

The Copyright Law of the United States (17 U.S.C. 101 et seq. as amended) limits the use of this material to **Instructional Use**.

Author _____

Title _____

Publisher _____

Course _____

Instructor _____

THE CODE OF CODES

Scientific and Social Issues in
the Human Genome Project

EDITED BY

DANIEL J. KEVLES
AND LEROY HOOD

HARVARD UNIVERSITY PRESS

Cambridge, Massachusetts

London, England

1992

WALTER GILBERT

A Vision of the Grail

3

The genome project is not just an isolated effort on the part of molecular biologists. It is a natural development of the current themes of biology as a whole. In the simplest sense, the idea of determining the sequence of the human genome is an attempt to define all of the genes that make up a human being. The information carried on the DNA, that genetic information passed down from our parents, is the most fundamental property of the body. To work out our DNA sequence is to achieve a historic step forward in knowledge. Even after we have made that step, we will still need to refer back to the sequence, to try to unravel its secrets more and more completely. But there is no more basic or more fundamental information that could be available.

The DNA sequence has a simple numerical expression: it is composed of three billion base pairs. That is enough information to code for about 100,000 to 300,000 genes, each gene being a region of DNA that can specify a protein or some other structure that carries out a function in the organism. Nobody knows how many genes are really involved, because we do not know the average size of a gene in the human body. Our estimate of 100,000 assumes that there are about 30,000 base pairs per gene, which is a reasonably good guess. But many genes are only 10,000 base pairs long, so perhaps there are as many as 300,000. Many of the most interesting of those genes have multiple RNA splicing

patterns; that is, the messenger RNAs transcribed from a single gene may splice together different parts of the DNA sequence of the gene. The functions of these patterns must be understood in order to study an individual human gene. So saying that a human is made up of 100,000 genes underestimates the complexity of the human being, because many of those genes may encode ten or twenty different functions in different tissues.

The three billion base pairs of the human genome contain the amount of information included in a thousand thousand-page telephone books. What we hope to learn by working out the sequence of these base pairs is a list of all the genes that make up a human being. This information raises three striking questions about the nature of humans. The first is the question asked by developmental biology: how does a human being develop from an egg? The best way of studying the beginnings of animal development is to study model systems, such as the worm or the fruit fly; modeling of this sort is a dramatic and central part of modern biology. The second question is: what actually specifies the human organism? What makes us human? This is what medical science is about—the *specific* ways in which we are different from animals. The third question one might ask is: how do we differ from one another? This is the question of population biology—the *variation* of humankind across the species. These three questions are posed in order of increasing complexity.

The human genome project answers the second question—not the third. It is directed toward a molecular biologist's view of a species rather than a population biologist's view. The latter views a species as the envelope of all possible variants that can breed together; the importance of that envelope is that different aspects of a species population will be drawn forth if you change the environment. Molecular biologists generally view the species as a single entity, sharply defined by a set of genes and a set of functions that makes up that entity. Both of these ways of viewing humans are correct and not wholly antithetical. One emphasizes the variation that evolution works upon, while the other emphasizes the essential underlying features that define a species. The population geneticist, or the classical biologist, in defining the species, can point to a type specimen, an organism, and say that it exemplifies the species. The molecular biologist's view is that this organism is defined by its DNA. That DNA molecule can be

sequenced to reveal the essential information that defines the type organism and hence the species.

Can we understand all the genes that make up a human being? Could we understand all their interactions and their differences across our species? These questions concern the totality of biology, and they are far beyond the human genome project. The human genome project cannot answer all of those questions, but it develops the human sequence as a research tool. The genome project also will determine the sequences of the genomes of simple model organisms. Together, the human genome sequence and those of the model organisms will be powerful tools for all biologists to use to approach a variety of fundamental questions.

The problem of working out the human genome can be broken up into three phases requiring inputs that differ by orders of magnitude. First the DNA itself—two meters in length—must be broken into ordered smaller fragments, a process called *physical mapping*. The best estimates of how long the mapping process should take are on the order of a hundred person-years. The second phase—actually determining the sequence of all the base pairs of all the chromosomes—will take three thousand to ten thousand person-years. The third phase—understanding all the genes—will be the problem of biology throughout the next century: about a million person-years, a hundred years of research for the world.

