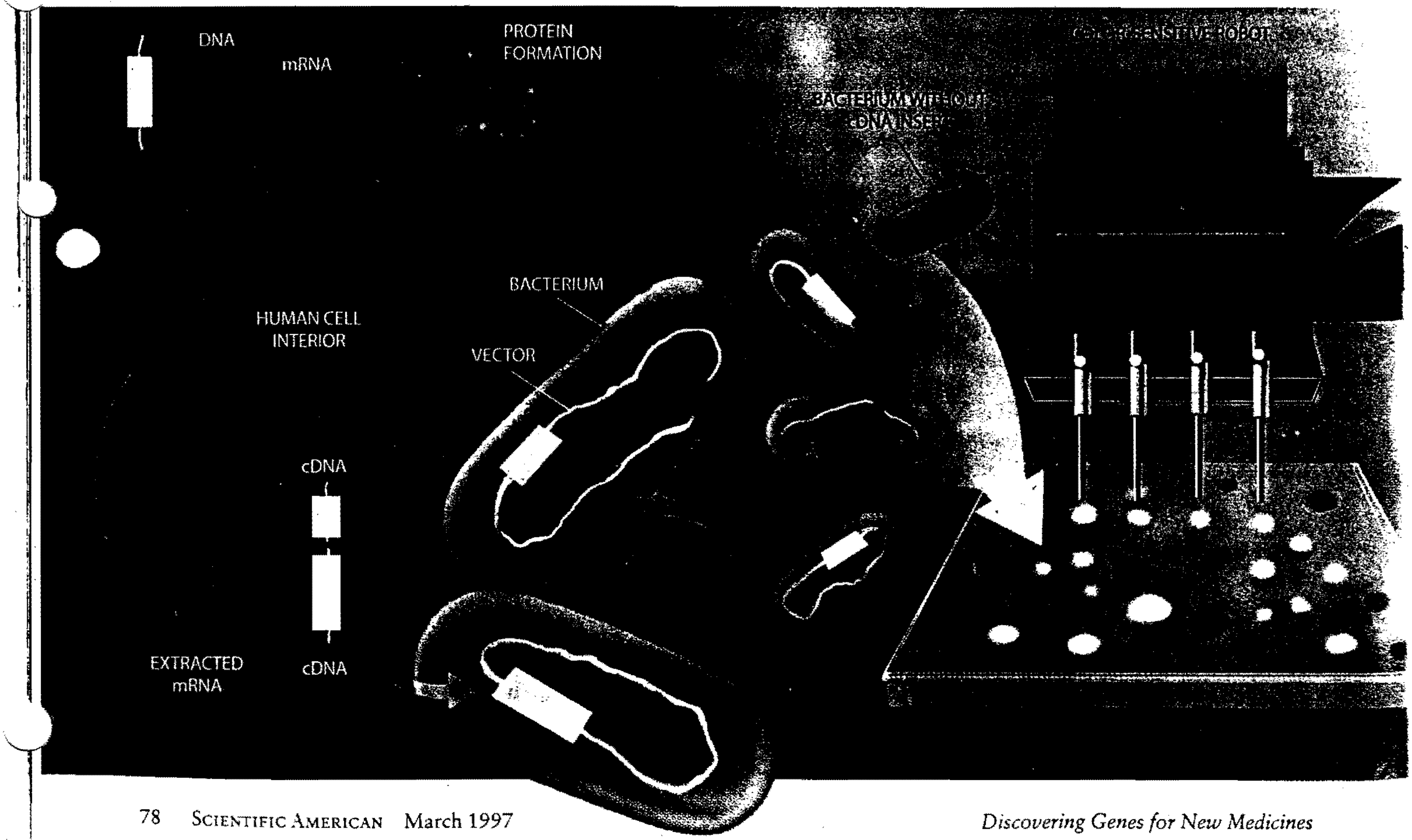
Discovering Genes for New Medicines

By identifying human genes involved in disease, researchers can create potentially therapeutic proteins and speed the development of powerful drugs

by William A. Haseltine

Most readers of this magazine are probably familiar with the idea of a gene as something that transmits inherited traits from one generation to the next. Less well appreciated is that malfunctioning genes are deeply involved in most diseases, not only inherited ones. Cancer, atherosclerosis, osteoporosis, arthritis and Alzheimer's disease, for example, are all characterized by specific changes in the activities of genes. Even infectious disease usually provokes the activation of identifiable genes in a patient's immune system. Moreover, accumulated damage to genes from a lifetime of exposure to ionizing radiation and injurious chemicals probably underlies some of the changes associated with aging.

