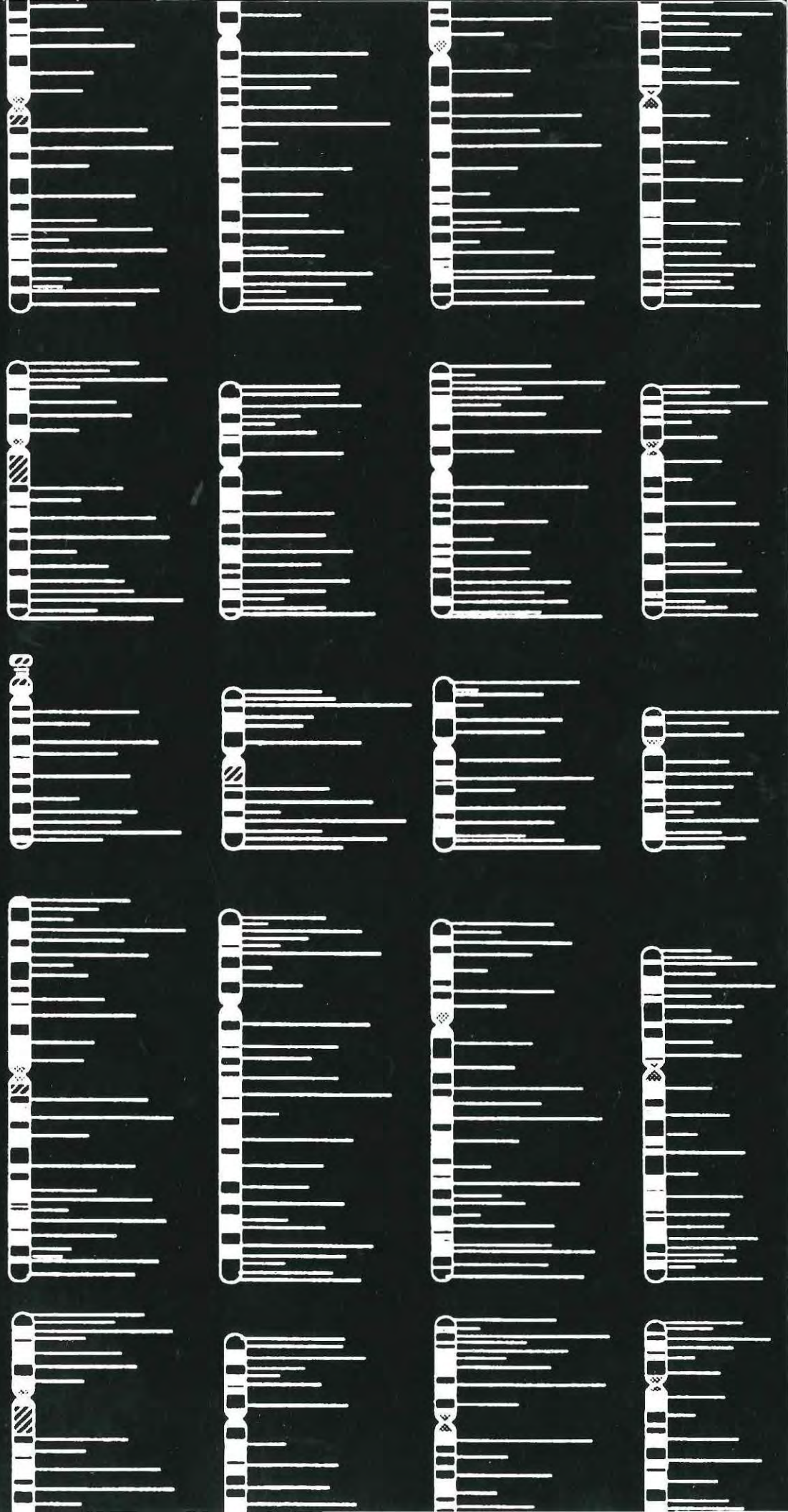
DOE/ER-0452P

Understanding Our Genetic Inheritance

The U.S. Human Genome Project:

The First Five Years FY 1991-1995



U.S. DEPARTMENT OF
HEALTH AND HUMAN SERVICES
Public Health Service
National Institutes of Health

U.S. DEPARTMENT OF ENERGY
Office of Energy Research
Office of Health and
Environmental Research

Available from the National Technical Information Service, U.S. Department of Commerce, Springfield, Virginia 22161

Price: Printed copy A05
Microfiche A01

Codes are used for pricing all publications. The code is determined by the number of pages in publication. Information pertaining to the pricing codes can be found in current issues of the following publications, which are generally available in most libraries: *Energy Research Abstracts (ERA); Government Reports Announcements* and Index (GRA and I); *Scientific and Technical Abstract Reports (STAR); and* publication, NTIS-PR-360 available from NTIS at the above address.

Cover:
Schematic representation of selected human chromosomes, the cellular structures that contain genes. Human cells contain 23 pairs of chromosomes. Alternating dark and light bands, the result of chemical staining, are used as visible landmarks for locating genes. Lines to the right of each chromosome mark positions where genetic markers have been mapped.

