Readings from

SCIENTIFIC
AMERICAN

The Molecules of Life

Readings from

SCIENTIFIC AMERICAN

THE MOLECULES OF LIFE

W. H. Freeman and Company
New York

THE COVER

The cover displays an end-on view of the DNA double helix, the molecule that encodes genetic information and has become the emblem of molecular biology, the theme of this book. The computer-generated image offers a wide-angle look along the axis of the *B* form of the double helix. The sugar and phosphate groups comprising the backbone of one of the two strands of the molecule are diagrammed in red, the elements of the other backbone in green. The strands are linked by paired bases: a purine (*blue*) on one strand pairs with a pyrimidine (*pink*) on the other strand. The swirling cloud of dots is the solvent-accessible surface: the outermost surface of the DNA molecule, defined by individual atoms, with which other molecules interact. Arthur J. Olson of the Research Institute of Scripps Clinic generated the image. He worked with the computer-graphics language GRAMPS (which he developed with T. J. O'Donnell), a molecular-modeling package (developed with Michael L. Connolly) called GRANNY and MS, a program for calculating dot surfaces, written by Connolly.

Library of Congress Cataloging in Publication Data

Main entry under title:

The Molecules of Life.

"Readings from Scientific American."
"The eleven chapters in this book originally appeared as articles in the October 1985 issue of Scientific American."—T.p. verso.
Bibliography: p.
Includes index.
1. Molecular biology—Addresses, essays, lectures.
I. Scientific American.
QH506.M665 1986 574.8'8 85-27598
ISBN 0-7167-1792-1
ISBN 0-7167-1783-2 (pbk.)

The eleven chapters in this book originally appeared as articles in the October 1985 issue of SCIENTIFIC AMERICAN.

Printed in the United States of America

1 2 3 4 5 6 7 8 9 0 KP 4 3 2 1 0 8 9 8 7 6

THE MOLECULES OF LIFE

FOREWORD

With molecular scissors called endonucleases, with invisible probes such as monoclonal antibodies and snippets of DNA, and with a host of other powerful tools, biologists have engineered the major scientific revolution of our era. They have learned to understand, manipulate, and transform the very stuff of life: the nucleic acids and proteins that are the elemental (but far from elementary) components of living things. The new molecular biology has implications far beyond the laboratory. To deal with them, the human species will need to gain ethical and moral maturity. A good starting point is an appreciation of the diversity, intricacy, and beauty this robust science presents to our understanding.

Biologists have dealt with molecules for 150 years or so, but perhaps the molecular era of biology can be said to have had its subtle and unnoticed beginning in 1944. Oswald Avery, Colin MacLeod, and Maclyn McCarty then identified the "transforming principle," a factor mediating inherited change in the pneumococcus, as a particular molecule: something called deoxyribonucleic acid. In other words, the genetic material is DNA. The news did not immediately transform biochemistry or genetics. That began after 1953, when the three-dimensional structure of DNA was resolved by James Watson and Francis Crick. In the ensuing years, the genetic code and the flow of information from DNA to proteins were clarified; the molecular basis of cell structure, the immune response, neural and hormonal signalling, and embryonic development began to be understood.

Within the past 10 years, the powerful new techniques for manipulating DNA have transformed the investigation not only of genes and their protein products, but also of a wide range of cellular structures and mechanisms. The central maneuver is the cloning of genes. The significance of cloning and other techniques and the broad implications of the new molecular biology are described in the first chapter.

The next three chapters deal with three major polymers of life: DNA, RNA, and proteins. Genetic information is stored and replicated as DNA, interpreted by RNA, and ultimately expressed as proteins. Expression is controlled primarily by the selection of particular stretches of DNA for transcription into messenger RNA. The DNA of genes is packaged with protein, coiled and supercoiled in a form called chromatin. The packaging is undone at one site or another, and then various protein molecules interact with particular nooks and crannies of the DNA to initiate transcription. The multiplicity and complexity of these interactions make it clear that the conformation of the double helix is almost as variable as DNA's linear sequences. Once transcribed into messenger RNA, the information specified by a gene must still undergo processing before it can be translated into protein. Evolutionary studies and what has recently been learned about the processing provoke speculation that RNA may have been the first genetic material.

Proteins are the primary products of genes; all other molecules—including sugars and fats and DNA itself—are the products of biochemistry conducted by the proteins called enzymes. Whether as enzymes, structural components, or the messengers and receptors of communications systems, proteins function by virtue of their complex shapes, which fit snugly with other molecules.

The molecules of particular cellular structures and systems are described in the next five chapters. The living cell is both set apart from and linked to other cells by a delicate but authoritative wall, the cell membrane, whose mobile elements maintain the integrity of the cell while at the same time admitting nutrients and a host of other molecules into the cell's cytoplasm. The cytoplasm and its metabolism are organized by a cell matrix: a skeletal framework whose composition and structure have been defined in detail by investigators working with fluorescent-labeled monoclonal antibodies. These proteins are highly specific laboratory versions of the antibodies a vertebrate animal deploys in defense against certain foreign invaders. Antibodies are but one component of the immune system; others are cell-surface receptors whose molecular structure, only now being deciphered, shows they are close cousins of the antibodies.

The individual cells of a multicellular organism need to communicate with one another. At the molecular level the two communciation networks, the nervous system and the endocrine system, turn out to be rather similar; they even depend on many of the same messenger molcules. The arrival of such messengers at the surface of a target cell is detected by specialized receptors. The detection triggers a series of reactions in the cell membrane that activate an internal messenger; it is this second messenger that prompts the cell to do what the external messenger ordained: to divide, say, or to change shape, or to synthesize and secrete a particular substance.

The last two chapters deal with the fresh contribution of molecular biology to two very different kinds of history: the embryonic development of an individual organism and the evolution of individual species. Cloning has been a major tool in the recent identification of several master genes that control the timing of developmental events and the spatial organization of the embryo. Evolution results from changes in genes, the result of mutation and natural selection. The molecular biologist's ability to directly compare the proteins and DNA of different species has provided a new quantitative measure of evolution.

The new biology described in this book (whose chapters first appeared as articles in the October, 1985, single-topic issue of *Scientific American*) is well characterized by Robert Weinberg in the first chapter: "The beauty and wonder of nature are nowhere more manifest than in the submicroscopic plan of life" the reader is about to encounter.

THE EDITORS*

October 1985

*PRESIDENT AND EDITOR: Jonathan Piel

CONTENTS

1

THE MOLECULES OF LIFE

The Molecules of Life

Introducing a volume about the new biology, which seeks to explain the molecular mechanisms underlying biological complexity. It has given rise to an industry, and to new ways of thinking about life

by Robert A. Weinberg

Biology in 1985 is dramatically different from its antecedents only 10 years ago. New investigative techniques have made commonplace many experiments that were previously far beyond the reach of even the cleverest experimental biologist. The new molecular biology has done much more than expand the repertoire of laboratory techniques. It has, with remarkable rapidity, established a biotechnology industry. More important, it has changed the ways people think about living things, from bacteria to human beings.