A few years ago I and some like-minded colleagues decided that knowing where and when different genes are switched on in the human body would lead to far-reaching advances in our ability to predict, prevent, treat and cure disease. When a gene is active, or as a geneticist would say, "expressed," the sequence of the chemical units, or bases, in its DNA is used as a blueprint to produce a specific protein. Proteins direct, in various ways, all of a cell's functions. They serve as structural components, as catalysts that carry out the multiple chemical processes of life and as control elements that regulate cell reproduction, cell specialization and physiological activity at all levels. The development of a human from fertilized egg to mature adult is, in fact, the consequence of an



orderly change in the pattern of gene expression in different tissues.

Knowing which genes are expressed in healthy and diseased tissues, we realized, would allow us to identify both the proteins required for normal functioning of tissues and the aberrations involved in disease. With that information in hand, it would be possible to develop new diagnostic tests for various illnesses and new drugs to alter the activity of affected proteins or genes. Investigators might also be able to use some of the proteins and genes we identified as therapeutic agents in their own right. We envisaged, in a sense, a high-resolution description of human anatomy descending to the molecular level of detail.

It was clear that identifying all the expressed genes in each of the dozens of tissues in the body would be a huge task. There are some 100,000 genes in a typical human cell. Only a small proportion of those genes (typically about 15,000) is expressed in any one type of cell, but the expressed genes vary from one cell type to another. So looking at just one or two cell types would not reveal the genes expressed in the rest of the body. We would also have to study tissues from all the stages of human development. Moreover, to identify the changes in gene expression that contribute to sickness, we would have to analyze diseased as well as healthy tissues.

Technological advances have provided a way to get the job done. Scientists can now rapidly discover which genes are expressed in any given tissue. Our strategy has proved the quickest way to identify genes of medical importance.

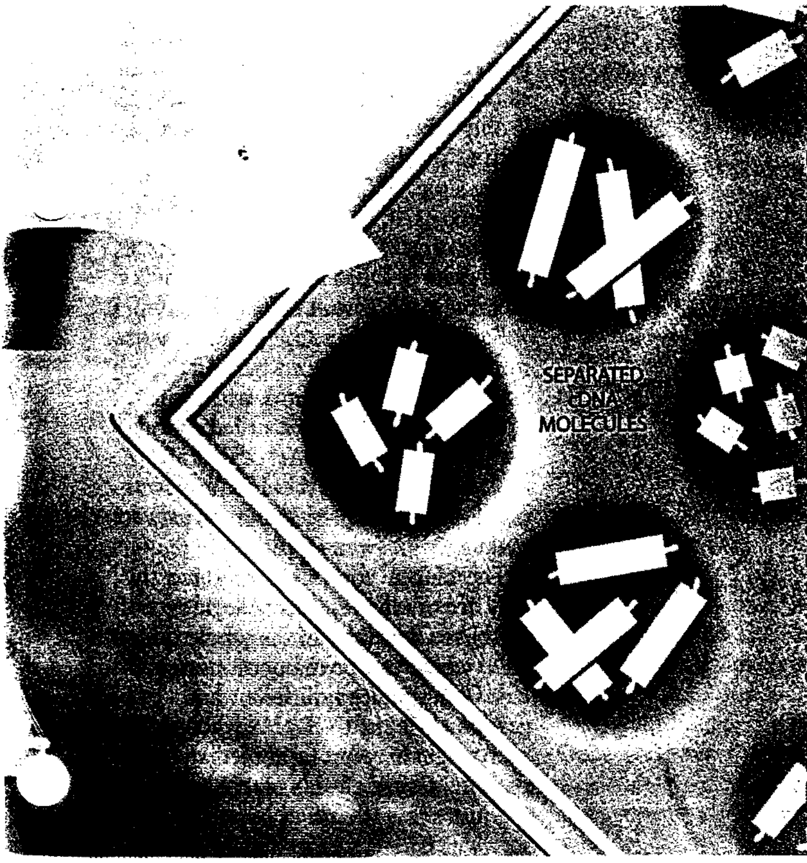
Take the example of atherosclerosis. In this common condition, a fatty substance called plaque accumulates inside arteries, notably those supplying the heart. Our strategy enables us to generate a list of genes expressed in normal arteries, along with a measure of the level of expression of each one. We can then compare the list with one derived from patients with atherosclerosis. The difference between the lists corresponds to the genes (and thus the proteins) involved in the disease. It also indicates how much the genes' expression has been increased or decreased by the illness. Researchers can then make the human proteins specified by those genes.