DOE/ER-0452P

Understanding Our Genetic Inheritance

The U.S. Human Genome Project:

The First Five Years FY 1991-1995

U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES
Public Health Service
National Institutes of Health
National Center for Human Genome Research

U.S. DEPARTMENT OF ENERGY
Office of Energy Research
Office of Health and Environmental Research
Human Genome Program

NIH Publication No. 90-1590
April 1990

Table of Contents

Understanding Our Genetic Inheritance

The U.S. Human Genome Project:

The First Five Years
FY 1991-1995

Executive Summary

The Human Genome Initiative is a worldwide research effort with the goal of analyzing the structure of human DNA and determining the location of the estimated 100,000 human genes. In parallel with this effort, the DNA of a set of model organisms will be studied to provide the comparative information necessary for understanding the functioning of the human genome. The information generated by the human genome project is expected to be the source book for biomedical science in the 21st century and will be of immense benefit to the field of medicine. It will help us to understand and eventually treat many of the more than 4000 genetic diseases that afflict mankind, as well as the many multifactorial diseases in which genetic predisposition plays an important role.

A centrally coordinated project focussed on specific objectives is believed to be the most efficient and least expensive way of obtaining this information. In the course of the project much new technology will be developed to facilitate a broad range of biological and biomedical research, bring down the cost of many experiments, and find application in numerous other fields. The basic data produced will be collected

in electronic databases that will make the information readily accessible in convenient form to all who need it.

This report describes the plans for the U.S. human genome project and updates those originally prepared by the Office of Technology Assessment (OTA) and the National Research Council (NRC) in 1988. In the intervening two years, improvements in technology for almost every aspect of genomics research have taken place. As a result, more specific goals can now be set for the project.

Five-year goals have been identified for the following areas, which together encompass the human genome project:

- Mapping and Sequencing the Human Genome
- Mapping and Sequencing the Genomes of Model Organisms
- Data Collection and Distribution
- Ethical, Legal, and Social Considerations
- Research Training
- Technology Development
- Technology Transfer

This plan sets out specific scientific goals to be achieved in the first five years together with the rationale for each goal. The specific goals will be reviewed annually and updated as further advances in the underlying technology occur.

The plan presented here was prepared jointly by the National Institutes of Health (NIH) and the Department of Energy (DOE), the two agencies that have received funding earmarked for the human genome project. Over the past two years, these agencies have developed a highly synergistic and well-integrated approach to carrying out this initiative, as evidenced by the adoption of this common plan.
The National Institutes of Health has a natural interest in the Human Genome Initiative in view of its long history of supporting research in genetics and molecular biology as an integral part of its mission to improve the health of all Americans. The Department of Energy has a long-standing program of genetic research directed at improving the ability to assess the effects of radiation and energy-related chemicals on human health.

To achieve the scientific goals set out in this report, a number of administrative measures have been put in place. In addition, a newsletter, an electronic bulletin board, a comprehensive administrative database, and other communications tools are being set up to facilitate communication and tracking of progress.

Research centers will be established to promote the collaboration of investigators from diverse disciplines on a major task of the genome program. DOE has already established three large centers in its National Laboratories and NIH will establish 10 to 20 additional centers over the next five years. The centers will become foci for collaboration with investigators at other locations and with industrial organizations that want to develop applications of the research results, thereby creating networks of interrelated projects.

Meetings and workshops will be organized to bring together investigators with common research objectives and to encourage collaboration, exchange of materials and use of common starting materials or protocols wherever these are appropriate. It is expected that mapping and sequencing groups will coalesce around individual human chromosomes or around particular model organisms.

NIH and DOE will continue their synergistic working relationship and will also interact closely with other interested agencies, as well as with genome mapping programs in other countries as they get organized. Close ties with industry and with the medical community have been established, and will continue to be encouraged, to ensure efficient technology transfer. The private sector is involved in this project at all levels from participation in the advisory committees to receipt of grants and contracts.

The overall budget needs for the effort are still anticipated to be the same as those identified by the OTA and the NRC, namely about $200 million per year for approximately 15 years. Fiscal years 1988 to 1990 have been a period for getting organized and getting research under way. The five-year goals specified in this plan are for the period FY 1991 through FY 1995 and assume the program will rapidly reach the level of funding specified above.

The First Five Years
FY 1991-1995

Introduction

The Human Genome Initiative is a worldwide research effort that has the goal of analyzing the structure of human DNA and determining the location of all human genes. In parallel with this effort, the DNA of a set of model organisms will be studied to provide the comparative information necessary for understanding the functioning of the human genome. The information generated by the human genome project is expected to be the source book for biomedical science in the 21st century. It will have a profound impact on and expedite progress in a variety of biological fields, including those such as developmental biology and neurobiology, where scientists are just beginning to understand the underlying molecular mechanisms. The analysis and interpretation of the information will occupy scientists for many years to come. Thus, the maximal benefit of the human genome project will only be achieved if it is surrounded by research efforts that are focussed on understanding and taking advantage of human genetic information.

The human genome project is expected to immensely benefit medical science. It will help us to understand and eventually treat many of the more than 4000 genetic diseases that afflict mankind, as well as the many multifactorial diseases in which genetic predisposition plays an important role. New technologies emanating from the genome project will also find application in other fields such as agriculture and the environmental sciences. They will be valuable for assessing the effects of radiation and other environmental factors on human genetic material.

It is anticipated that the private sector will derive great benefit from the trained manpower, the data, and the techniques developed by the human genome program and will develop many useful applications based on the new knowledge that is produced. Within a few years, DNA sequence information will undoubtedly be a major tool in most areas of basic and applied biological research.

As a result of the enormous strides in basic research on molecular and medical genetics in the last 30 to 40 years, technology has advanced to a stage of development at which such a project can realistically be contemplated. Because of the farsighted investment in basic research by the federal government over this time period, the United States is clearly the leader in this field. Pursuit of the human genome project will allow the United States to remain at the forefront of biomedical science and to train the scientific manpower that will be able to take advantage of the immense opportunities for research and innovation emanating from this project.

