

THE CODE OF CODES

Scientific and Social Issues in
the Human Genome Project

EDITED BY

DANIEL J. KEVLES
AND LEROY HOOD

CERMES
INSERM / CNRS
DOCUMENTATION

HARVARD UNIVERSITY PRESS
Cambridge, Massachusetts
London, England

LEROY HOOD

Biology and Medicine in the Twenty-First Century

6

During the past twenty years, brilliant advances in technology and fundamental new insights have led to a striking revolution in biology, which is slowly beginning to change medicine. This revolution will be accelerated as we move into the twenty-first century by even more far-reaching developments, especially the deciphering of the human genome, our blueprint for life. The human genome project is on its way to creating an encyclopedia of life, giving biologists and physicians direct computer access to the secrets of our chromosomes. The project is daunting in scope and scale, and accomplishing it will require still more advances in chemistries, techniques, instrumentation, and sophisticated computational hardware and software. If we succeed, the infrastructure of biology will be enriched, and the revolution that has begun in the practice of biology and clinical medicine will accelerate.

The human genome project is the first major biological initiative that takes the development of technologies as one of its major objectives. Some of these techniques are necessary to create and analyze the three types of maps essential to the genome project. We already know how to draw the genetic and physical maps, but improved technologies will increase enormously the rate at which they can be generated. We also must develop DNA sequencing techniques that are a hundred to a thousand times more rapid than what is currently available before we can seriously

Table 8 Genome sizes of model organisms

Organism	Million bases
<i>E. coli</i>	5
Yeast	15
Nematode (worm)	100
<i>Drosophila</i> (fly)	180
Mouse	3,000
Human	3,000

embark upon the task of sequencing the entire human genome. The development of hardware and software is required to organize the data from the three maps—genetic, physical, and sequence—of the human genome so that physicians and biologists can have computer access to this information for attacking fundamental problems in biology. Genomic analyses of model organisms—such as bacteria, yeast, worms, fruit flies, and mice—are also a part of the genome project (Table 8). These model organisms will provide valuable insights into how the genes shared with humans function, and the mouse genome (the only other mammalian genome in the list) will aid in defining human genes and regulatory regions through the identification of sequence regions conserved between the two species.

The timetable of the human genome project (see Table 9) divides into three five-year periods. During the first two five-year periods, the major focus will be technology development and creation of the genetic and physical maps. It is likely that only after the first ten years will the technology for large-scale sequencing have been developed to the point (faster than the current rate by a hundred-fold) where it will be feasible to do large-scale sequence analysis of the human and of many of the model organisms. Thus, the genome program proposes to carry out the bulk of the DNA sequencing only after appropriate rates of DNA sequencing are reached.

Just as the complex road system of the United States has transformed transportation in the country by permitting ready access to virtually any city, street, or house, so will the creation of genetic, physical, and sequence maps greatly facilitate our ability to access interesting genes. Currently, each time a new disease gene is to be isolated, a road must be built to that specific gene through

Table 9 Timetable for the human genome initiative

Time	Objectives
1-5 years	Technology: 5-10-fold improvement Informatics Crude genetic map Physical map for 5-10 chromosomes Sequence some biologically interesting regions (<1 percent) Model organisms: map and start sequence
5-10 years	Technology: 5-10-fold improvement More informatics Refined genetic map Physical map finished Sequence more biologically interesting regions (<5 percent) Model organisms: finish sequence
10-15 years	Technology: more Informatics: more Sequence: finished (95 percent) Additional model organisms sequenced

the techniques of recombinant DNA technology. Indeed, multiple roads are often independently built to interesting genes by competing groups. Once the three genome maps are available, the task of finding disease genes will be enormously simplified and the cost greatly reduced. Thus the maps of the human genome can be viewed as powerful tools that will significantly enrich the infrastructures of biology and medicine.

The benefits that will arise from having a complete sequence map of the human genome, sometime early in the twenty-first century, fall into four categories. First, the development of the requisite technologies necessary to accomplish the human genome project will revolutionize many other aspects of biology and medicine. Second, computer access to the genome maps will dramatically alter the practice of biology. Third, access to the genetic and sequence maps will fundamentally change the practice of clinical medicine. Finally, the information generated by the human genome project, as well as the new technologies that emerge from this endeavor, will ensure the United States a highly competitive position in the worldwide biotechnology industry.

The genome program will necessitate the development of more powerful technologies for DNA handling, mapping, sequencing, and analysis. There is potential for significant improvement in the physical and genetic mapping techniques and, indeed, the success in sequencing the human genome will require at least a hundred-fold increase in throughput for DNA sequencing. There are also challenging computational problems associated with the genome project. The improvement of technology impinges on four areas—the development of new techniques, automation, increased throughput, and increased sensitivity of analysis. In general, the key to technology development will be a multidisciplinary approach combining the powerful tools of applied mathematics and physics, chemistry, engineering, and computer science as well as biology.

Let me illustrate the power of this approach by describing the Science and Technology Center for Molecular Biotechnology that I head, in which we are developing an interdisciplinary group committed to the development of new technologies for biology. These interdisciplinary areas include expertise in protein chemistry, mass spectrometry, nucleic acids chemistry, large-scale DNA sequencing, genetic mapping, DNA diagnostics, and computational techniques (Figure 13). Cross-fertilization among these groups has led to the development of techniques and instrumentation that have had or will have a significant impact on the genome effort.

For example, in the early 1980s it became obvious that we needed to acquire the capacity to synthesize automatically small fragments of DNA (oligonucleotides), ten to fifty bases in length. These oligonucleotides or probes were useful for cloning genes and sequencing DNA, and later they served as primers for the polymerase chain reaction (PCR), a technique for amplifying any particular small region of DNA a million-fold or more. We automated a manual technique that attached the first DNA base in the oligonucleotide to a small, inert bead (solid support) and then carried out successive chemistries on this base-substituted solid support, adding one base at a time to the growing DNA chain (Figure 14).¹ The automation of this technique enormously increased the throughput for DNA synthesis, both by decreasing the cycle time (approximately five minutes) and by allowing multiple chains to be synthesized simultaneously (four-column machines).

ment of genomic DNA; it thus serves as a unique recognition marker for that region of genomic sequence. STSs are important for physical mapping for several reasons. First, they may be used to define uniquely each DNA clone, whether it be a yeast artificial chromosome (100,000–1,000,000 bp insert), a cosmid (30,000–45,000 bp insert), or a lambda clone (5,000–20,000 bp insert). Second, the STS can be used to identify other clones that share this unique DNA sequence and thus generate overlapping inserts for a physical map (see Figure 15). Third, the physical map may be stored in a computer as a series of STSs from overlapping clones. This information can be shipped electronically to distant investigators so that others may quickly re-create the physical map from their own genomic libraries using the PCR primer pairs as screening tools. Thus the need to store and ship large sets of DNA clones is entirely circumvented. Finally, the STS maps of one laboratory may be easily merged with those of other laboratories. Hence, the STS map of any chromosome is infinitely extendable, an advantage that arrays of physical clones lack. Thus the STS approach provides accountability in that it is easy to gauge the contributions of each investigator.