Once a protein can be manufactured in a pure form, scientists can fairly easily fashion a test to detect it in a patient. A test to reveal overproduction of a protein found in plaque might expose early signs of atherosclerosis, when better options exist for treating it. In addition, pharmacologists can use pure proteins to help them find new drugs. A chemical that inhibited production of a protein found in plaque might be considered as a drug to treat atherosclerosis.

Our approach, which I call medical genomics, is somewhat outside the mainstream of research in human genetics. A great many scientists are involved in the Human Genome Project, an international collaboration devoted to the discovery of the complete sequence of the chemical bases in human DNA. (All the codes in DNA are constructed from an alphabet consisting of just four bases.) That information will be important for studies of gene action and evolution and will particularly benefit research on inherited diseases. Yet the genome project is not the fastest way to discover genes, because most of the bases that make up DNA actually lie outside genes. Nor will the project pinpoint which genes are involved in illness.

In 1992 we created a company, Human Genome Sciences (HGS), to pursue our vision. Initially we conducted the work as a collaboration between HGS and the Institute for Genomic Research, a not-for-profit organization that HGS supports; the institute's director, J. Craig Venter, pioneered some of the key ideas in genomic research. Six months into the collaboration, SmithKline Beecham,

JENNIFER C. CHRISTIANSEN



How to Make and Separate cDNA Molecules

Cells use messenger RNA to make protein. We discover genes by making complementary DNA (cDNA) copies of messenger RNA. First we have to clone and produce large numbers of copies of each cDNA, so there will be enough to determine its constituent bases. Molecular biologists have developed ways to insert cDNA into specialized DNA loops, called vectors, that can reproduce inside bacterial cells. A mixture of cDNAs from a given tissue is called a library.

Researchers at HGS have now prepared human cDNA libraries from almost all normal organs and tissues, as well as from many that are diseased. To make multiple copies of a library, we add it to bacteria that take up the vectors.

All the bacteria are then spread out on a plate of nutrient gel and allowed to grow into colonies, so that each colony derives from a single bacterium. Next we use a robot that can automatically spot and pick off the gel those colonies that did successfully acquire a cDNA. The robot accomplishes this by color. The vectors we use are designed so that if they fail to combine with a cDNA insert, they produce a blue pigment. The robot, which picks as many as 10,000 colonies of bacteria every day, identifies those containing human cDNA by avoiding blue ones. The cDNA from each picked colony, now in analyzable quantities, is then robotically purified. *—W.A.H.*

one of the world's largest pharmaceutical companies, joined HGS in the effort. After the first year, HGS and SmithKline Beecham continued on their own. We were joined later by Schering-Plough, Takeda Chemical Industries in Japan, Merck KGaA in Germany and Synthelabo in France.

Genes by the Direct Route

Because the key to developing new medicines lies principally in the proteins produced by human genes, rather than the genes themselves, one might wonder why we bother with the genes at all. We could in principle analyze a cell's proteins directly. Knowing a protein's composition does not, however, allow us to make it, and to develop medicines, we must manufacture substantial amounts of proteins that seem important. The only practical way to do so is to isolate the corresponding genes and transplant them into cells that can express those genes in large amounts.

Our method for finding genes focuses on a critical intermediate product created in cells whenever a gene is expressed. This intermediate product is called messenger RNA (mRNA); like DNA, it consists of sequences of four bases. When a cell makes mRNA from a gene, it essentially copies the sequence of DNA bases in the gene. The mRNA then serves as a template for constructing the specific protein encoded by the gene. The value of mRNA for research is that cells make it only when the corresponding gene is active. Yet the mRNA's base sequence, being simply related to the sequence of the gene itself, provides us with enough information to isolate the gene from the total mass of DNA in cells and to make its protein if we want to.

