REPORT ON THE HUMAN GENOME INITIATIVE for the OFFICE OF HEALTH AND ENVIRONMENTAL RESEARCH

Prepared by the
Subcommittee on Human Genome
of the
Health and Environmental Research Advisory Committee
for the
U.S. Department of Energy
Office of Energy Research
Office of Health and Environmental Research

April 1987

Biomedical & Environmental Sciences Divisions
Lawrence Livermore National Laboratory

27 April 1987

Dr. Alvin W. Trivelpiece
Assistant Secretary
Office of Energy Research
U.S. Department of Energy
Washington, DC 20545

Dear Dr. Trivelpiece:

On behalf of the Health and Environmental Research Advisory Committee (HERAC), I am pleased to submit to you the enclosed *Report on the Human Genome Initiative*. This was prepared by a subcommittee under the chairmanship of Dr. Ignacio Tinoco, University of California, Berkeley, and is in response to a charge by you. It has been strongly endorsed by the parent committee.

The report urges DOE and the Nation to commit to a large, multi-year, multidisciplinary, technological undertaking to order and sequence the human genome. This effort will first require significant innovation in general capability to manipulate DNA, major new analytical methods for ordering and sequencing, theoretical developments in computer science and mathematical biology, and great expansions in our ability to store and manipulate the information and to interface it with other large and diverse genetic databases. The actual ordering and sequencing involves the coordinated processing of some 3 billion bases from a reference human genome.

Science is poised on the rudimentary edge of being able to read and understand human genes. A concerted, broadly based, scientific effort to provide new methods of sufficient power and scale should transform this activity from an inefficient one-gene-at-a-time, single laboratory effort into a coordinated, worldwide, comprehensive reading of "the book of man". The effort will be extraordinary in scope and magnitude, but so will be the benefit to biological understanding, new technology and the diagnosis and treatment of human disease.

It may seem audacious to ask DOE to spearhead such a biological revolution, but scientists of many persuasions on the subcommittee and on HERAC agree that DOE alone has the background, structure, and style necessary to coordinate this enormous, highly technical task. When done properly, the effort will be interagency and international in scope; but it must have strong central control, a base akin to the National Laboratories, and flexible ways to access a huge array of university and industrial partners. We believe this can and should be done, and that DOE is the one to do it.

Sincerely,

Mortimer L. Mendelsohn, M.D. Ph.D.
Chairman, Health and Environmental
Research Advisory Committee

Report on the Human Genome Initiative
Office of Health and Environmental Research

Prepared for Dr. Alvin W. Trivelpiece
Director, Office of Energy Research

by a Subcommittee of the
Health and Environmental Research Advisory Committee (HERAC)

Dr. Ignacio Tinoco, Jr. (Chairman)
University of California, Berkeley

Dr. George Cahill
Howard Hughes Medical Institute

Dr. Charles Cantor
College of Physicians and Surgeons
Columbia University

Dr. Thomas Caskey
Baylor College of Medicine

Dr. Renato Dulbecco
Salk Institute

Dr. Dean L. Engelhardt
Enzo Biochemicals, Inc.

Dr. Leroy Hood
California Institute of Technology

Dr. Leonard S. Lerman
Genetics Institute

Dr. Mortimer L. Mendelsohn
Lawrence Livermore National Laboratory

Dr. Robert L. Sinsheimer
University of California, Santa Cruz

Dr. Temple Smith
Dana/Farber Cancer Institute
Harvard University

Dr. Dieter Söll
Yale University

Dr. Gary Stormo
University of Colorado

Dr. Raymond L. White
University of Utah Medical Center

TABLE OF CONTENTS

REPORT ON THE HUMAN GENOME INITIATIVE

Advances in biology and medicine have reached the stage where it is now possible to acquire a thorough and very detailed understanding of human biology and inheritance at the molecular level. This understanding will require mapping and sequencing of DNA on a massive scale, a task which cannot be accomplished efficiently with current technologies.

Two major tools are needed:

a) The sequence of a reference human genome

b) Efficient methods for obtaining and interpreting the large amount of additional sequence data needed for a wide variety of biological and medical studies

Creation of these tools will require a broad interdisciplinary research effort that brings together technologies from the fields of biology, computing, materials science, instrumentation, robotics, physics and chemistry. This special focus on technological development is distinct from the current national effort in human biology and genetics and requires a new initiative.