Biology has traditionally been a descriptive science. The multitude of living organisms were catalogued, their traits enumerated and their structures examined on a gross or a microscopic level. In thus describing organismic traits, or phenotypes, biologists confronted only the consequences of biological processes, not the causative forces. An experimenter could watch a muscle contract or an embryo develop, but such observation alone could not provide the clues that were needed for any real understanding of underlying mechanisms.

The ability to observe was greatly extended by the development of microscopic techniques that made it possible to visualize cells and subcellular organelles. Electron microscopy pushed the limits of visualization even further: the fine structure of cells could be resolved with great precision. This advance led to the uncovering of still more structures and phenomena whose causative mechanisms remained unexplained. The explanations clearly lay with elements even smaller than the cellular components observed by microscopists.

It became apparent that the ultimate casual mechanisms behind many biological phenomena depend on the functioning of specific molecules inside and outside the cell. This *Scientific American* book describes how investigators think about biological systems in terms of their molecular components. The chapters that follow are permeated by the assumption that to describe biological phenomena is far less interesting than to elucidate the molecular mechanisms underlying them. The molecular biologists who present their work here manipulate things they will never see. Yet they work with a certainty that the invisible, submicroscopic agents they study can explain, at one essential level, the complexity of life.

The newly gained ability to describe and manipulate molecules means the biologist is no longer confined to studying life as the end product of two billion and more years of evolution. The new technology has made it possible to change critical elements of the biological blueprint at will, and in so doing to create versions of life that were never anticipated by natural evolution. In the long run this may prove to be the most radical change deriving from the power to manipulate biological molecules.

Among the many kinds of biological molecules in the living cell, three have attracted the greatest attention: protein, RNA and DNA. They are macromolecules, large molecules that are linear polymers built up from simple subunits, or monomers. It was the proteins that attracted the lion's share of attention until 20 years ago. The reason, in retrospect, is clear. Certain specialized tissues accumulate large amounts of only one kind of protein. Red blood cells have almost pure hemoglobin, cartilage consists largely of collagen and hair is largely keratin. Biochemists studied such proteins first because they could be isolated in pure form, purity being a prerequisite to further study.

As an array of sophisticated biochemical techniques emerged it became possible also to purify those proteins found only in trace amounts within the complex chemical soup of a living cell. Biochemists could now concentrate on proteins that function as enzymes, catalyzing the several thousand biochemical reactions that in the aggregate constitute the metabolism of living cells. This work went well, because many of the reactions could be easily reconstructed in a test tube containing the proper mixture of reactants and catalyzing enzymes.

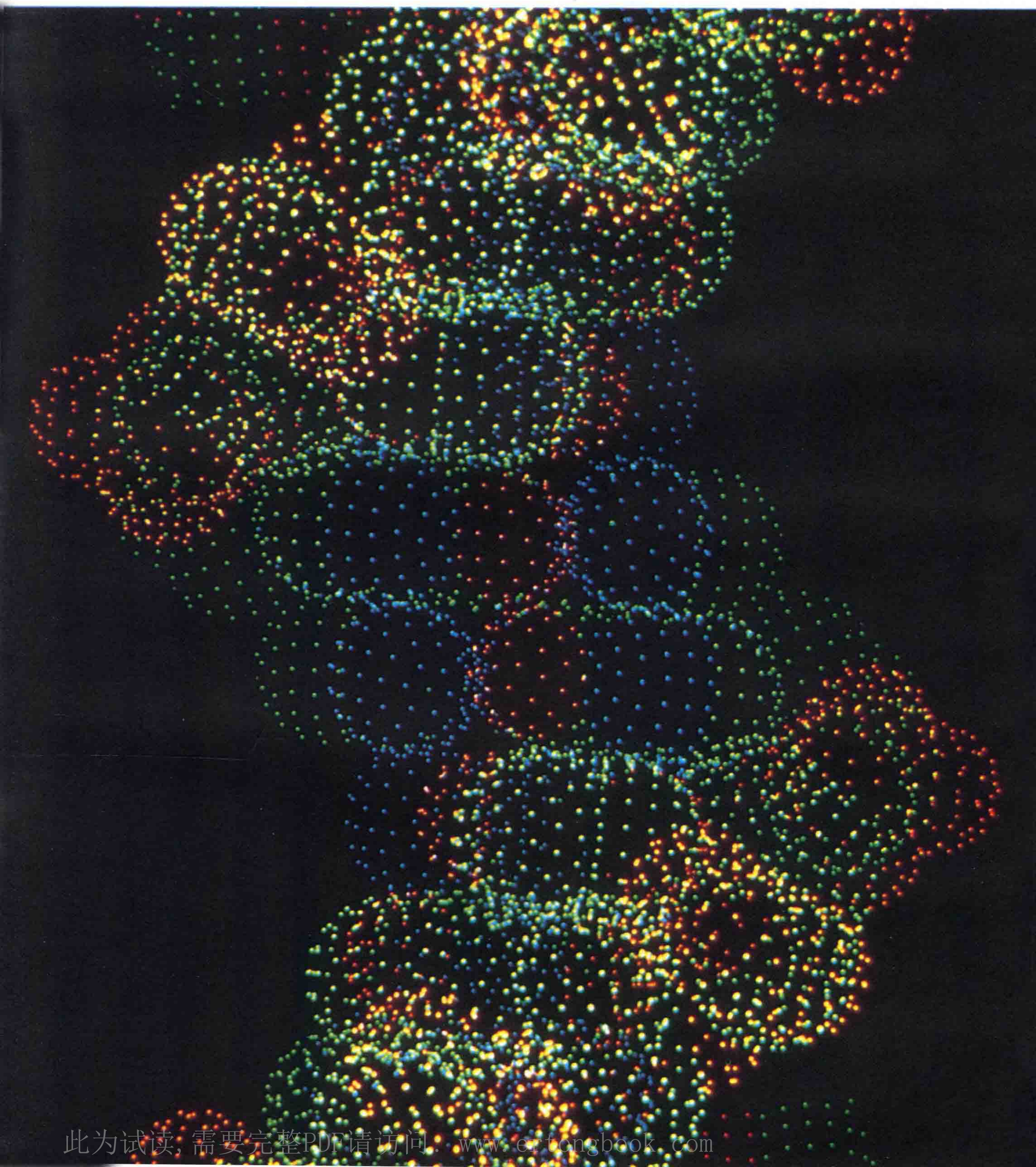
Yet in the past quarter of a century proteins have been gradually upstaged as objects of attention by the other macromolecules, first by RNA and more recently by DNA. There were two important reasons. The first one stems ironically from the great successes of protein biochemistry, which produced an avalanche of data on thousands of proteins and biochemical reactions. It soon became apparent that further study of individual trees gave little hope of understanding the entire forest. What was responsible for organizing and orchestrating this complex array of structures and processes? The answer lay not with the proteins