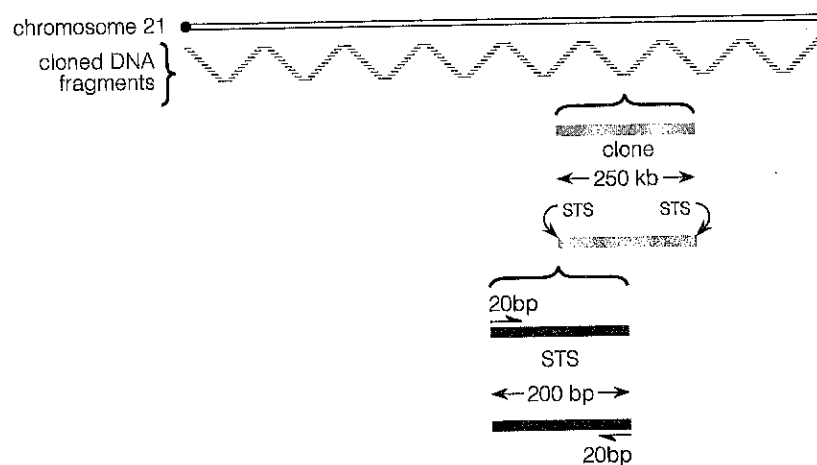


Figure 15 Sequence tagged sites (STSs) are short stretches of unique DNA sequence that will be used to create a physical map, each representing at least one clone in an overlapping set of clones for chromosome 21. In the illustration, the clones are an average of 250 kb in length, and each one begins and ends with an STS. The STSs are defined by unique PCR primers (these are 20 base pairs in length) and can be used to identify other clones that share the same DNA or STS sequence.

Genetic maps are created by following the segregation (passage from parents to children) of DNA polymorphisms in families. The genomes of humans are highly polymorphic: one base in five hundred will differ between any two individuals. If we are to develop a two-centimorgan map as a part of the genome project, more than 1,600 evenly spaced genetic markers have to be identified (the human genome is approximately 3,300 centimorgans in length). We've developed a technique that automates the analysis of DNA polymorphisms by using a robotic workstation that can handle plates with 96 small wells—thus 96 genetic markers can be analyzed simultaneously and automatically. This procedure enables us to: (1) amplify the segment of DNA that is to be examined for a polymorphism by PCR; (2) analyze the polymorphisms to determine which forms are present; and (3) automatically read and store the results directly in a computer.³ It has the capacity to increase enormously the throughput analysis of genetic markers—indeed, 1,200 assays can be carried out per day by a single technician using a robotic workstation (Figure 16). With it we will be able to analyze the markers necessary to create the genetic map and to determine rapidly the location of interesting new genetic markers without employing the time-consuming techniques of conventional genetic mapping, such as RFLP mapping, which are difficult to automate. Indeed, it uses polymorphic STS markers to create a genetic map, and these can in turn be employed to generate a physical map (Figure 15). Hence, this technique leads to a merging of the genetic and physical maps.

The heart of the genome program is the sequence analysis of the twenty-four different human chromosomes. The development of fully automated techniques for DNA sequencing is an imperative for the genome project. We have begun this process by developing an automated DNA sequencing machine that uses four different fluorescent dyes to color-code the four DNA bases.⁴ The sequence of the bases may then be read as colored bands migrating on an electrophoretic gel (Figure 17). This machine can analyze more than 12,000 base pairs of DNA sequence per day—the approximate amount of sequence a scientist in the early 1980s could determine in an entire year.

It is important to point out that large-scale DNA sequencing is a multi-step process.⁵ It is necessary to purify the DNA, fragment and map it, electrophorese the fragments, assemble each string of fragments into larger strings (ultimately the length of each chro-

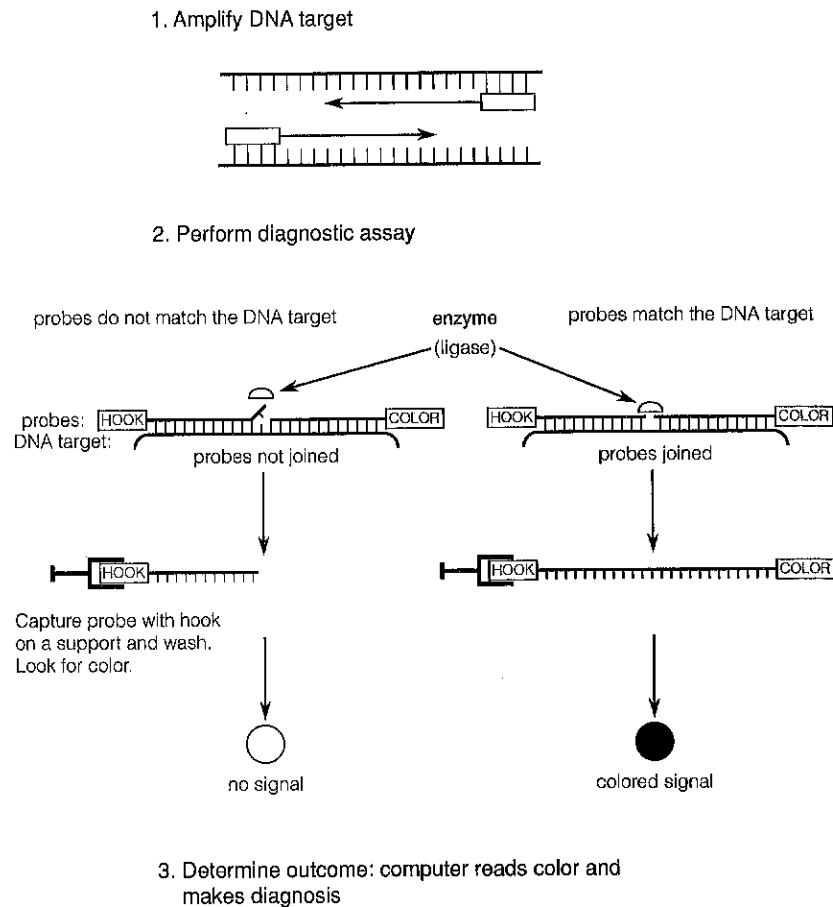


Figure 16 Three steps of one automated technique for genetic mapping of identified DNA polymorphisms: (1) The oligonucleotide site of the polymorphism is amplified by PCR. (2) The oligonucleotide ligase assay is carried out. Two adjacent DNA probes are synthesized. The left-hand probe has a hook (biotin) attached to its end, whereas the right-hand probe has a colored reporter group attached to its end. The base at the right-hand end of the hook probe is located at the polymorphic site. The two probes are hybridized to the target DNA. If the polymorphic base of the hook probe is complementary to the target base, then the enzyme DNA ligase can join the two probes, and when the hook is used to remove the left-hand probe from the reaction mixture, it also brings the right-hand probe (and color). Conversely, if the base of the hook probe is not complementary, then DNA ligase fails to join the two probes and removal of the hook brings with it only the left-hand probe (no color). In other words, only those samples that show a colored signal contain DNA with the target polymorphism. (3) A computer reads the presence or absence of color in each of the 96 wells and performs the genetic mapping calculations.