The possibility of initiating such a major and significant research program was extensively discussed in the scientific community during 1986 and 1987. In the spring of 1987, a Report on the Human Genome Initiative was prepared by the Health and Environmental Research Advisory Committee (HERAC) of the Department of Energy (DOE). In early 1988, further discussion culminated in the publication of two additional, widely circulated, influential reports[1]. The U.S. Congress' Office of Technology Assessment (OTA) report presented a comprehensive and detailed analysis of the scientific developments that had led to the promise of "mapping and sequencing" the human genome and presented a number of options as to how the United States might pursue such a project. The National Research Council (NRC) report recommended that the United States support the research effort and presented an outline for a multi-phase research plan for accomplishing the goal of sequencing human DNA over the course of the following two decades. A report to the Director of the National Institutes of Health (NIH) by the *Ad Hoc* Advisory Committee on Complex Genomes, also prepared in 1988, concurred with the NRC Report.

In fiscal year 1988, the Congress of the United States launched the human genome project by appropriating funds to both the DOE and the NIH specifically for support of research efforts to determine the structure of complex genomes. In the report accompanying the Senate appropriations bill for FY 1989, the Congress requested the NIH to prepare, by early 1990, a report on the optimal strategy for the conduct of the human genome program. The FY 1990 House Appropriations Committee report also asked the NIH for a comprehensive spending plan by the time of the FY 1991 appropriations hearings.

Prepared in response to those requests, the present report contains a summary of the progress that has been made in the field of genome research since the

---

[1]***Mapping Our Genes, The Genome Projects: How Big, How Fast***; Office of Technology Assessment, Congress of the United States; ***Mapping and Sequencing the Human Genome***; Commission on Life Sciences, National Research Council.

preparation of the OTA and NRC reports and presents a plan for the human genome program, with emphasis on the next five year period. Because the two agencies have been collaborating closely for the past two years in the management of the program, this plan was prepared jointly by the NIH and the DOE. The agencies plan to revise the plan approximately annually, based on the latest scientific developments.

Program Goals

It is generally agreed that the overall goal of the Human Genome Initiative is to acquire fundamental information needed to further our basic scientific understanding of human genetics and of the role of various genes in health and disease. The premise is that this can be done much more efficiently, and in a more cost-effective manner, as a targeted and coordinated program. Thus, we obtain valuable basic information in the least expensive way while increasing the "benefit-to-cost" ratio for genetics research in general.

As refined through the discussions over the last half of the 1980s and defined in the NRC report, the Human Genome Initiative has several interrelated goals:

- construction of a high-resolution genetic map of the human genome;
- production of a variety of physical maps of all human chromosomes and of the DNA of selected model organisms, with emphasis on maps that make the DNA accessible to investigators for further analysis;
- determination of the complete sequence of human DNA and of the DNA of selected model organisms;
- development of capabilities for collecting, storing, distributing, and analyzing the data produced;
- creation of appropriate technologies necessary to achieve these objectives.

At the time the NRC and OTA reports were written, the consensus of the scientific community was that state-of-the-art technology was sufficient for the development of detailed genetic and limited physical maps. That technology, however, was not

considered sufficient for completion of the physical map of a genome as large and complex as that of the human. Nor was the technology available for DNA sequencing considered to be adequate for the task of sequencing the 3 billion base pairs of human DNA. At that time, the largest continuous human sequence that had been determined was that of the human growth hormone gene, only 67,000 nucleotides long.

Thus, the NRC committee and others recommended a multi-phase program, in which the initial phase would consist of:

- expansion of the human genetic map to a resolution of one centimorgan;
- construction of complete physical maps of the DNA of certain model organisms and beginning the construction of physical maps of human chromosomes;
- development of new technology to increase the efficiency and accuracy, and lower the cost, of physical mapping and of DNA sequencing.

In these recommendations, the task of sequencing the complete human DNA was reserved to a later phase, one that would only be embarked upon if methods could be developed that would allow the sequence to be obtained at a reasonable cost. The overall program was expected to take at least fifteen years to complete. Technology development was to be an integral part of the project throughout.

This general plan is still appropriate, but some of the details must be changed as improvements in the technology have occurred in the past two years. In order to prepare the present report, advisors to and staff of NIH and DOE have joined forces to examine the state of the science and develop the plan to be followed over the next five years. This document represents the consensus of the two agencies regarding the conduct of genome research and will be updated periodically.

The rosters of the various advisory groups that participated in the plan's development are appended. These are the Health and Environmental Research Advisory Committee (HERAC) of the DOE (Appendix 1), and the Program Advisory Committee on the Human Genome (PACHG) of the NIH (Appendix 2). The primary working group for these committees was the joint subcommittee of the HERAC and the PACHG (Appendix 3) that was specified in the NIH-DOE Memorandum of Understanding (see below and Appendix 4), supplemented by additional experts (Appendix 5).