The Department of Energy, through the Office of Health and Environmental Research, has a mission to understand the health effects of radiation and of other harmful by-products of energy production. The Department has long supported work on human mutations, DNA damage and DNA repair. Now it is clear that the ability to determine quickly and accurately the sequence of a DNA is the most rapid and cost-effective way to assess DNA damage, and to protect the public health. Thus, the Department of Energy is poised for this initiative because of its research support and interest in human genetics, and its experience in developing large scale, long-term interdisciplinary projects. Development of these new technologies will place the United States in a commanding position in the biotechnology of the 21st century.

RECOMMENDATIONS

1. DOE should fund a major new initiative whose goal is to provide the methods and tools which will lead to an understanding of the human genome. Funding should start in fiscal year 1989 at $40 million and increase over a five year period to reach a level of $200 million per year. Appendix A provides details.

2. The early goals (first 5 to 7 years) of this program should be to:

a) Make a physical map of the human genome. A physical map consists of a complete set of segments of the DNA, arranged in order.

b) Locate genes and other markers on the map.

c) Produce and distribute cloned DNA sequences and other materials needed for using and improving the physical map.

d) Develop new techniques and improve existing methods for large-scale DNA mapping and sequencing (including applications of automation and robotics).

e) Develop new methods for characterizing and locating genes; both computational and cloning techniques are needed.

f) Establish computer facilities, and develop computer data bases for the storage, retrieval and dissemination of cloning, mapping, and sequence information (including cross-references to other relevant data bases). Improve and invent algorithms for analyzing DNA sequences, including methods for identifying coding regions, predicting protein structures and functions, and identifying genetic regulatory sites.

3. The major long-term goal is to obtain a base sequence for each of 24 reference human chromosomes, and to make DNA sequencing technology readily available to search for disease-related variations and to make biological comparisons. The improvements in technology listed in Recommendation 2 are necessary to attain this goal.

4. Work on these goals should take place in the National Laboratories, in universities and in industry. Both prospective and retrospective peer review should be used. Cooperation and collaboration among all groups is essential; in particular, all new map and sequence information must be placed promptly in a designated data base. Clones and cell lines must be made available for distribution to other qualified investigators.

5. Two scientific panels should be established immediately. One would develop policy, define overall strategy, and provide continuing oversight. The other would provide scientific review of proposals and programs for their technical merit and feasibility. The initial phase of the program should consist primarily of technological development in the areas of construction of large scale maps, automation, sequencing and the determination and analysis of sequence data. Because of the highly creative nature of this beginning phase, it is essential that the effort be widely distributed. The project should involve single-investigator-initiated proposals as well as multidisciplinary consortia that bring together the development of instrumentation and software, as well as biotechnology.

6. DOE should encourage wide collaboration at the scientific and managerial levels for the human genome project. Cooperation is needed with other agencies within the U.S. and with other countries throughout the world. Results should be open and in the public domain, within the constraints of technology transfer and the promotion of industrial involvements. Information transfer should be emphasized among the cooperating scientists, the scientific community and the public at large.

REPORT

THE ULTIMATE GOAL OF THIS INITIATIVE IS TO UNDERSTAND THE HUMAN GENOME

Knowledge of the human genome is as necessary to the continuing progress of medicine and other health sciences as knowledge of human anatomy has been for the present state of medicine. The DNA of the human genome contains complete instructions for construction of each human being, but we know only the crudest features. We each have two sets of 23 chromosomes with a total of about three billion base pairs per set. Each set consists of 22 autosomes plus one sex chromosome; thus there are 24 distinct chromosomes -- one female (X), one male (Y) and 22 autosomes. The chromosomes contain an unknown number of genes with estimates which range from 20,000 to 200,000. Presently only about 500 of these genes have been cloned and characterized. Our knowledge is equivalent to that of 15th century anatomists who knew about the major bones and organs, but knew very little about their functions. The significance of most vital organs, including obvious ones such as the liver and pancreas, or small ones such as the pituitary and the adrenals, was completely unknown. Most important, even the simplest concerted functions of the body, such as provided by the circulatory system, were not mapped.

We are at the same early state of knowledge with respect to the human genome. We do not know within a factor of ten how many genes there are, nor the range of functions performed by the gene products. We have very limited knowledge of how the expression of genes is controlled. What sequences of the DNA turn genes on and off at the right time for correct development and differentiation? We do not understand how the coordinated control of genes is accomplished. We expect that vital elements that exist in the human genome have not even been imagined. The human genome has been called the book of man; it contains the instructions that describe each human. It is time to obtain a copy of the book to begin to understand what the text means.