DOUBLE HELIX OF DNA, the molecule that is the repository of genetic information and so may be considered the fundamental molecule of life, is seen from the side in the computer-generated image on the opposite page. The spheres represent individual atoms: oxygen is red, nitrogen is blue, carbon is green and phosphorus is yellow. Diagonal regions of the image delineate the sugar-phosphate backbone of the ladderlike helix; the horizontal elements, made up of nitrogen, carbon and oxygen atoms, are the base pairs that cross-link the two strands of the helix. The computer program eliminates the backbone on the far side of the structure. The image, which depicts the *B* form of DNA, was generated by the Computer Graphics Laboratory of the University of California at San Francisco.

but with the study of genetics, and of the nucleic acids that carry genetic information.

The other reason nucleic acids, particularly DNA, have taken center stage is the advent of recombinant-DNA technology. In the course of the past decade biologists have learned to manipulate DNA in ways that (at least currently) are impossible for the protein chemist. DNA can be cut apart, modified and reassembled; it can be amplified to many copies; perhaps most telling, with DNA one can generate RNA and then protein molecules of wanted size and constitution. The central experimental maneuver in these manipulations is the cloning of genes, and it is cloning, more than any other single factor, that has changed the face of biology.

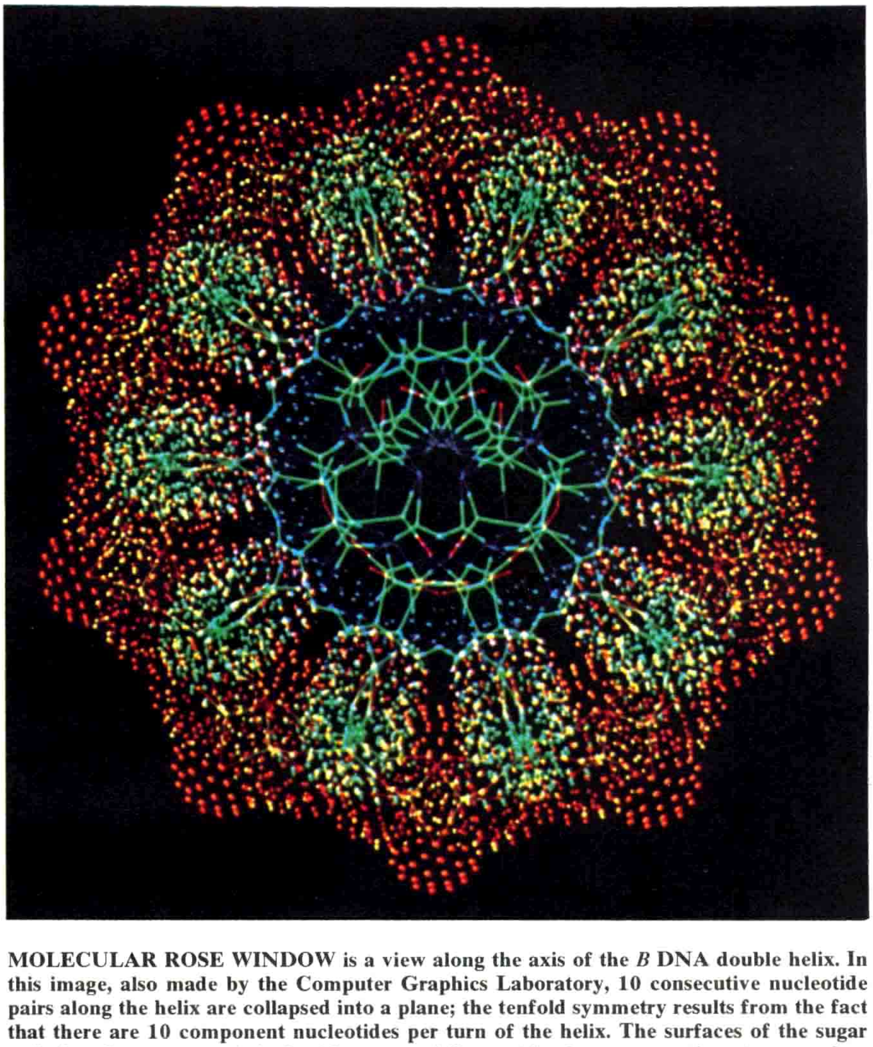
The groundwork for the cloning of genes was laid in 1953, when the



double-helical structure of DNA was perceived by James Watson and Francis Crick. A strand of DNA is a chain of nucleotides, each containing one of four bases: adenine (*A*), guanine (*G*), thymine (*T*) and cytosine (*C*). An *A* on one strand of the double helix pairs with a *T* on the other strand, and *G* pairs with *C*, so that the two strands are complementary. The sequence of bases specifies the order in which amino acids are assembled to form proteins. When the information in a gene is read out (expressed), its base sequence is copied (transcribed) into a strand of RNA. This messenger RNA (mRNA) serves as a template for the synthesis of protein: its base sequence is translated into the amino acid sequence of one protein or another.

The encoding of proteins is only a small part of DNA's function, and hence of its information content. To learn this and other simple facts it was necessary to first learn about the overall organization of DNA sequences and how the functional units of DNA—the individual genes—interact with one another in the total genetic repertoire of the organism, which is called its genome.

The genome of complex organisms resisted analysis until recently. Analysis of the gross biochemical properties of cellular DNA gave little hope of understanding the subtleties of genetic organization. The DNA content of even a bacterial cell is very large; the much larger genome of a mammalian cell carries some 2.5 billion base pairs of information arrayed along its chromosomal DNA. The base sequences are arranged in discrete compartments of information: the individual genes. There are between 50,000 and 100,000 genes in the genome of a mammal; each one is presumably responsible for specifying the structure of a particular gene product, usually a protein. Interest was therefore focused on studying individual genes, but that undertaking was doomed until recently by an inability to study single genes in isolation. In the absence of effective techniques of enrichment and isolation, individual cellular genes were abstractions. Their existence was suggested by genetic analysis, but their physical substance remained inaccessible to direct biochemical analysis.



MOLECULAR ROSE WINDOW is a view along the axis of the *B* DNA double helix. In this image, also made by the Computer Graphics Laboratory, 10 consecutive nucleotide pairs along the helix are collapsed into a plane; the tenfold symmetry results from the fact that there are 10 component nucleotides per turn of the helix. The surfaces of the sugar and phosphate groups of the backbones are delineated by dots representing atoms: carbon (*green*), oxygen (*red*) and phosphorus (*yellow*). In the center the dots are absent, and so the skeletal structure of the bases, largely nitrogen (*blue*) and carbon, is left exposed.