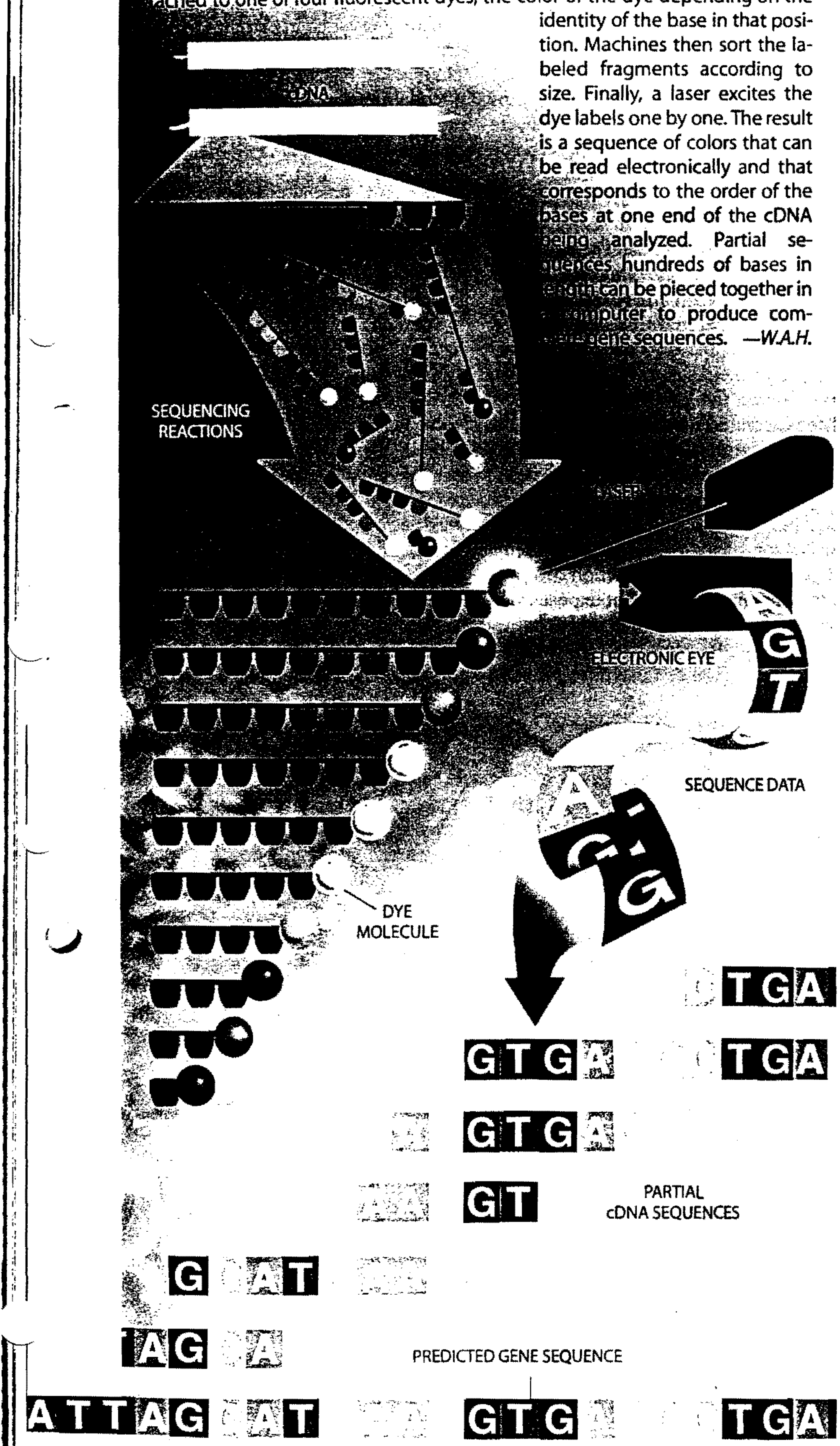
For our purposes, the problem with mRNA was that it can be difficult to handle. So we in fact work with a surrogate: stable DNA copies, called complementary DNAs (cDNAs) of the mRNA molecules. We make the cDNAs by simply reversing the process the cell uses to make mRNA from DNA.

The cDNA copies we produce this way are usually replicas of segments of mRNA rather than of the whole molecule, which can be many thousands of bases long. Indeed, different parts of a gene can give rise to cDNAs whose common origin may not be immediately apparent. Nevertheless, a cDNA containing just a few thousand bases still preserves its parent gene's unique signature.

How to Find a Partial cDNA Sequence

JENNIFER C. CHRISTIANSEN

Researchers find partial cDNA sequences by chemically breaking down copies of a cDNA molecule to create an array of fragments that differ in length by one base. In this process, the base at one end of each fragment is attached to one of four fluorescent dyes, the color of the dye depending on the identity of the base in that position. Machines then sort the labeled fragments according to size. Finally, a laser excites the dye labels one by one. The result is a sequence of colors that can be read electronically and that corresponds to the order of the bases at one end of the cDNA being analyzed. Partial sequences hundreds of bases in length can be pieced together in a computer to produce complete gene sequences. —*W.A.H.*



That is because it is vanishingly unlikely that two different genes would share an identical sequence thousands of bases long. Just as a random chapter taken from a book uniquely identifies the book, so a cDNA molecule uniquely identifies the gene that gave rise to it.