In addition to determining base-pair sequences, two types of maps of the genome will be made by the project (see Figure 10). Genetic maps will trace the inheritance of DNA regions in human populations and connect specific DNA regions to disease. Physical maps will provide DNA research material. The first effort to produce a complete genetic map of the human chromosomes was published a few years ago. It consists of approximately 150 polymorphic markers distributed along each of the human chromosomes; the markers are separated by about 20 megabases (20 million base pairs), about 20 centimorgans apart. As this map becomes finer in resolution over the next few years, more and more diseases will be pinpointed on it.

Physical maps also come in two types. One is created by measuring the distances along each chromosome in terms of the sequences at which restriction enzymes cut. That provides an abstract distance map for the size of the chromosomes and some points within them. The second type of physical map is called a

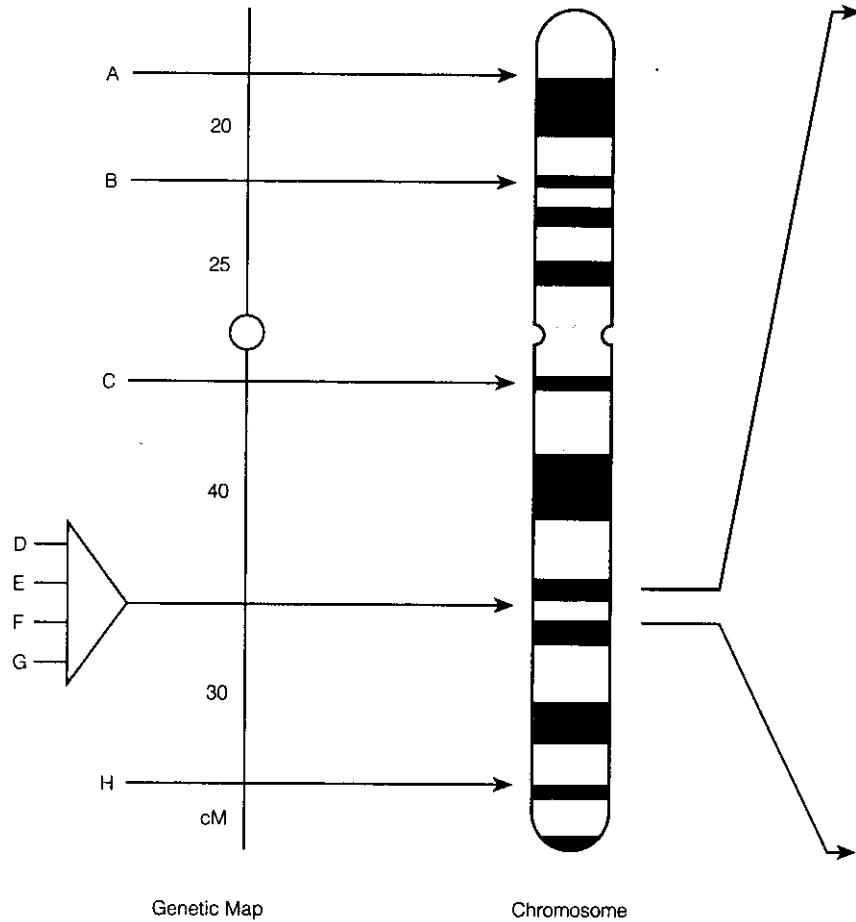
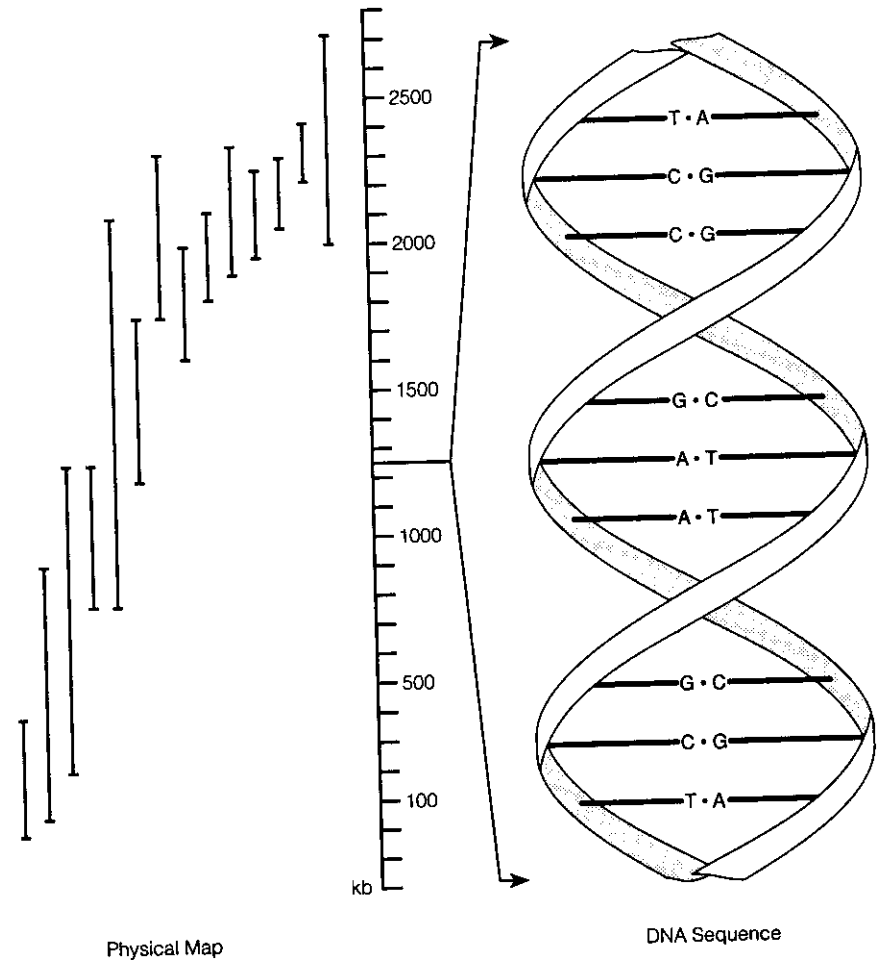


