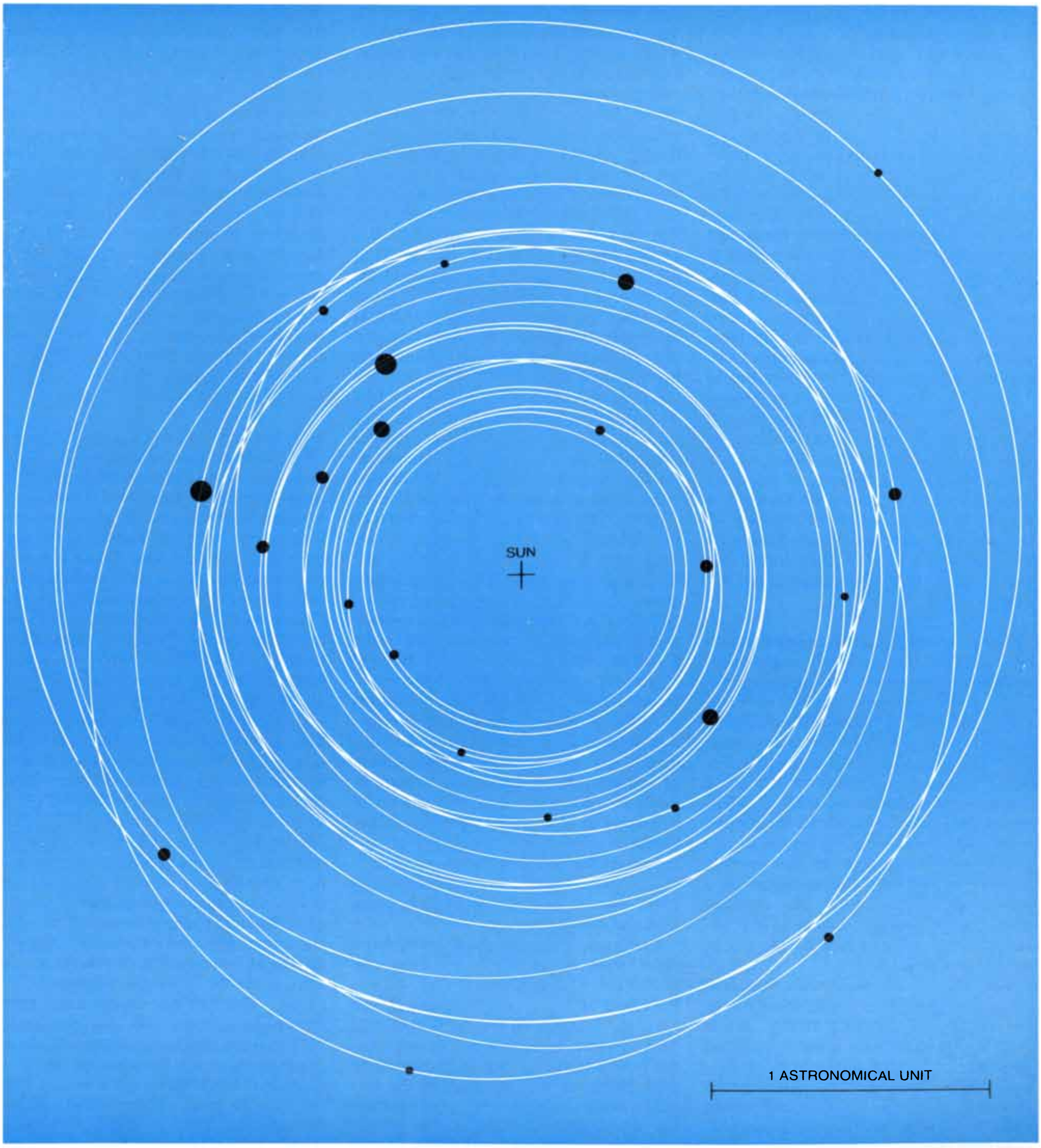
Some further reasoning now extends this line of thought. The impact velocity of two bodies is determined by the relative velocity of the bodies when they are distant from each other, augmented by the velocity they acquire as they are accelerated toward each other by their gravitational attraction. If the relative velocity when the bodies collide happens to be twice the escape velocity, the relative velocity when the bodies were at a large distance was $\sqrt{3}$ (approximately 1.7) times the escape velocity. If the ve-




**THREE-DIMENSIONAL SIMULATION of the formation of the inner solar system also began with 100 planetesimals. Each one was assigned a mass of $1.2 \times 10^{26}$ grams, so that the total mass ($1.2 \times 10^{28}$ grams) is that of the rocky planets Mercury through Mars, together with their moons. Again the initial parameters of the orbits were generated at random. The illustration shows the initial ellipses, but it does not suggest the inclination of the ellipses with respect to one another. Because the planetesimals are thought to have formed in a thin layer of dust in the central plane of the cloud from which the solar system condensed, the inclinations are less than five degrees.**

locity when the bodies collide is three times the escape velocity, the relative velocity when they were distant was $2\sqrt{2}$ (approximately 2.8) times the escape velocity. Thus the rule that bodies can coalesce only if their relative velocity at impact is less than two or three times the escape velocity also applies to the relative velocity of the bodies in their free trajectories well before their collision.

In some simulations of the formation of planets from planetesimals the relative velocity of any two planetesimals has been considered a free parameter: it has been assumed that the velocity can always be made low enough for the colliding planetesimals to coalesce and grow. On the other hand, the velocity can also be too low. The laws of planetary motion first formulated by Johannes Kepler specify that the orbit of a body around the sun is an ellipse. The velocity at any point in such an orbit is determined by the size and shape of the ellipse. It follows that the relative velocity of two bodies in orbit around the sun




**INTERMEDIATE STAGE of the simulation shows the developing inner solar system after 30.2 million years have passed and 22 bodies remain. Their orbits have become more markedly elliptical and their velocities with respect to one another are greater. In addition their range of distances from the sun has broadened. The innermost orbits are closer to the sun and the outermost orbits are farther away. The mutual inclinations among the orbits have greatly lessened the incidence of collisions compared with near-misses. Hence tens of millions of years of simulated time must pass before the process reaches an end. In the two-dimensional simulation 61,000 years sufficed.**

will tend to be great unless the orbits are similar in size, shape and orientation. As a result the choice of a low relative velocity confines the possibility of coalescence and growth to only a small fraction of the planetesimals, namely those whose orbits are similar to one another. The formation of planets will end when these planetesimals have combined, and the simulated inner solar system will resemble the rings of Saturn more than a set of four terrestrial planets in an otherwise empty expanse.

Clearly the choice of a relative velocity near the escape velocity would be the best strategy for a successful simulation. It would enable the colliding planetesimals to coalesce and at the same time it would maximize the number of potential collision partners. Another problem arises, however. As the bodies increase in size the escape velocity also increases: for a body of uniform density the escape velocity is proportional to the radius. If a planetesimal with a radius of one meter had an escape velocity of .15 centimeter per second, a planetesimal with a radius of one kilometer would have an escape velocity of 1.5 meters per second and a planetesimal with a radius of 1,000 kilometers would have an escape velocity of 1.5 kilometers per second. Each of these bodies can continue to grow by accretion if it collides with a body moving at a velocity of no more than two or three times the escape velocity. Hence the relative velocity of the planetesimals must keep step with their growth and with their increasing escape velocity if it is to stay in the optimum range.

The possibility that the relative velocity of the planetesimals might naturally keep step with the rising escape velocity was introduced in 1950 by L. E. Gurevich and A. I. Lebedinskii of the Institute of Applied Geophysics in Moscow. The idea was developed extensively by Victor S. Safronov of the same institute throughout the 1960's. Safronov's work was analytic: he developed mathematical expressions for the evolution of quantities such as the average relative velocity of the members of a swarm of planetesimals. His simplest model described the swarm in the absence of any drag caused by the gas of the solar nebula.

The relative velocities of the bodies in the swarm are determined by the balance of two competing effects. When bodies pass near one another without colliding, their gravitational attraction perturbs their orbits. The perturbation changes the relative velocity with which each body subsequently encounters other bodies moving in intersecting orbits. Sometimes the relative velocity after the near-miss is greater than it would have been otherwise; sometimes it is less. The net result, however, is an increase in the average magnitude of the relative velocity. Collisions have the opposite effect. They tend to make the orbits of the planetesimals more circular than they were and hence more similar; it follows that the relative velocities tend to decrease. At any given time, therefore, the pumping up of the velocity as a result of near-misses tends to be balanced by the damping of the velocity from collisions.

Each of these effects can be considered in greater detail. The Newtonian laws of motion contribute to this effort, together with the theory of the random walk, which describes the essentially random nature of the interactions of the planetesimals. For small perturbations caused by near-misses the change in the velocity of a small body (a "projectile") that encounters a larger body (a "target") turns out to depend on $\sqrt{M/D}$, where $M$ is the mass of the target and $D$ is the distance of closest approach. For very close misses the distance $D$ is approximately equal to $R$, the radius of the target. In addition the mass $M$ depends on $R^3$. The expression $\sqrt{M/D}$ can therefore be transformed into $\sqrt{R^3/R}$, or simply into $R$. In brief, therefore, the change in the projectile's velocity has precisely the same direct dependence on the radius that the escape velocity has. This relation implies that the pumping up of relative velocity that accompanies a number of near-misses proceeds in tandem with the increase in the escape velocity caused by growth of the bodies.

More distant encounters also perturb the orbits. Indeed, the encounters are significant at any distance from the target less than the radius of what might be called the target's gravitational sphere of influence. Within this volume the motion of the projectile is dominated by the gravitational field of the target. Beyond it the motion is dominated by the gravitational field of the sun. If the changes in velocity caused by distant encounters are averaged over all encounter distances out to the edge of the sphere of influence (at a distance of about 75 times the radius of the target), the result is a more or less fixed proportion of the changes caused by encounters at distances on the order of one target radius. Hence the velocity increments resulting from distant encounters also increase with the escape velocity.

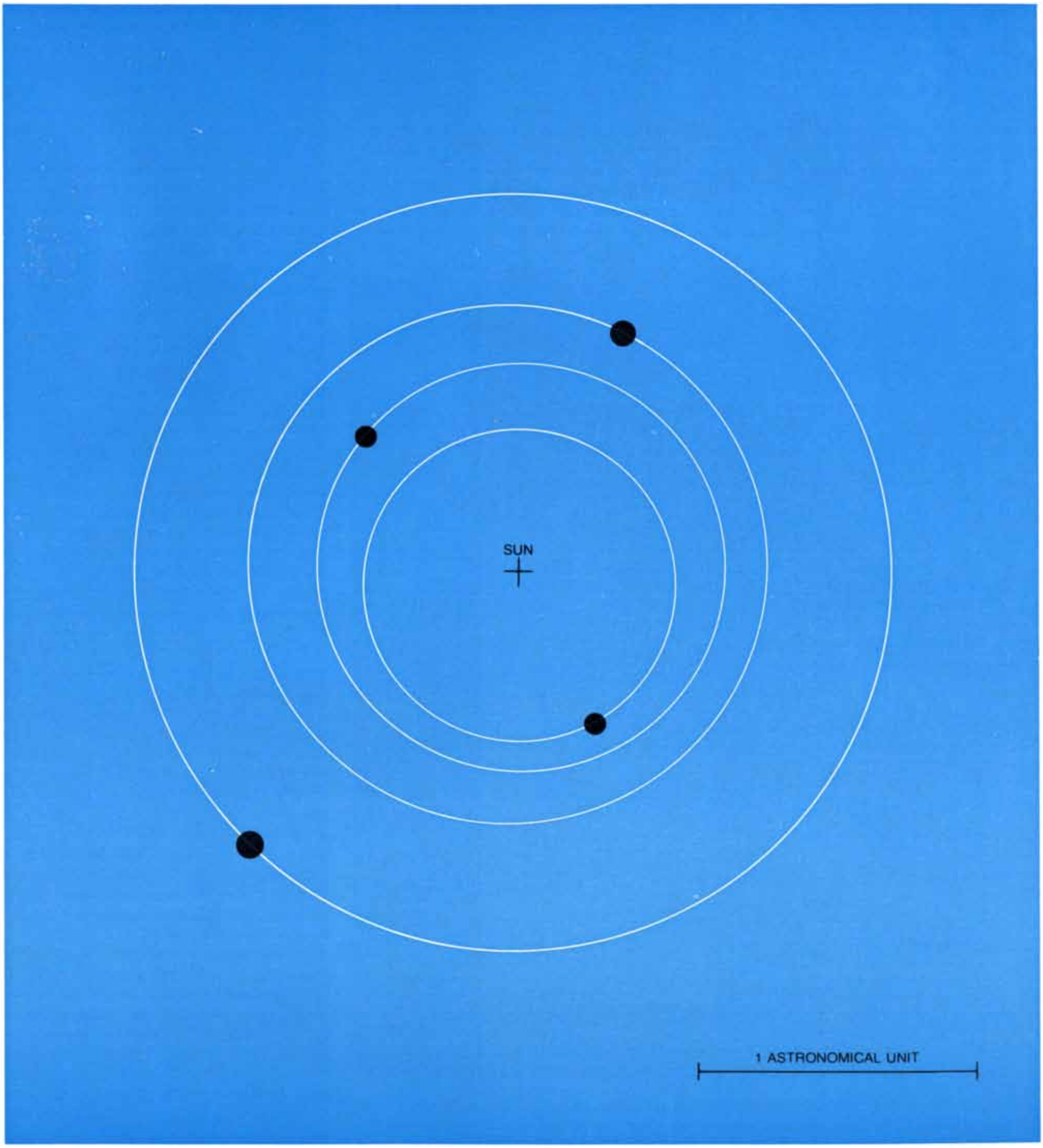
The pumping up of the velocity from encounters at all distances from one radius to 75 radii more than offsets the effect of collisional damping. To be sure, the increase in the cross-sectional area of each target as it grows increases the frequency of collisions. At the same time, however, the increase in the target's mass expands its gravitational sphere of influence, and so the probability of a perturbing approach is also raised. The end result is remarkable: the relative velocity of the swarm of planetesimals does indeed keep step with the escape velocity. Furthermore, for plausible values of the loss of energy in collisions the relative velocity is approximately equal to the escape velocity. This appears to be just what is needed if the planetesimals are to continue coalescing as the size of each body increases.

At the start of the process (as modeled by Safronov) the planetesimals are small. Each one has a radius of a few kilometers. The escape velocity is also small, and the relative velocity is near the escape velocity. Only bodies with similar orbits collide and coalesce. Nevertheless, the accumulation proceeds throughout the region of the solar system that the terrestrial planets will eventually occupy. As the bodies grow their escape velocity rises. The accumulation continues, however, because the relative velocity remains near the escape velocity. In accordance with Kepler's laws, the increase in relative velocity is associated with an increase in the eccentricity of the orbits. In other words, each orbit becomes a more pronounced ellipse. This trend allows each planetesimal to reach out to more distant bodies. The number of potential collision partners for any given planetesimal is thereby maintained, although the total number of bodies is decreasing and the bodies are becoming more widely spaced.

What remains to be understood before one can believe the earth and the other terrestrial planets may indeed have formed in this way? One might wonder how the small planetesimals formed, the ones that appear at the outset in Safronov's model. There is now a consensus that this is not a serious problem. Stuart J. Weidenschilling of the Planetary Science Institute has shown that the weak electrical attractions called van der Waals forces may well have caused dust grains in the solar nebula to stick to one another and form bodies about a centimeter in diameter. The nebula is assumed to have been rotating. The drag exerted on the bodies by the rotating gas would then have driven the bodies into the central plane of the nebula. (The central plane is perpendicular to the axis of the nebula's rotation.) The result would have been a thin central layer of dust. Some of the bodies would have grown there by collision. When the largest ones had attained a diameter of about a meter, the concentration of matter in the dust layer would have made the layer gravitationally unstable. The bodies would then have clumped together with the remaining dust to form planetesimals about a kilometer in diameter. The process has been described by a series of workers over the past 30 years. These are the planetesimals required by Safronov's theory. About $10^{13}$ of them are needed to make up the terrestrial planets.

Other problems are somewhat less tractable. For example, Safronov's simplest model describes a swarm of planetesimals that initially are all the same mass. In 1962 he extended his results to include the case of bodies of differing masses. He was able to show that the relative velocity matches the escape velocity of the largest bodies only if the largest bodies have most of the mass in the system. In that case the smaller bodies break up on collision but the larger ones grow. It is possible, however, that as the system evolves from a vast number of small planetesimals the mass will never come to be concentrated in the larger bodies. To deal with this question it would be necessary to calculate simultaneously the evolution of the relative velocity and the evolution of the distribution of mass. Safronov's theory is not adequate for such a calculation.

Richard J. Greenberg and his colleagues at the Planetary Science Institute have made calculations that are important in this respect. Their work is limited to the early stages of growth in a swarm of planetesimals, beginning at a time when most of the mass of the swarm is concentrated in planetesimals about a kilometer in diameter. The rela-



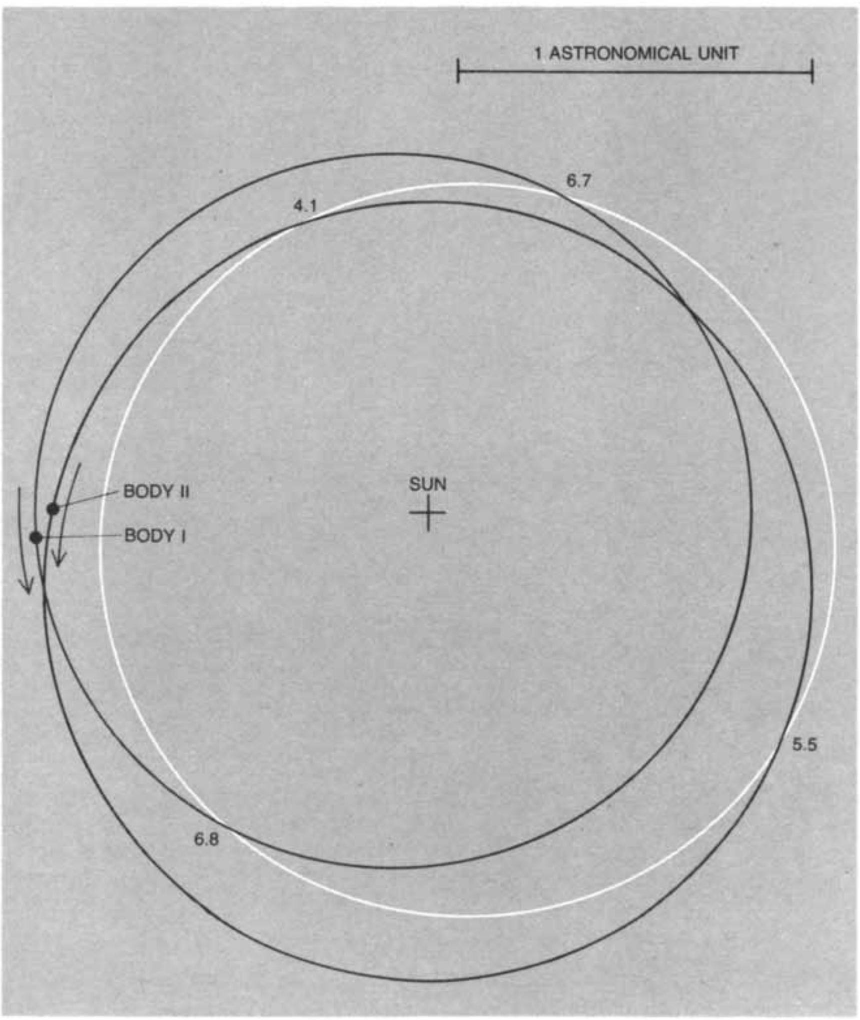

**FINAL STATE** of the simulation leaves four fully formed planets in isolated, almost circular orbits around the sun. The planetesimals have all coalesced. The largest planet, the fourth from the sun, is built up from 34 of the original bodies. Although the illustration shows the simulation after 441 million years, the process is almost complete much sooner. There are 11 bodies after 79 million years and six bodies after 151 million years. This suggests that the earth accreted in a time on the order of 100 million years. The simulation was made by the author with a computer at the Department of Terrestrial Magnetism of the Carnegie Institution of Washington.

tive velocity of these bodies is only a few meters per second, a value comparable to the escape velocity. According to Greenberg's calculations, some of the bodies collide with one another and coalesce, and after about 15,000 years a few of them have grown as large as 100 kilometers. Most of the mass, however, is still in the small planetesimals. As one would predict from Safronov's 1962 result, there are too few large bodies to pump up the relative velocity and overcome collisional damping.

How things progress at this stage is crucial. For bodies with a diameter as great as eight kilometers or so, it appears that the relative velocity is increasing. As long as the original low velocities persist, however, the first planetesimals to grow to a diameter as large as 100 kilometers may have an insuperable advantage over their smaller neighbors. In particular, an encounter between a small planetesimal and a large one at a velocity well below the escape velocity can perturb the trajectory of the small planetesimal to such an extent that the small planetesimal collides with the large one, even though it was not on what would appear to have been a collision course. The effect is called gravitational focusing. For the purpose of predicting collisions it increases the effective cross-sectional area of the larger body beyond its normal geometric dependence on the square of the radius to a dependence approaching the fourth power of the radius.

If gravitational focusing is sufficiently effective, the first large bodies that form capture the small bodies and thereby prevent the growth of bodies of intermediate size. The distribution of sizes among the planetesimals then has two peaks, one representing the large bodies and the other representing small bodies that have not yet been swept up. The large bodies move in almost circular orbits because nothing has happened to increase the eccentricity of the orbits. Indeed, the impacts of the small planetesimals on the large ones tend to make the orbits of the large planetesimals more circular.

If this process continues, the large planetesimals acquire most of the mass of the system. By then it may be too late for a planetary system like the solar system to form. The large bodies move in nonintersecting orbits, in which they gather up the last small planetesimals. The final result is a system of about 5,000 minor planets, each one about 1,000 kilometers in diameter. It is conceivable that the system would have a second chance to form a set of terrestrial planets. Perhaps the long-range gravitational interactions of the thousands of minor planets would increase the eccentricity of the orbits enough for growth to resume. But this would not be the straightforward evolution that recommends Safronov's theory.