Once we have made a cDNA, we can copy it to produce as much as we want. That means we will have enough material for determining the order of its bases. Because we know the rules that cells use to turn DNA sequences into the sequences of amino acids that constitute proteins, the ordering of bases tells us the amino acid sequence of the corresponding protein fragment. That sequence, in turn, can be compared with the sequences in proteins whose structures are known. This maneuver often tells us something about the function of the complete protein, because proteins containing similar sequences of amino acids often perform similar tasks.

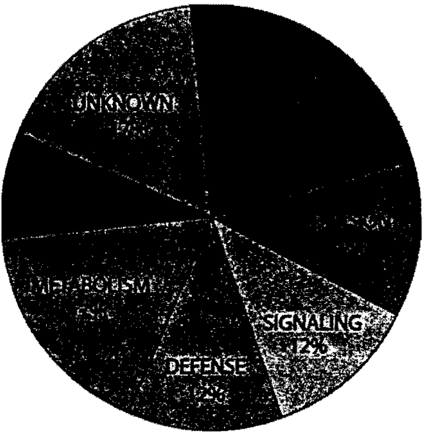
Analyzing cDNA sequences used to be extremely time-consuming, but in recent years biomedical instruments have been developed that can perform the task reliably and automatically. Another development was also necessary to make our strategy feasible. Sequencing equipment, when operated on the scale we were contemplating, produces gargantuan amounts of data. Happily, computer systems capable of handling the resulting megabytes are now available, and we and others have written software that helps us make sense of this wealth of genetic detail.

Assembling the Puzzle

Our technique for identifying the genes used by a cell is to analyze a sequence of 300 to 500 bases at one end of each cDNA molecule. These partial cDNA sequences act as markers for genes and are sometimes referred to as expressed sequence tags. We have chosen this length for our partial cDNA sequences because it is short enough to analyze fairly quickly but still long enough to identify a gene unambiguously. If a cDNA molecule is like a chapter from a book, a partial sequence is like the first page of the chapter—it can identify the book and even give us an idea what the book is about. Partial cDNA sequences, likewise, can tell us something about the gene they derive from. At HGS, we produce about a million bases of raw sequence data every day.

Our method is proving successful: in less than five years we have identified thousands of genes, many of which may play a part in illness. Other companies and academic researchers have also initiated programs to generate partial cDNA sequences.

HGS's computers recognize many of the partial sequences we produce as deriving either from one of the 6,000 genes researchers have already identified by other means or from a gene we have previously found ourselves. When we cannot definitely assign a newly generated partial sequence to a known gene, things get more interesting. Our computers then scan through our databases as well as public databases to see whether the new partial sequence overlaps something someone has logged before. When we find a clear overlap, we piece together the overlapping partial sequences into ever lengthening segments called contigs. Contigs correspond, then, to incomplete sequences we infer to be present somewhere in a parent gene. This process is somewhat analogous to fishing out the phrases "a midnight dreary, while I pondered" and "while I pondered, weak and weary/Over many a...volume" and combining them into a fragment recognizable as part of Edgar Allan Poe's "The Raven."



PERCENTAGE OF GENES devoted to each of the major activities in the typical human cell has been deduced from a study of 150,000 partial sequences. Similarities with human or other genes of known function were used to assign provisional categories of activity.