Figure 10 The ultimate goal of the human genome project is a determination of the sequence of base pairs that make up every human chromosome. The *sequence* will be the most detailed map of the human genome. Less detailed maps have been constructed since the beginning of this century; the maps being made today are essential not only to the sequencing program but also to other research goals—such as finding the genes responsible for hereditary diseases like Huntington’s chorea.

Genetic maps are representations of disease traits, physiological traits, or random polymorphisms that can be assigned to particular chromosomes and mapped relative to one another by following the transfer of alternative forms of these traits in families. This model of a genetic map shows the location of 8 markers (called here A–H), such



as genes or polymorphisms, along the chromosome. The triangle maps a small part of the chromosome in greater detail, as may be needed for an intensively studied part of a chromosome (perhaps a region suspected of being the location of a disease-causing gene). Distances in genetic maps are measured in centimorgans (cM, about 1 million base pairs).

Physical maps are not representations but overlapping collections of DNA fragments. DNA is snipped into fragments by the action of restriction enzymes, and then cloned and stored in a variety of forms—such as cosmids in bacteria or YACs in yeast. These tiny fragments (measured in kilobases, kb) may then be analyzed by various means to discover the base-by-base sequence of DNA.

cosmid map—it consists of DNA pieces each about fifty thousand bases in length, each cloned in a separate bacterial strain, and each overlapping other identified cosmids on either side. This map is actually a physical collection of bacteria, about a hundred thousand in number, containing the clones, known as cosmids, that span the entire genome. It is roughly equivalent to having the physical material in one's hand for each of the hundred thousand different human genes. Mapping by this method involves "fingerprinting" DNA pieces—recognizing sequence features that demonstrate that two cosmid DNAs share DNA sequences (or fail to) and hence overlap one another (or fail to). Assembled in a known overlap pattern, the cosmids provide the physical material to study the genome further.