A partial solution to this quandary came from the study of viruses. Their genome is very small compared with that of a cell and yet their genes are similar to those of the cells they infect. The DNA genome of one much studied animal virus, the SV40 virus of monkeys, has only 5,243 base pairs, in which are nested five genes. The analysis of an individual gene was therefore not confounded by a large excess of unrelated sequences. Moreover, the viral genome multiplies to several hundred thousand identical copies within an infected cell, and it was not hard to separate the viral DNA from the cellular DNA.

Once purified, the relatively simple viral DNA made it possible to study aspects of gene structure, the transcription and processing of messenger RNA and the synthesis of proteins that had previously been beyond reach. Still unresolved were the detailed structure and the base sequence of even the viral genome, whose 5,000-odd base pairs represented a daunting challenge to biochemists trained to take polymers apart one unit at a time. Then, in the mid-1970's, two revolutionary techniques became widely available that radically simplified the analysis of DNA structure.

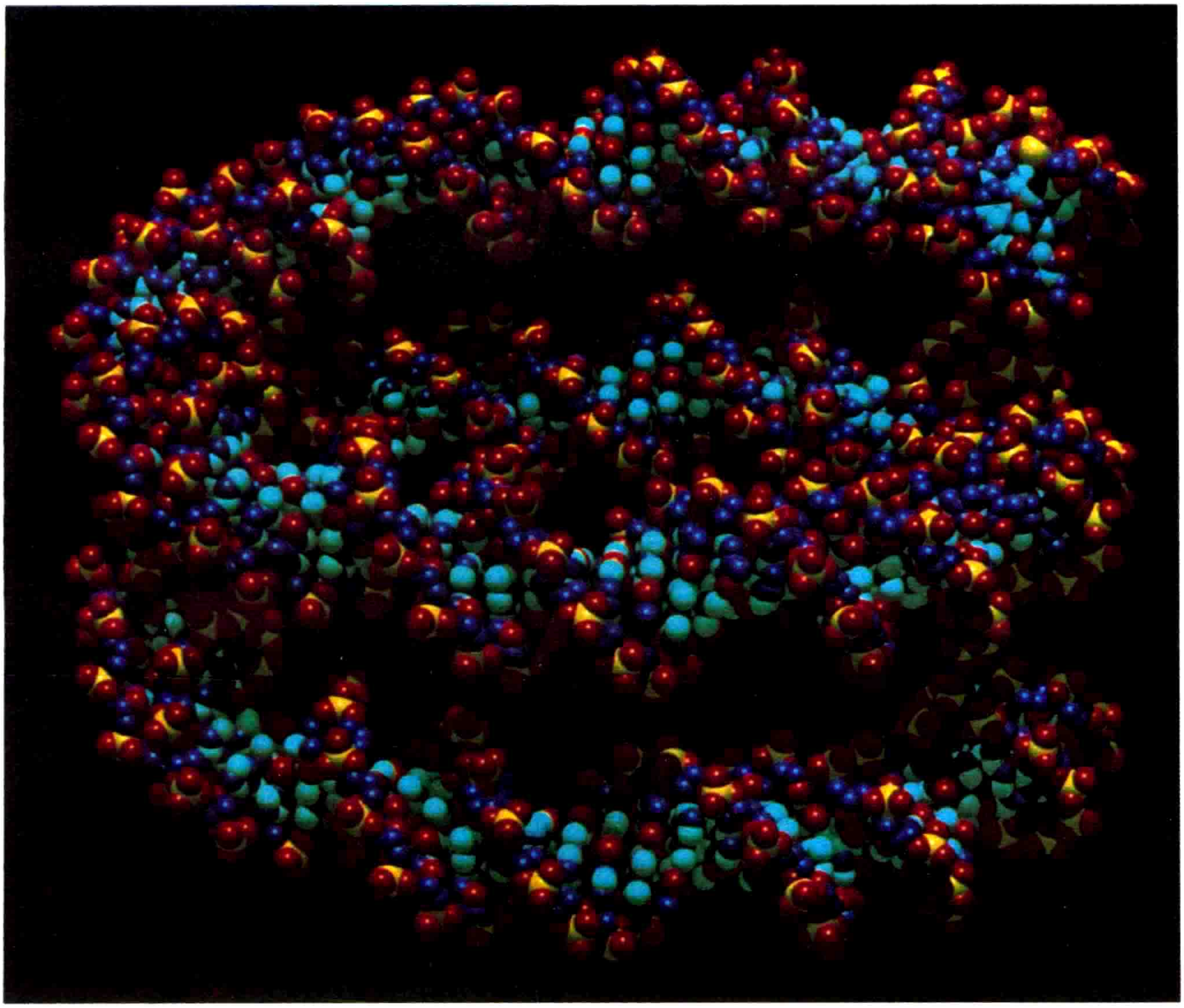
The first of the techniques stemmed from the discovery of DNA-cleaving enzymes called restriction endonucleases. These enzymes, extracted from bacteria, cut a DNA molecule only at specific sequences that occur here and there along the DNA double helix. The much used endonuclease *Eco*RI, for example, cuts wherever it encounters the sequence *GAATTC; Sma*I cuts at *CCCGGG*, and so on. The sequences forming recognition sites occur with a certain statistical probability in any stretch of DNA.

Restriction enzymes have become powerful tools for experimenters. They establish convenient, fixed landmarks along the otherwise featureless terrain of the DNA molecule. They allow one to reduce a very long DNA molecule into a set of discrete fragments, each of them from several hundred to several thousand bases long. The fragments can be separated from one another on the basis of their size by gel electrophoresis. Each fragment can then be subjected individually to further analysis.

The other technical revolution had to do with the sequencing of DNA. Several procedures were invented by which the entire base sequence of a segment generated by restriction-enzyme cleavage can be determined with remarkable rapidity. These methods made it possible, for example, to establish the entire nucleotide sequence of the SV40 genome by 1978. Because the genetic code for translating a DNA sequence into an amino acid sequence was already known, the base sequences in certain regions of the genome could be translated into amino acid sequences. In this way the structure of SV40's proteins could be deduced from its DNA structure. Previously protein structure had been determined by the painstakingly slow biochemical analysis of individually isolated proteins; now the rapid sequencing of DNA could determine protein sequences in a fraction of the time. DNA sequencing also revealed other regions of the SV40 genome that are involved not in encoding proteins but in regulating the expression of genes and the replication of DNA.