JENNIFER C. CHRISTIANSEN

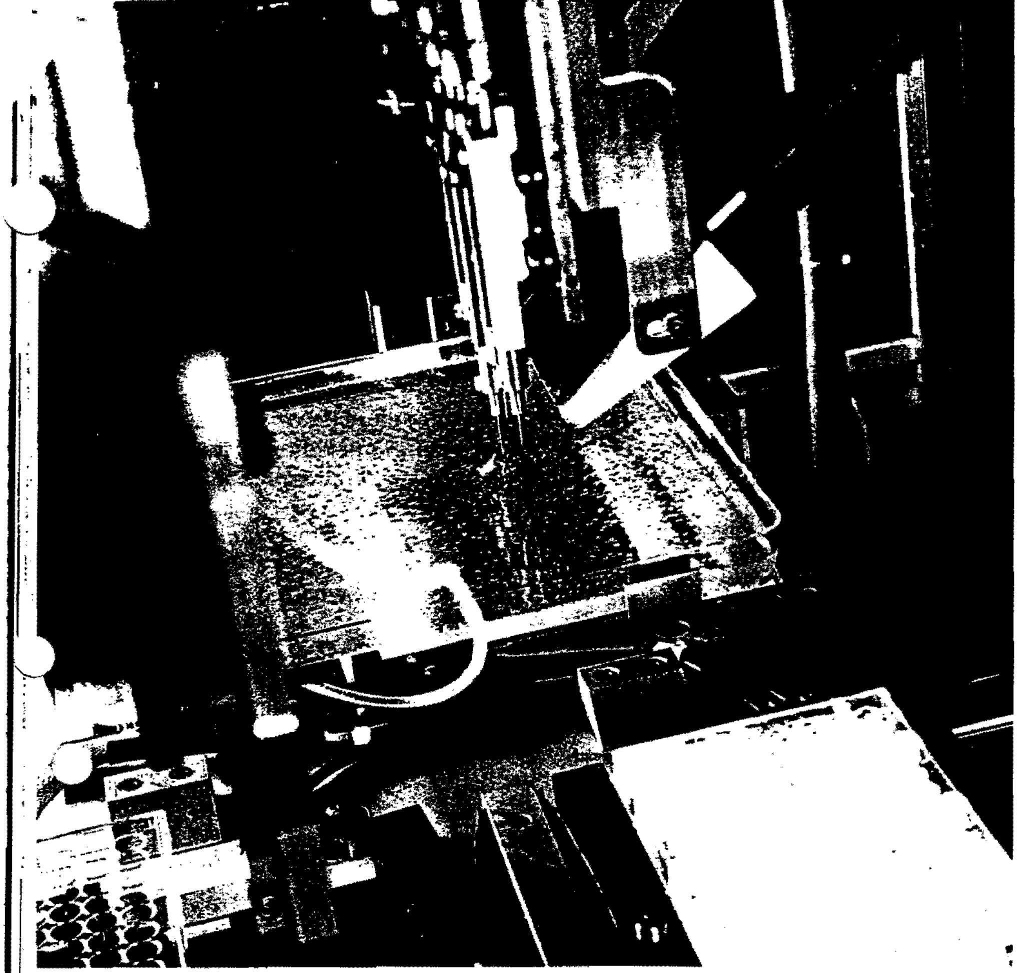
At the same time, we attempt to deduce the likely function of the protein corresponding to the partial sequence. Once we have predicted the protein's structure, we classify it according to its similarity to the structures of known proteins. Sometimes we find a match with another human protein, but often we notice a match with one from a bacterium, fungus, plant or insect: other organisms produce many proteins similar in function to those of humans. Our computers continually update these provisional classifications.

Three years ago, for example, we predicted that genes containing four specific contigs would each produce proteins similar to those known to correct mutations in the DNA of bacteria and yeast. Because researchers had learned that failure to repair mutations can cause colon cancer, we started to work out the full sequences of the four genes. When a prominent colon cancer researcher later approached us for help in identifying genes that might cause that illness—he already knew about one such gene—we were able to tell him that we were already working with three additional genes that might be involved.

Subsequent research has confirmed that mutations in any one of the four genes can cause life-threatening colon, ovarian or endometrial cancer. As many as one in every 200 people in North America and Europe carry a mutation in one of these mismatch repair genes, as they are called. Knowing this, scientists can develop tests to assess the mismatch repair genes in people who have relatives with these cancers. If the people who are tested display a genetic predisposition to illness, they can be monitored closely. Prompt detection of tumors can lead to lifesaving surgery, and such tests have already been used in clinical research to identify people at risk.

Our database now contains more than a million cDNA-derived partial gene sequences, sorted into 170,000 contigs. We think we have partial sequences from almost all expressed human genes. One indication is that when other scientists log gene sequences into public databases, we find that we already have a partial sequence for more than 95 percent of them. Piecing together partial sequences frequently uncovers entire new genes. Overall more than half of the new genes we identify have a resemblance to known genes that have been assigned a probable function. As time goes by, this proportion is likely to increase.

If a tissue gives rise to an unusually large number of cDNA sequences that derive from the same gene, it provides an indication that the gene in question is producing copious amounts of mRNA. That generally happens when the cells are producing large amounts of the cor-



ROBOT used to distinguish bacterial colonies that have picked up human DNA sequences is at the top. The instrument's arms ignore colonies that are blue, the sign that they contain no human DNA. By analyzing the sequences in the bacteria, researchers can identify human genes.