**INTERACTIONS OF TWO PLANETESIMALS are the simulated events that cause a swarm of planetesimals to become the inner solar system. Here the trajectories are plotted for two planetesimals that are about to interact. Each body has a mass of $10^{27}$ grams and a radius of about 4,000 kilometers. The orbit of a third planetesimal (*white*) represents the motion of the other bodies in the swarm. The velocity of each interacting body with respect to the third body's orbit is given in kilometers per second for points at which the orbits intersect. Two possible results of the interaction between two planetesimals are shown in the next two illustrations.**

Another difficulty with Safronov's model (and also with Greenberg's work) is that it treats the planetesimals as if they were moving in straight-line paths like molecules in the kinetic theory of gases. Specifically, Safronov examined the planetesimals in a region of space surrounding a given planetesimal. From this vantage the relative velocities of the planetesimals can be described as if the planetesimals were particles of a gas confined in a box. This method precludes an adequate treatment of any phenomena that depend on the distance of the planetesimals from the sun.

The most basic doubt about the simplifications implicit in Safronov's model concerns the extent to which the idea of a balance between the increase in velocity caused by mutual perturbations and the decrease in velocity caused by collisional damping is valid in a heliocentric theory. One might expect to find that it is not. After all, the velocity of a body in a heliocentric orbit depends on its distance from the sun. This introduces a coupling between the radial distribution of planetesimals and their relative velocities that has no counterpart in the kinetic theory of gases.

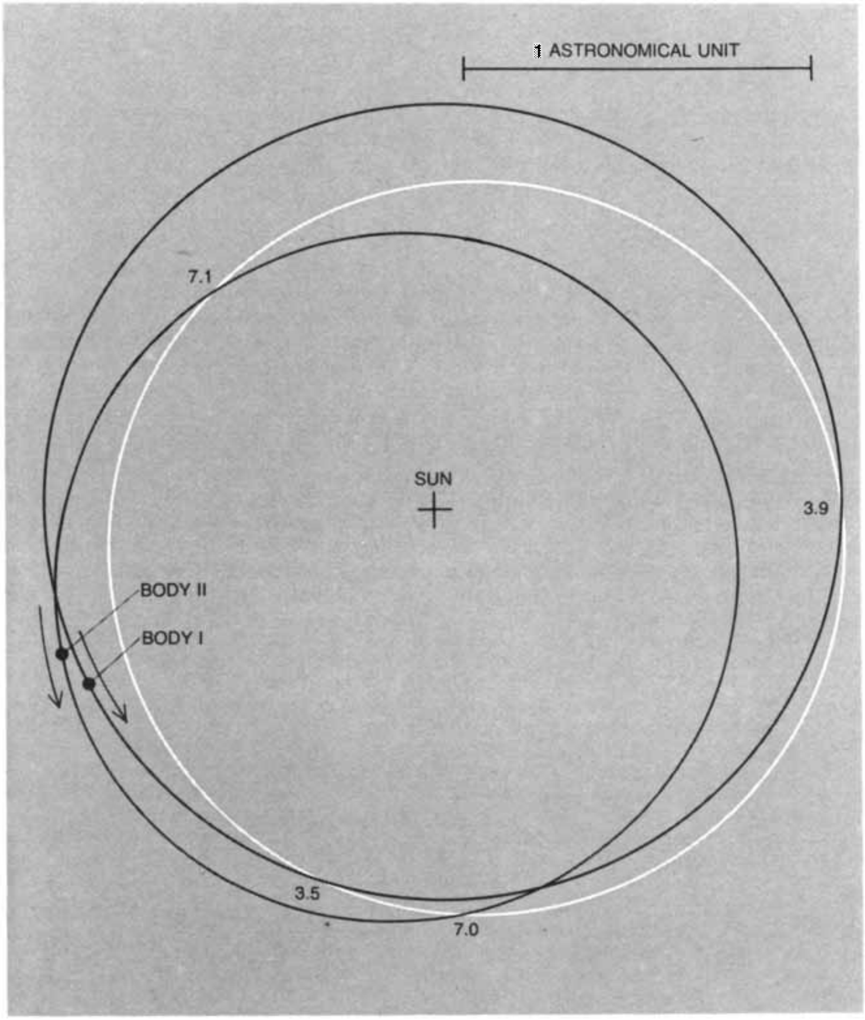
I have made computer studies that may help to settle this doubt. The studies begin with a swarm of 100 bodies of equal mass. At first the planetesimals follow heliocentric orbits of low eccentricity. The orbital changes resulting from encounters and collisions are small but numerous. Each change is calculated and new orbits are determined. The relative velocities and the eccentricities tend at first to increase rapidly, but then their rate of increase diminishes and at last they level off at values proportional to the escape velocity. This is the same effect Safronov found. When the initial orbits are chosen to have high eccentricity, the changes proceed in the opposite direction: the relative velocity is large at the outset and then decreases. It levels off at the same value that is attained in the low-eccentricity swarm.

The value of the steady-state velocity found in these numerical calculations is about two-thirds the value predicted by Safronov's theory. Still, the approximate agreement is satisfying. The relation between the relative velocity and the escape velocity is confirmed, and so is the tendency of the system to regulate itself so that the relative velocity lies in the range favoring coalescence.

The calculations do more, however, than confirm the earlier results. As the calculation proceeds, a swarm initially confined to a small range of distances from the sun spreads out to include a greater range of distances. Each planetesimal takes part in the increase: it moves about in the growing range of orbital radii, sometimes being near one extreme, sometimes near the other and most of the time near the middle. In effect the planetesimal takes a random walk in heliocentric distance. The wandering orbit of each planetesimal increases the number of opportunities for distant bodies to become collision partners. One need not rely entirely on the eccentricity of the orbits to bring about encounters between planetesimals. Although the effect does not eliminate the chance that minor planets too small and too numerous will form, it certainly tends to reduce that chance.

So far these calculations have been limited to the case of bodies of equal mass. They have also been limited to finding the steady-state distribution of velocities in a swarm of planetesimals that may collide but do not grow. (The computer calculates new orbits for bodies that collide but does not allow them to coalesce.) The extension of the calculations to include the case of planetesimals of differing masses is currently in progress.

A central problem is the need to understand the circumstances that cause a swarm of planetesimals to evolve into exactly four bodies—the terrestrial planets—rather than some other number. Again there are competing processes. Collisional damping always tends to terminate planetary growth by isolating the larger bodies in almost circular orbits. Gravitational perturbations work to the opposite effect: they tend to increase the orbital eccentricities and to move the orbits radially, allowing encounters and growth to continue. Sooner or later, when all the larger bodies are in nonintersecting orbits and all the smaller bodies that cross these orbits have been swept up, the tendency toward nearly circular orbits wins out. Unless long-range gravitational perturbations are then brought into the simulation, no influence remains that can bring the orbits of the large bodies into intersection. The question is how many large bodies then exist.




**NEAR-MISS of the two planetesimals perturbs their orbits, so that each one follows a new elliptical trajectory around the sun. The velocity of Body I with respect to the body that did not take part in the interaction has decreased; that of Body II has increased. (The velocities with respect to the third body are important because the two bodies that have interacted are unlikely to be the ones that interact next.) On the average, near-misses increase the relative velocities of the planetesimals. The near-miss whose result is diagrammed here brought the two interacting planetesimals at their closest to a separation of three times the radius of either body.**

Some simplified numerical calculations of the simultaneous growth of several planets have been made by Larry P. Cox of the Massachusetts Institute of Technology. In his work the orbits of 100 bodies of equal size are assumed at first to be randomly distributed throughout the region of the solar system now occupied by the four terrestrial planets. The bodies perturb one another and at the same time collide and grow. Eventually only bodies in nonintersecting orbits remain. To what extent do the spacing and the sizes of the final "planets" resemble those of the terrestrial planets?

Cox found that the answer depends on the range of initial eccentricities. When the range includes bodies in fairly elliptical orbits, a surprisingly close resemblance to the real terrestrial solar system evolves. In one simulation the number of final planets was six. One of them was a planetesimal quite near the sun that never collided with another body. The other planets might whimsically be called Mercury, Venus, the earth, the moon and Mars.

When lower initial eccentricities are chosen (in better agreement with Safronov's work), as many as 10 planets form. To some extent the dependence on eccentricity results from the fact that Cox's calculations are confined to two-dimensional motion. In this "flatland" solar system the orbits of the planetesimals often intersect. Therefore the probability of collisional damping (as opposed to gravitational perturbation) is greater than it would be in a three-dimensional system. My own three-dimensional simulations confirm this. Collisional damping would also be more important in a swarm whose members have differing initial sizes. A calculation that starts with equal-size bodies should thus introduce, one way or another, an augmented number of collisions in relation to near-misses. Such efforts are being made.

A complete understanding of the formation of planets from planetesimals will require investigation into still another matter: the formation of the planets in the midst of a gaseous interstellar medium. After all, the mass of the solar system consists almost entirely of the highly volatile gases hydrogen and helium, which are the major constituents of the sun and of Jupiter and Saturn. These gases pervaded the early solar nebula, and their role has already been included in the hypothesis describing how planetesimals one kilometer in diameter could have formed from dust grains sticking together.

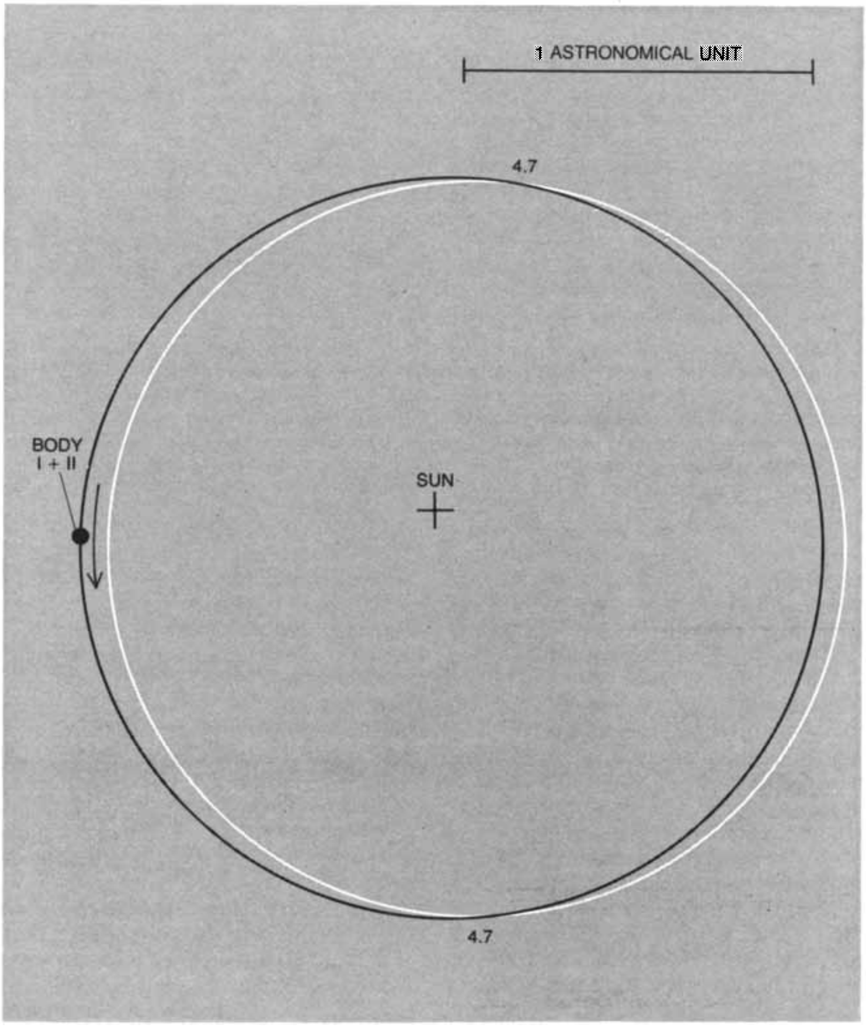
Since hydrogen and helium are not major constituents of the terrestrial planets today, they must have escaped from the terrestrial region of the solar system. Perhaps the early solar wind was more like a solar hurricane. In any event, the presence of the gases must have affected at least the earliest stages of the accumulation of the planets. Additional forces would then have acted on the planetesimals as they plowed through a resisting gaseous medium. The magnitude of the forces depends on the distribution of the medium, which depends in turn on the gravitational fields of the planetesimals and of the planets as they grow. The simulations are therefore more difficult than the ones that neglect the presence of gas. Still, the medium has been included in analytical models constructed by Chushiro Hayashi and his colleagues at the University of Kyoto. Their work embraces not only the early stages of planetesimal growth (the stages Greenberg describes) but also the later stages, in which bodies form that are the size of the moon and larger.

The drag caused by the gas varies with the surface area of the body plowing through it. The change in velocity that results from a given amount of drag is less, however, for a large mass than it is for a small one because the large mass has greater inertia. The effect of a gaseous medium is therefore more important for small bodies than it is for large ones. The effect on a body sufficiently small is that its orbit becomes almost circular. (A body following a circular path meets the least resistance from the gas.) In addition the body spirals slowly toward the sun. For a time the spiraling allows the planetesimals to collide and grow. Then the higher mass of the growing bodies makes the drag force less dominating. As the bodies gain mass, therefore, their spiraling diminishes.

The result would appear to be a swarm of bodies in isolated, almost circular orbits. This expectation, however, overlooks gravitational perturbations among the planetesimals, which increase as the bodies grow. The perturbations augment the velocity of the planetesimals with respect to the gaseous medium. They thereby increase the drag force; indeed, the force increases approximately as the square of the velocity. A planetesimal in a gaseous medium might thus continue to spiral and encounter other bodies until it is more than 1,000 kilometers in diameter.

In Hayashi's simulations bodies as large as 2,000 kilometers in diameter have formed within 10,000 years. It is not clear what would happen after that. To be sure, the gaseous medium has little effect on the first large bodies to form. They are isolated from one another in orbits of low eccentricity. All the rest of the mass of the swarm is then in smaller bodies that have eccentric, spiraling orbits. If these bodies were to collide and accrete, so that the resulting planetesimals had diameters ranging from 100 to 2,000 kilometers, the gravitational perturbations among the resulting planetesimals would increase their velocity with respect to the gaseous medium. The increased drag force would make their spirals more pronounced. It is conceivable that in this way the leftover planetesimals could migrate radially to intersect the orbits of the largest bodies: the embryos of the future planets. If the migrating planetesimals grew too large, however, they might become embryos themselves and take on isolated, almost circular orbits. The result would be an assemblage of planets too numerous and too small.

It has proved to be quite difficult to evaluate the probabilities of these events. Hayashi has predicted, however, the composition of the atmosphere




**COLLISION of the two planetesimals combines them into a single larger body, which takes up a more nearly circular orbit. The velocity of this larger body with respect to the body that was not involved in the collision is shown. On the average, collisions reduce the relative velocities of the planetesimals. In addition the planetesimals tend to become isolated: in their more circular orbits they are less likely to cross the orbits of other bodies and perhaps collide with them.**


**Find the case of C.C.® we hid along the Lewis and Clark trail.**

We retraced Lewis and Clark's historic expedition up the Missouri River into Montana. And where they found their roughest going, we hid a case of the smoothest whisky, Canadian Club.

Where Lewis and Clark had floated the unspoiled river, Canadian Club's rafts followed. We tested our nerves, as the explorers had, on the wild rapids of the Clark's Fork.

As we explored those historic Montana streams, we buried a case of Canadian Club overlooking the very site of one of the expedition's most important sightings. One clue: neither Lewis nor Clark made it.

**Discover the taste of the world's finest whisky.**

As you search along this historic trail, remember Canadian Club has a proud history of its own. For over 120 years it's been enjoyed by those who seek the very best. Canadian Club is lighter than Scotch, smoother than bourbon, and enjoyable on the rocks, in a sour or Manhattan, or with your favorite mixer.

So come search along the Lewis and Clark trail to discover why it's "The Best In The House," or simply venture down to your favorite bar and say, "Canadian Club, please!"


Canadian Club®

**"The Best In The House"® in 87 lands.**

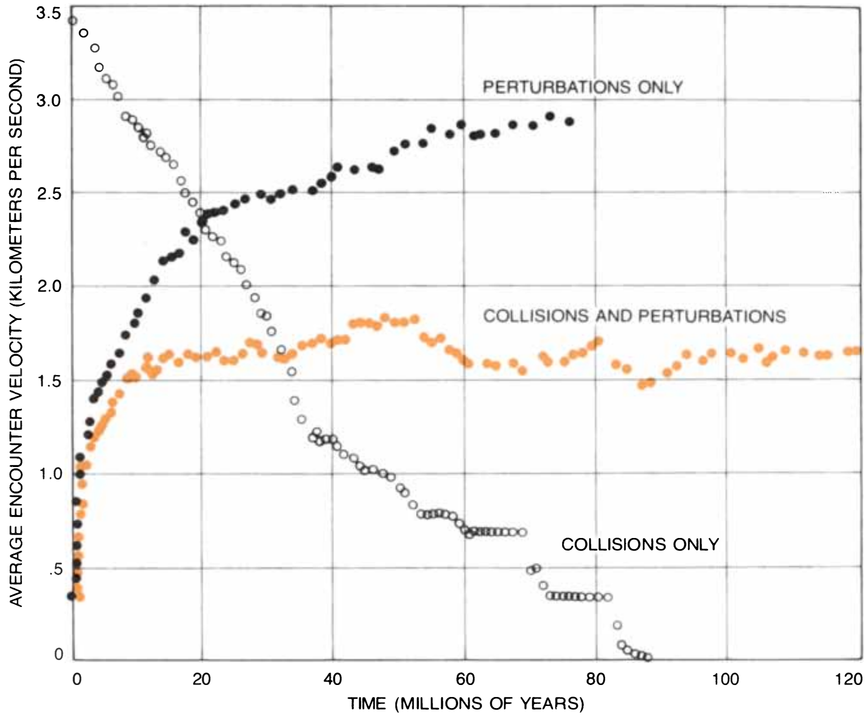
6 YEARS OLD. IMPORTED IN BOTTLE FROM CANADA BY HIRAM WALKER IMPORTERS INC., DETROIT MICH. 86.8 PROOF. BLENDED CANADIAN WHISKY. © 1981

that each newly formed terrestrial planet would have attracted from the surrounding gas of the solar nebula. The atmosphere includes a certain proportion of inert gases such as argon 36. The concentration of such gases can be measured today in the rocks of the earth. Hence it may be possible to learn whether the earth was formed in the midst of a gaseous medium not by theory and simulation but by a more straightforward gathering of data.

Most geologists would rather find out what the primitive earth was actually like than read an account of problems in physics and computer modeling. They will have to be patient. Even at the present stage of the theoretical work, however, it can be shown that some of the many conceivable initial states of the earth are more likely than others. The best example is the earth's initial temperature. Debates on whether the earth formed at a high temperature and was therefore mostly molten or whether it has always been a fairly cold and solid planet have gone on for generations. This makes it all the more remarkable that every current theory of planetary formation predicts a high initial temperature.

First consider the simulations of the accumulation of the terrestrial planets from planetesimals in the absence of gas. During the early stages of gas-free accumulation some of the kilometer-size planetesimals grow into bodies 100 and even 1,000 kilometers in diameter at the expense of their smaller neighbors. Such bodies grow throughout the region of space that will become the terrestrial solar system. None of them is more likely than the others to be the embryo of a planet. As Safronov describes the situation, "all planetesimals are created equal." It can be calculated that few of the planetesimals ever acquire a velocity that would allow them to escape from the solar system. Moreover, the terrestrial solar system today is empty except for the planets and their moons. Hence almost all the planetesimals become part of the planets, and the final stage in the coalescence of the planets is characterized by high-velocity impacts among very large bodies.

As Joseph Barrell, a geologist at Yale University, observed in 1918, the impact of such a body would leave much of the energy of the collision inside the growing planet. In particular the fragments of the collision would all become part of the growing planet; their kinetic energy would therefore be captured. Moreover, the heat the collision generates would take millions of years to pass through 1,000 kilometers of earth. (Even a few feet of earth can preserve a block of ice through a summer.) In effect the heat would be buried with the debris of the collision. Curiously, the only energy that could escape the collision immediately would be the relatively small amount of heat that fragments thrown far from the collision could radiate into space before they fell. Calculations of the thermal effect of the impacts suggest that accumulating bodies not much larger than the moon would melt to some extent, and that an accumulating earth-size body would be partially molten throughout.




**AVERAGE ENCOUNTER VELOCITY in a swarm of planetesimals at any given time is determined by the balance of two effects: near-misses tend to increase the relative velocities, whereas collisions tend to reduce them. Results are given here for a calculation of the velocities among 100 bodies. Each body was assigned a mass of $9 \times 10^{24}$ grams and a radius of 830 kilometers. The initial orbits were chosen to have low eccentricity. After each collision or near-miss the computer recalculated the orbits of the interacting bodies. The graph shows the average encounter velocity under the influence of near-misses and under the influence of collisions. When the effects of both kinds of interaction were combined, the velocity settled at a steady-state value of 1.6 kilometers per second. The calculations were done by the author.**

The accumulation of the earth in the midst of a gaseous medium also entails collisions among large planetesimals. Hayashi has shown in addition that in the presence of a gaseous medium each growing planet attracts a massive initial atmosphere. The compression of the atmosphere by the gravity of an accumulating earth would by itself lead to a surface temperature greater than the melting point of rocks long before the earth was fully formed.

The only alternative to the formation of the earth by the accumulation of planetesimals is the development of a major gravitational instability in the solar nebula, so that the earth falls together of its own weight. A. G. W. Cameron of the Center for Astrophysics of the Harvard College Observatory and the Smithsonian Astrophysical Observatory has shown that this process too entails a high initial temperature.

Over several decades a series of investigators including Harold C. Urey, William W. Rubey and A. P. Vinogradov have presented arguments supporting the opposite view. A more recent proponent of their position has been A. E. Ringwood of the Australian National University. The basis for the opposing view is that the earth can be thought of as a large chemical refinery. If a part of its interior attains a given temperature, certain substances become concentrated in melted rock and others remain behind as a residual solid. Similar chemical separations give rise to crystals of individual minerals whenever molten rock resolidifies. The chemical separations associated with melting and resolidification are the principal reason the composition of the continental crust differs from that of the ocean floor. If the earth had been more or less molten at first, one can argue that the separations that produced the continents should have been completed long ago. The continental crust should be more extensive than it is, and new crust should no longer be forming.