Further progress depended on procedures for isolating individual cellular genes. Here success came from studies of bacterial genetics that were initiated in the early 1970's. The procedures for gene isolation that grew out of this work are based ultimately on the similarity of molecular organization in all organisms, from bacteria through mammals. As a result bacterial and mammalian DNA's are structurally compatible; DNA segments from one life form can readily be blended with the DNA of another form.

This similarity in DNA structure extends to many of the bacteriophages (viruses that infect bacteria) and plasmids (small DNA circles that parasitize bacteria). Phages inject their DNA into a bacterium, cause it to be replicated many times over, package the newly replicated DNA in viral-protein coats and kill the bacterium; the progeny phages released from the cell go on to infect other susceptible hosts. The plasmids are even simpler, and they have a more symbiotic relation with



DNA SUPERHELIX, in which the double helix is itself wound into a coil, is depicted in a space-filling computer model made by Nelson L. Max of the Lawrence Livermore National Laboratory. This is thought to be the form in which DNA is actually packed into chromosomes in the cell nucleus, with two turns of the superhelix being wound around a complex of histone proteins (not shown here). The model is based on coordinates determined by Joel L. Sussman and Edward N. Trifonov of the Weizmann Institute of Science.

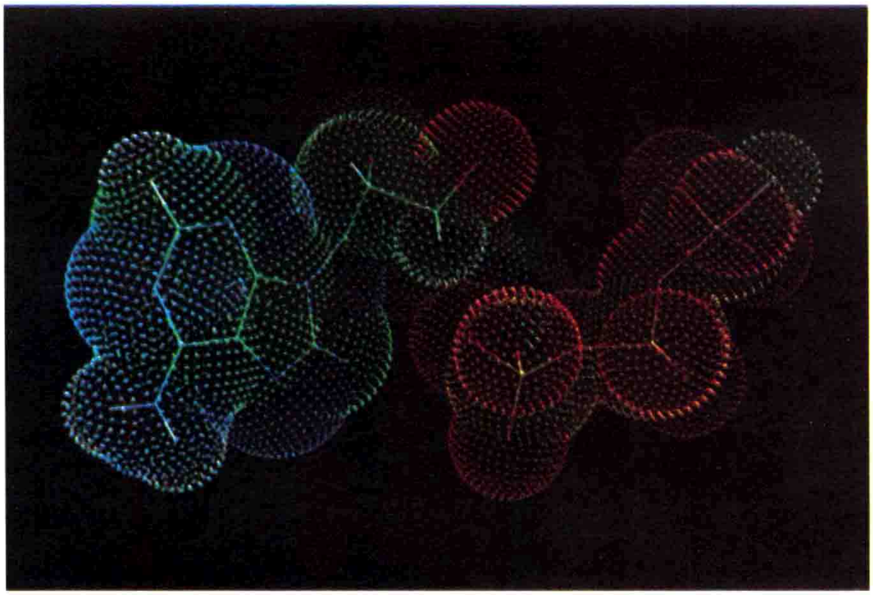
the bacterial cells in which they grow. They may carry genes that confer advantages on their host cell, such as resistance to an antibiotic. The host cell in turn allows the plasmid DNA to be replicated to a limited extent in the cell, thereby ensuring the continued presence of the plasmid in the daughter bacteria arising when a parent bacterium divides.

Some phage and plasmid DNA's are (like SV40 DNA) small in size, ranging in complexity from several thousand to 50,000 bases. Because of their small size they can be manipulated and restructured by a variety of recently developed tools. The molecules are easily isolated, unbroken and in large amounts. They can be cut at a number of defined sites with restriction enzymes and the resulting fragments can be rejoined with one another or joined to foreign DNA segments to reconstitute the original molecule or make a hybrid molecule. The rejoining is done with readily available enzymes of bacterial origin known as DNA ligases, which recognize the ends of DNA molecules and fuse them without leaving any trace of the joining.

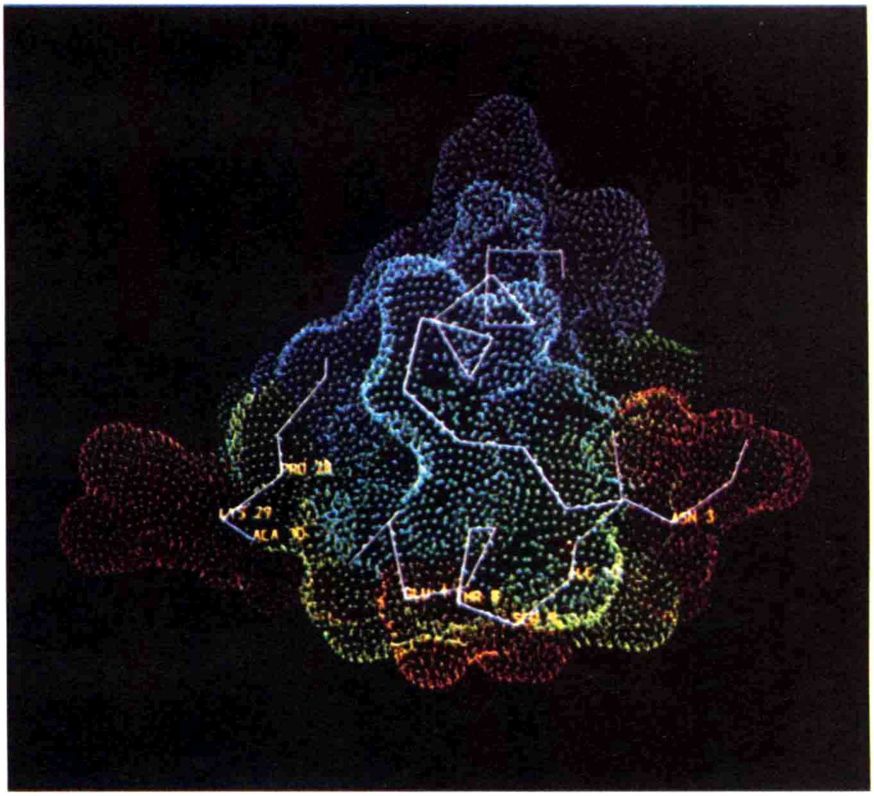
A hybrid DNA made of a plasmid fused with foreign (say mammalian) genetic material can replicate when it is introduced into a bacterial cell. This means the plasmid genome can serve as a "vector" for establishing and amplifying the foreign DNA in bacteria. A phage vector functions similarly, and it can serve as well to convey the foreign DNA from one bacterium to another. When the vector DNA is copied in the course of replication, the inserted foreign DNA is copied too.