BERND AUERS

responding protein, suggesting that the protein may be doing a particularly vital job. HGS also pays particular attention to genes that are expressed only in a narrow range of tissues, because such genes are most likely to be useful for intervening in diseases affecting those tissues. Of the thousands of genes we have discovered, we have identified about 300 that seem especially likely to be medically important.

New Genes, New Medicines

Using the partial cDNA sequence technique for gene discovery, researchers have for the first time been able to assess how many genes are devoted to each of the main cellular functions, such as defense, metabolism and so on. The vast store of unique information from partial cDNA sequences offers new possibilities for medical science. These opportunities are now being systematically explored.

Databases such as ours have already proved their value for finding proteins that are useful as signposts of disease. Prostate cancer is one example. A widely used test for detecting prostate cancer measures levels in the blood of a protein called prostate specific antigen. Patients who have prostate cancer often exhibit unusually high levels. Unfortunately, slow-growing, relatively benign tumors as well as malignant tumors requiring aggressive therapy can cause elevated levels of the antigen, and so the test is ambiguous.

HGS and its partners have analyzed mRNAs from multiple samples of healthy prostate tissue as well as from benign and malignant prostate tumors. We found about 300 genes that are expressed in the prostate but in no other tissue; of these, about 100 are active only in prostate tumors, and about 20 are expressed only in tumors rated by pathologists as malignant. We and our commercial partners are using these 20 genes and their protein products to devise tests to identify malignant prostate disease. We have similar work under way for breast, lung, liver and brain cancers.

Databases of partial cDNA sequences can also help find genes responsible for rare diseases. Researchers have long known, for example, that a certain form of blindness in children is the result of an inherited defect in the chemical breakdown of the sugar galactose. A search of our database revealed two previously unknown human genes whose corresponding proteins were predicted to be structurally similar to known galactose-metabolizing enzymes in yeast and bacteria. Investigators quickly confirmed that inherited defects in either of these two genes cause this type of blindness. In the future, the enzymes or the genes themselves might be used to prevent the affliction.

Partial cDNA sequences are also establishing an impressive record for helping researchers to find smaller molecules that are candidates to be new treatments. Methods for creating and testing small-molecule drugs—the most common type—have improved dramatically in the past few years. Automated equipment can rapidly screen natural and synthetic compounds for their ability to affect a human protein involved in disease, but the limited number of known protein targets has delayed progress. As more human proteins are investigated, progress should accelerate. Our work is now providing more than half of SmithKline Beecham's leads for potential products.

Databases such as ours not only make it easier to screen molecules randomly for useful activity. Knowing a protein's structure enables scientists to custom-design drugs to interact in a specific way with the protein. This technique, known as rational drug design, was used to create some of the new protease inhibitors that are proving effective against HIV (although our database was not involved in this particular effort). We are confident that partial cDNA sequences will allow pharmacologists to make more use of rational drug design.

One example of how our database has already proved useful concerns cells known as osteoclasts, which are normally present in bone; these cells produce an enzyme capable of degrading bone tissue. The enzyme appears to be produced in excess in some disease states, such as osteoarthritis and osteoporosis. We found in our computers a sequence for a gene expressed in osteoclasts that appeared to code for the destructive enzyme; its sequence was similar to that of a gene known to give rise to an enzyme that degrades cartilage. We confirmed

Protein	Activity	Possible Uses
Keratinocyte growth factor	Stimulates regrowth of skin	Healing wounds, stimulating hair growth, protecting against chemotherapy's side effects
Myeloid progenitor Inhibitory protein 1	Prevents chemotherapy drugs from killing bone marrow cells	Protecting against chemotherapy's side effects
Motor neuron growth factor	Prevents trauma-induced motor neuron death	Treating Lou Gehrig's disease, traumatic nerve injury, stroke and muscle atrophy in aging
Monocyte colony inhibitory factor	Inhibits macrophages	Treating rheumatoid arthritis and other autoimmune and macrophage-related diseases

JENNIFER C. CHRISTIANSEN

HUMAN PROTEINS made after their genes were discovered at Human Genome Sciences include several that demonstrate powerful effects in isolated cells and in experimental animals. These examples are among a number of human proteins now being tested to discover their possible medical value.

that the osteoclast gene was responsible for the degradative enzyme and also showed that it is not expressed in other tissues. Those discoveries meant we could invent ways to thwart the gene's protein without worrying that the methods would harm other tissues. We then made the protein, and SmithKline Beecham has used it to identify possible therapies by a combination of high-throughput screening and rational drug design. The company has also used our database to screen for molecules that might be used to treat atherosclerosis.