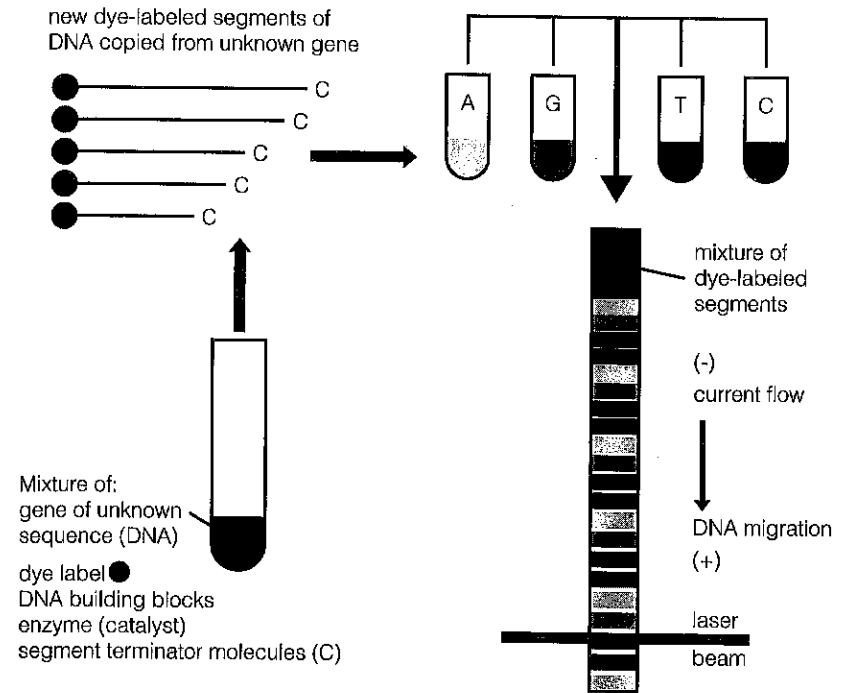


Figure 17 An illustration of the automated fluorescent DNA sequencing technology. The Sanger or enzymatic procedure for DNA sequencing (see Figure 6) is employed to synthesize four nested sets of DNA fragments all ending, respectively, in C, T, G, or A. The primer used to initiate synthesis for the C fragments is labeled with a red fluorescent dye, for the T fragments with a gold dye, for the G fragments with an orange dye, and for the A fragments with a green dye (here shown as varying shades of gray). These four DNA fragment mixtures are pooled and then fractionated by electrophoresis on a gel that has the capacity to resolve DNA fragments differing by one base. The laser beam activates the fluorescence of the dyes in each lane as the bands migrate past and this signal is picked up by a detector, which sends the information to a computer. Thus the color of the band identifies the end base on the DNA fragment, and the order of the colored bands as they migrate past the detector translates into the DNA sequence. The commercial form of this machine runs 24 sequences simultaneously, each reading out 450–500 bases.

mosome), and analyze the sequences. We need to automate virtually every step in this production line to eliminate the many potential bottlenecks to high-throughput DNA sequencing.

The chance is perhaps 50 percent that in ten years some entirely new approach to DNA sequencing will be employed—scanning-tip electron microscopy, mass spectrometry, or something else. However, the current approach for DNA sequencing has the potential to be improved a hundred-fold or more. I envision in ten

years instruments and/or strategies that will sequence one to ten million base pairs per day per technician.

The genome program poses striking problems for the computational sciences. Improvements are needed in signal processing, for example; if we can speed up our analysis of the fluorescent bands from the automated DNA sequencer, we can more than double the data output. The data bases will require advanced techniques for inputting, storing, and making readily accessible the three billion base pairs of genome sequence; they may also have to provide a hundred-fold more annotated description of this sequence. Another computational problem is string matching, the comparison of any new sequence generated against all of the sequences present in the data base to determine similarity of patterns.

To comprehend the string-matching problem, consider the following sequence:

TGCC TGGACTTCGCGCGACTATAGAGCGCGAGCGGCGTGAGC
 GAGACCAGTTTCGCAATGACTACGGTGCCACGCAAGGGTTCGTG
 CCTGGCTCAGCAAGGGTAGTCCTTAGTGAAGTGGCGGCTTAT
 GCCTGGACTTCGCGCGACTATAGAGCGCGAGCGGCGTGAGCG
 AGACCAGTTTCGCAATGACTACGGTGCCACGCAAGGGTTCGTG
 CTGGCTCACGAAGGGTAGTCCTTAGTGAAGTGGCGGCTTATG
 CCTGGACTTCGCGCGACTATAGAGCGCGAGCGGCGTGAGCGA
 GACCAGTTTCGCAATGACTACGGTGCCACGCAAGGGTTCGTGCC
 TGGCTCACGAAGGGTAGTCCTTAGTGAAGTGGCGGCTTATGC
 CTGGACTTCGCGCGACTATAGAGCGCGAGCGGCGTGAGCGGAG
 ACCAGTTTCGCAATGACTACGGTGCCACGCAAGGGTTCGTGCC
 GGCTCACGAAGGGTAGTCCTTAGTGAAGTGGCGGCTTATGCC
 TGGACTTCGCGCGACTATAGAGCGCGAGCGGCGTGAGCGGAGA
 CCAGTTTCGCAATGACTACGGTGCCACGCAAGGGTTCGTGCC
 GCTCACGAAGGGTAGTCCTTAGTGAAGTGGCGGCTTATGCC
 GGACTTCGCGCGACTATAGAGCGCGAGCGGCGTGAGCGGAGAC
 CAGTTTCGCAATGACTACGGTGCCACGCAAGGGTTCGTGCC
 CTCACGAAGGGTAGTCCTTAGTGAAGTGGCGGCTTATGCC
 GACTTCGCGCGACTATAGAGCGCGAGCGGCGTGAGCGGAGACC
 AGTTTCGCAATGACTACGGTGCCACGCAAGGGTTCGTGCC
 TCACGAAGGGTAGTCCTTAGTGAAGTGGCGGCTTATGCC
 ACTTCGCGCGACTATAGAGCGCGAGCGGCGTGAGCGGAGACCA
 GTTCGCAATGACTACGGTGCCACGCAAGGGTTCGTGCC
 CACGAAGGGTAGTCCTTAGTGAAGTGGCGGCTTATGCC
 CTTCGCGCGACTATAGAGCGCGAGCGGCGTGAGCGGAGACCAG
 TTCGCAATGACTACGGTGCCACGCAAGGGTTCGTGCC
 ACGAAGGGTAGTCCTTAGTGAAGTGGCGGCTTATGCC
 TTCGCGCGACTATAGAGCGCGAGCGGCGTGAGCGGAGACCAGT
 TCGCAATGACTACGGTGCCACGCAAGGGTTCGTGCC
 CGAAGGGTAGTCCTTAGTGAAGTGGCGGCTTATGCC
 TCGCGCGACTATAGAGCGCGAGCGGCGTGAGCGGAGACCAGT
 CGCAATGACTACGGTGCCACGCAAGGGTTCGTGCC
 GAAGGGTAGTCCTTAGTGAAGTGGCGGCTTATGCC
 CGCGCGACTATAGAGCGCGAGCGGCGTGAGCGGAGACCAGTTC
 GCAATGACTACGGTGCCACGCAAGGGTTCGTGCC

This stretch of sequence represents about one-millionth of the human genome. We must be able to extract from a sequence like this a variety of information, including the boundaries of genes, the presence of regulatory elements, and the presence of sequences that may relate to specialized chromosomal functions such as replication, compaction, and segregation. The key to extracting this information is the ability to compare this sequence against all preexisting sequences to test for similarities. We have approached the string-matching problem by the development of a specialized coprocessor, the Biological Information Signal Processor (BISP), which converts the Waterman-Smith algorithm, the most general approach for sequence similarity analysis, into a silicon chip (Figure 18). The BISP is about one centimeter square and contains 400,000 transistors; it is the most complex chip that the Jet Propulsion Laboratory at Caltech has ever designed. Its performance, measured against that of far more expensive computers, is strikingly rapid (Table 10). Clearly, close cooperation between biologists and computer scientists will be not merely advantageous but necessary to solving the complex and difficult problems inherent in the human genome project.