The process of cloning begins with whole cellular DNA of an organism such as a mammal. The DNA is cleaved into fragments of a size (from about 1,000 to about 30,000 bases) that can be accommodated by the carrying capacity of one or another vector. A complex genome such as the human one can be broken down into a few hundred thousand DNA fragments. Each fragment can be separately inserted into a vector DNA molecule. The process does not require painstaking molecule-by-molecule assembly by a patient technician. Instead millions of insert and vector DNA molecules are mixed together and the process is completed in minutes by the addition of DNA ligase. If the resulting collection of hybrid molecules is large enough, any single gene of interest will surely be found embedded in one or another of the DNA segments linked to the vector molecules.

Each of these hybrid molecules, part vector and part inserted mammalian



ATP (adenosine triphosphate) is the molecule that provides free energy for many biochemical reactions, including those required for the polymerization of DNA, RNA and protein. ATP is modeled in this image made by the Computer Graphics Laboratory. It is a nucleotide consisting of the base adenine (*left*), a ribose sugar and three phosphate groups (*right*). Energy is acquired when a third phosphate is added to adenosine diphosphate by the oxidation of fuel molecules or, in plants, by photosynthesis; energy is liberated when ATP is broken down, freeing this third phosphate group. The skeletal structure of the molecule is indicated by the lines; the dots delineate the effective surfaces of the constituent atoms.



INSULIN MOLECULE, a hormone that has multiple functions, is depicted in a computer-generated model. It was developed by Elizabeth D. Getzoff, J. A. Tainer and Arthur J. Olson of the Research Institute of Scripps Clinic with the same software that generated the image on the cover of this book. Insulin is a small protein hormone made up of two short folded chains of amino acids. The lines trace the backbone of the two amino acid chains; the dots delineate the solvent-accessible surface. Coloring reflects the relative mobility of constituent atoms: the atoms shown in red and orange are the ones most subject to excursion from

bacterial cell, where they are replicated many times over; each hybrid molecule spawns a separate progeny population, all of whose members are identical with the founder. Such a population is often called a clone to reflect its descent from a common ancestor and the identity of all its members.

The term "clone" has acquired another meaning. It is applied specifically to the bits of inserted foreign DNA in the hybrid molecules of the population. Each inserted segment originally resided in the DNA of a complex genome amid millions of other DNA segments of comparable size and complexity. When the manipulations described above are completed, the same segment is present in pure form within the confines of the particular clonal population, contaminated only by the associated vector DNA. The inserted DNA segment has been isolated from its previous surroundings and selectively amplified: it has been cloned, and so the purified DNA insert itself is often called a "clone."

The process of cloning requires one further step, which is usually the most challenging of all. The insertion and amplification process has given rise to hundreds of thousands of different clonal populations, each descended from a single hybrid DNA molecule. If the initial hybrids were properly diluted before being amplified, each descendant clonal population is physically separated from other populations carrying different inserted DNA's. Having established this large array (a "library") of distinct clonal populations, the experimenter is now faced with having to identify the one or several populations carrying the inserted DNA of interest.

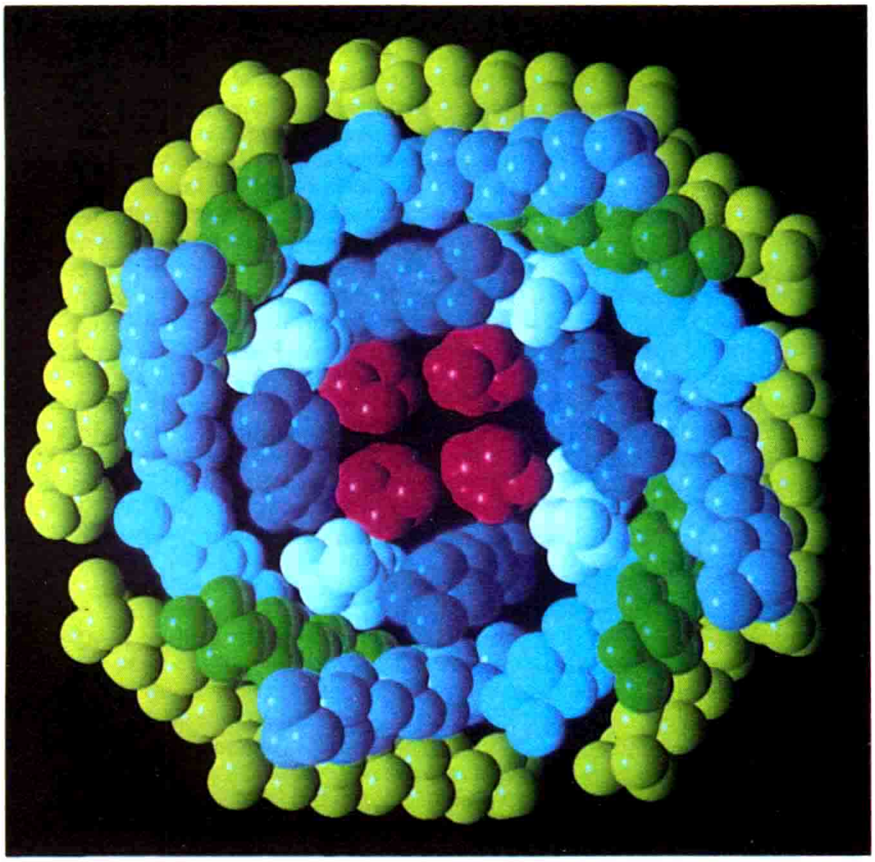
Identification can be simple if a related gene or DNA segment has been cloned before. The previously cloned DNA can be labeled with a radioactive isotope; under appropriate conditions the radioactive DNA will preferentially stick to the clone of interest (because complementary DNA strands "hybridize" by base pairing) and thus identify it. The most interesting cloning is done, however, to isolate genes that have never been cloned before, even in related form. A variety of clever strategies have been developed to address this challenge. The goal is to develop a specific probe with which to scan a library of clones and identify the clones of interest.

One strategy for probe development depends on the fact that some proteins are expressed at a high level in specialized cells. In the precursors of red blood cells, for example, globin (the protein component of hemoglobin) is synthesized in much larger amounts than any other protein. The mRNA that directs its synthesis is also present in large quantity, and there are ways to isolate it readily from other mRNA's in the same cell. The isolated mRNA, or a DNA copy of it, can serve as a probe that will hybridize with the corresponding gene sequence in a genomic library. Sophisticated versions of this strategy allow the mRNA encoding a protein of interest to be isolated selectively from a thousandfold excess of other mRNA molecules present in the same cell.

Often the protein whose gene is sought is rare, so that its mRNA cannot easily be isolated. In such cases a small amount of the protein is purified and the amino acid sequence of some part of it is determined. Knowing the genetic code, one can back-translate the amino acid sequence to learn what DNA base sequences are likely to be present in the gene encoding the protein. Small pieces of DNA corresponding to these derived base sequences can be synthesized by assembling off-the-shelf nucleotides. These man-made gene fragments serve as probes for identifying the clone of interest.