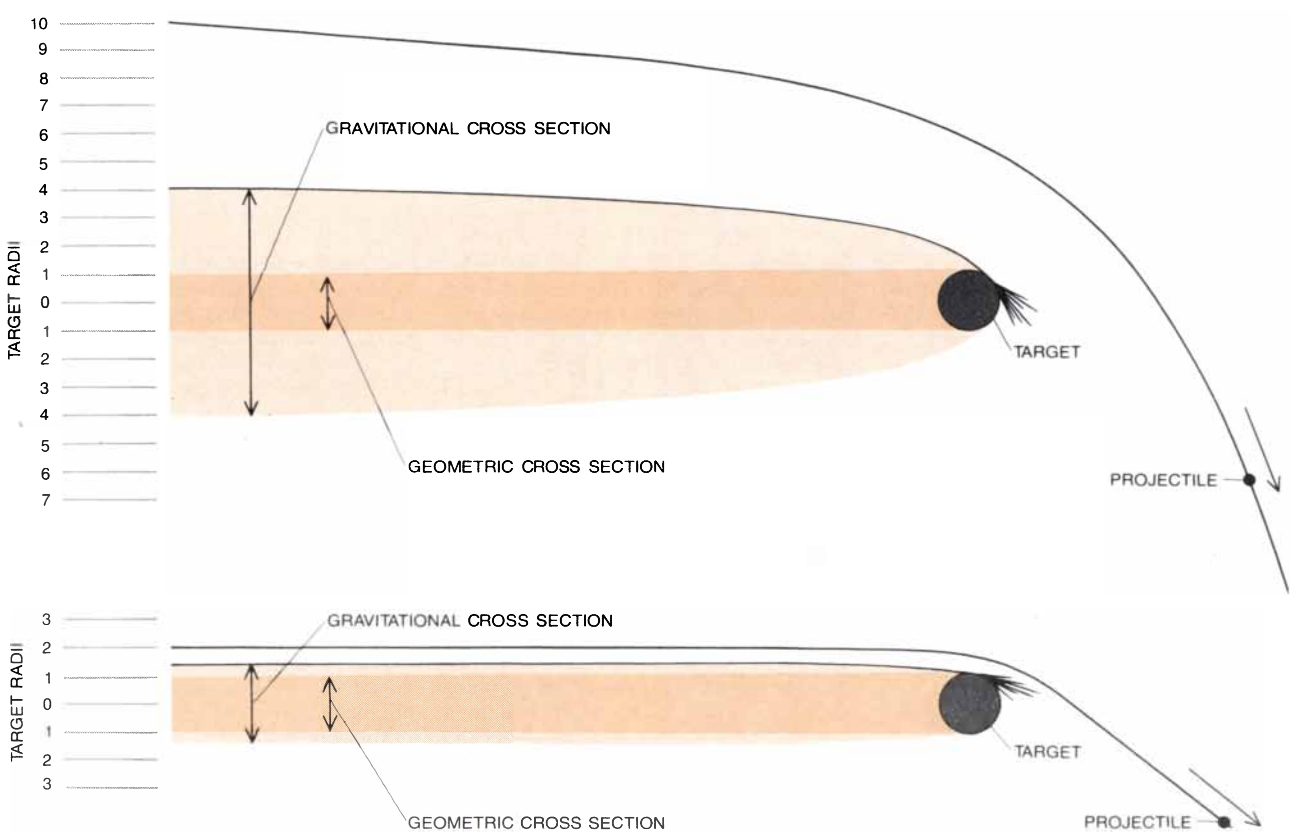
Why then is the chemical differentiation of the crust continuing today? The question deserves a good answer. For now, however, the answer will have to be vague. The heat generated in the earth today by the decay of long-lived radioactive nuclei is transported toward the surface primarily by convection in the mantle of the earth. Only a small fraction of the convecting rock is thought to melt. It is known that small degrees of melting favor chemical separation. Large ions, for example, readily escape from a partially melted crystal lattice. In contrast, the melting of an entire rock leaves the entire mass molten.

It has been calculated that the impact of planetesimals would have heated the coalescing earth at a rate 1,000 times greater than the rate at which radioactivity heats it today. At the time the earth formed, therefore, convection must have been much more rapid. A greater fraction of the rock would have been molten, and it is likely that the transport of heat by molten rock was as significant as transport by solid-state convection. Imagine that the velocity of the convection was rapid, the degree of chemical separation was small and the density of the rocks solidifying at the surface was not much less than the density of the underlying mantle. It then becomes conceivable that the newly generated crust of the earth was simply swept back into the interior. Perhaps, in other words, the tectonics of the earth must run slowly and the amount of melting must be small if the low-density rocks that characterize the continents are to draw apart from the rest and become isolated and stable.

Another quantity of geologic importance that can now be predicted from the theories is the time it took the earth to form. The formation of the earth from a gravitational instability in the solar nebula requires about 100,000 years. The formation of the earth from planetesimals in the absence of a gaseous medium requires 100 million years. Half of the mass of the earth is assembled, however, in some 20 million years. Although the formation of the earth from planetesimals in the midst of a gaseous medium is less well understood, the presence of the medium is unlikely to make the time scale much different.

In contrast with all these predictions, the relative concentrations of various isotopes of lead in the ores on the earth today suggest (on the simplest interpretation of the data, and assuming the earth was hot when it formed) that the formation took significantly longer. Indeed, they suggest that after 100 million years the earth was only half assembled. The theoretical models of the earth's formation all argue, therefore, against the simplest interpretation of the isotopic data. Perhaps a more complex but nonetheless plausible interpretation of the data will ultimately yield knowledge concerning the continuing chemical history of the earth. A healthy science is characterized by such interplay between theory and observations.

Until now most geologists have probably viewed theories of the formation of the earth as falling somewhere in a spectrum ranging from outrageous speculation to innocent entertainment. It is becoming serious business. Much geologic research, as well as the planning of space missions, requires assumptions about how the earth was formed. An understanding of the process is therefore coming none too soon.




**GRAVITATIONAL FOCUSING** causes a large body to capture a small one even though the small one did not seem to be on a collision course. In the upper drawing the velocity of two small bodies ("projectiles") with respect to a large body (the "target") is 1.45 kilometers per second. One of the small bodies collides with the target. In the absence of the target's gravitational field it would have missed the center of the target by 4.1 target radii. The second small body would have missed the target by 10 target radii. Instead its trajectory with respect to the target is bent into a hyperbola and it comes within five target radii. In the lower drawing the velocity of each projectile is 5.8 kilometers per second. The degree of focusing turns out to be less. In simulations of the formation of the inner solar system gravitational focusing can give rise to a final arrangement in which large numbers of relatively small planets have captured all the available planetesimals and taken up orbits that no longer intersect. The target in the illustration has a mass of $10^{27}$ grams and a radius of 3,970 kilometers. Each one of the projectiles is assumed to be small enough so that its gravitational perturbation of the target can be neglected.

# Lead and Silver in the Ancient Aegean

*The two metals are linked because they were smelted from the same ores. The characteristic pattern of the abundances of the isotopes of the lead shows that most of the ore came from only two sources*

by Noël H. Gale and Zofia Stos-Gale

In the eastern Mediterranean the Bronze Age, and particularly its later centuries, was a time of unprecedented splendor. From Troy on the Asian shore to Egypt and beyond to the great palaces of Crete and of Mycenae on the Greek mainland copper-based alloys transformed the arts of war and peace and vast treasures were accumulated, consisting notably of jewelry, vessels and other objects of silver and gold.

The archaeological evidence has presented several questions. What was the nature of the early mining and metallurgy that made these metals available? What role does the base metal lead play in early metallurgy? Were the necessary metallurgical skills developed independently in the Aegean or had they diffused into the region as a result of earlier inventions in the Near East? Finally, what were the sources of the metals? For example, it has often been asserted that Crete and Mycenae imported copper from Cyprus, gold from Egypt and silver from Asia Minor. Did they in fact do so?

The nature of the cultural and trading contacts between peoples is a subject of much interest in Aegean prehistory, but it is one that has been illumined only feebly by limited finds of the pottery of one culture in the domain of another or by often hotly disputed stylistic similarities between, say, daggers or gold and silver objects. The ability to determine the source of the ore of the metal in an artifact, however, together with the discovery and dating of ancient mines and the remains of ancient metallurgical installations, offers an opportunity for more decisive answers to these questions.

Studies along these lines done at the University of Oxford over the past five years have identified the sources of many silver and lead objects of the Aegean Bronze Age beyond reasonable doubt. Here we shall give an account of the work that has brought to the service of Aegean prehistory findings from a variety of disciplines: chemical and physical analysis, isotope analysis, geology, metallurgy and mining technology.

The age of metals, beginning in the Mediterranean world at some time in the middle of the fourth millennium B.C., saw great changes in the nature of early societies. There seems little doubt that the development of the ability to smelt metals from their ores was an important force for the change from the relatively unspecialized and essentially agricultural society of the late Neolithic into the society, based on craft specialization and becoming increasingly hierarchical, that in the Aegean begins with Phase I of the Early Bronze Age (designated EB I and extending from about 3500 to about 2900 B.C.) and develops ever faster in Phase II (EB II, from about 2900 to about 2100 B.C.).

Those who study the prehistory of metallurgy and its impact on society have been inclined to focus their attention on copper and its alloys, the arsenical and tin bronzes, for two reasons. The first reason is the assumed importance for social change of these utilitarian metals, suited as they are for the making of superior weapons and for allowing the development of craft specialization by the provision of new tools. The second reason arises from the presumption that the archaeological record supports the view that the first metal to be smelted from its ores was copper. Both views can be challenged.

In the first place the degree of social impact of metallurgy of any kind depends on the recognition of desirable qualities in the metal produced. Such qualities are of two types. One is utilitarian, a quality the bronzes certainly possess. The other might be called aesthetic: a recognition that certain metals are both attractive in appearance and rare. These are the qualities possessed by both gold and silver. Although neither metal was particularly useful for purely practical purposes, both soon gained esteem, and objects made of gold and silver became coveted possessions. Indeed, it can be argued that the emergence of concepts of wealth, power and hierarchical status associated with the possession of gold and silver probably did as much to change the nature of early societies as the widespread utilization of the copper alloys did.

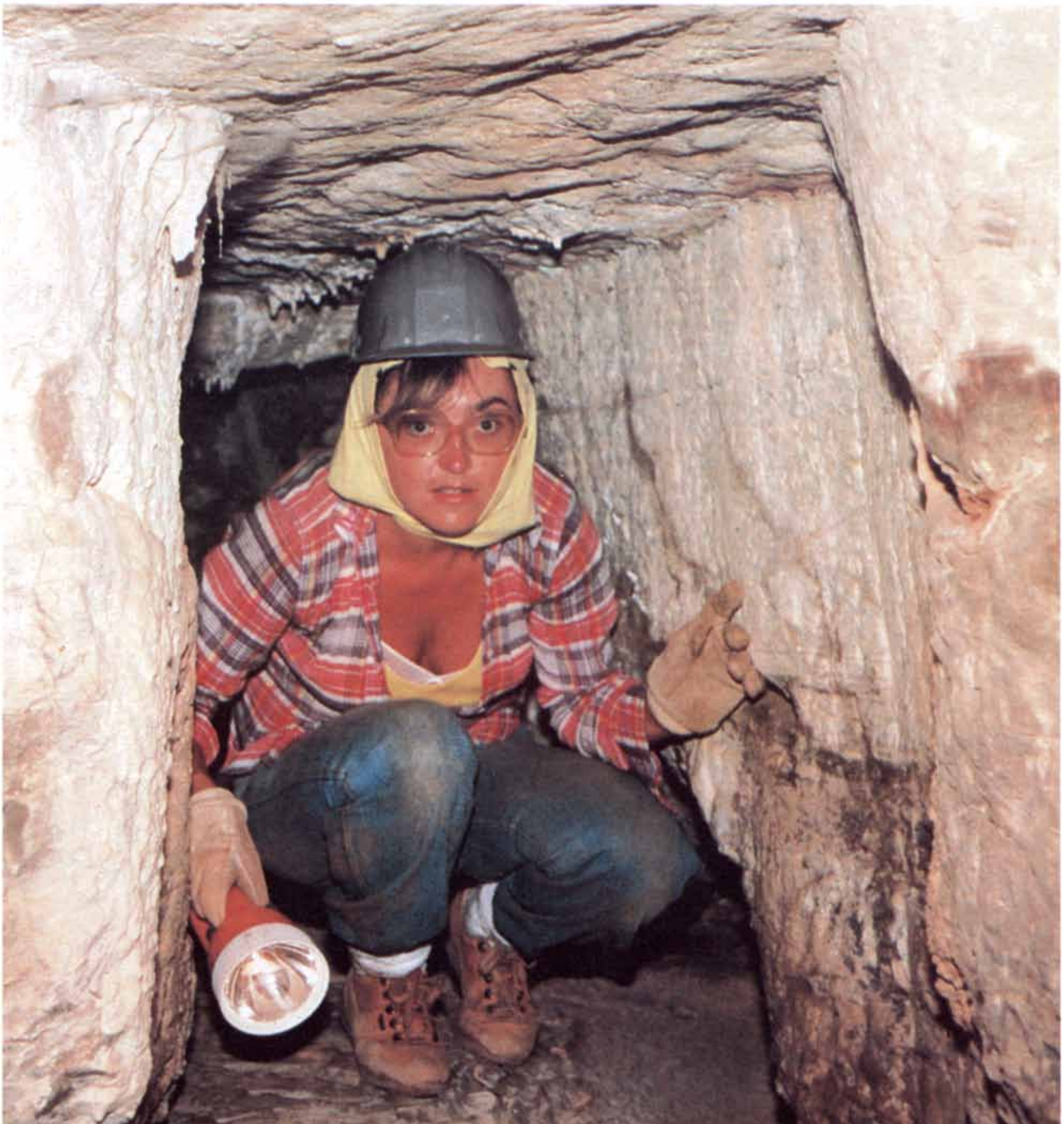
The reason the presumption that copper was the first metal to be smelted from its ores can be challenged is as follows. It is well known that substantial amounts of copper occur as native metal, and it is quite likely that this uncombined form of copper was the first copper utilized by man. Yet native copper does not seem to have been greatly prized for making weapons or tools or to have been exploited in large quantities, if one may judge from the relatively few proved examples in the archaeological record. If such was indeed the case, it is doubtful that the Early Bronze Age supplies of native copper would have been depleted to the point where it would have been necessary to turn to the smelting of ores.

Moreover, the incentive to do so probably would not have been strong. Tools and weapons made of native copper would first have had to compete with those made of flint, obsidian and other kinds of stone, and they would not have had any great advantage over them. In the Aegean the initial delay of a millennium between the first appearance of copper objects and the metallurgical revolution of the Early Bronze Age Phase II can probably be ascribed to the delay in the development of the ability to make arsenical or tin bronze alloys, with their superior metallurgical properties.

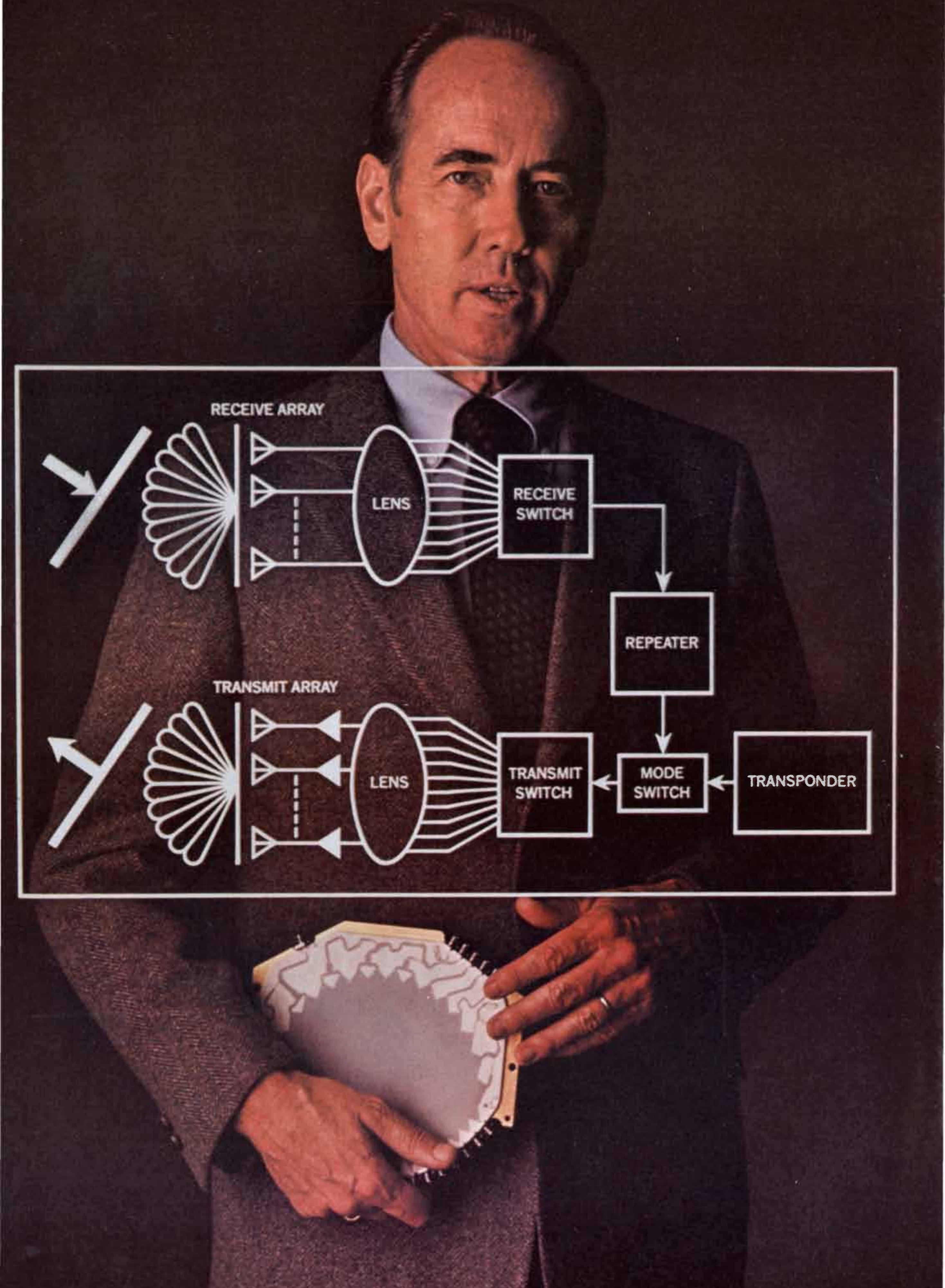
Arsenical bronze was almost certainly discovered accidentally when copper ores that contained arsenical impurities were smelted. The natural bronze so produced would soon have been found to be much harder even than work-hardened native copper. The intentional production of arsenical bronze with arsenic minerals may have followed quickly.

Smelting copper is relatively difficult. Unless the early users of metal had already gained experience in smelting more easily reduced ores, it is doubtful they would have attempted the process with such copper ores as green malachite or blue azurite or even brassy yellow chalcopyrite (which at least to some extent resembles native copper in color but has to be roasted before it can be smelted). Although malachite and azurite can be reduced to metal at temperatures well below the melting point of copper (1,083 degrees Celsius), the copper remains disseminated and unavailable until the temperature rises high enough to melt the copper and turn the gangue, or unwanted rock minerals, into a fluid slag; the result is two immiscible liquids at the bottom of the furnace.

In practice the smelting of all three of these common copper minerals calls for a temperature of about 1,200 degrees C. This fact disposes of the old



**MINE GALLERY in the Laurion area southeast of Athens was precisely chiseled through a matrix of marble, leaving a crawl space about a meter high and half a meter wide. When the gallery reached a vein or larger body of the lead-silver ore galena, the gallery was enlarged and the ore was removed. This is a gallery in Mine No. 5 at Thorikos. It was probably cut in the fifth century B.C. The mining of ores in the Laurion began thousands of years earlier, in the Bronze Age. The investigator in the picture is the coauthor, Zofia Stos-Gale.**


RECEIVE ARRAY
LENS
RECEIVE SWITCH
REPEATER
TRANSMIT ARRAY
LENS
TRANSMIT SWITCH
MODE SWITCH
TRANSPONDER

fancy about the discovery of copper smelting when a campfire was built accidentally on an exposure of copper ore; campfires do not attain such a temperature. In contrast, the commonest ore of lead, galena (lead sulfide), is so easily smelted that it would indeed be possible to extract lead from it in a campfire surrounded by a ring of stones. Experiments by Ronald Tylecote of the Institute of Archaeology of the University of London prove that lead can be smelted from galena in a fire burning charcoal or dry wood at temperatures below 800 degrees C. (This is a temperature well above the melting point of metallic lead: 327 degrees C.) Thus it is much easier to believe that the discovery of the smelting of lead from shiny metallic-looking galena was accidental.

What are the first evidences of lead smelting? Unlike native copper, native lead is rare, so that when lead metal is found on an archaeological site, it almost certainly represents the smelted product. At the city site of Çatal Hüyük in Asia Minor beads of lead have been uncovered that date back to about 6500 B.C. Lead has also been found in a sixth-millennium-B.C. context at Yarim Tepe in Iraq, at the fifth-millennium site of Arpachiyeh in Iraq and at the fourth-millennium sites of Anau and Hissar III in Iran and Naqada in Egypt. The finds suggest that lead smelting, probably on a small scale, began at least as early as the seventh millennium B.C.

The identification of definitely smelted copper is more difficult. Since native copper is not rare, finds of copper artifacts in early archaeological contexts are not necessarily evidence of the early smelting of copper from its ores. The artifacts may merely be hammered native copper. Neither do chemical analyses distinguish an artifact of a fairly pure native copper from one made of copper smelted from a fairly pure ore. Furthermore, copper ores such as malachite or azurite found in early archaeological sites are just as likely to have been used for cosmetics and pigments as for smelting.