One extremely rich lode of genes and proteins, from a medical point of view, is a class known as G-protein coupled receptors. These proteins span the cell's outer membrane and convey biological signals from other cells into the cell's interior. It is likely that drugs able to inhibit such vital receptors could be used to treat diseases as diverse as hypertension, ulcers, migraine, asthma, the common cold and psychiatric disorders. HGS has found more than 70 new G-protein coupled receptors. We are now testing their effects by introducing receptor genes we have discovered into cells and evaluating how the cells that make the encoded proteins respond to various stimuli. Two genes that are of special interest produce proteins that seem to be critically involved in hypertension and in adult-onset diabetes. Our partners in the pharmaceutical industry are searching for small molecules that should inhibit the biological signals transmitted by these receptors.

Last but not least, our research supports our belief that some of the human genes and proteins we are now discovering will, perhaps in modified form, themselves constitute new therapies. Many human proteins are already used as drugs; insulin and clotting factor for hemophiliacs are well-known examples. Proteins that stimulate the production of blood cells are also used to speed patients' recovery from chemotherapy.

The proteins of some 200 of the full-length gene sequences HGS has uncovered have possible applications as medicines. We have made most of these proteins and have instituted tests of their activity on cells. Some of them are also proving promising in tests using experimental animals. The proteins include several chemokines, molecules that stimulate immune system cells.

Developing pharmaceuticals will never be a quick process, because medicines, whether proteins, genes or small molecules, have to be extensively tested. Nevertheless, partial cDNA sequences can speed the discovery of candidate therapies. HGS allows academic researchers access to much of its database, although we ask for an agreement to share royalties from any ensuing products.

The systematic use of automated and computerized methods of gene discovery has yielded, for the first time, a comprehensive picture of where different genes are expressed—the anatomy of human gene expression. In addition, we are starting to learn about the changes in gene expression in disease. It is too early to know exactly when physicians will first successfully use this knowledge to treat disease. Our analyses predict, however, that a number of the resulting therapies will form mainstays of 21st-century medicine. ◘

To obtain high-quality reprints of this article, please see page 19.

The Author

WILLIAM A. HASELTINE is chairman of the board of directors and chief executive officer of Human Genome Sciences in Rockville, Md. He has a doctorate in biophysics from Harvard University and from 1976 to 1993 was a professor with appointments at the Dana-Farber Cancer Institute, Harvard Medical School and Harvard School of Public Health. Haseltine has won numerous scientific awards for his research on cancer and AIDS; he has also been awarded more than 50 patents. Since 1981 he has founded seven biotechnology companies. In 1988, together with Flossie Wong-Staal, Haseltine wrote in *Scientific American* about the molecular biology of the AIDS virus.

Further Reading

DEALING WITH GENES: THE LANGUAGE OF HEREDITY. Paul Berg and Maxine Singer. University Science Books, Blackwell Scientific Publications, 1992.

MUTATION OF A MUTL HOMOLOG IN HEREDITARY COLON CANCER. N. Papadopoulos in *Science*, Vol. 263, pages 1625–1629; March 18, 1994.

MUTATIONS OF TWO PMS HOMOLOGUES IN HEREDITARY NONPOLYPOSIS COLON CANCER. N. Nicolaides et al. in *Nature*, Vol. 317, pages 75–80; September 1, 1994.

INITIAL ASSESSMENT OF HUMAN GENE DIVERSITY AND EXPRESSION PATTERNS BASED UPON 83 MILLION NUCLEOTIDES OF cDNA SEQUENCE. Mark D. Adams et al. in *Nature*, Vol. 377, Supplement, pages 3–174; September 28, 1995.

A cDNA ENCODING THE CALCITONIN GENE-RELATED PEPTIDE TYPE 1 RECEPTOR. Nambi Aiyar et al. in *Journal of Biological Chemistry*, Vol. 271, No. 19, pages 11325–11329; May 10, 1996.

CATHEPSIN K, BUT NOT CATHEPSINS B, L, OR S, IS ABUNDANTLY EXPRESSED IN HUMAN OSTEOCLASTS. Fred H. Drake et al. in *Journal of Biological Chemistry*, Vol. 271, No. 21, pages 12511–12516; May 24, 1996.