The genetic map can be correlated with the cosmid map, because the genetic (or polymorphism) map specifies DNA regions of a known genetic distance apart that are detectable through DNA hybridization. Comparing the genetic and physical maps makes it possible to detect immediately whether any particular fragment or polymorphism lies on a particular cosmid. The correlation between the cosmid map and the genetic map provides a physical structure that makes it possible, if a polymorphism is near some disease gene, to determine where that disease gene lies on the cosmid map.

The simplest approach to sequencing the whole genome is to begin with a map—a cosmid map of the human being—and to sequence each of the cosmids separately. So, one strategy is to take a chromosome, which might be represented by a thousand cosmids, and simply sequence one cosmid after another to determine the entire structure of the chromosome. In the whole human sequence, the first 1 percent of the sequence—that is, the first thirty megabases (thirty million bases)—will probably be determined by sequencing regions near genes that are biologically or medically interesting. The next 10 percent (three hundred megabases) to be discovered will be the sequences of individual small chromosomes. The last 90 percent (2,700 megabases) will be everything else. As a result of further development of sequencing technology, these three tasks probably will each take a roughly equal amount of time.

Before 1976 it was essentially impossible to sequence DNA. When Allan M. Maxam and I worked out one of the first DNA

regions in 1971, it took us two years to do twenty base pairs. A throughput of this magnitude is impractical in terms of working out a whole gene. In 1976 Fred Sanger, in England, and Maxam and I discovered two rapid ways of analyzing DNA sequence, which made it possible for a single individual to decipher about five thousand base pairs a year—equivalent to the structure of a small gene. Fifteen years later that rate is somewhere between ten thousand and a hundred thousand base pairs of DNA a year, making it relatively easy to work out the structure of single genes.

Most of the time is spent not in sequencing the gene but in preparing DNA fragments suitable for sequencing. Currently, these procedures involve converting the genome into smaller fragments of DNA that are then cloned into suitable recombinant DNA vectors. Typical clones contain inserts of DNA ranging in size from 15,000 base pairs to 50,000 base pairs. These must be broken down, or subcloned, into smaller DNA fragments about 300–1,000 base pairs long, fragments that are suitable for DNA sequencing. How much work is done in the DNA sequencing depends on the strategy. Two alternative strategies can be employed—ordered and shotgun. In the ordered strategy, the DNA is sequenced in a linear and consecutive manner. In the shotgun strategy, a large piece of DNA is randomly sheared into smaller fragments, the smaller fragments are randomly sequenced, and then these small strings of sequence are assembled by computers into the final sequence. The shotgun process requires that each stretch of DNA be sequenced on average five or six times. If the sequencing process was focused and the preparation of the clones and the DNA made simple and routine, probably even today sequencing speeds of the order of a million bases per person-year would be possible. By directed effort, speeds of about 10 million bases per person-year could ultimately be achieved.

In dollar terms, the cost to sequence DNA at a million bases per year per person would run about ten cents per base. At that rate, a working group of about 300 people would take ten years to do the entire genome. This would be using the best of today's technology or the technology that is immediately on the threshold. The rate of DNA sequencing has gone from virtually zero ten years ago to about 20 million bases of DNA sequence in 1990. In December 1990, 50 million bases of DNA were collected in data

bases. The world rate of DNA sequencing is increasing very rapidly—accelerating at 60 percent per year. However, if one were to continue sequencing only at the current rate with no large, directed efforts, it would take several hundred years to sequence the human genome; only a few percent of all the genes have been sequenced at this point.

A million bases per year amounts to about five thousand base pairs per day. Two techniques can sequence at this rate. There are machines that can produce about ten thousand bases of rough sequence per day. A process called genomic sequencing can identify about thirty thousand base pairs per day of rough sequence. Because they both take a shotgun approach, these processes produce only about one-fifth as much final sequence.