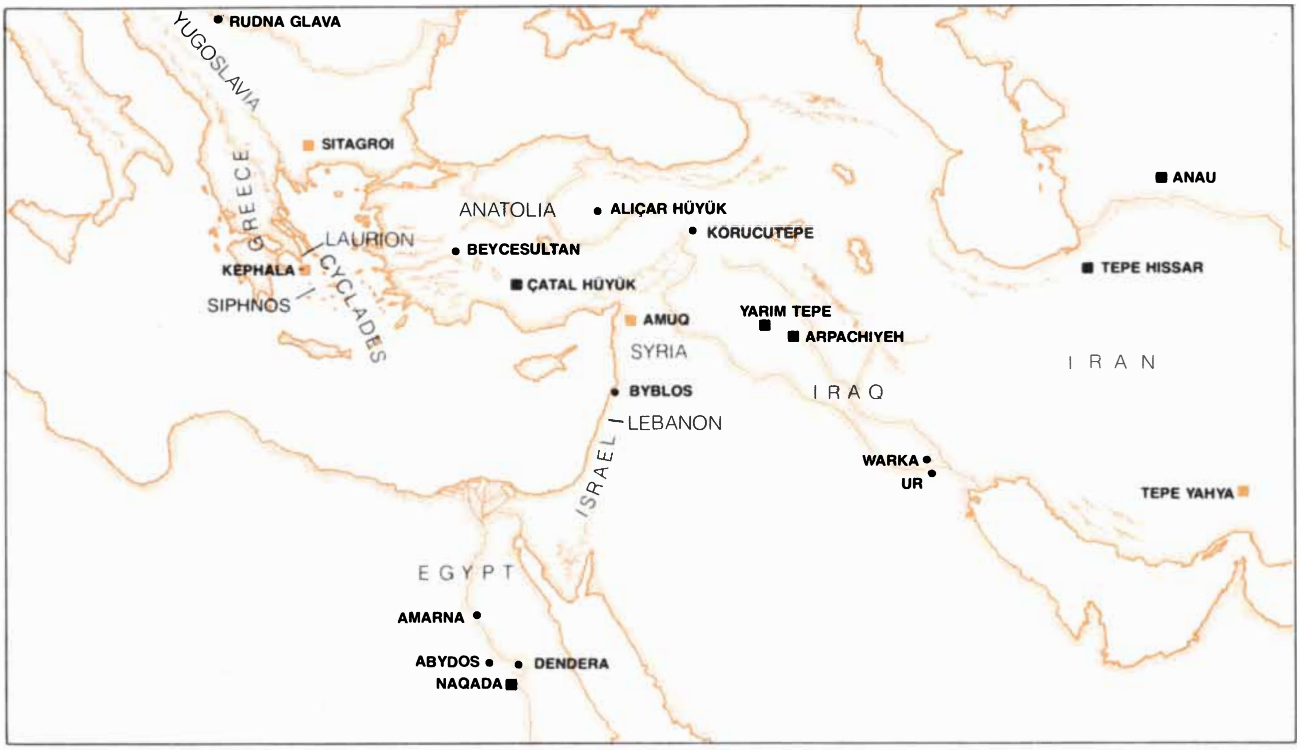
The only certain early evidence of smelted copper is either finds of artifacts of arsenical copper or finds of artifacts of pure copper associated with undoubted smelting slags. Very few finds of early copper artifacts pass either test. Those that do suggest the antiquity of smelted copper is substantially less than that of smelted lead. The oldest examples of smelted copper are from early fourth-millennium levels at Tepe Yahya in Iran. Copper mines at Rudna Glava in Yugoslavia are dated to about 3700 B.C., and smelted copper is found at the mid-fourth-millennium sites of Sitagroi III and Kephala in Greece and of Amuq in Syria. Hence it can be argued that the first metal smelted by man was lead, with a margin over copper that on the existing evidence approaches 3,000 years.

What has all of this to do with gold and silver? To focus on silver, native silver is not as rare as native lead, but it is only two-tenths of a percent as abundant as native copper. It is therefore too scarce to have been an important source of silver to ancient man. Silver is, however, present in lead ores such as galena and in various complex lead-antimony-silver minerals. Moreover, when such ores are smelted, the silver is carried off with the lead, whereas other elements in the ore, such as iron, manganese, silicon, calcium and aluminum, pass chiefly into the slag.

The silver is separated from the lead by the process known as cupellation. The lead-silver alloy is melted in a crucible and held at a temperature of about 1,100 degrees C. as air is blown over it. The air oxidizes the lead, transforming it into litharge (lead monoxide). Impurities such as copper, tin, antimony, arsenic and bismuth are largely oxidized with the lead; the silver (containing a trace of gold) is not oxidized, and when the litharge is absorbed into the walls of the crucible (or removed by mechanical means), a molten globule of silver remains. Silver obtained by cupellation always contains some residual lead; the amount ranges from 2 percent down to .05 percent.

The technique of cupelling lead to win silver may have been discovered about as early in the prehistory of metallurgy as the technique of smelting copper, perhaps even earlier. A cemetery site of the mid-fourth millennium at Byblos in Lebanon has yielded more than 200 silver artifacts. None, however, has yet been analyzed to determine if it was made from cupelled silver. The same is true of a few silver objects of the late fourth millennium that have been unearthed in Palestine, at Ur and Warka in Mesopotamia and at Beycesultan, Aliçar Hüyük and Korucutepe in Asia Minor. In Egypt at least 26 silver artifacts have been discovered that are of Predynastic age, that is, earlier than 3000 B.C. We have analyzed one of them, a silver box lid from Naqada, made in about 3600 B.C. It has a lead content of .4 percent and is therefore definitely an example of cupelled silver.

The discovery that silver could be obtained by cupelling silver-rich lead must have stimulated an increased interest in the smelting of lead. Lead in itself is soft and malleable, tarnishes easily and was at first of no great utility. Indeed, it may be no accident that the appearance of quantities of silver artifacts in the fourth millennium coincides with or somewhat predates the evidence for the start of copper smelting. An increase in the smelting of one group of ores could be expected to lead to an increase




**EVIDENCE OF SMELTING** early in the age of metals has been found at the sites indicated here. Black squares represent the smelting of lead and colored squares that of copper. The oldest definite evidence of copper smelting, early in the fourth millennium B.C., is from Tepe Yahya in Iran. The oldest evidence of lead smelting, in the mid-seventh millennium B.C., is from Çatal Hüyük in Turkey.


**AEGEAN ORE DEPOSITS,** many of them unmined in the Bronze Age or in Classical Greek times, are indicated. Black squares designate lead ores and colored squares gold ores. Solid black triangles designate ancient lead mines and open black triangles evidence of processing activity such as slag heaps and litharge (lead monoxide). If the solid triangle surrounds a colored dot, it indicates modern mining.

in the efficiency of smelting practices in general.

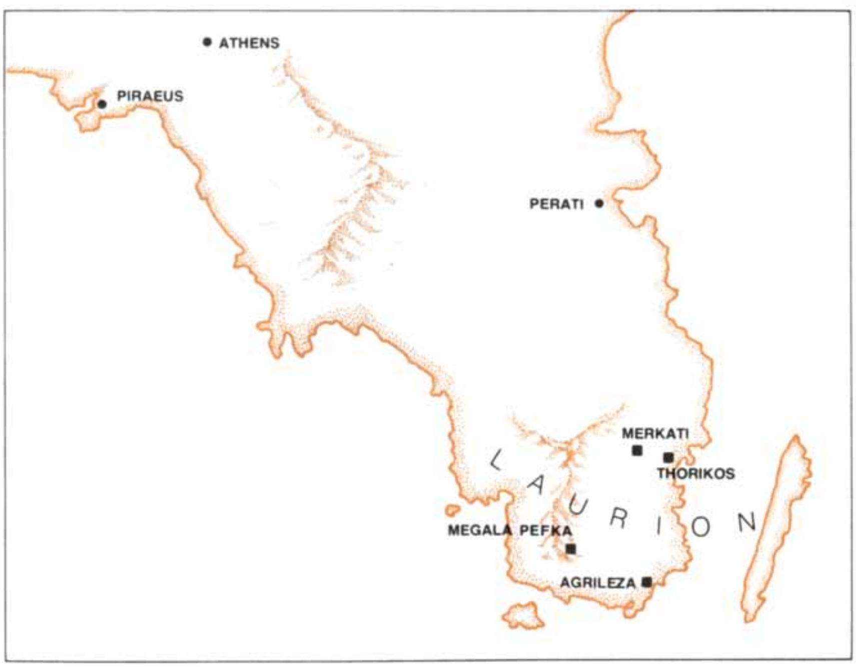
The earliest lead smelting was probably the relatively easy low-temperature smelting of pure galena ore. The stimulus provided by the demand for silver, however, would have made it necessary to smelt impure ores in which the galena was mixed with gangue. The separation of lead from such impure ores would have necessitated a smelting temperature of about 1,200 degrees C. and also the experimental addition of other minerals as fluxes to promote the separation of the gangue as a molten slag. Similar experiments with other minerals of promising appearance and with higher smelting temperatures may have ushered in the Bronze Age.

The fact that cupelled silver contains small amounts of lead allows the fingerprinting, so to speak, of ancient silver artifacts. Ordinary chemical analyses of metal objects may reveal something of how the metal was extracted from the ore or pinpoint the first appearance of arsenical or tin bronzes, but with lead, silver and unalloyed copper artifacts such analyses are of limited usefulness in determining the kind of ore or the location of the mine. The reason is that the processes of smelting and cupellation—with the addition of fluxes, the partition of elements between the slags and the metal and the differences in the rate at which different metals volatilize—disrupt the pattern of major, minor and trace elements characteristic of the original ore.

Isotope analysis, on the other hand, yields accurate data on the proportion of the four lead isotopes (Pb-204, Pb-206, Pb-207 and Pb-208) in the metal. The chemical processes of smelting, refining and corrosion leave the original isotope composition unaltered. Hence the proportions of the lead isotopes in an artifact can be compared directly with those of the ores from various ancient mines. The lead-isotope compositions of different lead ores or lead-and-silver ores differ because some of the atoms of lead 206, lead 207 and lead 208 are formed respectively by the radioactive decay of uranium 238, uranium 235 and thorium 232. Different ores will have a different lead-isotope composition depending on their geologic age and on the relative amounts of uranium and thorium initially present in the ore-forming fluids.

We have now spent half a decade in a comprehensive study of early lead, silver and gold metallurgy in the Aegean world (Egypt included). The earlier phases of our research were undertaken in collaboration with workers at the Max Planck Institute for Nuclear Physics at Heidelberg, who helped us in field work, in dating and in some of the chemical analyses. Our objectives were


**ISLAND OF SIPHNOS in the Cyclades was famous in Classical times as a source of silver. The Siphnian ores are now known to have been a mixture of lead, antimony and silver. They were mined at a number of coastal and inland sites. Both the carbon-14 dating and thermoluminescence dating show that the ores at Agios Sostis, on the northeast coast, were being mined early in the Bronze Age. They were mined for their silver, but the lead was also utilized.**



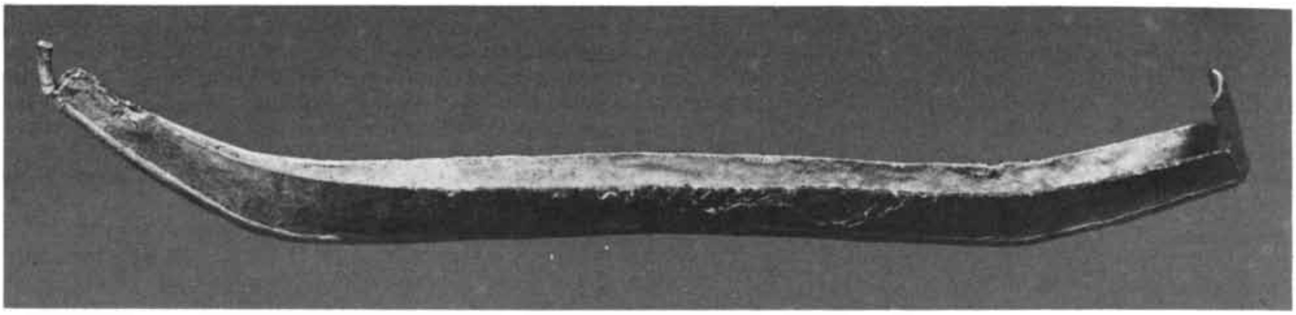

**PENINSULAR LAURION, with its numerous lead mines (*squares*), was the most famous source of silver in Classical times. Both isotope studies of ancient artifacts and evidence from Mine No. 3 at Thorikos show that Laurion ores were first exploited early in the Bronze Age.**

several. We intended to find, investigate and establish the age of various ancient Aegean mines, to analyze the mining techniques of the time and to locate and study furnace remains, slags, litharge and other remnants of smelting and cupellation. By chemical and lead-isotope analyses of ancient artifacts and of ores from the mines we further expected to learn the mineralogical and geographic sources of lead, silver and gold available to various Aegean cultures from the Bronze Age up to Classical Greek times. These investigations in turn would yield information on cultural contacts and trade routes, subjects of importance both to the prehistorian and to the student of early mining and metallurgy.

In the Aegean area there are numerous deposits of gold and silver-rich lead. Several of them are recorded in Classical literature, for example by the dramatist Aeschylus (525–456 B.C.), the historian Herodotus (490–425 B.C.) and the geographer Strabo (63 B.C.–A.D. 21). The most famous are the rich lead-silver deposits of the Laurion region near the tip of the peninsula of Attica, some 40 kilometers southeast of Athens. In the fifth century B.C. Laurion silver was the main source of the power of the Athenian city-state.

Other mines mentioned by Herodotus that we have found and explored are the gold mines on the southeast coast of the island of Thasos (on the west coast of which there are many ancient lead-silver mines), some of the ancient mines of the Pangaeon region in Macedonia (on the mainland to the north of Thasos) and the lead-silver mines of the Cycladic island of Siphnos (90 kilometers southeast of the Laurion region). We have also sampled ores from modern mines and from unexploited mineral deposits around the Aegean and have collected samples of slag and litharge for analysis.

Of the 31 regions of lead-silver mineralization we have studied many were probably not worked in the Bronze Age and some not even in Classical times. Even where we have seen traces of "ancient" mining or heaps of "ancient" slag the period of activity has in most instances not yet been established and may be later than the Bronze Age. Attempts to assign a date to a mine on the basis of the kinds of tools found in it or, in the absence of tools, the kind of toolmarks on the mine walls are of doubtful value because advances in ancient mining technology were extremely slow. Reliable dating calls for the time-consuming search of the mine galleries for charcoal that can be dated by carbon-14 analysis or for potsherds that can be dated on a stylistic basis or by thermoluminescence analysis. The charcoal is usually left from the ancient method of breaking rock by building a fire against it and then dashing water on it, or perhaps from the method of ventilating a



**LEAD BOAT MODEL is one of a group found in an Early Bronze Age grave on the Cycladic island of Naxos. It is now in the Ashmolean Museum in Oxford. Isotope analysis of the lead indicates that it came to Naxos from Siphnos early in the third millennium B.C.**

mine by building a fire at the bottom of a vertical shaft to create an updraft.

The important mines of the Laurion contain galena relatively rich in silver (between 500 and 5,000 grams of silver per ton). Evidence of exploitation of the Laurion by the Athenians and other Greeks in the fifth and fourth centuries B.C. has been intensively investigated by geologists and archaeologists since the 19th century. The region is honeycombed with ancient mines that worked ore lying mostly at the contacts between schist and marble. The ores were partly sorted from gangue inside the mine and then were carried out of the mine. There they were broken into smaller pieces and were further separated on rectangular or helicoidal washing tables of ingenious design. The water for the washing was stored in large cisterns lined with an impervious hydraulic cement made with the litharge obtained as a by-product of cupellation. One of the most important centers in the Laurion seems to have been one at Thorikos; it has been excavated by Herman Mussche and Paule Spitaels of the University of Ghent. The site included a temple, a theater of the sixth century B.C., public buildings, metallurgical workshops, ore washeries, mines and a harbor.

Herodotus writes that the Siphnos mines had made that island the richest in all the Cyclades, so rich that a tithe of the silver was sufficient to furnish a richly decorated marble treasury in the shrine of Apollo at Delphi. Pausanias records the legend that later the Siphnos mines were flooded and the flow of silver from the island was ended because Apollo became angry when the Siphnians failed to pay their tithes.

A zone of iron-manganese mineralization strikes across Siphnos from north-northeast to south-southwest in shattered Tertiary marble. It is within the iron-manganese mineralization zone that the ancient mines are found. Mining for the iron ore in the 19th century did considerable damage to the galleries dug by the ancient miners, who ignored the iron but were so efficient in their pursuit of the lead-silver deposits that today it is difficult to find any trace of lead ore.

Toolmarks are abundant in these narrow, irregular galleries. Niches for oil lamps are cut here and there in the walls. A feature of the mining technology at Siphnos (and also observed in the mines on Thasos) is that most of the waste rock cut away to extend the galleries was not carried to the surface but was piled up on the spot in the form of waste walls.

The mines at the north end of the island, on the peninsula of Agios Sostis, are at sea level, and some galleries that extended out under the sea are now flooded, in accordance with the legend recorded by Pausanias. Ancient galleries are also found high up in the interior of the island at Agios Sylvestros, Vorini, Kapsalos and Xero Xylon. There are also traces of ancient mining at Agios Joannis, the site of a modern mine, and slag and litharge are found at Plati Yialos and Kapsalos.

In our first two seasons of field work on Siphnos we found no samples of ancient ores. Eventually small amounts were discovered at most of the Siphnian sites we surveyed. They proved to be highly oxidized lead-antimony-silver ores: yellowish, heavy, brittle and fine-grained. Lead-isotope analyses of these ores and of samples of slag and litharge have enabled us to define the characteristic "field" of lead isotopes in the ores of Siphnos. The isotope field of the Siphnian ores differs from that of any other island in the Cyclades and also differs from the field of the ores of the Laurion.

At about this time, following a suggestion by Colin Renfrew, who is now at the University of Cambridge, our Oxford group extended its studies beyond the Classical period and began to search for the sources of lead and silver exploited by the Bronze Age cultures of the Aegean region. As a start we examined four models of boats, fashioned from lead, that had been found in a burial of Early Cycladic age (3400–2100 B.C.) on the island of Naxos and had then been presented to the Ashmolean Museum in Oxford by that preeminent figure in Minoan archaeology, Sir Arthur Evans. To our surprise we found that the isotope field of the lead boat models, and their high antimony content as well, suggested that the metal had come from Siphnos. The conclusion was inescapable: mining on Siphnos must have begun as long ago as the Early Bronze Age.

Intensive investigation of the Agios Sostis mines on Siphnos has now confirmed this conclusion. The entire east flank of the peninsula is underlain by a labyrinth of ancient mine galleries. The galleries have no regular pattern; they drive through the zones of iron-manganese ore in search only of the relatively scarce pockets of lead-antimony-silver ore.

Over the millenniums the stones in many of the waste walls inside the galleries have been naturally cemented by the percolation of calcareous fluids. One such wall was carefully dismantled in order to determine the stratigraphic position (and therefore the relative age) of any foreign material found in it. The more loosely cemented upper part of the wall contained a fragment of pottery finished with a black slip. Tests of the age of the fragment, based on the thermoluminescence method, yielded two dates of what historians refer to as Archaic times: 660 $\pm$ 100 B.C. and 530 $\pm$ 140 B.C. In the more tightly cemented lower part of the wall were found two more potsherds, of a coarse ware, together with some charcoal. The thermoluminescence ages of the sherds were respectively 2590 $\pm$ 440 B.C. and 2780 $\pm$ 400 B.C. The calibrated carbon-14 ages of the charcoal samples were 2970 $\pm$ 180 B.C. and 2610 $\pm$ 50 B.C.

This evidence that the Agios Sostis mines were being exploited early in the Bronze Age is given additional weight by other finds in the area. The stone hammers found in the galleries are of an Early Bronze Age type. Mixed with the surface scatter of slag and litharge are quantities of obsidian flakes, including one perfect tanged arrowhead; all could be Early Bronze Age artifacts. Finally, among the surface finds of potsherds is one sherd with a mat impression, characteristic of the late Neolithic and Early Bronze Age pottery of the region, and another with the impressed herringbone pattern typical of the period known as Early Cycladic I (3400–2900 B.C.).

The Bronze Age people of the Aegean had a number of uses for lead. The metal has even been found in late Neolithic contexts on the islands of Makronisi and Kea just off the coast of Attica opposite the Laurion. Lead is part of the Early Bronze Age record on eight Cycladic islands (Antiparos, Dhespotikon, Kea, Syros, Amorgos, Ios, Melos and, as we have noted, Naxos). It is also known in Early Bronze Age contexts in Cycladic settlements on the Attic coast (Raphin and Agios Kosmas), on Crete, on Lesbos, on Lemnos and at Troy.

Lead was made into spindle whorls, sinkers for fishnets and wire (often found in burials) and also served for repairing broken pottery. It was employed decoratively for rings and bracelets and for human figurines and ship models. In later Bronze Age times the metal was formed into balance weights that had as their uniform base a unit of about 61 grams. These weights have been found throughout the Aegean in Middle and Late Bronze Age contexts. Notably large numbers of them were unearthed at Akrotiri on the island of Thera near Crete and on Kea.

Silver in the Bronze Age was used to make bowls and cups, but above all it was used to make jewelry: diadems, necklaces, bracelets, rings and pins for dresses. The more elaborate pieces, including some of lead, were buried with their owners, along with such other grave goods as pottery and bronze weapons and tools. This regular removal from circulation of so many desirable objects must have sustained a demand for silver that might otherwise have been satisfied, at least in part, by the passing on of heirlooms. (From the viewpoint of modern research the burial practices of the Bronze Age are fortunate; they make it far less likely that the silver jewelry might have been melted down for recycling along with other silver objects from diverse sources, thereby destroying the isotopic record of its origin.)

Comparison of the finds from various Cycladic graves yields insight into the emergence of social stratification early in the Bronze Age. At first such stratification is evident only as a slight variation in the quantity and quality of the grave goods. Later it is marked by a concentration of precious objects in graves that are conspicuous by the absence of the commoner kinds of pottery and that sometimes hold such princely objects as silver diadems and bowls. The trend was to culminate late in the Bronze Age, exemplified by the stupendous wealth of gold and silver in the shaft graves at Mycenae and the riches of lesser tombs elsewhere, such as the *tholos* (beehive-shaped) tomb at Vapheio near Sparta in Laconia.