Yet another strategy that begins with a protein depends on antibodies directed against the protein whose gene one seeks to clone. Bacteria infected by a phage carrying the wanted gene synthesize small quantities of the protein, and so a phage library can be screened by the proper antibody, which binds to the protein and thereby identifies the gene-carrying clone.

With these and other experimental strategies available, the technology is now at hand to clone any gene whose protein product is known and can be isolated in even a small amount. Given sufficient interest, any of the genes encoding the many hundreds of enzymes that have been studied by biochemists can be isolated. The genes for the important structural proteins of the cell,



SODIUM CHANNEL, a large protein molecule embedded in the membrane of nerve cells, has been modeled by H. Robert Guy of the National Cancer Institute; this image of one model was computer-generated by Richard J. Feldmann of the National Institutes of Health. The protein admits sodium ions to the neuron, thereby supporting the action potential, the voltage pulse that ultimately triggers the release of neurotransmitter. The protein has four homologous domains; each domain includes eight distinct protein substructures. Similarly colored groups of spheres represent the four homologous versions of each substructure. (Two substructures in each domain are very similar and are both shown in pale green.)

including those that determine cellular architecture, have been cloned. Other genes encoding such intercellular messengers as insulin, interferon, the interleukins and a number of growth factors have been isolated. Indeed, genes are being cloned and their sequences deciphered faster than the new data can be fully interpreted. Most of the sequences are now being stored in computer banks; perhaps future generations of biologists, aided by new analytical procedures, will be able to interpret them fully.

The genes specifying known proteins account for only a small part of a complex organism's total genetic repertoire. Most of the remaining genes probably encode proteins too, but so far their existence is implied only by the effects they exert on cellular and organismic structure and function. Some of them specify biochemical conversions in the cell, others govern complex developmental processes that create shape and form in a developing embryo and still others may specify behavioral attributes of an organism. Such genes remain elusive because the means of identifying them in genomic libraries are limited.

The flow of genes from genome to gene library makes more things possible than the detailed description of DNA and protein structure. Once cloned, a gene can be inserted into a foreign cell, which can be forced to express it. The cell then synthesizes the protein the gene specified in its original home.

The gene to be expressed is excised from the vector in which it was cloned and subjected to important modifications. The modifications are necessary because a mammalian gene carries regulatory sequences that promote its transcription into mRNA in its home cell, not in a bacterial cell. These need to be replaced by bacterial regulatory sequences. The modified gene is then introduced into an "expression vector": a plasmid designed to facilitate the expression of the gene in a foreign cellular environment. The mammalian gene (or a similarly engineered plant gene) carrying bacterial regulatory sequences is then introduced into a selected foreign host, usually a bacterial or yeast cell.

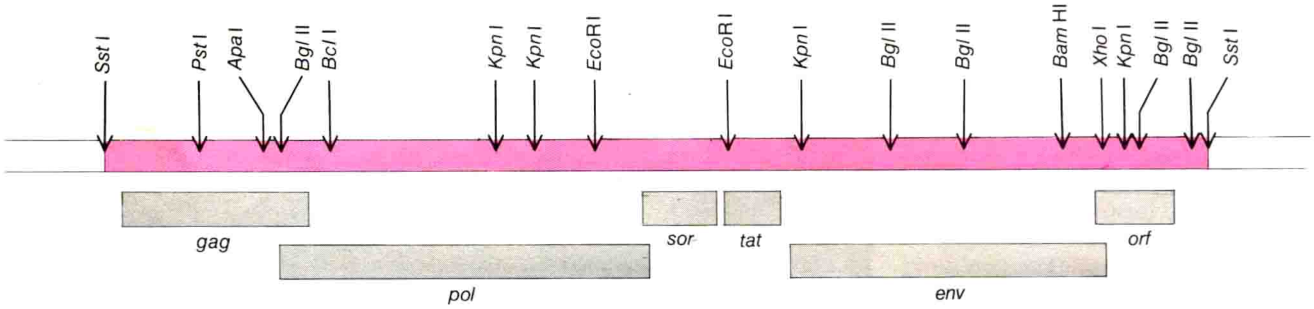
A protein that is synthesized only in limited amounts in its normal host can be produced in large quantity when its gene is redesigned for high-level expression in bacteria or yeast. This can confer great economic advantage and represents a cornerstone of the biotechnology industry. Microorganisms bearing cloned genes can be grown quite cheaply in large volume in fermentation chambers, leading to an enormous scale-up in protein production. Among the products currently being manufactured or being considered for manufacture are insulin, interferon (for combating infections and perhaps tumor growth), urokinase and plasminogen activator (for dissolving blood clots), rennin (for making cheese from milk), tumor-necrosis factor (for possible cancer therapy), the enzyme cellulase (to make sugar from plant cellulose) and viral peptide antigens (for creating novel and safe vaccines).

In a different version of gene insertion, mammalian genes that have been cloned in bacteria are introduced into mammalian cells grown in culture rather than into microorganisms. Although cultured mammalian cells cannot be grown economically in the large numbers that characterize a bacterial or yeast culture, they have the advantage of being able to make minor but significant modifications to proteins encoded by mammalian genes. For example, certain mammalian proteins function better when sugar and lipid side chains have been attached to their amino acid backbone. The addition takes place routinely in mammalian cells but not in bacteria.

Cloned genes can now be inserted not only into microorganisms or cultured mammalian cells but also into the genome of an intact multicellular plant or animal. Here the motives are quite different from those governing the genetic engineering of unicellular microorganisms to achieve large-scale production of desirable gene products. Plants and animals can be modified genetically in an effort to alter such organismic traits as growth rate, disease resistance and ability to adapt to novel environments.