It should also be clear that understanding the human genome is a very long-range task. Once the gross features of a human genome are mapped, it will be important to identify and localize all the genes. The control elements must be identified which determine when and where each gene is expressed, and thus program our development from a single cell to a complex structure. The study of single-gene defects in humans has already been extremely beneficial for the diagnosis and treatment of some diseases. Although genes may account for only ten percent of the human genome, complicated chromosomal changes and aberrations, which are not simply dependent on DNA sequences in genes, are also heavily implicated in genetic diseases. Thus, Down's syndrome is caused by an extra copy of chromosome 21, Cri-du-chat is caused by a deletion -- a loss of a segment -- in chromosome 5,

and many birth defects and congenital defects have a chromosomal basis. The part of the human genome whose function is not yet known or even imagined must be characterized and understood. Searching analysis will continue to be required to discern differences among human genomes that correlate with sickness and health.

Accomplishing these goals obviously requires sequencing a large fraction of the genome. However, some genomic regions, such as long stretches of repetitive DNA, may not need detailed sequencing. As the details of the genome unfold, it should be possible to set priorities and make rational decisions about what should and should not be done.

A. Concerted efforts in several different areas should be supported.

1) A first step is to map the human genome -- to arrange in order large segments of DNA (in size from 100 to 1,000 kilobases); there are 3,000 to 30,000 of these pieces. As a prerequisite to sequencing the human genome, it is necessary to have pure DNA fragments from known locations on the genome. These DNA fragments constitute an ordered clone bank. At present 30 to 50 kilobase fragments of DNA (cosmid clones) can be prepared routinely and partially ordered; these fragments are vital for current progress. However, methods for preparing and separating larger fragments are becoming available. Large DNA fragments can be formed with restriction enzymes or reagents specific for sequences which are eight or more base pairs long. They can be separated by new methods of electrophoresis. Unique identification of these DNA fragments can be obtained with probes or restriction enzymes; the fragments can be characterized by a complete set of restriction sites with known intervals. Practical methods to determine their order must be worked out.

As human map and sequence data accumulate, many investigators will be able to apply this knowledge to problems of medical and biological importance. They will need access to large numbers of biological samples including cloned DNA fragments and human cell lines. Methods for the efficient production and distribution of these materials need to be developed. Effective quality control for the identity and purity of the samples is essential.

2) Genes should be assigned to the fragments as each fragment is identified. There are standard methods available for locating genes whose gene products (a protein or nucleic acid) are known. These include genes whose defects are responsible for blood diseases such as certain hemophilias, alpha and beta thalassemias and sickle cell anemia. Genes for enzymes with known activities are particularly easy to find. There are essential enzymes whose absence causes death, but the deficiency of other enzymes may only lead to illness, or the predisposition to certain diseases. An enzyme deficiency genetic disease is phenylketonuria which causes mental retardation, but can be treated by removing excess phenylalanine from the diet. A defective anti-trypsin gene produces lungs very susceptible to injury and requires extra care with smoke or other lung irritants. When gene products are not known, as in many human diseases, the process is more difficult. Here the methods which have been successful for Huntington's disease, retinoblastoma, cystic fibrosis, and Duchenne muscular dystrophy can be used. A genetically linked marker for the disease must first be found; this is a sequence of DNA which is located near the disease gene and serves to track the inheritance of the gene. Many genes may only be found from analysis of the DNA sequence; identification will lead to the gene product and its function.

Genes for all the enzymes involved in metabolism, in biosynthesis and in repair need to be localized. Structural proteins, proteins of the immune response, transport proteins and the RNAs of protein synthesis are all important. The genes for hormones, which act in very small amounts, need to be identified and entered in the genome map. The largely unknown control proteins, which orchestrate differentiation, development and senescence, may be the most important to characterize. As more genes are identified and their gene products determined, the polygenic disorders like heart disease, hypertension, diabetes, schizophrenia, manic depression and even some symptoms of aging can be attacked. It will become possible to develop methods for early diagnosis and effective treatment.

There are currently many projects, sponsored primarily by the National Institutes of Health and the Howard Hughes Medical Institute, involved in locating genes on the human chromosomes. This research is extremely valuable and is primarily aimed toward medically important genes. The DOE initiative will facilitate rather than compete with those projects; it will provide a valuable resource for the projects. Even so, the success of all those other projects would locate only a few percent of the total number of human genes. The tools and methods developed through this project will greatly speed the finding and understanding of the total complement of human genes, a task far beyond the scope of any current research efforts.