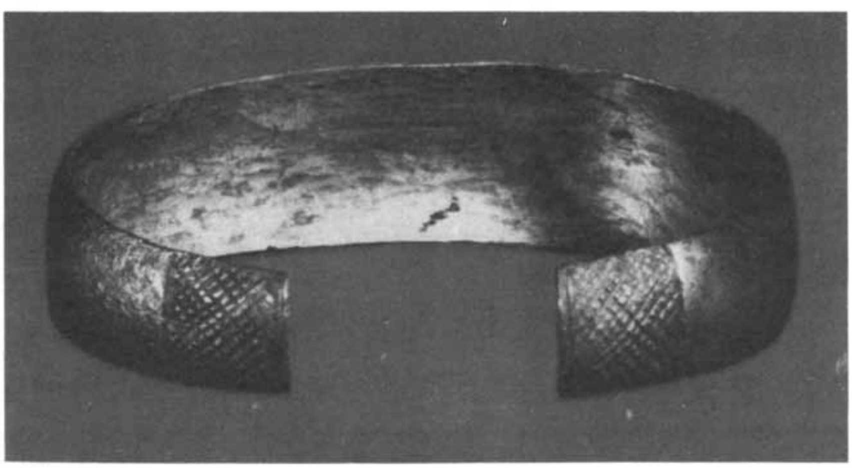
Up to now the sources of lead and silver for the Bronze Age cultures of the Aegean have been quite uncertain, although many archaeologists have suggested that they were to be sought in Asia Minor or far to the west on the Iberian peninsula. Keith Branigan of the University of Sheffield has suggested that the lead and silver first came from sources in the Aegean itself, but that these sources may have become exhausted by the Middle Bronze Age; then, by the Late Bronze Age, they may have been replaced by sources in the western Mediterranean, in Asia Minor or perhaps in central Europe. We began to investigate this question first with respect to the Early Cycladic archaeological record.

From the Early Cycladic period the known silver objects include two diadems, at least 13 bracelets, seven dishes, two necklaces and at least five silver pins. The lead objects include three figurines, the four Naxos ship models, three weights, a seal, several rivets for pottery repair and various fragments. This material has been found chiefly on the islands of Amorgos, Naxos, Syros, Antiparos, Dhespotikon, Ios, Makronisi and Pholegandros. Most of it has been found in graves together with pottery, characteristic Cycladic marble figurines with folded arms, marble bowls, bronze objects and other grave goods. The artifacts provide a precious record of the vitality and originality of the Early Bronze Age culture of the Cyclades, the achievements of which are one of the high points of prehistoric art.

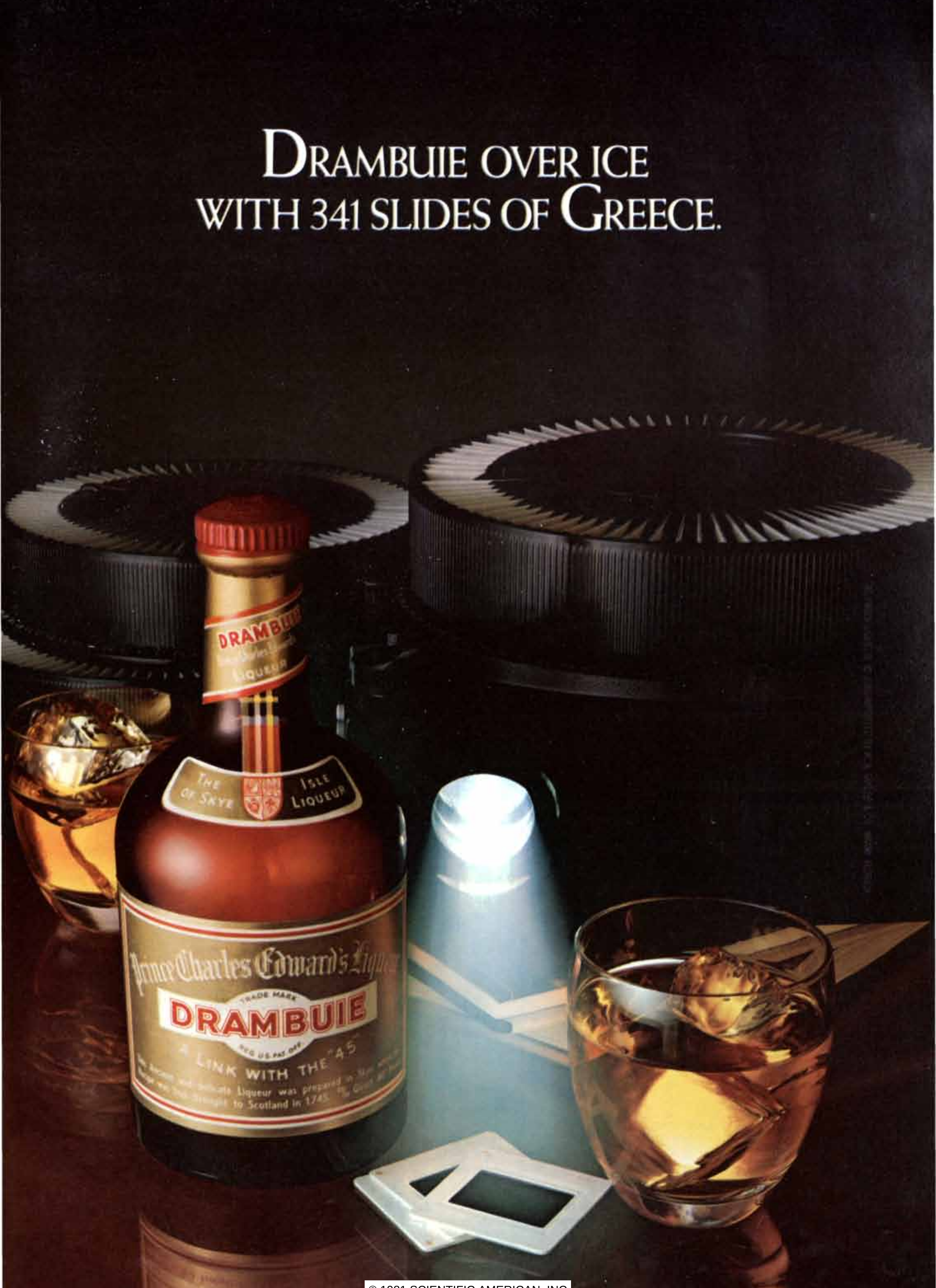
Through the generosity of museum curators in Oxford, London, Liverpool and Athens we have been able to analyze 16 of the objects, or about a third of the total. The picture that emerges is a surprisingly simple one. Six of the 16 objects are of metal from mines in the Laurion and eight are of metal from mines on Siphnos. The remaining two are of metal from a source not yet identified with certainty. We have found lead-silver ore deposits on at least six Cycladic islands other than Siphnos; on two of the islands, Syros and Seriphos, not only are the ores rich in silver but also there are signs of ancient mining activities. Thus the result of our analyses was not altogether unexpected.

Was this pattern of metal sources maintained later in the Bronze Age? In a preliminary attempt to answer the question we have analyzed 24 lead artifacts excavated at Akrotiri on the island of Thera. Akrotiri, the Pompeii of the Bronze Age, was an extensive settlement when it was buried under the ash of the volcanic explosion that rent Thera in about 1500 B.C. The mansions, frescoes, architecture, street plans, sanitation and drainage arrangements, pottery and other artifacts preserved in the ash suggest that Theran society was rich, highly civilized, dependent on maritime trade and marked by a lively and original appreciation of the arts.

The status of Akrotiri with respect to Minoan Crete is uncertain. It is not known whether Akrotiri developed independently of Crete, while enjoying a flourishing trade with it and remaining politically unaffiliated, whether it was conquered and governed directly by some Cretan palace center or whether the Cretans had set up a self-governing colony there. The presence of a large quantity of lead, some silver and some litharge suggests the production of both lead and silver on Thera. Our chemical and isotopic analyses of this material show the same simple dichotomy of source we found in the objects from



**SILVER BRACELET of the Early Bronze Age, found in a grave on the Cycladic island of Amorgos, is one of some 30 silver artifacts that are representative of the period. Isotope analysis of seven of the artifacts indicates that one of them is metal from Siphnos and five are metal from the Laurion. The Amorgos bracelet is one of those made of silver from the Laurion.**


DRAMBUIE OVER ICE
WITH 341 SLIDES OF GREECE.
THE ISLE OF SKYE LIQUEUR
Prince Charles Edward's Liqueur
TRADE MARK
DRAMBUIE
REG. U.S. PAT. OFF.
A LINK WITH THE "45"

elsewhere in the Cyclades but an overwhelming shift in proportion. Only one of the Akrotiri lead pieces is of metal from Siphnos; the other 23 are from the mines in the Laurion.

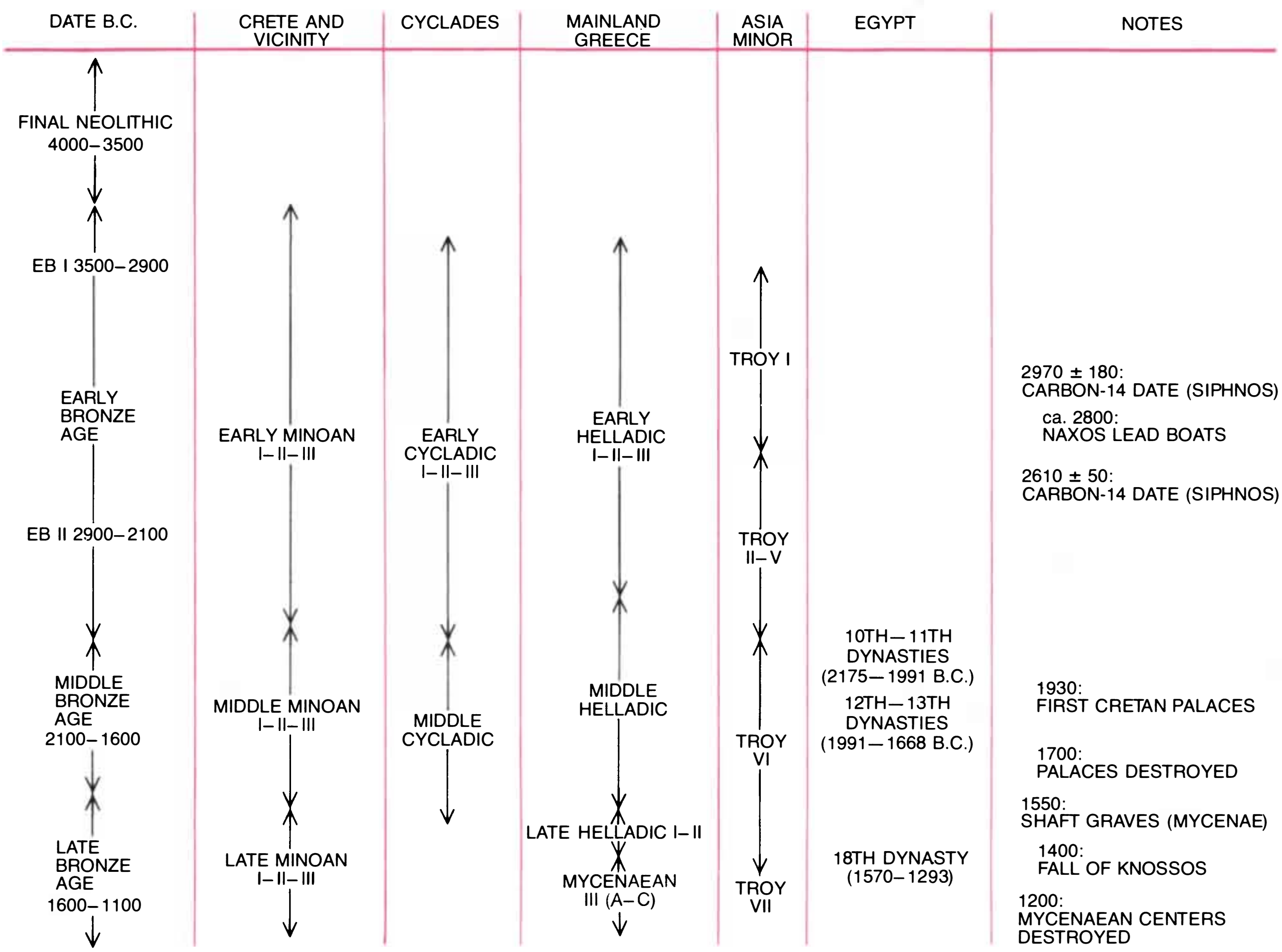
What about Crete? In Early Bronze Age times in the Cyclades artifacts of lead and silver account for about 40 percent of all metal finds, and gold is exceedingly rare. In Crete this state of affairs is reversed. Of all published metal finds dated to the Early Bronze Age on Crete only about 3 percent are lead and silver, and gold is relatively common. The finds of lead and silver are concentrated in the north of the island, between Heraklion and Mochlos, an area where gold and copper objects are rarer. This area directly faces the Cyclades, and its Early Bronze Age artifacts show evidence of contact with Cycladic culture. A suggestion by Branigan that the inhabitants of northern Crete may have obtained lead and silver from the Cyclades therefore seems reasonable. We have not yet been able to put it to the test, but we hope soon to extend our work on Cretan material to include artifacts of the Early Bronze Age.

So far we have examined lead and silver artifacts from only two Cretan sites, Knossos and Kommos, and only for Middle Minoan to Late Minoan times (about 2150 to 1300 B.C.). In the Middle Bronze Age, Crete evolved as a rich hierarchical civilization, distinguished by great palaces, flourishing towns, brilliant craftsmanship and wide connections overseas. The destruction of the first palaces at Knossos and Phaistos at the end of the Middle Bronze Age (about 1700 B.C.) in no way signaled a diminution of Minoan culture. The archaeological record holds no signs of newcomers, and the rebuilding of the palaces ushered in the greatest episode in Minoan civilization: the Second Palace Period (1700–1450 B.C.). The island's craftsmen then did their finest work; the pottery, gemstones, jewelry and frescoes of the period are of a level of aesthetic sensibility and technical skill that secures them a firm place in the history of artistic achievement. At the same time the Minoan script Linear A was developed.

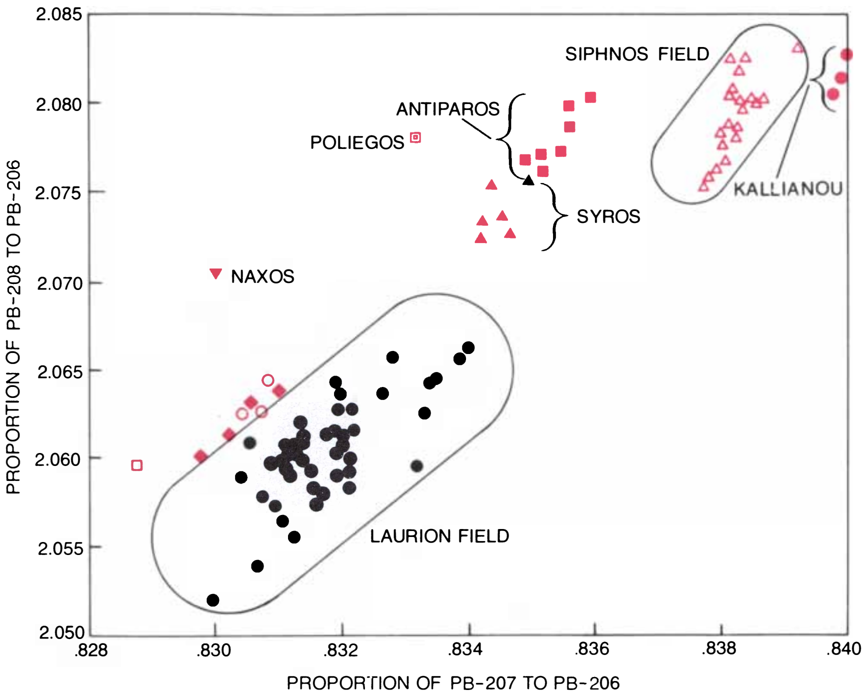
There is abundant evidence of Minoan trade (or other kinds of contact) outside Crete in the Second Palace Period. There was traffic with the Near East, with Egypt and, closer to home, with the Cyclades and with Argolis on the mainland of Greece. Hence lead and silver, which are virtually unavailable on Crete itself, could have come from any of these places.

In about 1450 B.C. almost every major site on Crete was destroyed; Minoan civilization never again rose to its former splendor. Knossos and its environs were occupied by the Mycenaeans from about 1400 to 1340 B.C. If the Mycenaeans then had access to the mines of the Laurion, one might expect to find Laurion lead and silver at least in the levels at Knossos dating to that period (Late Minoan III A).

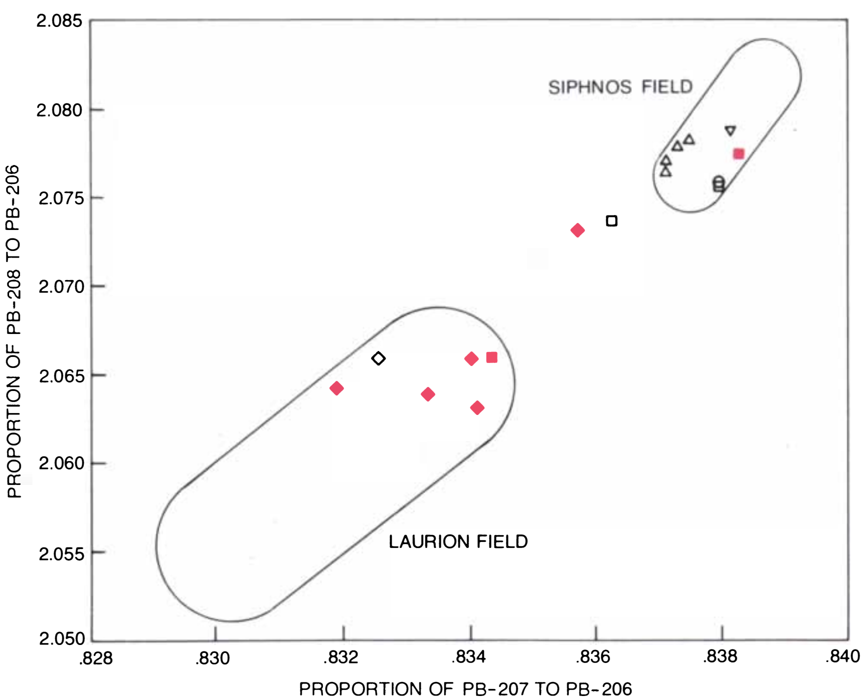
Our preliminary work shows in fact that the predominant source of lead artifacts from Knossos, all the way from the start of Middle Minoan times down to the Mycenaean occupation, was the Laurion. Of the artifacts we have analyzed 80 percent are of Laurion lead. Only 10 percent are of lead from Siphnos, even though Siphnos is much closer to Crete. The remaining 10 percent come from a source that has not yet been identified. At Kommos the paucity




**CHRONOLOGY OF THE BRONZE AGE in the Aegean appears in this table; all the dates are approximate. Comparable periods in Asia Minor and Egypt are included, and key Aegean developments are noted. The earliest Aegean lead mining may predate 3000 B.C.**




**ISOTOPE ANALYSES plot the ratio of lead 207 to lead 206 on the abscissa and of lead 208 to lead 206 on the ordinate to determine the relation between lead ores from mines in the Aegean. All ore samples from Siphnos (*color*) lie within the field at the upper right. All ore samples from the Laurion (*black*) lie within the field at the lower left. The ore samples from Anaphi (*colored circles*), Seriphos (*colored diamonds*) and Kythnos (*open colored square*) border on the Laurion field, but none of the ore samples from four other Cycladic islands or from a mainland mine, Kallianou in Euboea, come significantly close to the Siphnos or the Laurion field.**




**LEAD-ISOTOPE RATIOS of 16 Early Cycladic artifacts, nine of lead (*black*) and seven of silver (*color*), place 14 of them within either the Siphnos or the Laurion field. The four lead boat models found at Naxos (*open triangles*), a lead figurine from Antiparos (*circle*), a fragment of lead from Cheiromylos (*inverted triangle*), a lead weight (*square*) and a silver diadem from Syros all fall within the Siphnos field, whereas four silver bracelets from Amorgos (*solid diamonds*), a silver pin from Syros (*square*) and a lead object from Amorgos (*open diamond*) fall within the Laurion field. A silver bowl from Amorgos and a lead rivet from Syros are of uncertain origin, perhaps a mixture of Laurion and Siphnos metals or metal mined near Troy.**

of the finds suggests that lead was hard to come by. We have analyzed the only three Bronze Age lead artifacts excavated at that site. Two are of Laurion lead. The third is from an unidentified source. Thus the overall dominance of the Laurion as a source for Cretan lead in the Middle and Late Bronze Age is similar to the pattern we found for Thera. The finds from unidentified sources suggest, however, that Crete did have limited access to sources of lead other than the Laurion and Siphnos.

Our analyses of lead and silver artifacts from the Cyclades, from Thera and from Crete yield clear evidence that the Laurion mines were exploited in the Bronze Age from at least 2900 B.C. (Early Cycladic II) to about 1350 B.C. (Late Minoan III A). Further confirmation of mining activity in the Laurion over this span could be provided by analyses of lead and silver artifacts found at Bronze Age sites in Attica itself. In the hope of such confirmation we have analyzed lead vessels and wire from Mycenaean levels of 1400 to 1150 B.C. (Mycenaean III A to III C) at the Acropolis and the Agora in Athens, lead and litharge from an earlier context at Thorikos (1700 to 1550 B.C., or late Middle Helladic) and later lead and silver artifacts from nearby Perati (1190–1100 B.C.). All proved to have the isotope composition (and the chemical characteristics) of lead from the Laurion.

Evidence that the output from the Laurion mines reached more distant mainland points is provided by the analysis of four lead balance weights from Laconia. The weights, which we have proved to be of Laurion lead, were found in the same *tholos* tomb at Vapheio that held famous gold cups embossed with images of Cretan bulls.