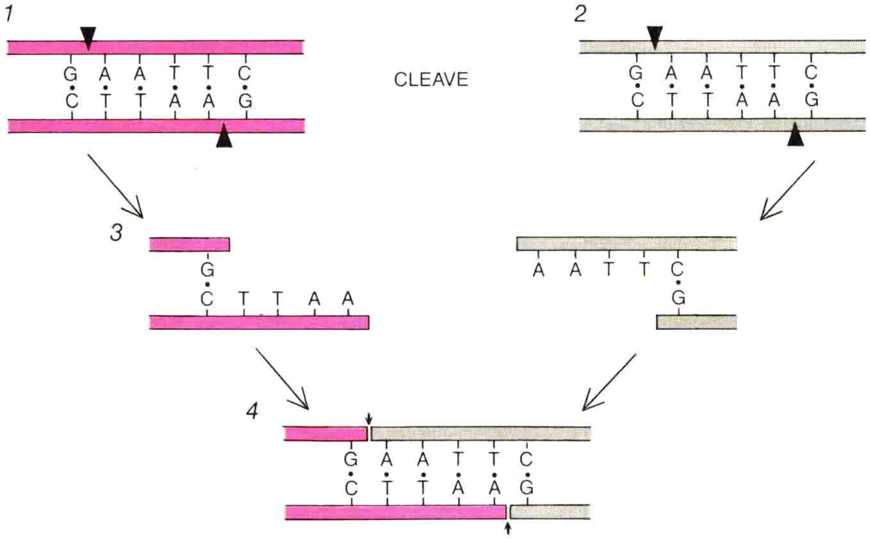
Gene insertion into a multicellular organism is a quite different project from gene transfer into a single cultured cell. The introduction of a cloned gene into most types of cells in a plant or an animal can alter the behavior of only those few cells that acquire the gene. Obviously it is of far greater interest to imprint the change on an entire organism and on the organism's descendants. That calls for gene insertion specifically into germ cells (sperm or eggs), which transmit genetic information from parent to offspring.

Techniques are indeed now available for achieving germ-line insertion into mammals, flies and certain plants. It is done either by direct physical injection of a cloned gene into the early embryo or by the use of a viral vector to carry the gene into the cells of an embryo. Again the resulting animal (or plant) carries the inserted gene in only some of its cells, but now one can hope the gene is in some of its germ cells. The presence of the gene in germ cells may allow some of the organism's offspring to inherit the inserted gene along with other parental genes, so



RESTRICTION-ENZYME MAP of the genome of the virus HTLV-III, the AIDS agent, was developed in the laboratory of Robert C. Gallo of the National Cancer Institute. Such a map is the primary means by which molecular biologists depict the organization of a stretch of DNA. The DNA is cleaved with a restriction endonuclease, an enzyme that cuts DNA at specific sites. The size of the fragments is known from the distance they travel through an electro- Cleaving a genome with a number of different restriction enzymes provides additional mileposts. This relatively simple map of HTLV-III shows the sites where several restriction enzymes cleave the HTLV-III DNA (*top*) and the locations of the various genes (*bottom*). For example, the surface antigen of the virus is encoded by the *env* (for envelope) gene, and the enzyme (reverse transcriptase) that copies the RNA of the virus into DNA is encoded by *pol* (polymerase).

that it will be present in all their cells. Thus incorporated into the germ line of the progeny organisms, the gene is passed on to, and affects, the descendants of those organisms.



CLONING IS FACILITATED by "sticky ends" generated by some restriction enzymes. The enzyme *Eco*RI, for example, makes a staggered cut in the sequence *GAATTC.* When genomic DNA (*1*) and a vector DNA (*2*) are cut with *Eco*RI, the ends of the resulting fragments have single-strand projections of complementary bases (*3*). When the fragments are mixed, hydrogen bonds (*dots*) form between those bases, reversibly joining genomic and vector DNA's (*4*). The joint is sealed irreversibly with the enzyme DNA ligase (*arrows*).

Techniques for germ-line insertion are still limited in important ways, and they may forever be. They cannot direct the foreign DNA to insert itself (to "integrate") into a particular chromosomal site; the locus of insertion is random. They cannot supplant an existing gene in the organism by knocking it out; rather, they merely add incrementally to the existing genome. Moreover, inserted genes do not always function precisely like resident genes, which are turned on and off at appropriate times in development.

Germ-line insertion is nonetheless powerful. Mice have been developed that carry and transmit to their offspring the genes for extra growth hormones. Giant mice (half again as large as normal) ensue; cattle with altered growth properties will soon follow. Flies have been developed that carry a variety of inserted genes, leading to novel insights into the molecular biology of fly development. Plants are being developed that carry genes conferring resistance to herbicides. As gene-insertion techniques are improved and as additional genes are cloned, the possibilities for altering organismic traits will expand enormously. The molecular biologist will no longer confront living forms as the finished products of evolution but will be an active participant in initiating organismic change.

For experimental biologists, cloning and its associated techniques have attained and will retain a preeminent role. Cloning makes it possible to analyze a biological system at three levels. First, the genes relevant to a particular biological problem can be isolated, the sequence of the DNA can be elucidated and the functioning and regulation of the DNA can be revealed. Second, once the DNA of a gene has been cloned the RNA transcribed from it can be produced in large amounts for study. The RNA can act in many ways to modulate the expression of genes; RNA structure and processing are central to a full understanding of gene function. Third, what is perhaps the greatest advantage of cloning stems from the analysis of the proteins encoded by a gene. How do various proteins act to elicit myriad responses in the cell? Proteins that formerly were available for study in minute amounts can now be made in great quantities once their gene has been isolated. In sum, all the major

logical system can now be made available in large amounts, in pure form.

Equally important is a newly gained ability to perturb biological systems. Genes and their encoded proteins can be redesigned so that new functions can be imparted to DNA and proteins. The relations among the interacting components of a biological system can be altered to generate novel and often revealing behavior by the system as a whole. The redesigning of genes is accomplished by changing DNA sequences through what is termed site-directed mutagenesis. This may involve the replacement of one restriction-enzyme fragment with another in the midst of a cloned gene. Alternatively, chemically synthesized DNA segments may be stitched into a gene, replacing or adding to existing sequence information. Single nucleotides can be substituted as well to create point mutations, the most subtle changes a gene can undergo. Genetic changes that have accumulated in a gene over hundreds of millions of years of natural evolution can be mimicked and superseded by several weeks' manipulation in the laboratory.

Genes altered by these techniques can then be reintroduced into the biological systems with which they normally interact. An enzyme having a low affinity for the substrate on which it acts can be engineered to associate avidly with the substrate or even to redirect its attention to novel compounds. A protein that normally is transported to one cellular compart-

the cell. A gene that normally is stimulated to expression by one agent can be made to respond to a completely new signal. In short, by altering the genes that organize a biological system the molecular biologist can change the usual relations between its elements in ways that show how the system normally works. Many biologists of the future will think of a biological system in terms of a series of well-defined mechanical parts that can be dismantled, engineered and reassembled under the guidance of the molecular mechanic.

It is still far from clear that attempts to reduce complex systems to small and simple components, pushed to an extreme, can provide adequate insights for coming to grips with the great problem biologists confront today: describing the overall functioning of a complex organism. Can the biology of a mammal be understood as simply the sum of a large number of systems, each controlled by a different, well-defined gene? Probably not. A more realistic assessment would be that the interactions of complex networks of genes, gene products and specialized cells underlie many aspects of organismic function. Each gene in an organism has evolved not in isolation but in the context of other genes with which it has interacted continuously over a long period of evolutionary development. Most molecular biologists would concede that they do not yet possess the conceptual tools for understanding entire complex biological systems or processes having multiple