Interactive environments like the Science and Technology Center for Molecular Biotechnology, where many different disciplines can be focused on the development of the wide spectrum of techniques needed, are key to the success of the genome project. The human genome project needs to attract talented scientists from computer science, applied physics, applied mathematics, engineering, and chemistry, as well as many disciplines within biology itself. Scientists in these disciplines may be momentarily interested in biological problems such as the human genome project, but it is difficult to persuade them to make a long-term commitment. A critical question is, How can more scientists from other disciplines be brought into these efforts?

One approach to the problem is to create a new kind of biologist—mainly by establishing Ph.D. programs in biotechnology that build bridges to other disciplines. Such programs would select students who wish to major in one area of biology, such as molecular biology, and in another discipline, such as computer science. The student would have a mentor in each area and take appropriate qualifying examinations in each. The objective would be to choose, for example, a fundamental problem in molecular

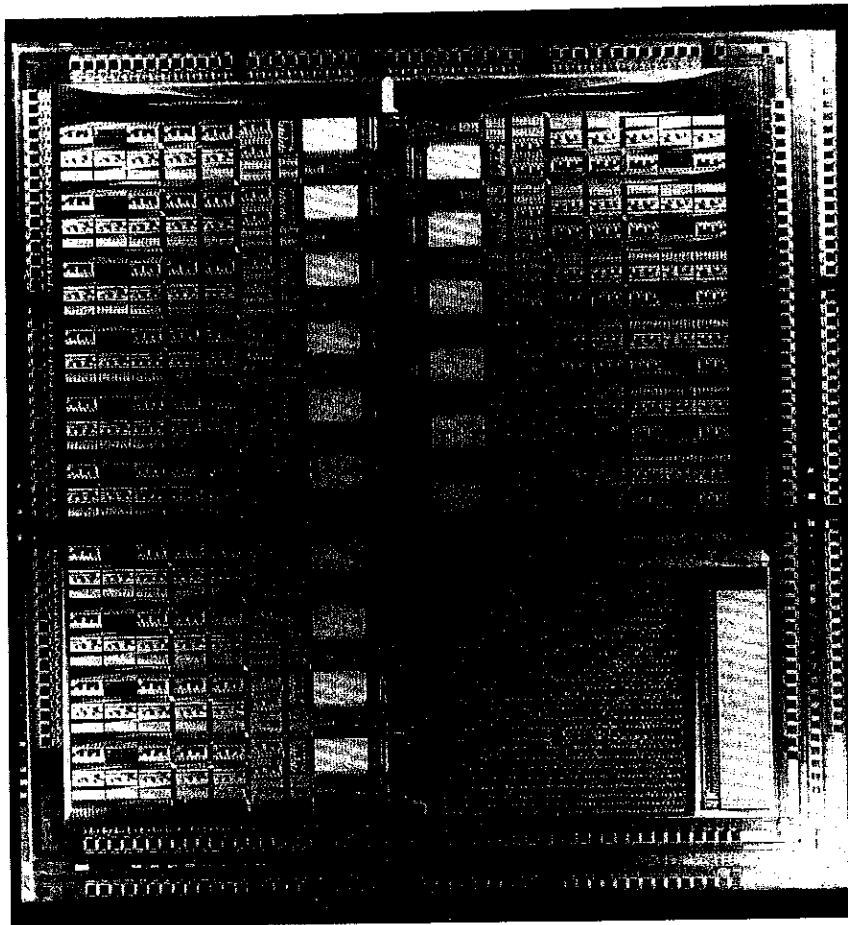


Figure 18 The Biological Information Signal Processor: a BISP chip is about one centimeter square and contains 400,000 transistors.

Table 10 The superior performance of the Biological Information Signal Processor: completion times for four systems comparing a 500-base sequence against a data base of 40 million bases (using a Smith-Waterman dynamic programming algorithm)

Computer	Time
Sun Sparcstation 1	5 hours
Cray 2	12 minutes
Connection Machine 3	1 minute
BISP	3.5 seconds

biology and then develop and apply a tool in computer science that could be applied to it, thus bringing computer science into biology through the student. This program would develop interdisciplinary scientists, those with expertise in biology and other disciplines and the ability to forge interdisciplinary collaborations. Moreover, these students will be the channels through which biologists and scientists from other disciplines may actively collaborate on developing biologically oriented techniques. I believe interdisciplinary scientists will play a major leadership role in biology and medicine of the twenty-first century.

Interdisciplinary collaboration will be essential to the progress of biology in the next century. The future of biology will depend upon the analysis of complex systems and networks that may involve molecules, cells, or even arrays of cells. If we are ever to understand such systems, the individual elements in the network must be defined, as must the nature of their connectivity. Computer models will be required to explore network behavior when individual elements are perturbed. Finally, the modeled behavior will then have to be tested on real biological systems. The living systems may be whole organisms or appropriately reconstructed subsystems of organisms. The human genome project will take a big step forward in identifying key elements of the complex system responsible for human growth and development by delineating the elements of the 100,000 human genes.

Once the sequence of the entire human genome is known, various computational and biological approaches can be taken to determine the location of the 100,000 genes. Several computer programs have combined the various general features of genes so that they may be identified among raw DNA sequence data—by looking for special base compositions of coding regions, for example, or for special sequences at exon-intron boundaries. Another approach will compare newly analyzed sequence data against all preexisting gene sequences from humans or model organisms, the idea being that sequence similarities may help reveal gene boundaries. Finally, the sequence of the human and mouse genomes will be compared. The mouse contains most human genes. The coding regions (and regulatory elements) are far more highly conserved during evolution than the intervening DNA that sur-

rounds the genes. Accordingly, an important element in the genome project will be the comparative sequence analysis of human and mouse DNA to aid in the identification and analysis of genes. The identification of coding regions is made more difficult by the fact that many genes show alternative patterns of RNA splicing; from the same gene sequence on DNA, several different messenger RNAs may be transcribed that splice together different combinations of exons or place particular exons at different sites. In the end, to define all the alternative forms of particular genes, one may have to study carefully the messenger RNAs in appropriate tissues. Nevertheless, the identification of most of the 100,000 human genes will provide biologists with an enormously powerful tool for exploring many aspects of contemporary biology.