The human genome project can be viewed as a purely technological effort to obtain the DNA sequence, put it into a computer data base, and study it. Today we know a variety of techniques for analyzing that DNA sequence. In fact, if one is presented with an arbitrary DNA sequence, there are techniques that will, broadly speaking, identify the gene structure with approximately 90 percent accuracy. What does that mean to biology? For example, if we were given a sequence data base of the human genome today, could we understand anything from it? The answer, I think, is yes; we could understand a tremendous amount. Today we learn a great deal about the functioning of genes by looking at the sequence of proteins they produce. For example, there is a set of about a hundred genes called oncogenes: each one identified as a DNA fragment, isolated from a tumor line or tumor cell, that will endow a normal cell with the ability to grow indefinitely. The normal form of each of these genes, a proto-oncogene, influences some aspect of cell growth. At first these genes were just a random collection of names with only an ability to control cell growth in common. But by looking at the protein sequences predicted from the DNA sequences, a lot of information about their functions can be inferred.

For example, we can recognize, by sequence comparison, that one oncogene is related to a receptor for a hormone. We can see that other oncogene products are slightly altered growth factors. We can recognize that some of the oncogene products bind to DNA and probably influence genes by determining the way the DNA is transcribed. Thus oncogenes may be divided into distinct

categories by sequence, from which we infer their function. All of these insights, which at one level suggest experiments and biological relationships, appear immediately from sequence comparisons.

Another example is a characteristic amino-acid sequence motif called a "zinc finger," which was first recognized in a protein regulating the transcription of a particular gene. This motif is a string of amino acids with two cysteines and two histidines in a specific relationship that enables them to bind zinc. Scientists studying this short sequence of amino acids recognized that it denotes the ability of proteins containing this motif to bind to RNA or DNA. Then biologists realized that this sequence appears in a number of transcription factors. Many genes that control pattern formation in *Drosophila* specify proteins that have this same amino-acid motif, and one could therefore infer that these gene products function by producing proteins that bind to DNA. When a novel gene is sequenced, the presence of the amino-acid motif predicts that this new, and otherwise uncharacterized, gene product is a transcription factor. So, we *can* learn a great deal just by looking at sequence of genes or proteins.

Many of the genes in our bodies are members of large gene families. The ability to recognize gene-family relationships comes from an analysis of the gene sequences. There is now about a 50 percent probability that when we isolate a new gene we will see that it is related to something that has been previously identified. As the human genome project progresses, cross-relationships among gene products are going to become more evident, opening up the possibility of postulating functions for new genes and then proposing biological experiments to test these ideas.

We would like to know the three-dimensional structure of these gene products, the proteins, but that is not a problem for the human genome project. The structure-function problem is a very well defined problem in biology, and it is the crucial problem that underlies our understanding of proteins. We can go from the DNA sequence to the amino acid sequence by computation, but can we go from an amino-acid sequence to a three-dimensional structure and a function? So far the answer is no. But this is a well-posed theoretical problem with two approaches for its solution. One is to try to devise better computer programs to fold amino-acid sequences by energy calculations that seek stable con-

figurations (the lowest-free-energy forms). With this approach protein structure could be obtained from first principles. The second approach is simply to have enough three-dimensional structures from known proteins to be able to recognize in a novel protein the small motifs that serve as building blocks and then to predict its structure as a melange of motifs whose structures are known. This very powerful approach is just now beginning to work and will probably lead to the solution of the protein-folding problem in the next five years.

If that happens, and if we also have a data base of DNA sequence, we can expect that we will be able to predict not only the protein sequences encoded by the genes but also the three-dimensional structures of the proteins. It is here that a theoretical biology will emerge. It will be a science of pattern recognition—extracting from the genome sequence the identity of human genes, their interrelationships, and their control elements. This information will be used to predict how the genes and their proteins function. A scientist will then use laboratory procedures to test these conjectures. Thus the future data base will enable us to approach human biology in an entirely new manner. Today we cannot identify which genes are expressed in the brain or in the heart. We know that the body has mechanisms to express one set of genes in the heart and a second set in the brain. These mechanisms define molecular addresses, the control elements on the DNA that direct the tissue-specific expression of the genes. In the future we will be able to use these molecular addresses to classify genes and organs. One can immediately think of global questions that can be asked if one has an appreciable fraction of the total information on human genes. But one cannot ask these questions today. Our present technology only allows us to go after genes one at a time, and to work out the relationships of the new genes to previously discovered genes.