3) Current methods for mapping and determining base sequences in DNA need to be increased in speed by orders of magnitude and radical new methods should be encouraged. Automation of current methods has begun. There is a Japanese national project which is trying to develop automated equipment based on current sequencing methods to determine sequences at the rate of 300 kilobases to one million bases per day. Even if the Japanese effort is successful, a thorough sequencing of a human genome will not be possible because methods for preparing, purifying and ordering DNA fragments are not available to provide the necessary fragments for sequencing. However, advances in technology are being developed which will allow complete and cost-effective sequencing. These new methods need to be automated to provide a reference sequence and also to provide the ability to make comparative studies both within the human population and between humans and other animals. Close collaboration between engineers and molecular biologists can provide efficient, reliable methods that use the capabilities of automated instrumentation to fullest advantage. It should be possible to reduce the total cost of sequence determination to one tenth, one hundredth, or less of the cost of current manual methods. Appendix A gives details.

4) Computer facilities to organize, disseminate and interpret the sequence of the human genome must be supported. At present there are several organizations which act as repositories for sequence data and human gene information (such as Genbank at Los Alamos, the National Library of Medicine, the Yale human gene map, the European Molecular Biology Organization, the Japanese National Institute of Genetics).

Easy cross-reference and cross-access between data bases must be assured. Algorithms and programs to interpret the data are in very early stages of development. We need programs to identify accurately DNA sequences corresponding to genes and their control. At present we cannot identify unambiguously the signals to start messenger RNA synthesis, to start protein synthesis, to remove introns and thus to provide a protein sequence. When a DNA sequence predicts a protein sequence, we need to be able to predict the protein shape, its function and possible cellular or extracellular sites for its location. Algorithms to identify the control elements for expression and regulation of the genes (enhancers, repressors, etc.) are needed. DNA sequences involved in chromosome organization, recognition and regulation must be understood. Appendix B provides further details.

B. DOE can and should organize and administer this initiative.

The Department of Energy, extending back to its predecessor the Atomic Energy Commission, has successfully managed many long-term and complex technological programs. DOE has a history of coordinating such projects through contracts with industries, universities and its own laboratories. The size, interdisciplinary nature and long-term scale of the human genome project, with the many technologies involved, fits these experiences of DOE well. In addition, within DOE the mission of the Office of Health and Environmental Research (OHER) is to understand the health effects of radiation and other by-products of energy production. This requires fundamental knowledge of the effects of chemical and physical damage to the human genome.

The OHER mission in human genetics has led to the initiation and support of a number of research and technological developments which are closely linked to the human genome mapping and sequencing project. These include basic research on radiation and chemically-induced damage of DNA and on the repair of DNA damage. Risk analyses of the effects of the deleterious agents on cancer and genetic diseases have also been done. DOE-supported studies in genomic mapping, chromosome isolation, and sequence data management and analysis are even more directly related. Thus, this initiative is a natural outgrowth of current DOE-supported research. Furthermore, the initiative will make important contributions to other DOE missions, including environmental waste control, improving energy production, producing and utilizing biomass, and so forth.

The National Laboratories can be an important resource for the genome project. They are currently furthering the goals of the project by providing sorted chromosomes, genetic probes and clone libraries. Genbank at Los Alamos is presently supported by NIH, DOE and other agencies as a computer facility for organizing and disseminating DNA sequence information. The National Laboratories are experienced in providing technical and engineering support for large projects, and for efficient development of technological tools. The completion of a physical map of the human genome, the organization of associated clone libraries and the production of a reference sequence produce a tool. This tool can be the most powerful technological resource available for the understanding of biology and medicine.

The Office of Health and Environmental Research seeks a fundamental understanding of the health effects of radiation and of energy-related chemical toxicants, so as to apply its findings to the protection and improvement of human health. The complete sequence of a human genome provides a reference base against which perturbations induced by the environment will be recognized and measured. A long-term interest has been the monitoring of somatic cell and germ cell damage caused by radiation and by other toxic agents such as chemical mutagens.

Americans receive exposure to various mutagens, including mutagens from energy sources such as the combustion of fossil fuels. The exposure levels of paramount importance to society are low, and there is enormous individual heterogeneity in susceptibility to exposure. Rapid and cost-effective methods are needed to assess exposures and risks to large numbers of people. The definitive measure of mutation is the sequence of DNA. The ability to determine quickly and accurately the sequence of any DNA is the ultimate way to assess immediate and cumulative damage by many agents. Thus DOE has unique capabilities to manage this initiative, and the initiative is central to its mission.