The plan addresses specific scientific goals to be achieved in the next five years in the following areas:

- Mapping and Sequencing the Human Genome
- Mapping and Sequencing the DNA of Model Organisms
- Data Collection and Distribution
- Ethical, Legal, and Social Considerations
- Research Training
- Technology Development
- Technology Transfer

Also presented are the implementation strategies that will be used to achieve these goals with respect to:

- Administration of Research
- Roles of NIH and DOE
- Roles of Other Federal Agencies
- International Collaboration

Finally, the report addresses budget projections. The five year period covered by the plan will begin with FY 1991. It is assumed that funding levels for the combined NIH and DOE programs will rapidly reach the level recommended by the NRC report, approximately $200 million per year, adjusted for inflation. Although five-year goals are presented with some specificity, and although substantial progress has been made in technology development in the past two years, it must be stressed that the next five years will still be a time in which rapid advances in methods and strategies will be necessary if the program is to meet the goals outlined. Extreme flexibility and diligence on the part of program management, for both the research and its administration, will be needed during this period of experimentation and technological development.

Scientific Goals

1. Mapping and Sequencing the Human Genome

The human genome consists of 50,000 to 100,000 genes located on 23 pairs of chromosomes. One chromosome in each pair is inherited from the mother, the other from the father. Each chromosome contains a long molecule of DNA, the chemical of which genes are made. The DNA, in turn, is a double-stranded molecule in which each strand is a linear array of units called nucleotides or bases. There are four different bases, called A,T,G, and C. The bases on one DNA strand are precisely paired with the bases on the other strand, so that an A is always opposite T and G opposite C.

The order of the four bases on the DNA strand determines the information content of a particular gene or piece of DNA. Genes differ in length, ranging in size from roughly 2,000 to as many as 2 million base pairs. Mapping is the process of determining the position and spacing of genes, or other genetic landmarks, on the chromosomes relative to one another. There are basically two types of maps, genetic and physical, which differ in the methods used to construct them and in the metric that is used to measure the distance between genes. Sequencing is the process of determining the order of the nucleotides, or base pairs, in a DNA molecule.

Although mapping of human genes began early in the 20th century, it has been intensively pursued only for the past two decades. For most of this period the methods that were developed, though original and ingenious, have been inadequate for comprehensive mapping and have only allowed the construction of relatively crude maps with very little detail. Recently, much more effective technology has been introduced. To date, about 1,700 of the estimated 50,000 to 100,000 human genes (less than 2 percent) have been mapped.

A frequently asked question is: whose genome will be sequenced? The answer is, no one's. The first complete human genome to be sequenced will be a composite of sequences from many sources, most of these being cell lines that have existed in laboratories all over the world for some time. The sequence will be a generic sequence representative of humans in general and not of any particular individual. The complete sequence will provide a standard against which other partial sequences can be compared. It has been suggested that, due to the great variability between individual human beings, a single sequence would not be very useful.

While it is true that much valuable insight will come from comparing many different human sequences, the presumption is that functionally important DNA is conserved among humans, just as it is between humans and mice in those areas that have been studied. DNA regions of particular interest, such as genes involved in genetic diseases, will be sequenced from many individuals in the course of research on those diseases. As more information about the extent of genetic variation accumulates from these and other studies in the next few years, it will be evaluated to determine the impact on strategy for the human genome project.

Genetic Map

Genetic maps have many uses, including identification of the genes associated with genetic diseases and other biological properties. Genetic maps also form the essential backbone or scaffold needed to guide a physical mapping effort.

Genetic maps are constructed by determining how frequently two "markers", such as a physical trait, a particular medical syndrome, or a detectable DNA sequence, are inherited together. Genes that lie close together on a chromosome have a much higher chance of being inherited together than do genes that lie farther apart. Genetic studies of families, to determine how frequently two traits are inherited together, lead to the production of "genetic maps" in which distance between genes is measured in centimorgans (in honor of the American geneticist Thomas Hunt Morgan). Two markers are one centimorgan apart if they are separated one percent of the time during transmission from parents to children. The physical or molecular distance to which a centimorgan corresponds varies a great deal, but the genome-wide average distance for a centimorgan is believed to be roughly 1 million base pairs.

The development of genetic mapping tools is prominent among the technical advances that led to the Human Genome Initiative. The introduction of DNA markers, such as restriction fragment length polymorphisms, or RFLPs,

to detect genetic variation among individuals has been one of the most important innovations. Such markers are relatively easy to find in large numbers and have been used to construct genetic maps. In the past two years, advances have continued in this area. New types of DNA markers have been defined, and techniques, such as denaturing-gradient gel electrophoresis, have been adapted to detect subtle variations in DNA sequences. As a result, the number of useful markers has increased in the past two years.

It is estimated that 3000 well-spaced and informative markers will be needed to achieve a completely linked map, with markers an average of one centimorgan apart as recommended by the NRC. For the first five years, the genome program has set as its goal the creation of a 2 to 5 centimorgan map, which would require 600 to 1500 such markers. Each marker should be identified by a sequence-tagged site (STS) as defined in the section on physical mapping. A working group has been established to develop a plan for achieving this goal.

5 YEAR GOAL: Complete a fully connected human genetic map with markers spaced an average of 2 to 5 centimorgans apart. Identify each marker by an STS.

Physical Map

The distance between sites on physical maps is measured in units of physical length, such as numbers of nucleotide pairs. Physical maps can be constructed in a variety of different ways. They are used as the basis for the isolation and characterization of individual genes or other DNA regions of interest, as well as to provide the starting material for DNA sequencing. The ability to construct physical maps derives from recombinant DNA techniques that allow the isolation and cloning of DNA fragments, the identification of specific sequence markers on DNA, and the determination of the order of and distance between such markers on a chromosome.