The genome project is an application of scientific technology to produce a certain end—the information content of the genome. Science is going to change quite drastically over the next ten years in ways that we have not even begun to realize. A major change is occurring in the relationship of molecular biology to the rest of biology. Over the last decade it has become clear that molecular techniques are a powerful way of studying almost any question in biology, ranging from questions of development to those of

evolution and population biology. All sorts of questions are now being studied by finding a gene and seeing what it does to the organism, or by deducing a feature about the pattern of inheritance.

As a successful science, molecular biology has become a set of cookbook techniques. Its very success is producing an odd sort of reaction: all of these wonderful techniques can simply be looked up in a handbook, and biologists seem to spend their time reading techniques and then cloning genes, or reading techniques and then sequencing DNA. Where is the biology? We are witnessing the last stage of the development of a technology and, as has happened a number of times before, many of the techniques of molecular biology will very soon leave the research laboratories entirely. We will purchase them externally as services; they will not be performed by research scientists.

Thirty or forty years ago students were taught how to blow glassware because they were expected to make their own apparatus. Today we buy plastics and throw them away, and we would not expect a student to know how to make a condenser out of glass. About fifteen to twenty years ago, when restriction enzymes had just been discovered, every graduate student in my laboratory made restriction enzymes. We wanted to work with DNA, we had to have those proteins, so every student made one or more restriction enzymes, knew how to purify a protein, and had to maintain the supply of associated materials. Now we buy the enzymes and even the bacterial strains needed for making clone libraries. Research scientists no longer have to grow bacteria themselves in order to generate competent hosts. Now they are beginning to buy ready-made clone libraries; soon they will supply a probe and order a specific clone from a commercial source. Over this next decade, DNA sequencing and many recombinant DNA techniques will move out of the laboratories. Five years from now, instead of struggling with cloning particular genes, we will simply buy the clone or the sequence. DNA sequencing will become centralized in very large-scale service organizations that will sequence DNA on demand. The science will have moved on to the problem of what a sequence *means*, what the gene actually does. Biologists who recognize and adapt to these changes will find that biology will continue to be active and stimulating, although it will be different.

Sometime in the 1990s, the world sequencing rate will reach about a billion bases a year, and we will have completed the human sequence as well as a variety of model sequences. At the end of the genome project, we will want to be able to identify all the genes that make up a human being. For example, we will compare the sequences of the human and the mouse and be able to determine the genes that define a mammal by this comparison, because the regions of DNA that code for protein are very well conserved over evolutionary time whereas the regions that do not have important functions are not well conserved. So by comparing a human to a primate, we will be able to identify the genes that encode the features of primates and distinguish them from other mammals. Then, by tweaking our computer programs, we will finally identify the regions of DNA that differ between the primate and the human—and understand those genes that make us uniquely human.

The possession of a genetic map and the DNA sequence of a human being will transform medicine. One immediate change, which will emerge over the next decade, will be a knowledge of genes that cause rare genetic disease. More important, however, will be the identification of genes for common diseases. When we have a detailed genetic map, we will be able to identify whole sets of genes that influence general aspects of how the body grows or how the body fails to function. We will find sets of genes for such conditions as heart disease, susceptibility to cancer, or high blood pressure. Along with many other common afflictions, these will turn out to have multiple genetic origins in populations, as will such mental conditions as schizophrenia, manic-depressive illness, and susceptibility to Alzheimer's disease. A whole variety of human susceptibilities will be recognized as having genetic origins.

One of the benefits of genetic mapping will be the ability to develop a medicine tailored to the individual: drugs without side effects. The side effects of drugs are often due to truly different responses of an individual to the chemical agent, because the variation between individuals is great enough to produce a different biochemistry. For example, there is a recessive gene in European populations that controls a sensitivity to medications for high blood pressure. The 5 percent of the population that shows this gene can use blood-pressure medications only in amounts

one-hundredth or so of the normal dose. Genetic typing for such differences will produce novel medicines suited to specific patients.