Other Federal and private agencies have a major interest in this initiative. The National Institutes of Health, in particular the National Cancer Institute and the National Institute of General Medical Sciences, are already heavily committed to support research on DNA sequence and function. This support deserves to be increased. The Howard Hughes Medical Institute supports an increasing number of projects on human genetic diseases. Important work is also being done in Europe and Japan. It has become clear to everyone that the tools to map and to sequence the human genome can now be developed; what will be accomplished depends on the effort and commitment.

DOE should develop general methods and provide tools useful to all the other molecular biology projects. Instrumentation, automation, computation and other multidisciplinary approaches should be emphasized. DOE should foster cooperation among all the organizations involved, both national and international. However, it should not delay implementation of its plans or defer to some other organization. Thorough communication should ensure that there is no duplication of facilities and waste of resources. We strongly encourage continuing cooperation among the various agencies.

C. Major advances in diagnosis, prevention and treatment of disease will result.

Research to understand the human genome is taking place, so why is it necessary to have a new initiative? The answer is that the results of this initiative are so valuable to humanity that it is essential to proceed as fast as possible. Consider diabetes, for example. One in 300 American children take daily insulin injections by age 18. About half of these will have kidney failure within 30 years. Today about half of all people on kidney dialysis (at a cost of about $1 billion annually) are diabetics. The disease is genetic, associated with factors on chromosome 6, thus children at risk can be identified. Knowledge of the precise genetic basis of the disease by appropriate sequencing may allow reversal of the autoimmune process which leads to diabetes.

The major killers in this country -- cancer, cardiovascular disease, hypertension and stroke -- all have significant genetic components. The ability to respond to these diseases before they strike will save lives. The immune system controls the body's intrinsic defenses and is responsible for autoimmune diseases and other degenerative diseases such as arthritis. Analysis of the genes of the immune system will allow effective stimulation of the defenses and appropriate therapy for the diseases. Detailed sequence information will lead to methods for more exact matching of donor and recipient in transplantations. Monitoring changes in the DNA sequence of one tissue in one person will reveal damage caused by environmental factors. Many more examples could be given. However, the analogy of knowing human anatomy and knowing the human genome is apt. We could not cure heart disease as soon as we understood blood circulation, but it was a necessary first step. It is also well to emphasize that we do not need to recognize and order all genes for success. Each new fragment of DNA sequence can bring human benefits.

It is now practical to locate genes, to sequence DNA, to supplement some of the deficiencies caused by missing or defective genes. A major effort will bring immediate and continuing benefits. Each new gene identified and mapped will allow certain diagnosis of any diseases associated with this gene. Recent examples include Duchenne muscular dystrophy, chronic granulomatous disease, cystic fibrosis, Alzheimer's disease and Huntington's disease. The identification of genetic risk factors for common diseases such as diabetes and premature coronary disease are further examples where genetic map information could lead to methods of risk modification for an entire population. The recent identification of genes which lead to abnormal development emphasizes a relatively unexplored health problem -- birth defects. Alteration of

genes following birth is well documented in the development of the body's immune defenses. Abnormal alteration (mutation) of genes is responsible for numerous cancers. Thus the knowledge of the human genome -- the genes, their regulation and their abnormal function -- will have the greatest impact on health maintenance yet experienced in medicine. No individual will be untouched by this initiative.

We cannot afford, nor do we have foundation support for, individual and redundant efforts on the 3500 inherited diseases presently known. Many laboratories are presently working in parallel to obtain DNA fragments and sequences near important human genes. Progress has been made on particular diseases because of foundations dedicated to them, but much of the effort has been redundant. Although the gene for Huntington's disease has been localized to a region on the short arm of chromosome 4 for three years, and the gene for cystic fibrosis has been localized to a small region of chromosome 7 for over a year, overlapping DNA fragments which span these regions are yet to be developed. The high cost of these important studies would be markedly reduced by the development of much faster and comprehensive sequencing and mapping studies. A reference sequence would thus provide rapid, and much more economical, discovery and identification of human disease genes.