Some biologists have argued that the copy DNAs (cDNA) of the mRNA should be sequenced rather than the DNA of the genome. The cDNA provides a direct read-out of the coding regions of genes, and these sequences, termed *expressed sequence tags* (ESTs), could also be used as markers spread throughout the genome to facilitate obtaining DNA fragments for a physical map. Since each automated DNA sequencer could readily define 5,000 ESTs per year, ESTs for many of the 100,000 or so human genes could readily be obtained early in the genome program.⁶ This truly represents a biological gold rush—and it has many fascinating implications. ESTs will permit a rapid assessment of the human genes through similarity analyses and raise fascinating patent questions (see Chapter 14). The regulatory elements will not be expressed in ESTs, nor will many other sequences important for general chromosomal functions. Moreover, for a variety of technical reasons, not all human genes can be identified by the EST approach (see below). Hence genomic and cDNA sequencing are both important for the genome program.

Each gene has regulatory elements or special DNA sequences that generally fall within 500 to 5,000 base pairs of the boundary of the gene itself (Figure 19). The regulatory elements function by virtue of the fact that specific DNA binding proteins interact with them. Called transactivating factors, they have three distinct functions. They control the temporal (developmental time) and spatial (tissue site) modes of expression and thereby coordinate a gene's expression in particular cells with that of thousands of other genes. They also control the amplitude of expression. For exam-

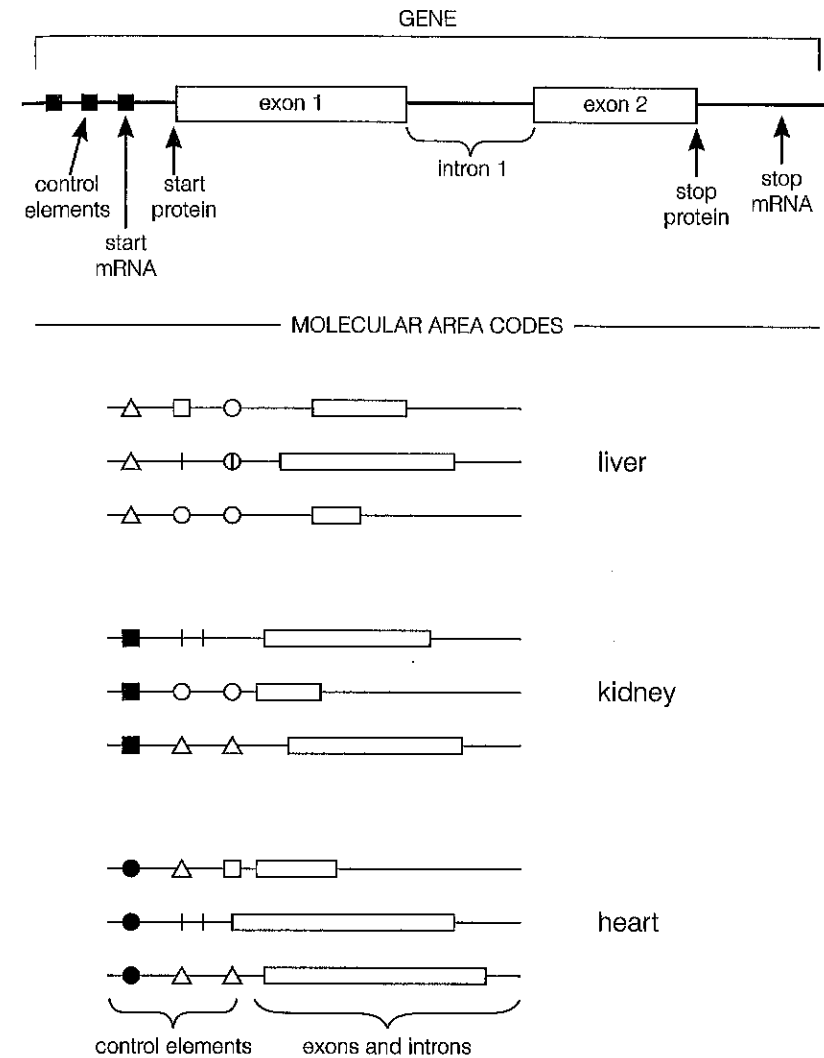


Figure 19 Only some parts (the exons) of the sequence that makes up a gene are transcribed into mRNA, the actual instructions for making the gene products. In addition, the gene also has noncoding regions (the introns) and control elements that regulate the expression of each gene. Also indicated here are the special control elements that serve as start and stop sites for mRNA and protein. The control elements for each gene constitute a molecular area code address. These regulatory addresses will someday permit us to identify, by computer analyses of their regulatory sequences, the temporal and spatial sites of expression as well as the amplitude of expression for each gene. Hence, they will permit us to determine in which organ or cell type particular genes are expressed.

ple, the regulatory elements and DNA binding proteins that control the expression of albumin dictate that it is expressed only in liver cells, that it is expressed late but not early in human development, and that it is expressed at perhaps thousands of times greater concentrations of mRNA than the average gene. These three functions may be written as a "molecular area code" (see Figure 19). The idea is that the three elements of gene expression will be dictated by specific DNA sequences and that these can be deciphered, just as we decipher a regular telephone number: suppose the first three digits determine the spatial location of a particular sequence; the second four the temporal location, and so on. In other words, defined regulatory elements may serve as molecular area codes to determine in which cells a gene is expressed, when it is expressed during development, the magnitude of its expression and, perhaps most interesting of all, the other genes with which it will be coordinately expressed. Molecular area codes will be an important tool for identifying the individual members of the biological network, and they will therefore be a part of the regulatory network that the genome project will delineate.

The regulatory elements or molecular area codes will in general be found in precisely the same way that the genes themselves are found. Comparisons with other known regulatory elements will be carried out by computational analysis and, in time, the general sequence properties of regulatory elements can probably be used to create specific computer programs to recognize these elements. In addition, cross-comparisons of the putative regulatory regions in mouse and human DNA sequences will be useful in delineating the regulatory elements, because they, as with their gene counterparts, will be highly conserved. Indeed, the first mammalian regulatory element ever identified was found because it is highly conserved in both human and mouse DNA.

The study of individual proteins in biology has traditionally started with the identification of a particular function, the development of an assay for this function, and the use of the assay to purify the protein that carries out the function. After sequencing the protein (that is, determining the order of its amino-acid subunits), the genetic code dictionary is used to translate the protein into DNA sequence; DNA probes are then synthesized and the gene is cloned by conventional recombinant techniques. The ge-

nome project will reverse this approach. In the future, we will know the 100,000 human genes and will have to develop new approaches and tools for ascertaining their functions. Indeed, the genome project will empower us to analyze genes that are inaccessible to the contemporary techniques of molecular biotechnology. For example, more than half of our genes are expressed in the brain, and many of them are expressed for such a short time during development and in so few cells that virtually no contemporary techniques would permit us to identify them. Perhaps we will be able to identify some of them only through direct sequence analyses of genomic DNA.