DNA-based tests for all diseases, including the neurological ones, will result in subdivisions of those diseases into many categories, which will require different treatment programs. Consider badly defined entities such as mental diseases—for instance, schizophrenia. We will probably be able to identify a set of genes that lead to similar mental states. The ability to test for these genes will mean both a sharper diagnosis of the condition and a sharper prognosis of what happens to people who have it. Our knowledge of specific genes with brain-related functions—for example, a dozen genes that specify receptors for a single neurotransmitter—will affect treatment as well. The ability to isolate the multiplicity of receptors for a given neurotransmitter means that we can try to discover specific drugs that identify and affect in turn each of those receptors separately. These drugs will target specific receptors on specific cells. In general, knowledge of common features will lead to replacement medicines that will supply natural components of the body to enhance a natural function of the body.

Gene typing and genetic mapping could also have very strong social effects. However, the problems posed by the knowledge are not insurmountable and can be dealt with in a democratic society. First, we are going to have the ability to recognize defective genes in the embryo by simple techniques. That will mean a constant improvement and extension of prenatal diagnosis, which will lead to the elimination of much human misery. But better and more intrusive prenatal diagnosis will exacerbate the abortion discussion, with which society is wrestling at the moment.

What about genes in the workplace? What will happen if individuals who are susceptible to toxic chemicals can be identified as well as individuals who are resistant? Should society permit or resist genetic analysis of workers for environmental or work-associated susceptibilities? Medical insurance is already creating problems in this regard. We have examples of insurance companies arguing that a congenital illness is a previous condition and is therefore not covered by medical insurance. Should we permit that use of genetic analysis? Both of these examples suggest that our society adopt attitudes or pass laws that preserve the privacy of the individual. But will society do that? The problem of testing

is already with us today in the matter of infection with the AIDS virus, HIV. Should one test for it? And what, if anything, should one do with that knowledge? These questions will be raised again and again as we gain more and deeper genetic knowledge.

Racism is a danger. Will the ability to analyze the genetic structure of individuals be used to try to define improved individuals and thus fan the flames of racism? Or will society recognize in a healthy fashion the worth of each individual? All human beings share an underlying DNA structure necessary for functioning. One of the ways of ensuring that that structure is not misconceived is implicit in the way in which the human genome project will be done. As we understand human DNA by sequencing the chromosomes of different individuals from around the world, we will put together a sequence that represents an amalgam of the underlying human structure. It will reflect our common humanity.

I think there will also be a change in our philosophical understanding of ourselves. Even though the human sequence is as long as a thousand thousand-page telephone books, which sounds like a great deal of information, in computer terms it is actually very little. Three billion bases of sequence can be put on a single compact disk (CD), and one will be able to pull a CD out of one's pocket and say, "Here is a human being; it's me!" But this will be difficult for humans. Not only do we look upon the human race as having tremendous variation; we look upon ourselves as having an infinite potential. To recognize that we are determined, in a certain sense, by a finite collection of information that is knowable will change our view of ourselves. It is the closing of an intellectual frontier, with which we will have to come to terms.

Over the next ten years, as a consequence of the advance of our biological knowledge, we will arrive at new understandings. We will understand *deeply* how we are assembled, dictated by our genetic information. Part of that understanding is, of course, to realize that genetic information does *not* dictate everything about us. We are not slaves of that information. We must see beyond a first reaction that we are the consequences of our genes; that we are guilty of a crime because our genes made us do it; or that we are noble because our genes made us so. This shallow genetic determinism is unwise and untrue. But society will have to wrestle with the question of how much of our makeup is dictated by the

environment, how much is dictated by our genetics, and how much is dictated by our own will and determination.

One consequence of the human genome project is that we will see more and more clearly how connected all life really is. Research in early development tells us that the genes that form our bodies are similar to the genes that produce worms and fruit flies and every complicated organism. Those genes were created before the branching off of any of the higher organisms that are on the earth today. The data base of the human genome, coupled with our knowledge of the genetic makeup of model organisms, promises to reveal patterns of genes and to show us how we ourselves are embedded in the sweep of evolution that created our world.