There are several kinds of physical maps, which can be categorized into two general types. The cytogenetic map describes the order and spacing of markers on a DNA molecule. Based on microscopic analysis, cytogenetic maps record the location of genes or DNA markers relative to visible landmarks on the chromosomes. This is the oldest type of physical map and the resolution (precision in locating markers) is rather low, on the order of

10 million base pairs. Nevertheless, the cytogenetic map is still an extremely valuable tool and markers continue to be mapped in this way. At the recent 10th Human Gene Mapping Workshop, the number of mapped markers was reported to be 4362, as opposed to 2057 only two years ago. Another example of this type of physical map is the long-range restriction map, which records the order of and distance between specific sequences, known as restriction sites, on chromosomes. The resolution of long-range restriction maps is between 100,000 and 2 million base pairs.

The second type of physical map consists of a collection of cloned pieces of DNA that represent a complete chromosome or chromosomal segment, together with information about the order of the cloned pieces. There are a variety of techniques for cloning DNA and a number of methods for determining the order of the clones. The technology for constructing overlapping clone sets (known as "contigs") is continually improving. At present, a collection of ordered clones is typically the starting material for sequencing. However, novel approaches that do not require cloning, but still allow the investigator access to the DNA to be sequenced, are under development.

In the past two years, improvements in several techniques have made the initial stages in the construction of physical maps of large genomes significantly easier and more rapid than was predictable at the time of the NRC recommendations. These techniques include pulsed-field gel electrophoresis, yeast artificial chromosome cloning, the polymerase chain reaction (PCR), fluorescence *in situ* hybridization, and radiation hybrid analysis. Currently, the U.S. government supports research projects to physically map the DNA of all or parts of 11 of the 24 human chromosomes (there are 23 pairs of chromosomes, but the X and Y sex chromosomes are not like each other, resulting in 24 different chromosomes).

NIH is supporting, through its extramural grants program, projects for physical mapping of three chromosomes (3,4,18). The DOE is supporting projects in the Los Alamos and Livermore National Laboratories to produce complete overlapping clone maps of two others (16,19), and the two agencies are funding separate but complementary physical mapping efforts on another six chromosomes (5,11,17,21,22,X). These projects involve the construction of physical maps of both types, using both state-of-the-art techniques and new methods under development. The DOE also supports the preparation of clone libraries representing the various chromosomes under study at Los Alamos and Livermore.

There are still several technological barriers to the rapid, inexpensive, and routine construction of physical maps. One is the relatively short length of DNA over which a continuous, or uninterrupted, set of overlapping clones can be readily established. Contigs are typically small, consisting of between two and six cosmid clones (a cosmid is a type of vector that can carry a maximum of 40 thousand base pairs). To be more than minimally useful, the length of DNA over which the physical map shows continuity, or "connectivity," must be considerably longer.

A challenging but reasonable goal for physical mapping research projects is to extend to about 2 million base pairs the length of a DNA segment that can be covered by a single contig or spanned by a set of closely spaced, ordered markers. If physical mapping of human chromosomes is to be achieved within the next five years, it is important that current physical mapping efforts give their highest priority to the problem of completing maps, i.e. of achieving uninterrupted continuity of physical mapping data over large regions of DNA.

Another difficulty faced by those trying to assemble physical maps of chromosomes has been the inability to compare the results of one mapping method directly with those of another and to combine maps constructed by two different techniques into a single map. This problem is addressed by the recent proposal of a new concept or definition of a useful physical map[2]. According to the proposed system, data from any of a variety of physical mapping techniques can be reported in a common "language." In this system, each mapped element (individual clone, contig, or sequenced region) is defined by a unique "sequence-tagged site" or STS, which is basically a short DNA sequence that has been shown to be unique. A map is then constructed showing the order and spacing of the STSs.

The STS system, as proposed, appears to have several advantages. The STS map can be represented electronically and stored in a database that is publicly available and contains sufficient information to enable any scientist to recover *de novo* any mapped chromosomal region in his/her own laboratory. Thus, the proposed STS system will facilitate the scientific community's access to the human physical map. Quality control and project accountability will also be improved because the mapping results reported by any individual laboratory can readily be checked elsewhere.

---

[2]Olson *et al.*, *Science* 245:1434 (1989). The authors of this paper were members of the original NRC Committee on Mapping and Sequencing the Human Genome.

Access to mapped DNA through the information in the STS database will obviate the need for an expensive, long-term, centralized repository of clones, although it will not eliminate the need to generate and map such clones nor the need to store them in and distribute them from the laboratory in which they are produced. The proposed STS system will also facilitate the integration of results from different laboratories, regardless of the methods used, to produce a single, useful physical map and will establish a uniform criterion for determining how complete the map of a particular region is. Finally, an STS map may in the future be the appropriate starting point for DNA sequencing.

The STS proposal is still under discussion in the scientific community and few, if any, mapping projects have started to use the STS system. Another uncertainty is the additional cost of generating STS markers. NIH and DOE have established a joint working group to develop more detailed plans for testing and implementing the STS approach to physical mapping.

Over the next five years, in addition to generation of STS maps, efforts should be continued to generate complete contig maps of large regions of the human genome. Because current technology is not yet sufficient for this task, however, it is unclear what fraction of the genome can be cloned and ordered during this time. An STS map, with one STS characterized approximately every 100,000 base pairs, is an achievable goal. Such a map will assist continued efforts to isolate the intervening DNA.

5 YEAR GOAL: Assemble STS maps of all human chromosomes with the goal of having markers spaced at approximately 100,000 base-pair intervals.

Generate overlapping sets of cloned DNA or closely spaced unambiguously ordered markers with continuity over lengths of 2 million base pairs for large parts of the human genome.