How might one go about ascertaining the function of newly discovered genes? First, one can do a search through existing data bases to see whether other genes of known function exhibit similar sequence characteristics. Second, the molecular area codes provided by the regulatory elements will generate insights as to the temporal, spatial, and coordinate expression of genes that may be useful in speculating on gene functions. Third, the genes may also provide information as to where in the cell the functions of corresponding genes are localized, once again providing insight into function. Finally, many genes may be present in model organisms whose genomes will be sequenced by the genome project. Thus, if a gene that corresponds to an unknown human gene, is found in the fly or the nematode, the model organism may be used for experimentation to discover the function of the gene in humans.

The sequences of all human genes will permit us to identify the corresponding proteins. This information will in turn allow us to find the motifs and domains that are the building blocks of proteins (Figure 20). Domains are the individual functional units within the protein; motifs are the building-block components for each domain. Indeed, a protein might be likened to a train. The domains would be the individual cars in the train, each different type of car—the flat car, the engine, the caboose—carrying out a different function. The motifs of a domain would be the individual components of cars, such as the wheels, the car sides and windows. A protein may have from one to fifteen or even more domains. For example, the antibody molecule that protects humans against foreign invaders (such as viruses and bacteria) folds into six domains, two of which are involved in recognizing the invad-

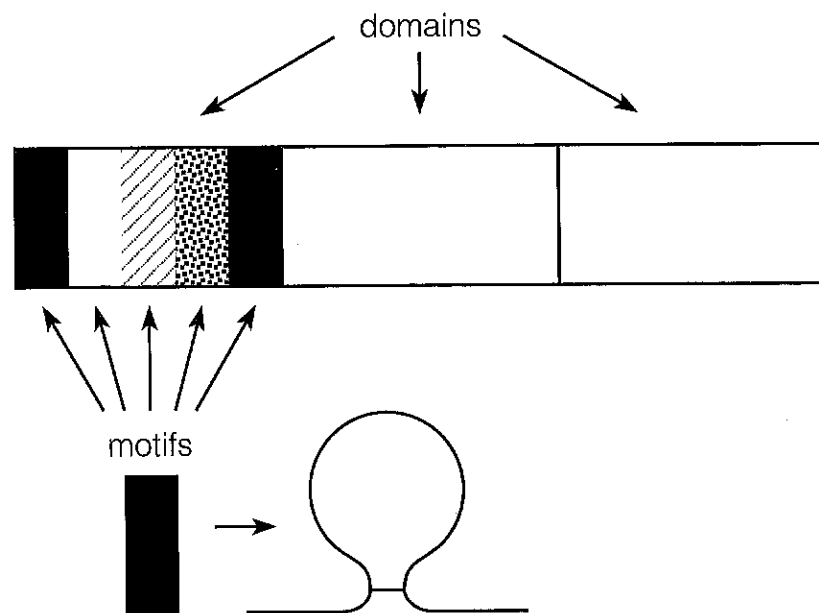


Figure 20 Proteins are composed of smaller building blocks, motifs and domains. Different domains have different functions, such as identification or elimination of a particular molecule. Motifs are the structural components of the protein molecule.

ers and four of which are involved in destroying or eliminating them. Each domain is composed of smaller motifs, called “ β -pleated sheets.” Having the sequence of all human proteins will allow us to use computational techniques to define domains and motifs. Indeed, if we identify the possible 100 to 500 motifs that are the fundamental building-block components of proteins, we will have a valuable tool for understanding the functions of the protein and how the order of the amino-acid subunits determines its three-dimensional structure. This is the so-called protein-folding problem.

The protein-folding problem is one of the major unsolved mysteries of contemporary biology. Within the next fifteen to twenty years it may be possible to decipher the folding rules so that, given the primary amino-acid sequence of a particular protein, we can predict what its three-dimensional structure would be. It is clear that the protein motifs may play a fundamental role in this process: that is, once the structure of a particular motif has been

determined, then all variant forms of this motif expressed in different proteins will have very similar structures. If we could determine the 100 to 500 basic structures of the protein motifs, then we would have a structural alphabet for understanding how proteins are assembled in three dimensions. Other approaches will also facilitate solutions to the protein folding problem. I have in mind theoretical calculations such as energy minimization, *in vitro* mutagenesis to alter a gene’s DNA sequence rationally so as to determine how the corresponding protein structure changes, as well as an examination of many additional proteins with high-resolution three-dimensional structures.

Once we can predict how a protein will fold in three dimensions, still another task will remain: to predict, from first principles, the function of the protein and understand how its structure and function are related. It is interesting to note that there is still no protein in contemporary biology for which we understand completely how its structure enables it to carry out its function. The step from structure to an understanding of function is a challenging one. Once again, new tools and approaches will have to be developed to take it.

The genome project in the twenty-first century will have a profound impact on medicine, both for diagnosis and therapy. The development of automated instrumentation for examining DNA polymorphisms raises the possibility of identifying the polymorphic forms of genes that cause disease or predispose individuals to disease. The ability to recognize particular DNA sequences by molecular complementarity between probe and target DNA is known as DNA diagnostics (Figure 21). This technology will figure in the diagnosis of genetic diseases whose single-gene defects have been identified; in determining the presence of dominant or recessive oncogenes that may predispose an individual to cancer; in the identification of infectious agents, such as the AIDS virus; and in forensics—that is, the use of DNA fingerprints to identify the donor origins of any tissue or blood sample. Perhaps the most important area of DNA diagnostics will be the identification of genes that predispose individuals to disease. However, many such diseases—cardiovascular, neurological, autoimmune—are polygenic; they are the result of the action of two or more genes.

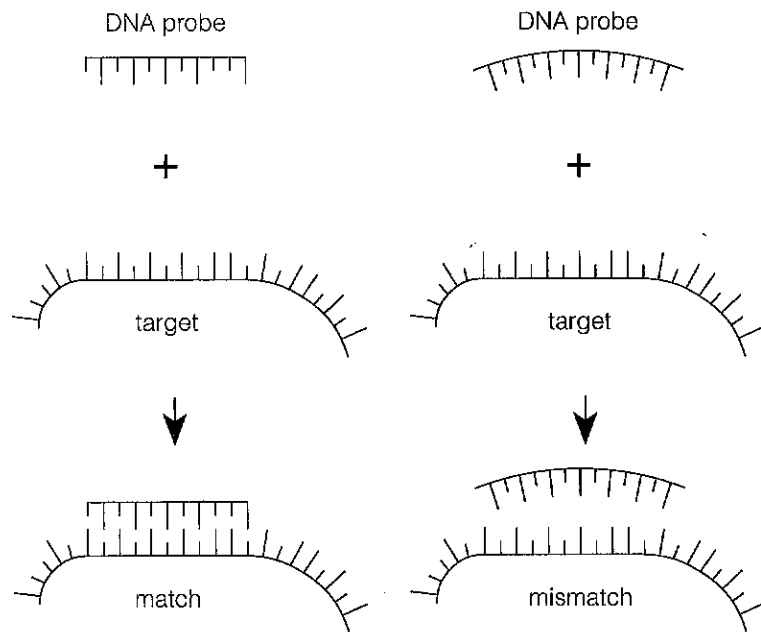


Figure 21 DNA diagnostics employs a stretch of known DNA sequence (a probe) to determine whether a gene from a patient's (target) DNA is complementary or not. If the probe is complementary to a normal gene, then a mismatch will indicate a mutation in the patient's DNA.