The development of rapid, cost-effective methods for sequencing may be the greatest benefit. The more efficient technologies that will be developed for the human genome project will be directly applicable to all sequencing problems. It is appropriate to ask whether we can afford not to develop such improved technology given the level of resources already going into sequencing. We do not know what sequence information is the most valuable. It is likely that the most significant applications to medicine cannot be foreseen at the present time, but the ability to determine DNA sequences routinely will allow immediate application of that knowledge.

D. The process will produce advances in biotechnology.

The long-range goal of this initiative is to understand the human genome. This will require improved technology in many other fields. It will automatically further fundamental advances in molecular biology. It will encourage correct theories which relate DNA structure and function, RNA structure and function, and protein structure and function. The ability to organize, manipulate, correlate and retrieve large amounts of data must be improved. Fast and accurate robots that can clone, purify and sequence DNA need to be developed. The advances in all these areas will be applicable to the use of biological materials in industry and agriculture. For example, more efficient production of biomass for energy production should result. Important environmental goals that are of major importance to the Department of Energy will be furthered, such as protection of plants by improving their resistance to environmental stress, and neutralization of toxic wastes by using genetically engineered microbes. Development of the new technologies for this initiative in the fields of biology, chemistry, physics, instrumentation, automation and computing will place the U.S. at the forefront of the biotechnology of the 21st century.

New knowledge about the human genome also means new knowledge about all other genomes. Fundamental knowledge about DNA structure applies to all organisms. Even more directly, sequences of some genes are similar from animals to plants to bacteria. Studies on other organisms, where genetic experiments can be done, will help progress in the human genome. Also maps of other species will greatly increase the validity of applying the results of experiments on other organisms to human health problems. Thus, the human genome project will complement all the other biological research being done on humans and other organisms to increase the rate at which we understand human biology. Now is the appropriate time to begin the direct examination of the human genetic system.

E. Fundamental knowledge in biology will result, and young scientists will be trained to be able to make new discoveries.

The many practical applications of this initiative have been discussed, but we must stress that the most important result will be new knowledge. We cannot predict what new insights we will obtain, but we are certain to learn completely new patterns of biological organization, structure and control. The discovery of large numbers of currently unknown genes will further our knowledge of all biological processes. The human genome sequence will serve as a reference library that will stimulate and coordinate the next century of biological research. The graduate students and other young investigators who work on this initiative will obtain the background and training to attain the goals of 21st century initiatives. Their exciting research findings should also encourage more entering college students to choose the fields of biological and physical sciences and engineering.

F. Deleterious effects on other programs must be prevented.

A major new initiative, no matter how worthy, must not disrupt or hinder ongoing worthwhile programs. This initiative deserves the highest priority. However, the most efficient progress will occur if research in all aspects of the relation of DNA to RNA to protein and to health are strongly supported. This requires major increases in funding for the human genome.

It is also important that effort not be shifted from current projects on the genetics of other organisms to study the human genome. The human genome is the emphasis of this initiative; it is not its only component. Everyone must realize the similarity among genes and the utility of transferring knowledge from one organism to another. Furthermore, the DOE initiative will involve people from a wide range of disciplines, including biology, chemistry, engineering, physics and mathematics. There is a large pool of scientists and engineers available.

There is some fear that a large influx of money into a field will distort and disrupt current research. However, there is good precedence that this is not necessarily so. The Howard Hughes Institute increased its biomedical funding from $3 million in 1975 to more than $200 million in 1986. There has been a significant beneficial effect.

A large and increasing financial commitment should be made to support this initiative. It should be distributed among the National Laboratories, Universities and Research Institutes; industry contracts may be used when appropriate. Both small science and large science projects should be supported. Peer review should be used for initial funding, and continuing funding should require further review. Flexibility and innovation should be fostered. It is particularly important in this rapidly developing field not to start any large, inflexible organizations whose direction would be hard to change. A large part of the challenge of this initiative is to think of new ideas and to develop relevant technology. A wide range of funding mechanisms will be needed and a wide variety of organizations must be supported.

Appendix A. Analysis of Costs.

General. The total cost of sequencing the human genome will certainly fall in the billion dollar range, although it is important to stress that the actual cost will be very sensitive to the state-of-the-art technologies associated with DNA sequencing, and the related requirements for automation of procedures for cloning, mapping, data handling and data analysis. As an example, compare the current and projected future costs for DNA sequencing and their corresponding implications for sequencing the human genome.