DNA Sequencing

Three decades ago when Francis Crick and James Watson elucidated the double helix structure of DNA, there was no way to determine the sequence of even short DNA molecules. Only years later, with the advent of recombinant DNA technology in the early 1970s, was it possible to think of

isolating individual genes. That breakthrough, combined with the development of powerful DNA sequencing techniques, provided the technological basis for the Human Genome Initiative.

To date, the only organisms for which a complete DNA sequence has been determined are viruses. The largest published viral genome sequence is that of the Epstein-Barr virus, a sequence of 170,000 base pairs. Scientists are now attempting to sequence the DNA of certain bacteria, approximately 4.5 million base pairs long. The size and complexity of human DNA, however, still makes the sequencing of the human genome awesome to contemplate. Although many short stretches of human DNA have been sequenced--slightly more than 5 million base pairs altogether--the human genome comprises about 3 billion base pairs of DNA and is nearly 1,000 times larger than that of a bacterial genome.

If such a large amount of DNA is to be sequenced, a substantial increase in the speed and reduction in the cost of sequencing technology will be required. The current cost of DNA sequencing, in laboratories that do it routinely, is estimated to be about $2 to $5 per base pair of finished sequence, that is, sequence whose accuracy has been adequately confirmed. The costs of DNA preparation, salaries and overhead are included in these figures. In laboratories that sequence DNA only occasionally, the costs are much higher. These costs must be reduced below 50 cents a base pair before large scale sequencing will be cost effective.

Sequencing technology has improved significantly in the past two years. Machines that automatically identify the order of base pairs in appropriately prepared DNA samples are now readily available. In the most advanced laboratories it is possible, using these machines, for one individual to generate about 2000 base pairs of finished DNA sequence per day per machine, starting with properly prepared cloned DNA.

One approach to lowering the cost of DNA sequencing is further automation. The maximum reduction in cost of current sequencing technology will come from the creation of a fully automated assembly line for rapid DNA sequencing. Efforts are underway in both DOE and NIH-sponsored projects, as well as in private companies, to automate most of the preparatory steps in the sequencing process through the development of high-speed robotic work stations for sample handling.

During the next five years, pilot projects will be undertaken in order to test strategies and develop technologies for larger sequencing projects, with the aim of reducing costs to well below $1 per base pair by the end of the first

five-year period. These projects should analyze biologically interesting regions in the size range of 200,000 to 1 million base pairs. In these developmental efforts, it will be more important to complete the sequence of chosen segments rather than to merely obtain a very high number of base pairs of sequence comprising many smaller segments. This approach will maximize the possibility of successfully identifying and developing the technology needed to proceed with large-scale genomic analysis.

In addition, the amount of biological information obtained in the sequencing of human DNA in the course of these developmental research programs will be significantly increased if parallel efforts to sequence equivalent regions in the mouse are undertaken. Such comparative approaches will be encouraged.

In order to keep the costs of the human genome project within the original estimates, the cost of routine large-scale sequencing will ultimately have to be reduced to well below 50 cents per base pair. Therefore, sequencing projects larger than these pilot projects, such as the sequencing of an entire human chromosome, will not be considered until the cost of sequencing is reduced to that level. The cost of sequencing will be assessed in five years and a recommendation made as to further technological developments needed before large sequencing projects are undertaken.

It is by no means certain that enhancement of current technology, as described above, will bring the cost of sequencing down sufficiently. Therefore, entirely new approaches to DNA sequencing will also be encouraged. There are a number of techniques that hold some promise, including the use of capillary gel electrophoresis, the use of stable isotopes and mass spectrometry, and new imaging techniques, such as scanning tunneling or atomic force microscopy and X-ray imaging. Projects of this sort are being pursued under support from the DOE and the NIH, as well as in private industry.

5 YEAR GOAL: Improve current methods and/or develop new methods for DNA sequencing that will allow large scale sequencing of DNA at a cost of 50 cents per base pair.

Determine the sequence of an aggregate of 10 million base pairs of human DNA in large, continuous stretches in the course of technology development and validation.

2. Model Organisms

Experience has shown many times over that information derived from studies of the biology of model organisms is essential to interpreting data obtained in studies of humans and in understanding human biology. Research involving microbial, animal, and plant models will continue to provide a basis for analyzing normal gene regulation, genetic diseases, and evolutionary processes. For this reason, the human genome program will support mapping and sequencing of the genomes of a select number of non-human organisms.

Research projects that use model organisms will also be valuable to technology development. Since the genomes of these organisms are smaller and simpler than that of the human, they represent excellent systems for the development and testing of procedures needed for the much more complex human genome.

A number of organisms have already been identified as particularly useful models for comparative genetic analyses because a large amount of information about their genetics and molecular biology has already been accumulated. These organisms are bacteria (*Escherichia coli*), yeast (*Saccharomyces cerevisiae*), the fruit fly *Drosophila melanogaster*, the worm *Caenorhabditis elegans* and the laboratory mouse. However, it is fully expected that research projects involving other model organisms will also contribute significantly to the Human Genome Initiative.

Complete physical maps, both long-range restriction maps and an overlapping clone set, are already available for *E. coli*. Long range restriction maps are available for several other bacteria, and overlapping clone sets are being assembled. Extensive overlapping clone sets have also been assembled for both *S. cerevisiae* and *C. elegans*. Projects to sequence *E. coli* DNA have been initiated in both the United States and Japan. Sequencing of the DNA of another bacterium, *B. subtilis*, has begun in a consortium of European laboratories. Another European consortium and an American-Japanese collaborative project have each begun to sequence one of the chromosomes of *S. cerevisiae*. Finally, a collaborative project, involving a laboratory in the United States and one in the United Kingdom, is planned to begin the sequencing of *C. elegans* DNA.