Human genetic mapping will permit the identification of specific predisposing genes and DNA diagnostics will facilitate their analysis in many different individuals.

To illustrate, Table 11, abstracted from a recent review in the *New England Journal of Medicine*, gives the factors that predispose one to cardiovascular disease, the leading killer in the United States today. These factors fall into two categories—modifiable and nonmodifiable. The majority of the nonmodifiable factors are genetic in origin. In the future, we will be able to identify the various genes that code for cardiovascular function (Table 12) and, through DNA diagnostics, to identify whether an individual possesses genes that predispose him or her to cardiovascular disease. A second example is the observation made by ourselves and others that two and possibly three different immune receptor genes—one on chromosome 6, one on chromosome 7, and one on chromosome 14—predispose certain humans to the autoimmune

Table 11 Risk factors for cardiovascular disease

Genetic predisposition	} Unmodifiable
Male	
Aging	
Elevated levels of low-density lipoprotein cholesterol	} Modifiable
Low levels of high-density lipoprotein cholesterol	
Smoking	
Arterial hypertension	
Physical inactivity	
Obesity	
Type II diabetes	

Table 12 Polygenic factors potentially contributing to cardiovascular disease

<u>Cells</u>	<u>Factors</u>
Endothelial	Variety of growth factors and chemoattractants
Platelets	
Monocytes/macrophages	
Vascular smooth muscle	
Fibroblasts	
<u>Genetic differences</u>	
	<ul style="list-style-type: none"> • Production of growth factors or chemoattractants • Responses to these factors • Host of paracrine and autocrine factors • Production of thromboxane in platelets or prostacyclin in endothelial cells

disease multiple sclerosis. Moreover, therapeutic approaches will be designed to circumvent the limitations of these defective genes. Approaches to circumvention might include new techniques in molecular pharmacology, special manipulations of the immune system (immunotherapy), appropriate avoidance or manipulation of environmental factors such as smoking, and, in the future, genetic engineering to replace defective genes in certain tissues with good genes.

The diagnosis of disease-predisposing genes will alter the basic practice of medicine in the twenty-first century. Perhaps in twenty years it will be possible to take DNA from newborns and analyze

fifty or more genes for the allelic forms that can predispose the infant to many common diseases—cardiovascular, cancer, autoimmune, or metabolic. For each defective gene there will be therapeutic regimens that will circumvent the limitations of the defective gene. Thus medicine will move from a reactive mode (curing patients already sick) to a preventive mode (keeping people well). Preventive medicine should enable most individuals to live a normal, healthy, and intellectually alert life without disease.

It has been estimated that the identification of the gene for cystic fibrosis cost approximately \$150 million dollars. If the genetic and sequence maps of the human genome were known, it would be possible to identify a particular disease or predisposing genes for perhaps \$200,000. In the future, we will use our detailed genetic maps to localize a particular disease or predisposing gene to a specific chromosome; indeed, we will find it within a two-centimorgan region within that chromosome. Then we will use the sequence information for this smaller region to identify the specific sequence responsible for the particular disease gene. Thus, the identification of disease genes will become a simple, straightforward, and inexpensive process.

Once the human and mouse genomes have been determined, we will be able to model human gene defects in the mouse. Techniques are now being developed whereby genes can be precisely placed in their appropriate location in the chromosomes of embryonic stem cells, and these cells in turn can develop into mice. Accordingly, once the mutation for Huntington's disease has been identified, the corresponding gene defect could be created in the homologous mouse gene. The mouse would be a model for studying means of circumventing the disease, at least until genetic engineering might be able to correct the consequences of this tragic mutation. In this manner, we will be able to model a variety of different human diseases in mice to determine appropriate therapeutic approaches.

Once the 100,000 human genes have been identified, they will be used as therapeutic reagents for dealing with all aspects of human disease. If we can use molecular area codes to identify all of the genes expressed in a particular cell, such as the lymphocyte, then we can begin to model, to experiment, and hence to understand in some detail the interactions of genes that generate this unique cell phenotype. These studies are beyond the genome pro-

gram, but as with protein folding, the identification of all human genes will provide key insights for subsequent analyses. Likewise, if one can query the computer for "heart" and obtain a listing of the genes that are expressed in the heart, so one can begin to model, experiment, and understand in detail the physiology of this organ and its pathology as well. Likewise, the genome project should have a major impact on our understanding of the brain. Our ability to understand how networks of neurons interact with one other and transmit information may be facilitated by an understanding of the most basic building-block components of these networks, the genes that specify the proteins that are active in the brain. Once we comprehend the normal physiology of various organs and organ systems, then we can begin to comprehend in detail the consequences of subtle disease pathologies and design appropriate therapeutic responses.

The benefits of the human genome project to industry are likely to be enormous, both through the information available from sequence and genetic maps and from the development of new techniques and instrumentation. Knowledge of the 100,000 human genes will provide a vast therapeutic repertoire with which the pharmaceutical industry can attack fundamental aspects of human disease. The spectacular success of erythropoietin (EPO), a hormone that promotes the development of red blood cells, and granulocyte-colony stimulating factor (G-CSF), a hormone that promotes the development of white blood cells to fight infections, is evident in therapies for chronic anemia and cancer, respectively. In the future, we can expect to have literally hundreds, if not thousands, of additional proteins that will facilitate the development of therapeutic approaches to a variety of different diseases.

DNA diagnostics and the identification of genes that cause and predispose to disease will place enormous pressure on the pharmaceutical industry to come up with appropriate therapeutic strategies. The gap between the ability to diagnose and the ability to treat genetic diseases could well be five to twenty or more years.

One striking new approach to the control of gene expression is the use of anti-sense nucleic acids. This entails the use of nucleic acid probes that can bind to nuclear RNA and block its processing or exit from the nucleus or that can bind directly to the gene to

prevent its transcription into RNA. These approaches are in the earliest stages of exploration, but if they are successful anti-sense therapy will be strikingly specific, in that it will enable the regulation of specific genes to be precisely controlled. These approaches may have important implications for many major human diseases including cancer, cardiovascular disease, immune diseases such as allergies, and autoimmune diseases. Clearly the delineation of the 100,000 human genes will provide vital DNA sequence information for the anti-sense strategies.

As the protein-folding problem is solved, exciting new possibilities for therapy will arise. It will be possible to design new therapeutic proteins of virtually any desired shape. For example, the genes in tumor cells often may express unique tumor-specific molecules, or antigens (Figure 22). Once a particular tumor antigen (or gene) has been sequenced, its three-dimensional structure can be deduced. A recognition unit can then be designed that is complementary to the tumor antigen and that has a killing domain attached to the recognition unit, which will function to destroy any tumor cell when the recognition unit attaches. In this manner, individually specific therapeutic reagents can be designed for many different tumors. If it is to succeed, this strategy requires the identification of unique or highly specific tumor antigens, an objective that should be reached within the time required to solve the protein-folding problem, which will likely come within the next fifteen to twenty years. The ultimate molecular engineering objective for the pharmaceutical industry is to design small organic molecules that have long half-lives and that may be taken orally as a replacement for protein therapeutic reagents. What the genome project will provide is 100,000 three-dimensional shapes (proteins) that execute the functions of life; these shapes can be used to engineer appropriate small molecules with diverse therapeutic potentials.