Estimated Cost for Determining the DNA Sequence of a Human Genome (Given Unique Fragments)*

SOURCE	COST	GENOME COST
Current commercial laboratories	$1/base	$6 billion
Japanese sequencing machines	$0.17/base	$1 billion
Future cost with automation	$0.01/base	$60 million

*This estimate does not include the cost of isolating and ordering the fragments; it only includes sequencing each DNA strand, or 6 billion bases. Sequencing both strands provides a check on the accuracy of the sequence.

This table illustrates the importance of making substantial initial investments in technology. We emphasize that the above estimates do not include costs for cloning, mapping or data analysis. Thus our proposal for sequencing the human genome would necessarily be staged. The first 5 years would focus on three general objectives: 1) mapping the human genome, 2) development of technology, 3) sequencing of selected chromosomal regions.

Advances in technology are a necessary first step in sequencing the human genome. These advances will make large-scale sequencing and subsequent comparative studies practical and cost-effective. At present the only automated sequencing machines are based on Sanger's method. There are probably distinct advantages to be gained from automating the Maxam-Gilbert method. A detailed comparison of the two approaches should precede a major investment in one of them. Both approaches can also benefit from considerable optimization. A twofold increase in the length of sequence accessible on a single gel lane would cut the cost of sequencing by considerably more than a factor of two. A number of ways to increase this sequencing range, such as pulsed-field techniques, are very promising and need to be tested.

Multiplex sequencing techniques such as those being developed by George Church are still in their infancy. However, their potential attractiveness is so great that a careful evaluation and refinement of such methods is surely warranted before one embarks on large-scale sequencing. Direct physical approaches to sequence determination such as mass spectrometry or scanning tunneling microscopy are speculative, but their potential impact must not be overlooked. Such approaches should be critically tested in the next few years.

Current strategies for using any of the existing sequencing methods are mostly shotgun approaches which sequence random fragments of DNA. These are quite inefficient since they require sequencing the same region many times over. Sequencing of overlapping fragments is needed to determine the order of the fragments; this is called a bottom-up approach. Phased, or top-down approaches, including systematic ordering and mapping, linked library construction, and optimized production of DNA fragments will all result in far less redundancy in the sequencing. These preliminary steps probably represent half of the final cost and require more than half of the skilled labor. Each of these preliminaries to the actual acquisition of sequence data needs full exploration, refinement and optimization. Most of these preliminaries can and should be automated. Very exciting developments, like methods for cloning or purifying large DNA fragments, and schemes for orderly generation of nested sets of DNA pieces are so new that their potential cannot yet be evaluated. However, it is inevitable that some of these methods will have to be incorporated into any effective large scale sequencing effort.

Once the speed, error rate, and cost are appropriate then one can begin the organized and coordinated effort to sequence a reference human genome. The technologies will then be sufficient to sequence other genomes and to examine human polymorphisms. The wide range of technologies that must be developed for this project are outlined below.

Technologies Required for Sequencing the Human Genome

1. Production of DNA fragments containing 100 to 1000 kilobases

    a. Chromosome separation
    b. Sequence-specific chemical and enzymatic scissors (restriction enzymes)
    c. Separation and purification of large fragments
    d. Large-insert cloning

2. Automated DNA handling, mapping and sequencing

    a. DNA preparation
    b. DNA cloning
    c. Physical, restriction fragment, and genetic mapping
    d. Chemical, physical and enzymatic sequencing

3. Data storage and analysis

    a. Immediate data entry with uniform notation
    b. Efficient searching with cross-referencing and access to other data banks
    c. Rapid data distribution
    d. Parallel or concurrent processing
    e. New algorithms for analyzing and interpreting DNA and protein sequences

4. Detection and analysis of DNA, RNA and protein at very low levels

    a. Single molecule analytical methods
    b. Methods for detecting large numbers of DNA fragments simultaneously (multiplexing)

We estimate that the cost of the development of all of these technologies will be about $500 million dollars. The total cost will be near $1 billion and completion of the project will take many years. However, each advance in technology will produce immediate benefits to medicine, agriculture and industry.

Strategy. A substantial effort directed at technology, mapping and pilot-project sequencing can begin immediately. The committee recognizes that implementation of this initiative by DOE has already begun, and it praises the speed and thrust of the effort. $11.5 million has been requested for fiscal year 1988; an amount double this would be more appropriate. Funds spent early in this project will save money later, because each advance in technology will make all the following steps more efficient and less costly. Support of $40 million dollars the first year (fiscal year 1989) and increasing linearly to $200 million dollars by the fifth year (fiscal year 1993) could be used very effectively. We envision three types of grants -- to individual investigators, to centers with 3 to 10 senior investigators and to a few large centers that will include mapping, sequencing and interpreting the human genome. In addition to the principal investigators, each project will involve junior scientists and engineers, and students. A total of 2500 professional people might be working on the initiative by 1993. The professional personnel will include molecular biologists, chemists, engineers, physicists, computer scientists and so forth.