While the mouse genome is not simpler than that of man, it is particularly useful for comparisons because of the many biological similarities between the mouse and man. The genetic map of the mouse, based on morphological markers, has already led to many insights into human genetics. There is every reason to believe that a physical map of the mouse genome will be equally useful. In order to prepare a physical map of the mouse, a genetic map based on DNA markers will need to be

created. This will then lead to the development of a physical map that can be directly compared with the human physical map.

The general methodology used in studying model organisms will be similar to that described under the previous sections on mapping and sequencing. The need to achieve long range continuity in physical mapping and sequencing projects also applies to model organisms, as do the requirements for reducing costs.

5 YEAR GOAL: Prepare a genetic map of the mouse genome based on DNA markers. Start physical mapping on one or two chromosomes.

Sequence an aggregate of about 20 million base pairs of DNA from a variety of model organisms, focusing on stretches that are 1 million base pairs long, in the course of the development and validation of new and/or improved DNA-sequencing technology.

3. Informatics: Data Collection and Analysis

The direct product of the Human Genome Initiative will be genome maps and DNA sequences. For maximum utility, it will be critical to develop appropriate computer tools and information systems for the collection, storage, and distribution of the immense amounts of mapping and sequencing data that will be generated in the course of the program.

At present, it is not clear whether the most useful product of the Human Genome Initiative will be a single large database or a distributed set of smaller, networked databases. It is also unclear how genome databases will be structured in the future and whether existing databases can be adapted to meet the overall, long-term needs of the Human Genome Initiative, or whether new systems will have to be developed. However, it is certain that genome databases will need to be comprehensive and up to date, and, if there are several databases, it will be imperative that they effectively link with one another.

In addition to database development, it will be vital to develop new methods and tools for the analysis and interpretation of genome maps and DNA sequences. Successfully addressing both of these areas of genome informatics will require the development of a coordinated national program to make the information and

analysis tools from this project readily available to the widest possible range of scientists and physicians in the most useful, timely, and cost-effective manner.

While it is currently possible to describe the informatics goals of the Human Genome Initiative in broad terms, considerable refinement will be necessary as this program develops and informatics technology improves over time. A Joint Informatics Task Force (JITF) has been established by NIH and DOE to help the agencies develop detailed informatics programs. The report recommending the establishment of the JITF is included as Appendix 6.

The responsibilities of the JITF will include identification of the uses to which the data will be put and establishment of priorities for both technical objectives and policy areas. Specific issues to be addressed will include: genome database structures, management, and services; development of algorithms, software, and hardware for organization and analysis of data; data exchange standards; electronic networks for collection and distribution of genome information; training and education of informatics personnel; and coordination of genome informatics activities among laboratories and agencies. The JITF will also serve as a national focus for interaction with international activities related to genome informatics.

The challenge will be not only to design databases to meet the growing needs for access and for increasingly sophisticated search capabilities, but also to keep up with the voluminous amount of information that will be produced at ever faster rates. A number of research efforts are in progress to improve database design, software for database access, and data entry procedures.

Recently, the National Center for Biotechnology Information (NCBI) was established at the National Library of Medicine to create automated systems for knowledge about molecular biology, biochemistry and genetics, and to pursue research in biological information handling, particularly with respect to human molecular biology. Thus, the mission of the NCBI supports, in part, that of the Human Genome Initiative. Consequently, the efforts of the NCBI will be closely coordinated with the human genome program through the JITF and by frequent staff interactions with the NIH and the DOE.

5 YEAR GOAL: Develop effective software and database designs to support large-scale mapping and sequencing projects.

Create database tools that provide easy access to up-to-date physical mapping, genetic mapping, chromosome mapping, and sequencing information and allow ready comparison of the data in these several data sets.

Develop algorithms and analytical tools to interpret genomic information.

4. Ethical, Legal, and Social Considerations

The plan to map and sequence the entire human genome is predicated on the belief that humankind will benefit immensely from attendant advances in medicine, biological research, and biotechnology. Yet, as with any new technology, controversial usage of the information and capabilities that will flow from the Human Genome Initiative also may emerge. Ethical, legal, and social issues arise in regard to ways of ensuring that this information is used in the most responsible manner.

Some of the questions that must be considered concern individual privacy and confidentiality. Should information about an individual's genetic makeup become available to others without that person's knowledge and permission? How can we assure that genetic information does not lead to stigmatization or to discrimination in areas such as insurance or employment?

Concerns also arise in connection with the medical applications resulting from the genome program, such as the anticipated ability to predict a person's future health. Initially, at least, there will be a time lapse -- in many cases of years -- between the ability to diagnose certain genetic disorders and the ability to treat them. How will an individual cope with a devastating diagnosis when no treatment is available? What issues does such a situation raise?

These questions are not new. Physicians and counselors are facing them today when treating patients with genetic and other diseases. However, the greatly increased flow of information about human genetics will make the need to deal with these issues more compelling. The NIH and DOE human genome programs will support studies that investigate concerns such as these. About 3 percent of the genome budget will be available for activities that address ethical, social, and legal issues related to the project.

A series of specific recommendations for the research agenda and related activities in the ethics component of the human genome program has been developed by a joint DOE-NIH working group on ethics. These recommendations will guide the program over the next five years and will continue to be refined as the program proceeds. A complete report of the ethics working group is attached (Appendix 7).