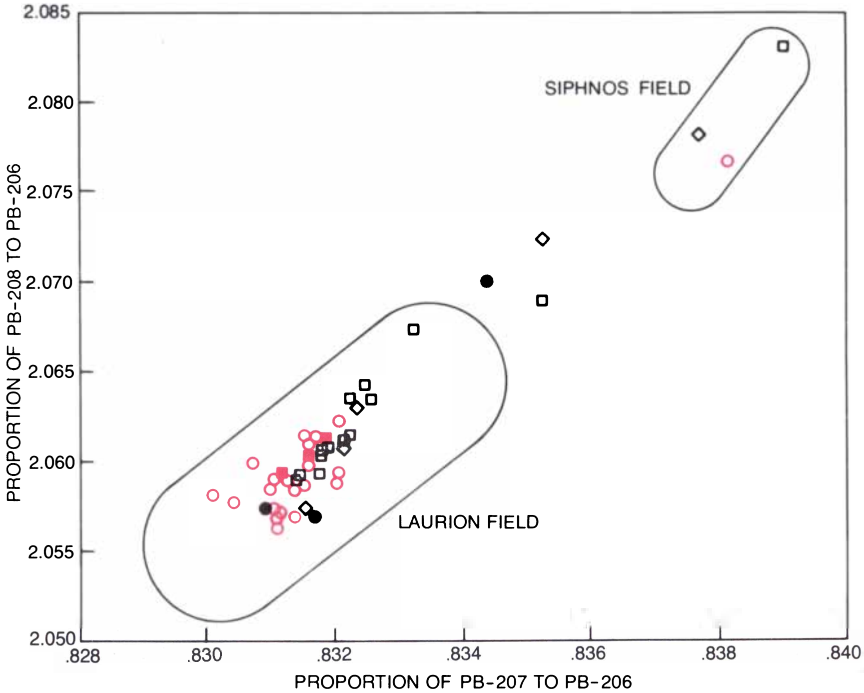
In the 1870's Heinrich Schliemann found a fabulous treasure in the celebrated Shaft Graves at Mycenae that date to about 1550 B.C. (Late Helladic I). Attention is usually focused on the gold artifacts, but the graves also contained massive amounts of silver and some lead. What was the origin of the treasure, and what is the explanation for the rise of Mycenae at the time of the Shaft Graves?

Sir Arthur Evans was of the opinion that the Minoan character of many objects in the graves attested to the occupation of Mycenae by a conquering Cretan dynasty. This view, however, was refuted by a contemporary of Evans', the Austrian scholar Georg Karo, and today the work of Ellen Davis of the City University of New York indicates that most of the goldwork in the Shaft Graves is Mycenaean. Karo himself sought a reason for the sudden Mycenaean acquisition of wealth and for the character of the treasures in the hypothesis that many of the objects found in the graves had been plundered from Crete by raiders from Mycenae; such raiders could not have acquired their wealth by commerce because, according to Karo, they lacked natural resources of their own. Our own work on this question is not yet far advanced, but we have nonetheless demonstrated that Laurion lead was reaching Mycenae in Shaft Grave times. Schliemann found three lead caldrons in Shaft Grave IV, their rims cased in bronze. Analysis shows that the lead is from the Laurion. Similar analyses of four silver objects from Shaft Grave IV show that their silver also comes from the Laurion.

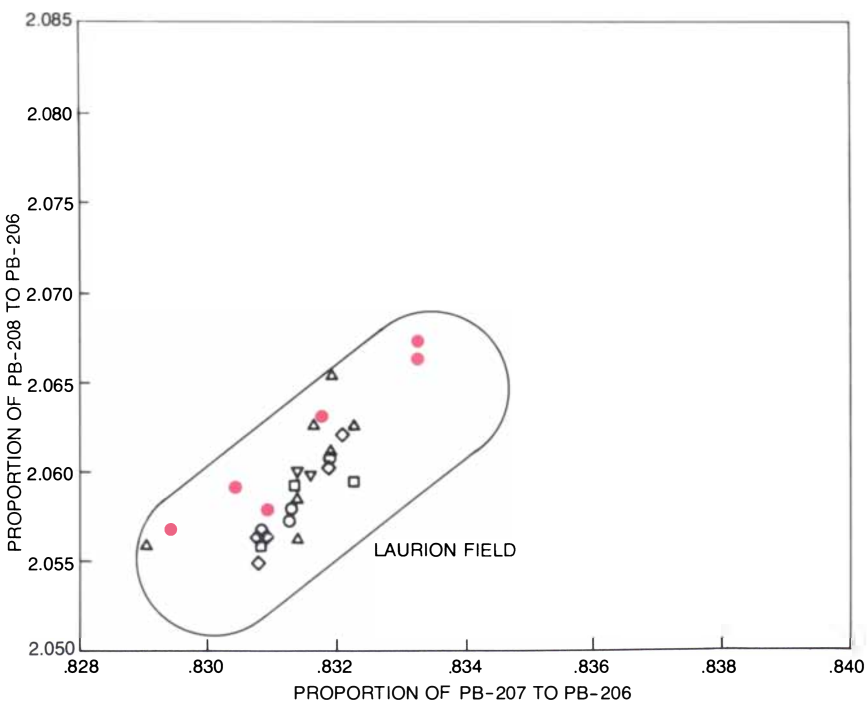
Egypt was noted in ancient times for its riches in gold, but it entirely lacked indigenous sources of silver. As a result silver was highly prized by the Egyptians. Might some of the gold in the Shaft Graves at Mycenae have come from Egypt in exchange for Laurion silver? Our preliminary analyses of lead and silver artifacts from Dynastic Egypt have so far uncovered three of each metal that are of the characteristic Laurion isotope composition. The three lead artifacts are all of 18th Dynasty age (1570–1293 B.C.); they were found at Amarna and Abydos. Two of the three silver objects are respectively of 10th Dynasty (2175–2135 B.C.) and 11th Dynasty (2134–1991 B.C.) age; they are from Dendera. The third, from Abydos, is dated to the 12th or 13th Dynasty (1991–1668 B.C.). Although these results prove nothing directly about the origin of the Mycenaean gold treasure, they do demonstrate that Laurion lead and silver were coming to Egypt both before and after the Shaft Grave period, just as our analyses of Shaft Grave artifacts show that the Mycenaeans of that period had access to Laurion lead and silver.

Before our work little was known with any certainty about the sources of the lead and silver used by the Bronze Age peoples of the Aegean. Now chemical analysis, isotope analysis, thermoluminescence dating, carbon-14 dating, the exploration of ancient mines and the uncovering of the remains of ancient metallurgy have established a number of facts. It is evident that the islanders of Early Cycladic times got these metals about equally from Siphnos and the Laurion. In later times the source of the metals for Thera was overwhelmingly the Laurion, with minor recourse to Siphnos. On Crete in the Late Bronze Age the major source was also the Laurion, although a little metal came from Siphnos and undetermined sources. In mainland Greece the evidence so far obtained from Thorikos, Perati, Athens, Vapheio and Mycenae indicates the Laurion as an exclusive source, and Laurion silver and lead in Egypt have been identified in artifacts of the 10th, 11th, 12th and 18th dynasties.

The Bronze Age exploitation of Siphnos, inferred from the lead-isotope analyses of artifacts, has been substantiated




**LEAD ARTIFACTS** from Thera (*color*) and Crete (*black*), dating from the Middle and Late Bronze Age, prove to be primarily from the Laurion. The objects from Thera belong to the period between 1550 and 1500 B.C. They include 21 lead weights, pottery rivets, fishnet sinkers and doorpost sockets from Akrotiri and three samples of litharge (*solid squares*). Only one artifact was of Siphnian metal. The 22 objects from Crete belong to the period between 2050 and 1330 B.C. at Knossos (*open squares and diamonds*) and to the period between 1550 and 1200 B.C. at Kommos (*solid circles*). Only two of the Knossos artifacts are of Siphnian lead. Two are of metal of unknown origin, as is one of the three pieces unearthed at lead-poor Kommos.




**LEAD AND SILVER** from Attica, Argolis and Laconia (*black*) and from some Egyptian artifacts (*color*) proved to be of Laurion origin. The open squares show the isotope ratios of lead from three Mycenaean caldrons found in Shaft Grave IV by Heinrich Schliemann. The open circles show the ratios of four lead weights found in a *tholos* (beehive-shaped) tomb at Vapheio. The open triangles show the ratios of three silver rings and three lead-wire samples from Perati; the two inverted triangles show the ratios of lead and litharge of Middle Helladic age from Thorikos, and the six diamonds show the ratios of lead from Athens. The six Egyptian artifacts (*colored dots*) include three of lead from the 18th Dynasty, one of silver from the 10th Dynasty, one of silver from the 11th Dynasty and one of silver from the 12th or 13th Dynasty.

by intensive investigation of the Siphnian mines and by direct dating of the mining activities at Agios Sostis on the island. Meanwhile recent studies by Mussche and Spitaels of Mine No. 3 at Thorikos in the Laurion show that there mining began at least as early as Early Helladic III times (about 2700 B.C.). This confirms our evidence from the analysis of the artifacts.

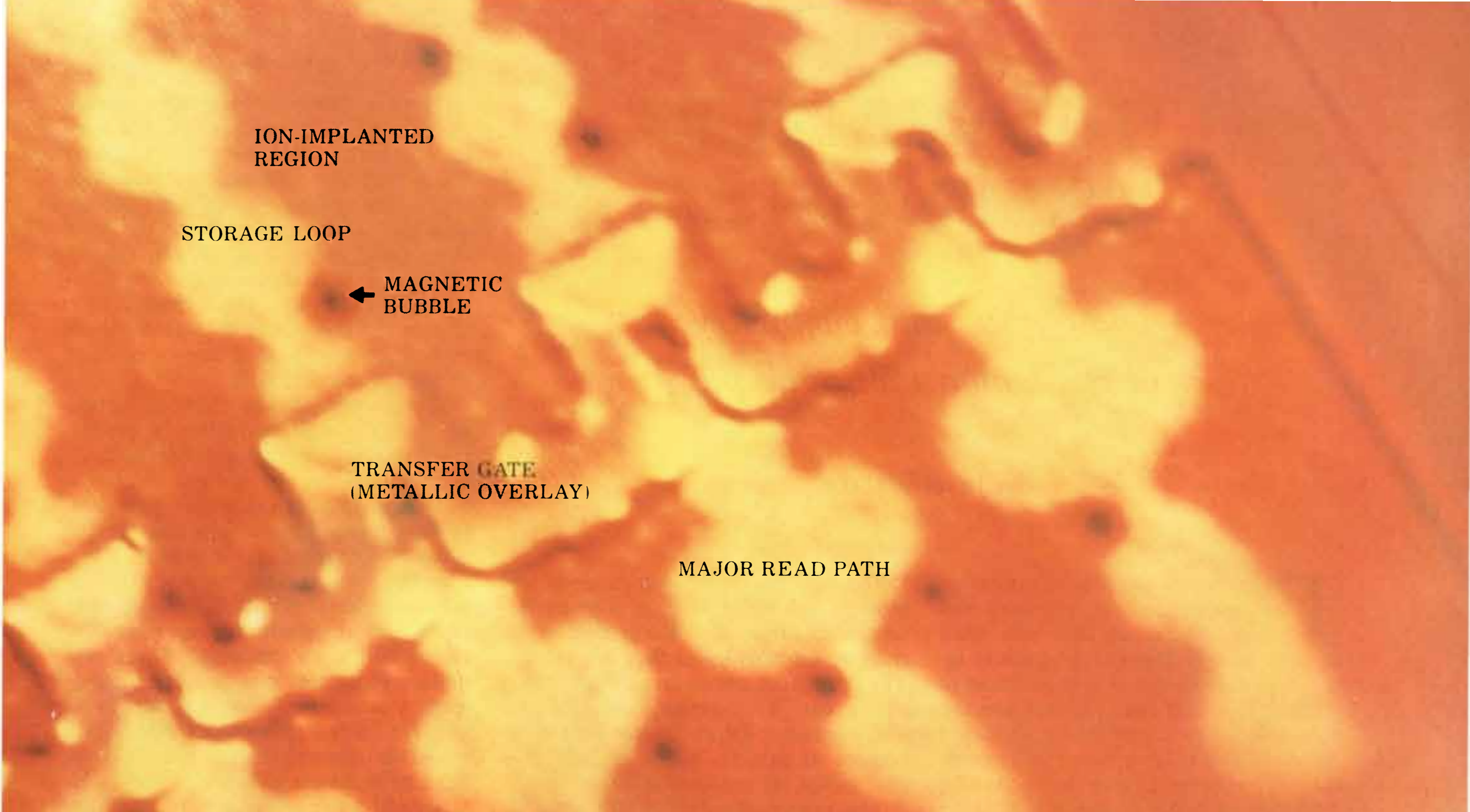
A straightforward picture emerges: in the Bronze Age both the mainland Greeks and the people of the Cyclades had essentially only two sources of lead and silver. There are many other sources of these metals in the Aegean; why should the Laurion and Siphnos have been singled out? It is very probable that it was because the metal chiefly sought was silver, lead being essentially a by-product. Assays of the ores from the mining sites in the two areas show that they contain the highest content of silver (about 5 percent in both cases) of any in the Aegean. When one considers that in ancient times mining, smelting and cupelling called for heavy labor and large amounts of wood, it is hardly surprising that prehistoric miners and metalsmiths searched out the richest ore deposits they could find.

It has often been posited that lead and silver came to the Aegean civilization from the western Mediterranean, but we have no evidence for it. Moreover, there is little evidence for silver from Asia Minor (and that evidence only for one artifact from the Cyclades). This accords with Renfrew's view that in the later Bronze Age, Aegean trade with foreign lands was minimal compared with the internal flow of goods and materials.

The work of Karl M. Petruso of Boston University on the lead balance weights that have been found in such abundance at Akrotiri and on Kea, and in lesser numbers at some 20 sites on Crete, in the Cyclades and in mainland Greece, has given strong support to Renfrew's position. It is Petruso who has shown that these weights, together with a few not made of lead, are all in multiples of the basic 61-gram unit; the uniform system is apparently a Minoan one that developed without influence from outside the Aegean. Such a system would make sense only if it was applied to the valuation of some commodity or commodities on a common basis at different places, all employing the same system.

The need for such a common system in turn can be understood only in terms of an Aegean trading network and the peaceful exchange of goods. Recently Jack Davis of the University of Illinois has drawn attention to archaeological evidence strongly suggesting that there were frequent, perhaps regular, contacts between Crete, the islands of the western Cyclades (particularly Thera, Melos and Kea) and the mainland (Attica and the northeastern Peloponnesus) from at least Middle Minoan I to Late Minoan I A times. Our own evidence for Laurion lead and silver in Late Minoan I, Late Helladic I and later times in Crete, Thera, Kea (data obtained as we completed this article) and the mainland sites of Thorikos, Perati, Mycenae and Vapheio suggests that an important constituent of the trade along this "Western String" was lead and silver.

**BLOCKS OF MARBLE at Demoliaki in the Laurion once formed a sloping circular sluice some 20 feet in diameter, used in Classical times to separate lead ore from lighter waste rock. The cups cut into the sluiceway retained the heavy ore as a stream of water carried off the waste.**




ION-IMPLANTED BUBBLE MEMORY. MAGNIFIED 3000X.

# Magnetic bubbles: Packing them in by the millions.

The more magnetic bubbles we can put on a chip, the more attractive this technology becomes as a mass memory. This means an inexpensive, reliable, and nonvolatile alternative for storing large amounts of data in telecommunications and information management systems.

We recently stored 8,798,406 bits in a single experimental chip. This indicates the potential for storing the equivalent of a 600-page novel in an area the size of a postage stamp. We did this by using simple, easily resolved patterns to guide the bubbles. To create these patterns, we implant high-energy ions directly into the magnetic bubble material through a mask. The result: smaller patterns, which permit smaller bubbles, and thus more bubbles per chip than in commercial bubble memories.

We're also studying how to manipulate bubbles by means of two wafer-thin conducting sheets instead of a pair of external coils. With this technique, we'll be able to cut the overall size of bubble devices by a third, move the bubbles 10 times faster, and also put far more bubbles on a chip than in today's commercial devices.

A group of Bell Labs scientists and engineers working on magnetic bubble technology holds over 150 patents in the field. About half of them have been awarded to Andrew H. Bobeck, a co-holder of the basic patent for the 1966 invention. Among our inventions:

- Magnetic bubble concepts and devices
- Garnet materials for bubble devices
- Method of growing epitaxial garnet films from a supercooled solution
- Basic technology for device manufacture
- Ion implant method of propagating bubbles
- Dual-conductor sheet method of propagating bubbles

In the Bell System, bubble memories are already at work in equipment that provides recorded voice announcements and in systems that administer and test digital networks. Eventually, the memories could be used in electronic switching systems, and in advanced home and business telephones.

Our goal is to make this technology even more economical and versatile for storing data electronically. Ultimately—through our partnership with AT&T, Western Electric and the Bell telephone companies—this translates into better service to Bell System customers.

*For information about employment, write:*

*Bell Laboratories, Room 3C-303, 600 Mountain Avenue, Murray Hill, N.J. 07974*

*Please include a resume.*

*An equal opportunity employer.*

**Bell Laboratories**

# THE AMATEUR SCIENTIST

## *The physics and chemistry of the lemon meringue pie*

by Jearl Walker

The scientific aspect of cooking, encompassing the physical and chemical changes that take place in food as it is cooked, is virtually ignored in cookbooks. It was an effort to explore this field that led me into the mysteries of sauce béarnaise, which I described in this department for December, 1979. Now I have been making meringue in order to learn something about the processes affecting egg whites as they are beaten and heated. Perhaps my limited achievements will inspire others to experiment in the field.

A meringue dish raises several questions about the protein chemistry of eggs and the physics of heat transfer. For example, why do egg whites stiffen when they are beaten? Why does excessive beating separate water from the whites and ruin the dessert made with them? Why does the addition of a mild acid such as vinegar strengthen the whites? Why is salt undesirable if it is added before they are beaten?

Suppose the objective is a lemon meringue pie, in which the meringue is a fluffy topping consisting of beaten and cooked egg whites to which a bit of sugar has been added. The first step in the making of the meringue is to separate the whites from the yolks. Crack an egg and drop its contents into an egg separator. The central depression of the separator holds the yolk; the white slides off the yolk, over the rim of the separator and into a bowl. Yolk must be rigorously excluded from the meringue, and the surest way to do so is to first separate the white of each egg into a small bowl. If it is apparent that no yolk went along, the contents of the bowl can be added to a larger bowl containing the egg whites that have already been separated.

Now the egg whites are beaten with a metal whisk, an eggbeater or an electric mixer. The aim is to stiffen the whites and to work in bubbles of air. Keep beating until the material is opaque, glistening and fairly firm. At this point the sugar can be beaten in gradually, after which the mixture can be laid carefully on the lemon base and the pie can be put in the oven.

The heat of the baking expands and hardens the white topping. The expansion results from the enlargement of the trapped air bubbles as they are heated. The hardening comes from the cooking of the egg. The meringue is light and fluffy because the air beaten into it decreases its density, which declines further as the air expands in the oven.

The reason the egg whites must be free of yolk is that if they contain even a small amount of it, the beating operation will not add enough air. The resulting meringue will be flat and dense. The meringue can also fail from excessive beating. In my zeal to stiffen the whites I often overbeat them, causing the water to separate out. This error results in a flat and hard meringue.

Recently, after yet another embarrassing failure with a lemon meringue pie, I began to wonder what really happens to the egg whites when I beat them. What makes them change from a clear fluid to an opaque white mixture of egg and air bubbles? I also wondered precisely what happens to the mixture when I overbeat it.

These questions led me to wonder about other procedures. Some cooks add a small amount of acid, either vinegar (acetic acid) or cream of tartar (tartaric acid), to aid in stiffening the whites. Salt, on the other hand, is avoided because it is said to harden the whites. Sugar must not be added until the whites have been fully beaten, and then it must be put in slowly with a minimum of additional beating. Finally, some cooks insist that the best meringue is made with whites beaten in a copper bowl.

To explain what is happening to the whites in any of these procedures I shall first describe the protein in the egg white. What a good cook is doing is properly preparing that protein for the oven. What I am doing in my lack of experience is ruining the protein. The success of the meringue depends on how the protein behaves when it is beaten and combined with mild acids.

The whites of chicken eggs are composed of several types of protein, principally ovalbumin, globulin and ovomucin. Proteins are huge molecules consisting of a chain (or several chains) of amino acid units linked by peptide bonds. These bonds are strong, covalent ones between a carbon atom in one amino acid and a nitrogen atom in the next amino acid in the chain. A generalized protein chain is shown in the top illustration on page 197. The symbol *R* represents the characteristic side chain of each amino acid. Ovalbumin consists of 386 amino acid units.

The particular sequence of amino acid units in a protein chain is the primary structure of the protein molecule. Such a chain is not straight but has a three-dimensional secondary structure. In some types of protein the chain forms a helix; in others the structure looks something like a pleated sheet. Ovalbumin and globulin may have a mixture of the two, along with regions where the chains coil more randomly. These secondary structures can be folded into a tertiary structure. Ovalbumin and globulin are folded so complexly that they resemble spheres rather than chains. Such proteins are called globular.

Several types of force hold a protein in its secondary and tertiary structures. Disulfide bridges are covalent bonds between sulfur atoms at separated points along the protein chain. Some points along the chain are joined by an ionic bond: an electrostatic attraction between two oppositely charged sites.

Neutral (uncharged) sites can attract each other through the force that is called the hydrophobic bond in biology and the van der Waals force in physics. One site forces a small separation of the centers of the positive and negative charges in the other site. The separation causes the positive charges of one site to be attracted to the negative charges of the other. A similar interaction takes place in polar groups. In such a group there is a permanent separation between the centers of positive and negative charge. The positive end of one polar group is attracted to the negative end of another polar group. Commonly the positive end is a hydrogen atom, in which case the attraction is called a hydrogen bond.

Of these interactions the peptide and disulfide bonds, which are both covalent, are the strongest. The ionic bonds are considerably weaker, and the hydrophobic and hydrogen bonds are weaker still. When a cook forces a whisk through egg whites, shearing the fluid, some of the weaker bonds are ruptured and parts of the tertiary structure of the proteins are destroyed. The cook does not totally disrupt the proteins because the forces holding them in their primary and secondary structures are comparatively strong. As a result of the whisking a protein is gradually unraveled from its initial spaghetti-like sphere. Any such altering of the structure of protein is called denaturing. One of the objectives

in beating egg whites is to denature the proteins.