Recommended funding levels are:

FISCAL YEAR	$ MILLION	TOTAL
1988	20	20
1989	40	60
1990	80	140
1991	120	260
1992	160	420
1993	200	620
1994	200	820
1995	200	1,020

Reasonable goals to attain by the end of seven years of support at the level requested (by the end of 1995 with $1 billion spent) are:

1) The United States should have the capacity to sequence ten million bases per day.

2) The complete map of each chromosome and an essentially complete sequence of at least one human chromosome should be finished.

Attainment of these goals will prove that the U.S. has the capabilities to continue the process to obtain all the benefits promised. We assume that equivalent progress will have been made in computer algorithms to analyze the sequences, and to characterize medically important genes.

Appendix B. The Need for Computer Resources: A Data Bank for the Future

As physical map data are gathered they must be stored in a way that facilitates the cross comparisons required to construct a complete map. Programs that do these functions already exist, but they may be inadequate for this project, because the human genome is about 1000 times as large as the largest current map (E. coli). Inefficiencies that are tolerable on small projects will be major problems on projects the size and complexity of the human genome.

It is particularly important to include in the data base references to other data bases and to facilitate communication between data bases. Specifically, it is necessary to be able to locate physical fragments with respect to any known genetic markers or to restriction fragment length polymorphisms. This is essential for the project to fulfill its promise of facilitating our understanding of human diseases. The entire set of data bases on Genomic Resources (the human gene map at Yale, the mouse gene map at Jackson Laboratories, etc.) which the Howard Hughes Medical Institute is helping to make cross-referenceable, contain data relevant to the human genome project. There are major nucleic acid and protein sequence data banks in the U.S., Europe and Japan which have agreed to collaborate closely. This effort must be supported and further developed. A coordinated effort must be established to maximize the interaction between these data bases, to reduce duplication of effort and to improve speed of data collection. Furthermore, there will undoubtedly be new discoveries, such as the introns which were discovered a decade ago, therefore it is important that the data bases be designed to absorb such changes gracefully.

Sequence Analysis. If the human sequence were magically made available, much of its interpretation would still remain obscure. Research performed now could unlock a substantial amount of the hidden information as the sequence becomes available. For instance, one of the key pieces of information included in the DNA sequence is the protein sequence, but that requires knowledge of the locations of transcription and of the splice junctions. Splice-junction information is usually obtained by sequencing both the genomic DNA and the messenger RNA. This requires substantially more work than would be needed if we could recognize the splice junction from the DNA sequence. However, the best current methods are correct only about 85% of the time in predicting splice junctions in genomic sequences. That means that all the junctions of a three-intron mRNA would be properly recognized only about 40% of the time. If a human gene resembles this example, then it is likely that with current methods we would know

considerably less than half of the coding sequences even if we knew the entire sequence and the locations in the DNA sequence of the primary transcripts. This is an area where focused research could greatly improve the outlook, even without new data. The use of current data with a good expert system (an expert system is a computer program that uses all the information that an expert would have to solve a problem) could significantly increase identification of splice-junction sites. New data will continue to enhance the performance of such programs. Although such programs will never be perfect, they provide predictions that are easily tested. Effective programs would avoid the necessity of sequencing the entire messenger RNA for a protein.

An expert system approach can be used on other patterns as the data emerge. For example, the system used to find splice-junction sites could also be used to identify promoter regions when more data are available to define them. The interaction between computer-aided predictions and experimental results is important. The results will improve the predictions, and the predictions should direct the experiments. An investment begun now in computer applications research will maximize the return, in the short term as well as the long term.

There are many more areas of sequence analysis that will benefit the human genome project. Current search and comparison programs should be made more efficient to handle the enormous size of the data base. We should also do more to understand the biological significance, in contrast to statistical significance, of finding sequence homologies. Research into general pattern identification methods would prove valuable.

Equally important to locating the proteins on the DNA sequence and determining their regulation is to understand their functions. Recent years have produced improvements in our ability to predict protein structures from their sequences. More research is needed to be capable of reliably predicting both structures and functions. That would provide an additional major key to unlock the information of the genome.