New industrial opportunities will arise from DNA diagnostics. They will encompass those aspects of medicine previously discussed and many additional applications. DNA fingerprinting may be used to identify personnel in the armed services. Applied to animals, DNA diagnostics will unequivocally delineate the parentage of prize dairy cattle or race horses. Genetic maps will be created for the major crop plants and be employed to identify and

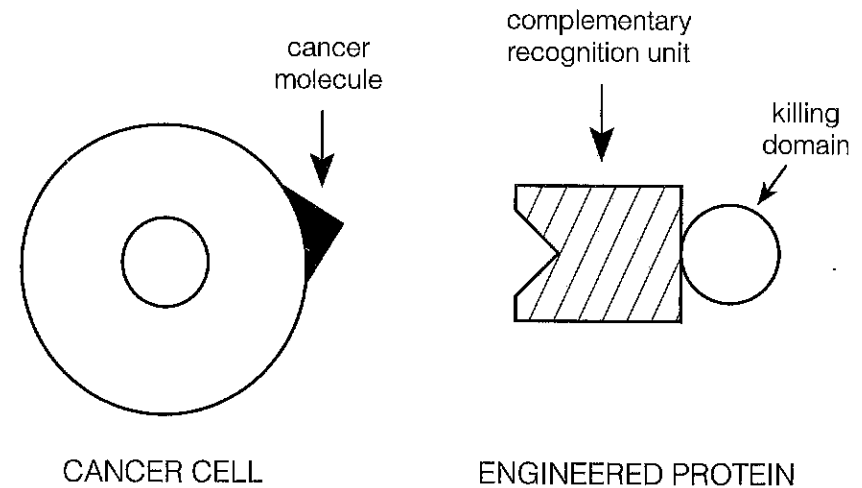


Figure 22 A solution to the protein-folding problem will offer great therapeutic benefits. Once the DNA sequence that codes for a cancer molecule is known, for example, the three-dimensional structure of the molecule (shown here as a triangle) may be discovered. Then it will be possible to engineer a protein that will attach to the molecule (obviously a round shape would not work in this example) and destroy it.

ultimately to engineer for desirable polygenic traits, such as a higher protein content or better taste.

By spawning the development of new technologies and instrumentation, the genome project will obviously create opportunities for companies that now produce biological instrumentation. For example, chemical and biological robots will be needed for routine tasks such as cloning, mapping, or sequencing. Opportunities will arise for companies to offer commercially many services that are currently provided primarily by molecular biologists. These include genetic mapping, DNA sequencing, cloning, and gene transfer to cells or organisms, to name only a few.

There will be striking future industrial opportunities in biocomputing. New software will be needed for the signal processing and image analysis associated with a wide variety of analytic and preparative instruments: DNA sequencers, chemical and biological robots, DNA mappers, mass spectrometers, NMR machines, X-ray crystallography, and so on. The combinatorial problems of biology—string matching, for example—will require the development of new algorithms, the development of new hardware such

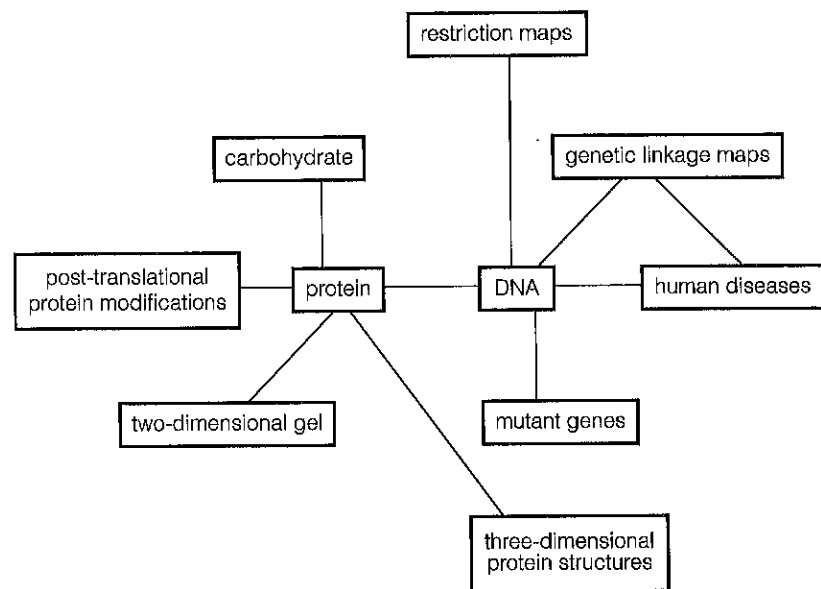


Figure 23 A model of some of the biological data bases and network connections that will be needed in the future.

as specialized coprocessors, and, increasingly, the use of parallel computers. In the future, there may be more than one hundred distinct biological data bases (Figure 23). It will be an enormous challenge to maintain these data bases as well as to make all of them readily accessible to the biologist or physician user. The development of new object-oriented data bases, which can organize information in keeping with its functional attributes, provide interesting new possibilities for instantaneous accessibility. It is also clear in the future that biologists are going to be quite dependent on the computer modeling of complex systems and networks to create new hypotheses that can then be tested in biological systems or living organisms. The opportunities in biological computer modeling will be enormous.

The United States is now the undisputed world leader in biotechnology. The genome project will help to ensure that we retain this lead. An important question is, To what extent can U.S. industry take advantage of this leadership? Without a national commitment to the support of long-term research endeavors and of their possible commercial opportunities, the outlook is uncertain.

The human genome project is unique in several regards. Since it is one of the first major biological initiatives that has technology development as a centerpiece, the need is tremendous for an interdisciplinary attack on challenging mapping, sequencing, and informatics problems. These problems will require the application of leading-edge techniques and instrumentation from applied mathematics, applied physics, chemistry, computer science, and biology. Moreover, the genome project, if successfully executed, will enormously enrich the infrastructure of biology by providing biologists and physicists with computer access to genetic, physical, and sequencing maps. For example, identifying the molecular addresses encoded in the regulatory elements of human genes will make available powerful data for attacking fundamental problems in developmental biology. Likewise, identifying the lexicon of perhaps 100 to 500 protein motifs may lead to valuable insights for attacking the protein-folding problem. Neither developmental biology nor protein folding is a problem inherent to the genome project; rather, the genome project will provide new tools for attacking these fundamental problems in other areas of biology. This infrastructure will fundamentally alter the practice of biology and medicine as we move into the twenty-first century. It will also secure the leadership of the United States in biotechnology and present U.S. industry with a wealth of opportunities.

To some, much of this discussion may appear fanciful science fiction. Yet the pace of biological discovery and technological advances continues to accelerate. This is truly the golden age of biology. Twenty years ago few of us could have imagined where we would be today. I suspect that, if anything, I have greatly underestimated the magnitude of the changes that will come, in part, as a consequence of the human genome project. I believe that we will learn more about human development and pathology in the next twenty-five years than we have in the past two thousand.