Once the protein molecules are partially unraveled they begin to attach themselves to one another to form a three-dimensional mesh, or gel. This attraction between proteins is unlikely before denaturation because the proteins are globular and relatively few of their sites for possible bonds are exposed. Moreover, the sites suitable for hydrophobic bonding are initially tucked away inside the globular protein. With part of the protein unraveled more sites are exposed for attachment to other proteins. As the whites are beaten air bubbles—air surrounded by a thin film of water—are formed and trapped in the gel mesh. The film is held together by the surface tension of the water and strengthened by the mesh of proteins in the film. Those proteins are not only attached to one another but also bonded (by hydrogen bonds) to water molecules in the film.

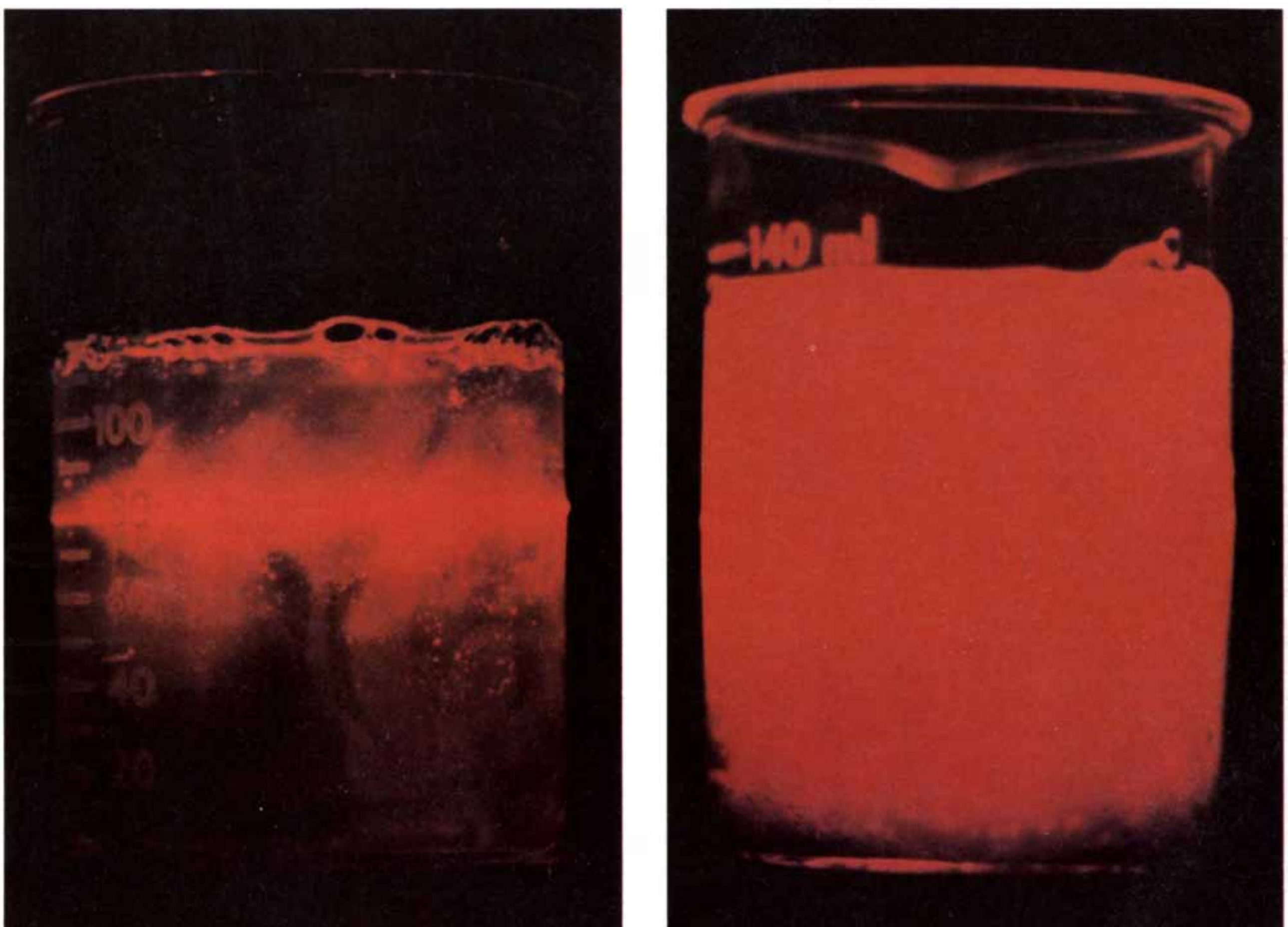
When the mixture is heated, the air bubbles expand and thereby expand the meringue. Moreover, the heat further denatures the proteins, unraveling them more and thus enabling the mesh to stretch as the air expands. The bonding between the proteins also becomes more extensive, coagulating the whites into a firm structure. Once the expansion has been achieved the trapped air bubbles are no longer required because the heat has hardened the whites in their expanded structure. The meringue is further heated only to dry it so that it is slightly crisp.

Excessive beating ruins the meringue because it unravels the proteins too much. As a result the mesh becomes too firm and water separates from the proteins. One sign that the mesh is too firm is that stiff peaks are formed when the beater is pulled out. If the mesh is too firm, the whites cannot expand properly during the cooking to match the expansion by the trapped air bubbles. The bubbles burst and the meringue collapses before the heat has had time to coagulate the protein. By the time the coagulation is achieved the meringue has become flat and the cook has nothing but baked egg whites on a lemon base.

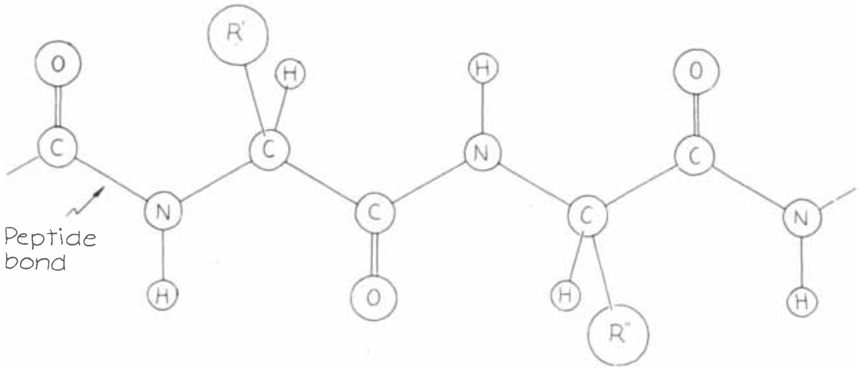
A good cook knows when the whites are on the verge of being overbeaten by carefully watching the surface, waiting for the critical moment when the whites begin to lose their sparkle and water droplets are about to form. The unraveled protein is soluble in water because on the outside of the glob it forms are many hydrophilic (water-attracting) sites to which water molecules can bond with hydrogen bonds. Most of the hydrophobic (water-repelling) sites on the protein are tucked away inside the glob. As the protein is unraveled more of the hydrophobic sites are exposed and fewer hydrophilic sites are available for bonding water molecules.

Eventually, when the protein has been unraveled too much, the bonding of water to it is so greatly reduced that the protein is no longer soluble. If a small amount of protein was in water, the protein would precipitate out. In the bowl of egg whites it is the water that can be seen to separate and form droplets. They tell the cook that the protein has been denatured too much. They also reveal that the water is no longer as capable of holding small air bubbles in the gel mesh, because the water molecules are no longer attached to the proteins forming the mesh. Thus the separation of water in the whites is a sign that the meringue will fail.

Fresh egg whites are transparent, but once their proteins denature and form a mesh they become opaque. I arranged a demonstration of the change by putting a beaker of egg whites on a hot plate and passing a thin beam of light from a he-




*The scattering of laser light by egg whites before heating* **(left)** *and afterward* **(right)**




*A chain of amino acids and peptide bonds in egg-white protein*

lium-neon laser through them. Initially the beam was sharply visible in the whites partly because some of the light was scattered by the globular proteins. Additional scattering was evident because the whites were not homogeneous. Once the hot plate was warm and began to make the proteins coagulate into a mesh the beam became less distinct. The mesh scattered the light more, illuminating much of the interior of the whites in the beaker. When the coagulation was complete, most of the light was absorbed in the whites.

Research has shown that meringues can be made from mixtures containing only one or two of the proteins normally found in egg whites. With ovalbumin alone the meringue has almost the same volume but displays a coarser texture. It also takes longer to beat ovalbumin to the proper foam. With globulin and ovomucin alone the beating takes less time but the meringue collapses after baking, presumably because the mesh is not strong enough to support the weight. Globulin alone yields a foam with small bubbles, thereby imparting to the meringue a smoother texture than is achieved with ovalbumin alone.

I do not know why ovalbumin and globulin give rise to different kinds of foam, but I would guess it has to do with the ease with which the proteins are denatured. Since ovalbumin takes longer than globulin to produce a good foam, the ovalbumin is probably harder to unravel by beating. Most likely this quality is due to a tertiary structure that is somewhat stronger in ovalbumin than it is in globulin. Since globulin appears to unravel easier, a globulin mixture builds up unraveled protein faster than an ovalbumin mixture. With more unraveled globulin the gel mesh will be finer.

When a cook beats natural egg whites, the fine texture of the foam is due more to the globulin than it is to the ovalbumin. The ovomucin does not foam, but its mesh appears to help stabilize the foam that results from the denatured globulin and ovalbumin. Many cooks know that duck eggs make comparatively poor meringues. The reason may be that duck eggs contain less globulin than chicken eggs.

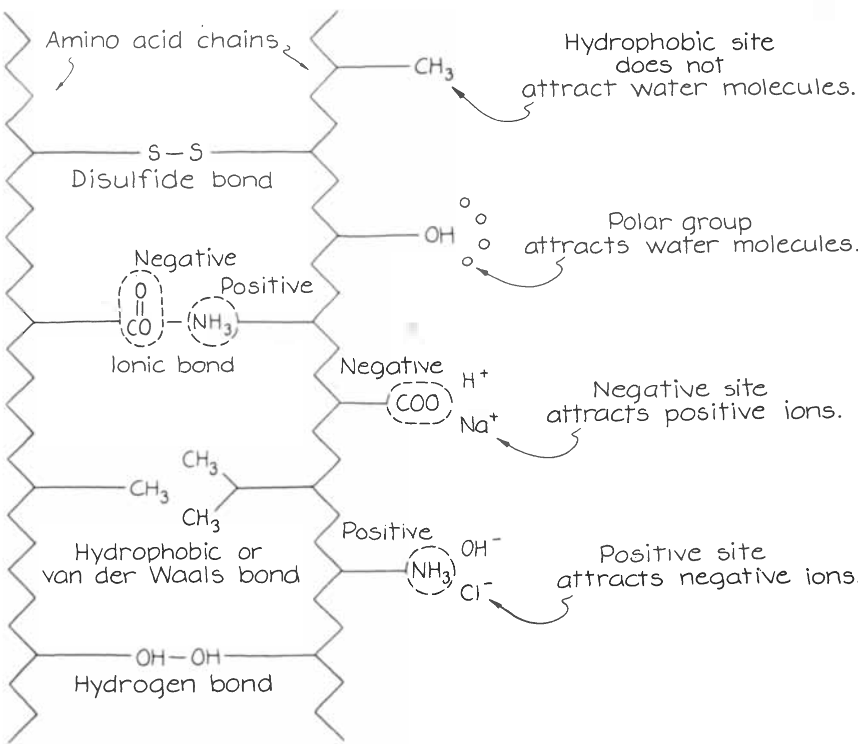
Before the whites are separated from the yolks the eggs should be at room temperature. Most cookbooks caution against working with eggs straight from the refrigerator; the whites from cold eggs make a puny foam. The fault lies in their viscosity. When they are cold, the viscosity is high and the beater cannot properly mix air bubbles into the fluid.

When yolk is accidentally added to egg whites, the lipid (fat) molecules in it attach themselves to the hydrophobic sites on the protein by means of the van der Waals force. Hence those sites are no longer available for bonding with other proteins to create a gel mesh. It is then harder to trap air, and the volume that can be attained by beating the whites is smaller. When I beat egg whites properly, the volume they fill expands by five or six times because of the trapped air. When I add a tiny amount of yolk (or any other source of fat) to the same amount of whites, however, the expansion is dramatically reduced.

To demonstrate this effect yourself beat the whites from three Grade A eggs. When the surface is glistening suitably, mark the height of the whites on the side of the bowl. Then rinse and dry the bowl and separate the whites from three more eggs. This time add about 10 percent of one yolk to the whites and beat them as before. You will find that after prolonged beating the volume is about half that of the preceding batch and the mixture is noticeably runny, revealing the lack of a good gel mesh.

I have seen another sign of a poor or missing mesh. When I beat egg whites without any trace of yolk, the mixture climbs the central shaft of the beater of my electric mixer. This is known as the Weissenberg effect; it indicates that the fluid being stirred is non-Newtonian. (I described the effect in this department for November, 1978.) A good mesh is caught by the turning shaft and rises on it. When the whites contain yolk, the mesh is poorer and climbs the shaft weakly or not at all.

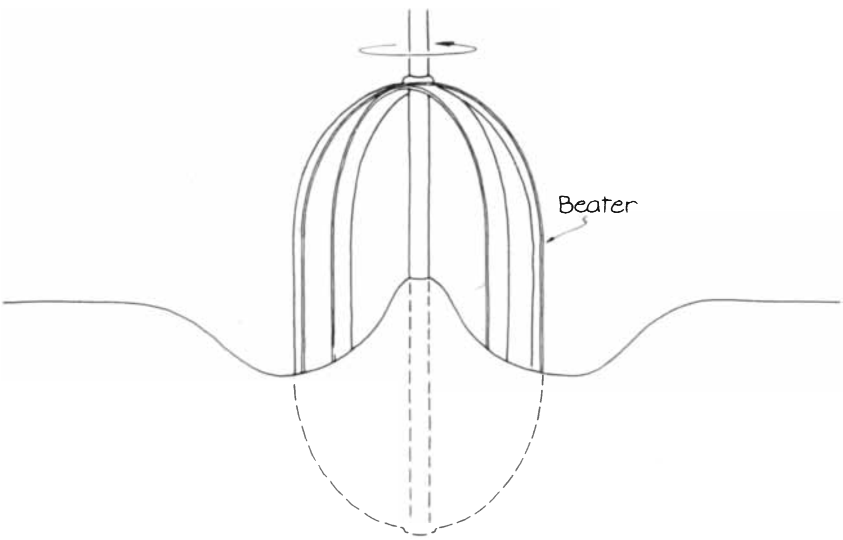
Some cooks believe adding a mild acid makes the job of beating the whites into a foam easier. The acid is either added directly in the early stages of mixing or is wiped on the interior of the bowl before the whites are put in. The belief is probably valid. When the acid is added, its molecules dissociate into positive and negative ions. The positive ions tend to collect around the nega-



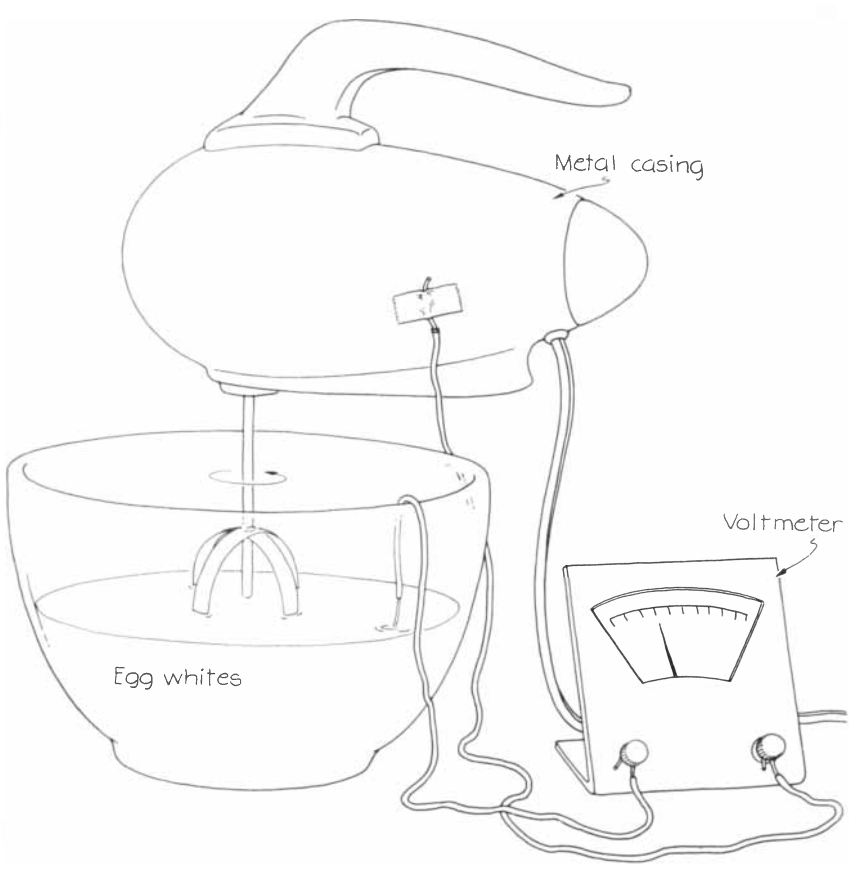

*Types of bonds in protein*

tively charged sites on the protein. Suppose those sites are involved in holding the protein in its raveled structure. Now, with a cluster of positive ions around them, they no longer can play a role in the structure and the protein partially unravels. Similarly, the negative ions may cluster around the positively charged sites on the protein and also help to unravel it.

The addition of acid has a more important effect, which is critical in the formation of a gel mesh. The proteins in freshly prepared whites have a net negative charge. Hence they form a mesh with some difficulty because of their mutual repulsion. When acid is added, the net charge on each protein is reduced because of the clustering of the ions around the protein's charged sites. The repulsion between the proteins is thus reduced and they can more easily bind to one another to form a mesh.




*The Weissenberg effect as egg whites are beaten*




*An arrangement for testing whether an electric current develops as egg whites are beaten*

When salt (NaCl) is added to the egg whites before they are beaten, two things can happen. The salt dissociates into positive ($Na^+$) and negative ($Cl^-$) ions, which cluster around charged sites on the proteins. The sodium ions cluster around the negative sites, the chlorine ions around the positive sites. If those sites were partly responsible for the tertiary structure of the protein, the protein begins to unravel. The effect is small, since most proteins are not denatured by the salt.

The clustering of ions around the charged sites can also reduce the net charge on the proteins. Their mutual repulsion is then reduced and a mesh is easier to form.

In spite of these possible benefits from the addition of salt, its major effect is undesirable. It can remove water from the protein. As a result the water cannot form a film around the air beaten into the mixture and so is no longer attached to the mesh of entwined proteins you are attempting to create. Air is not trapped and a poor foam results. If salt is needed in a mixture of egg whites, it should be added only after the mixture has been beaten into a foam. Even then anything more than a small amount will tend to make the foam collapse.

What about the copper-bowl technique? There may be something to it, but I am not sure what it is. Aluminum works poorly because it discolors the whites. Copper might help if it somehow decreases the net charge on the proteins, making the formation of a mesh easier, but I am skeptical. You might experiment to see whether a copper bowl makes beating the egg whites easier, increases the volume of properly beaten whites or increases the volume of the meringue.

Many cookbooks maintain that egg whites are stiffened partly because of an electric field that develops when they are beaten. Recipes give instructions on how best to create such a field, or at least on how not to destroy one. For example, I have read that one should avoid beating the egg whites in a wet bowl lest the film of water eliminate the electric field.

The shearing effect of a moving whisk or beater might create an electric field much as one builds up when cellophane is pulled off a roll. If an electric field does develop in egg whites, it could help to unravel the proteins by breaking some of the weaker bonds of the tertiary structure.

I tried to find an electric field in egg whites that I beat with my electric mixer. I attached one probe of a voltmeter to the casing of the mixer and the oth-

# Who will be first to see into the depths of space?

It could be you and Hughes Electro-Optical and Data Systems.

We built the first long wavelength infrared sensor ever launched into space, and a cryogenic rocket-borne infrared telescope that's sighted hundreds of new celestial objects. Looking earthward, also, our scanners map the globe every eighteen days. In fact, we're making a wide range of electro-optics, fiber-optics, thermal imaging and infrared systems that are like second sight on earth for mankind's benefit and the defense of the nation.

Today, Hughes is one of the nation's largest employers in virtually every engineering, scientific, computer and technical discipline — with 1,500 projects and a $6 billion backlog. Yet we're decentralized to give you the kind of environments that stimulate innovation and individual recognition.

And our new facility in El Segundo will be a 140-acre city of research and development, manufacturing and testing excellence, to help thrust your vision into the future.

Who will be first with the optical and data systems of the future? It could be you and Hughes.

At Hughes Electro-Optical and Data Systems, we'll introduce you to people, ideas and jobs that could change your world. And maybe ours.

Call or send resume to:

Hughes Electro-Optical
& Data Systems
P.O. Box 913, Dept. AS-6
El Segundo, CA 90245
(213) 616-3374

Current openings:

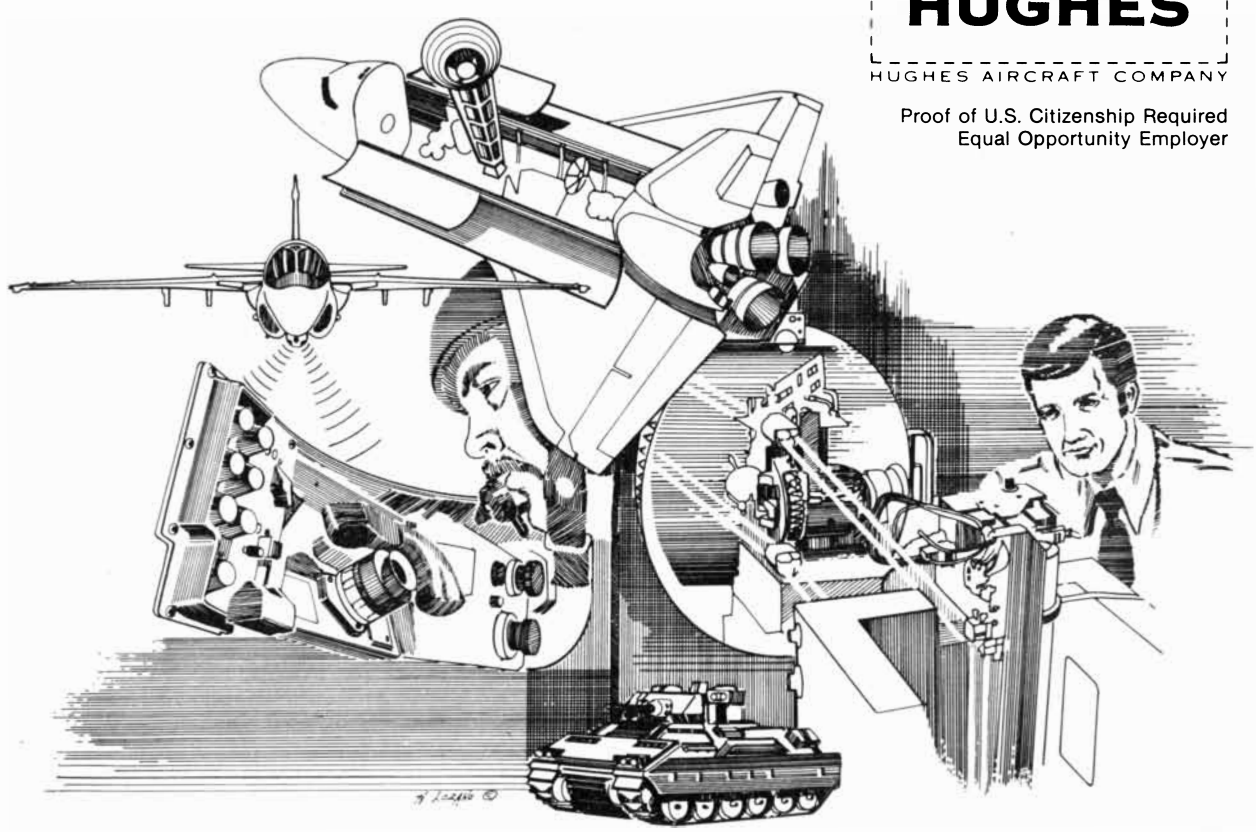
IR Sensors/Imaging
Circuits/Servos Design
Image Processors
Microprocessors
Analog/Digital Design
Electromechanical Packaging
CAD/CAM/CAT
Materials & Processes
Missions Analysis
Cryogenics
Control Systems Design & Analysis
Electro-Optical/Optical Design
Laser Systems
Components & Materials
Computer Science/Software

## It could be you and Hughes **Electro-Optical and Data Systems.**

HUGHES

HUGHES AIRCRAFT COMPANY

Proof of U.S. Citizenship Required
Equal Opportunity Employer



er to a small length of wire I put in the whites. The probes would record any electric field that might lie between the beaters and the wire buried in the whites. No voltage registered on the meter, even at its most sensitive setting of .6 volt. (If you try this experiment, be sure to keep the wire from getting tangled in the beaters as they turn.)

Next I attached a wire to the first probe and lowered it into the whites close to the other wire. Again I watched in vain for a voltage difference between the two probes. If the mixing does generate any voltage, it must be less than .05 volt and must be localized along the whites actually being sheared at any given instant. As far as I can tell no field is set up throughout the mixture. The cook does not have any control over the field (if one exists) and so should not worry about it.

Sugar should be added to egg whites only after they have been properly beaten into a stable foam. Sugar added before the protein gel mesh has had a chance to form removes water from the protein and thickens the fluid, so that aerating the mixture is more difficult. Consequently good cooks add the sugar only after the foam is built up by the beating, and even then they add it slowly while continuing to beat the mixture. Ordinary granulated sugar will not do; one needs superfine sugar that will dissolve quickly into the water walls surrounding the air bubbles. Ordinary granulated sugar takes too long to dissolve and will form clumps in the meringue. Taste the mixture to see how the dissolving has progressed. A gritty texture reveals undissolved sugar. If necessary, a small amount of water can be added to complete the dissolving.

If any undissolved sugar remains in the meringue, droplets of syrup will appear on the surface as the meringue is baked. Such a pie, called a weeping meringue, also brings tears to the eyes of the cook who labored for a better product. Once the pie is baked it must be stored in a container or the sugar in the meringue will absorb water from the air, causing water droplets to form on the surface of the pie. The droplets not only add weight that can make the meringue collapse but also soften the protein gel mesh supporting the meringue. I have given up trying to make a lemon meringue pie during the humid midsummer of Cleveland. There is so much water vapor in the air that the meringue begins to fall almost as soon as I take it out of the oven.

The lemon base for a lemon meringue pie is made from egg yolks, sugar, cornstarch, salt, butter, lemon juice and grated lemon peel. Its preparation is begun by heating the cornstarch, the sugar and the salt in a small saucepan and then stirring in water. When the mixture is smooth and hot, the pan is removed from the heat source and half of the liquid is poured into a mixing bowl with the yolks, which should already have been lightly mixed. After a quick mixing the stuff is poured back into the pan. The entire mixture is then stirred and heated over medium heat until it has boiled for about a minute. Then the juice, the peels and the butter are stirred in and the mixture is poured into a baked pie shell.

What I have described is the standard procedure for making the base quickly. A good cook takes more care, particularly with the texture of the base. Let me simplify the mixture by concentrating on the yolks and the sugar. The cornstarch thickens the base and the lemon juice and the peels of course give the pie its flavor. The secret of the base lies in the yolks and the sugar.

When the yolks and the sugar are beaten together, they form a syrup. The beating lightens the mixture by denaturing the proteins in the yolks. Although the fat in the yolks interferes with the formation of a gel mesh of these denatured proteins, something of a mesh does form. The beating also whips air bubbles into the mixture. The sugar absorbs the water from the yolks and forms a syrup. The product is thick because of the syrup but is lightened somewhat because of the air beaten into the denatured proteins. The expert can ascertain whether the mixture is properly beaten by examining how it flows from the whisk when the whisk is lifted from the mixture. If the mixture forms a ribbon on itself as it flows back into the bowl, it is ready for the pie shell.

Cookbooks contain a wide range of instructions for baking the pie. Some cookbooks suggest baking at a relatively high temperature of 400 degrees Fahrenheit for a short time, from seven to nine minutes. Others call for a much lower temperature (175 degrees F.) and a cooking time of several hours, after which the oven is turned off and the pie is left in it overnight. Still other books recommend intermediate temperatures and cooking times.

In general a lower temperature yields a crisper meringue, a higher temperature a chewier one. Those cookbook descriptions are a bit vague in my view; it would be more precise to describe the elasticity of the meringue. The elasticity is determined by the temperature. At a low temperature and a longer cooking time the meringue is slowly stretched by the expanding air bubbles and is also dried. It is not very elastic, partly because of the drying but mostly because the gel mesh is stretched for such a long time that it becomes rigid. At a higher temperature and a shorter cooking time the meringue is quickly stretched and perhaps is not as thoroughly dried. Then it is more elastic because the mesh was not stretched for enough time to become rigid.

All recipes call for heating from the bottom of the oven, never from the top, otherwise the top of the meringue would absorb the direct infrared radiation and quickly turn brown (or black). According to some cooks, the pie should be baked until the top begins to turn brown; others say browning should be avoided.

I wish I could say more about meringues, particularly about the forces that are most responsible for the protein gel mesh and how added ingredients affect it. In a future article I plan to continue the discussion of other egg dishes. In particular I shall turn to the secrets behind the superb omelets, soufflés and scrambled eggs turned out by my grandmother (who is also good at meringues and lemon meringue pies). Each of these concoctions is a challenge to my clumsy hands.

9-inch baked pie shell

Meringue

- 4 egg whites (1/2 cup) at room temperature
- 1/4 teaspoon cream of tartar
- 1/2 cup sugar

Base:

- 1/3 cup cornstarch
- 1 1/2 cups sugar, superfine
- 1/4 teaspoon salt
- 4 egg yolks, slightly beaten
- 1/4 cup lemon juice
- 2 tablespoons grated lemon peel

1 teaspoon = 5 milliliters
1 tablespoon = 15 milliliters
1 cup = 237 milliliters

*The ingredients of a lemon meringue pie*

# BIBLIOGRAPHY

*Readers interested in further explanation of the subjects covered by the articles in this issue may find the following lists of publications helpful.*

## MATHEMATICAL GAMES


Symmetry. Hermann Weyl. Princeton University Press, 1952.

Rotations and Reflections. Martin Gardner in *The Unexpected Hanging and Other Mathematical Diversions.* Simon and Schuster, 1969.

Symmetry in Science and Art. A. V. Schubnikov and V. A. Koptsik. Plenum Press, 1974.


## GAS-COOLED NUCLEAR POWER REACTORS


Atomic Energy: A New Start. David E. Lilienthal. Harper & Row Publishers, Inc., 1980.

Nuclear Energy—the Next Phase. Peter Fortescue. International Scientific Forum on Geopolitics of Energy, University of Miami, Center for Theoretical Studies, 1980.

Operational and Maintenance Exposure Trends for Nuclear Power: Proceedings of the 1980 Utility/User Conference on the HTGR. John L. Helm. Gas-Cooled Reactor Associates, 1980.


## THE DECAY OF THE PROTON


Hierarchy of Interactions in Unified Gauge Theories. H. Georgi, R. Quinn and S. Weinberg in *Physical Review Letters,* Vol. 33, No. 7, pages 451–454; August 12, 1974.

Baryon and Lepton Nonconserving Processes. Steven Weinberg in *Physical Review Letters,* Vol. 43, No. 21, pages 1566–1570; November 19, 1979.

From Atoms to Quarks: An Introduction to the Strange World of Particle Physics. J. S. Trefil. Charles Scribner's Sons, 1980.

Varieties of Baryon and Lepton Nonconservation. Steven Weinberg in *Physical Review D,* Vol. 22, No. 7, pages 1694–1700; October 1, 1980.

Conceptual Foundations of the Unified Theory of Weak and Electromagnetic Interactions. Steven Weinberg in *Science,* Vol. 210, No. 4474, pages 1212–1218; December 12, 1980.


## THE BETA-LACTAM ANTIBIOTICS


Antibiotics. H. W. Florey, E. Chain, H. G. Heatley, M. A. Jennings, A. G. Sanders, E. P. Abraham and M. E. Florey. Oxford University Press, 1949.

The Birth of Penicillin. Ronald Hare. George Allen and Unwin, Ltd., 1970.

The Chemistry of Penicillin. Edited by Hans T. Clarke, John R. Johnson and Sir Robert Robinson. Princeton University Press, 1970.

Cephalosporins and Penicillins: Chemistry and Biology. Edited by Edwin H. Flynn. Academic Press, 1972.

Beta-Lactamases. Edited by J. M. T. Hamilton-Miller and J. T. Smith. Academic Press, 1979.


## ARCHAEBACTERIA


Archean Microfossils Showing Cell Division from the Swaziland System of South Africa. Andrew H. Knoll and Elso S. Barghoorn in *Science,* Vol. 198, No. 4315, pages 396–398; October 28, 1977.

Thermophilic Microorganisms and Life at High Temperatures. Thomas D. Brock. Springer-Verlag, 1978.

Archaebacteria. C. R. Woese, L. J. Magrum and G. E. Fox in *Journal of Molecular Evolution,* Vol. 11, No. 3, pages 245–252; 1978.

Methanogens: Reevaluation of a Unique Biological Group. W. E. Balch, G. E. Fox, L. J. Magrum, C. R. Woese and R. S. Wolfe in *Microbiological Reviews,* Vol. 43, No. 2, pages 260–296; June, 1979.

Respiration in Methanogenic Bacteria. R. S. Wolfe in *Diversity of Bacterial Respiratory Systems,* edited by C. R. Knowles. CRC Press, Inc., 1980.

The Phylogeny of Prokaryotes. G. E. Fox, E. Stackebrandt, R. B. Hespell, J. Gibson, J. Maniloff, T. A. Dyer, R. S. Wolfe, W. E. Balch, R. S. Tanner, L. J. Magrum, L. B. Zablen, R. Blakemore, R. Gupta, L. Bonen, B. J. Lewis, D. A. Stahl, K. R. Luehrsen, K. N. Chen and C. R. Woese in *Science,* Vol. 209, No. 4455, pages 457–463; July 25, 1980.


## THE ALLOCATION OF RESOURCES BY LINEAR PROGRAMMING


Linear Programming. William W. Cooper and Abraham Charnes in *Scientific American,* Vol. 191, No. 2, pages 21–23; August, 1954.

Linear Programming and Economic Analysis. Robert Dorfman, Paul A. Samuelson and Robert M. Solow. McGraw-Hill Book Company, 1958.

The Theory of Linear Economic Models. David Gale. McGraw-Hill Book Company, 1960.

Linear Programming and Extensions. G. B. Dantzig. Princeton University Press, 1965.

Applied Mathematical Programming. Stephen P. Bradley, Arnoldo C. Hax and Thomas L. Magnanti. Addison-Wesley Publishing Company, 1977.

A Course in Linear Programming. V. Chvátal. W. H. Freeman and Company, in press.


## THE REGULATION OF TEMPERATURE IN THE HONEYBEE SWARM


Schwarmbienen auf Wohnungssuche. M. Lindauer in *Zeitschrift für verleichende Physiologie,* Vol. 37, pages 263–324; 1955.

The Social Behavior of Bees: A Comparative Study. Charles D. Michener. Harvard University Press, 1974.

The Natural History of the Flight of Honeybee Swarms. T. D. Seeley, R. A. Morse and P. K. Visscher in *Psyche (Journal of Entomology),* Vol. 86, pages 103–113; 1979.


## THE FORMATION OF THE EARTH FROM PLANETESIMALS


Origin of the Earth and Moon. A. E. Ringwood. Springer-Verlag, 1979.

Formation of the Terrestrial Planets. George W. Wetherill in *Annual Review of Astronomy and Astrophysics,* Vol. 18, pages 77–113; 1980.


## LEAD AND SILVER IN THE ANCIENT AEGEAN


Die Schachgräber von Mykenai. Georg Karo. F. Bruckmann AG, 1930–33.

The Emergence of Civilisation. Colin Renfrew. Methuen & Co., 1972.

A History of Metallurgy. R. F. Tylecote. The Metals Society, 1976.

Art and Culture of the Cyclades. Edited by Jürgen Thimme. Müller, 1977.


## THE AMATEUR SCIENTIST

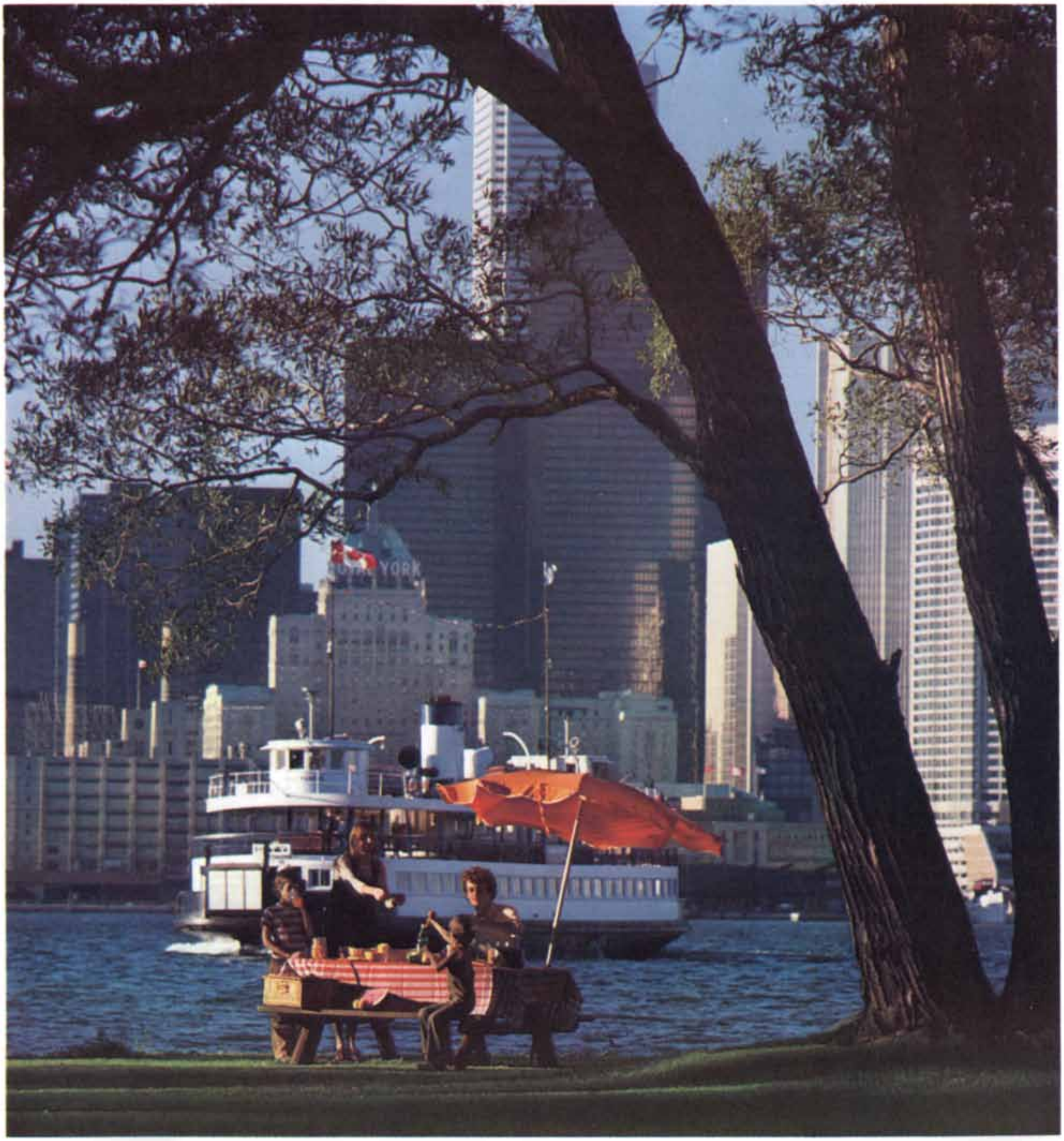

The Functional Properties of the Egg White Proteins. L. R. MacDonnell, R. E. Feeney, H. L. Hanson, Agnes Campbell and T. F. Sugihara in *Food Technology,* Vol. 9, No. 2, pages 49–53; February, 1955.

Proteins. Vicki Cobb in *Science Experiments You Can Eat.* J. B. Lippincott Company, 1972.

Proteins: Three-dimensional Conformation. Albert L. Lehninger in *Biochemistry.* Worth Publishers, Inc., 1975.

Miracles in a Shell (Egg Cookery). Madeleine Kamman in *The Making of a Cook.* Atheneum Publishers, 1978.

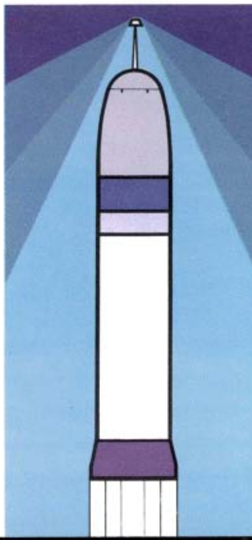
"green & gracious"



Vibrant, cosmopolitan, international in flavour, Toronto plays host with first-class hotels, haute cuisine & haute couture, great little shops, galleries, theatres, cabarets – all that dazzles, plus the Science Centre, Cinesphere, C.N. Tower, our super zoo. And all of summer's simple pleasures – bicycle paths & boardwalks, island parks & playgrounds, are yours to discover in Toronto, Ontario/Canada.

Call us COLLECT (416) 965-4008 or write: Ontario Travel, Dept. G.G., Queen's Park, Toronto M7A 2E5.

ONTARIO
yours to discover!




at the front of the aerospike diverts airflow and creates the aerodynamic illusion that Trident is bullet-shaped. This spike, along with other advances, enabled Lockheed to meet the Navy's requirement of doubled range.

## The Airlifter Afterbody Champion.

Airlifters are a special breed. To airdrop large equipment, they need wide rear doors that create a flat, two-dimensional surface. To provide fast, easy loading, their cargo decks must


almost hug the ground—and this means airlifters need a highly upswept fuselage afterbody to let the tail clear the ground on takeoff, particularly short field takeoff. Both the flat rear surface and upswept afterbody create aerodynamic drag that you do not encounter in passenger jetliners.

This caused the first great airlifter, Lockheed's C-130 Hercules, to have 10% of its overall drag in the afterbody. Shrewd linking of structural, mechanical and aerodynamic efforts enabled Lockheed to reduce afterbody drag on the second great airlifter, the C-141 StarLifter, to 3%. In the third great airlifter, the huge C-5 Galaxy, Lockheed engineers chiseled afterbody drag down to an incredible 1%.

## The Aerodynamic Lift/Drag Champion.

In aerodynamics, lift is good, drag is bad. The more lift and the less drag, the more efficient the aircraft. And that's where the U-2 reconnaissance aircraft stands out. Those long, slender wings give it the highest lift-to-drag ratio of any operational powered aircraft in the world. But there's even more to the U-2's stellar record. The wings are so successful, in part, because they are extremely light in weight. Manufacturing those feathery wings was as great a triumph as designing them.

The U-2 story is far from over. An advanced version, the TR-1 recently entered production at Lockheed.

What's next? Lockheed's past and present record says new advances are coming, in many forms of flight.

**Lockheed**

Seagram's
IMPORTED
Seagram's
V.O.
CANADIAN WHISKY
A BLEND
OF CANADA'S FINEST WHISKIES
THIS WHISKY IS SIX YEARS OLD
86.8 PROOF
BOTTLED IN CANADA
Taste!
You be the judge. Discover why more V.O. is bought than any other imported distilled spirit in America. More than any Scotch, Canadian, Rum, Gin, Vodka, Tequila.
VO
Enjoy our quality in moderation.
CANADIAN WHISKY. A BLEND OF CANADA'S FINEST WHISKIES 6 YEARS OLD. 86.8 PROOF. SEAGRAM DISTILLERS CO., N.Y.C.