The purpose of the ethics component of the human genome program is to:

- o address and anticipate the implications for individuals and society of mapping and sequencing the human genome;
- o examine the ethical, legal and social sequelae of mapping and sequencing the human genome;
- o stimulate public discussion of the issues;
- o develop policy options to assure that the information is used for the benefit of the individual and society.

The program will endeavor to anticipate problems before they arise and develop suggestions that would forestall adverse effects. The approach to accomplishing these objectives will be to:

- o stimulate research on the issues through grants;
- o refine the research agenda through workshops, commissioned papers, and invited lectures on specific topics selected by the working group on ethics;
- o solicit public testimony from the community at large through town meetings;
- o support the development of educational materials for all levels;
- o encourage international collaboration in this area.

5 YEAR GOAL: Develop programs addressed at understanding the ethical, legal and social implications of the human genome project.

Identify and define the major issues and develop initial policy options to address them.

5. Research Training

The Human Genome Initiative is creating the need for a considerable number of scientists and other trained personnel who have the skills to pursue the research goals and apply the information generated by the program. The ability of the U.S. research establishment and industry to take advantage of the products of the human genome project will require highly trained individuals.

Scientists with diverse expertise are required: geneticists and molecular biologists, as well as investigators from fields such as physics, chemistry, engineering, mathematics, and computer science. Critically needed are scientists with interdisciplinary skills -- those who understand the biological problem at hand and can find solutions by applying skills from other disciplines. Many more technicians also will be required to operate the large amount of technology that the genome program will employ.

The NIH *Ad Hoc* Program Advisory Committee on Complex genomes recommended that research training be an integral part of the human genome program. This recommendation has been reinforced by the current Program Advisory Committee on the Human Genome. In response to these recommendations, the following initiatives have been put in place:

Pre-doctoral training grants in genome research will support training of scientists with the skills needed to carry out basic and applied research related to the goals of the Human Genome Initiative and to apply that knowledge to solve important biomedical research problems. The focus of this training will be interdisciplinary, intended to give students a deeper understanding of how the methods and principles of one or more of the non-biological sciences can interact with those of biology to address research problems related to genome analysis.

Post-doctoral fellowships in genome research will provide support for training at the post-graduate level. In addition to the customary training for Ph.D. and M.D. degree holders in molecular biology and other areas relevant to genomics research, there will be an effort to attract individuals who wish to pursue interdisciplinary training. Candidates for these grants who are trained in mathematics, computer science, chemistry, physics, or engineering and who want to augment their skills in those fields with training in biological science to enable them to pursue genome research, will be encouraged. Conversely, biologists who want to acquire research training in biocomputation, instrumentation, biophysics, or other areas related to genome research will be desirable candidates. There also will be fellowship support for individuals interested in the ethical, legal, and social implications of genome research.

Senior fellowships will be available to experienced investigators in physics, mathematics, engineering, and biological, chemical or computer science who want to acquire training and experience in another discipline. It is expected that these senior fellows subsequently will use this additional training to develop and broaden their research interests to include problems related to genome analysis.

Training at National Laboratories will be supported by DOE and will be available for both pre-doctoral and post-doctoral individuals who want to learn techniques of genome research.

Short courses will also be needed to provide in-depth training in a defined area. These courses could address the need of individuals to enhance their skills in molecular techniques, computational sciences, and ethical or legal studies. The NCHGR and the DOE are currently studying the best ways to meet such needs.

5 YEAR GOAL: Support research training of pre- and post-doctoral fellows starting in FY 1990. Increase the numbers of trainees supported until a steady state of about 600 per year is reached by the fifth year.

Examine the need for other types of research training in the next year.

6. Technology Development

Although considerable strides have been made in technology development since the publication of the NRC and OTA reports, there is still a need for further innovation to adapt the technology to large-scale projects and to bring costs down. During the next five years, there will be an emphasis on technology development in all areas of the program.

Automation, optimization, cost reduction and other improvements will be supported in areas such as cloning technology, robotics, DNA sequencing, gel technology, software tools, and instrument development. Equally important will be the support of completely novel approaches, such as the use of scanning tunnelling microscopy or mass spectrometry for sequencing. The technology that ultimately will be used to sequence the human genome may turn out to be a method that is still on the drawing board.

5 YEAR GOAL: Support innovative and high-risk technological developments as well as improvements in current technology to meet the needs of the genome project as a whole.

7. Technology Transfer

Rapid transfer of the technology developed under the human genome program to industries that can develop economically and medically useful applications is a major goal of the project. This will occur in a variety of ways ranging from direct federally funded research at private companies to expedited transfer of new technology into the private sector. The human genome project is certain to spawn and nurture parallel efforts on a host of plant and animal genomes that are of direct commercial interest. Rapid provision of technology and trained personnel will play a most critical role in driving these efforts.

Industry will benefit directly from the availability of scientists trained by the human genome project and by the availability of databases that provide access to the data generated by the project. These databases will be used in many diverse ways to design products for medical and industrial applications.

In the coming year, a plan will be developed for technology transfer with respect to inventions produced by the genome project. A variety of mechanisms will be explored for facilitating this transfer, for improving information flow, and for identifying potential blocks to efficient transfer. The DOE National Laboratories are already working with private sector interests to establish cooperative ventures. The NIH intramural laboratories have similarly developed a system of cooperative research and development agreements with industry.

The biotechnology industry in the United States is strong and innovative and has very close ties to scientists doing genome research. Indeed, this industry will be a strong participant in all aspects of the project from the beginning. Representatives from industry sit on the advisory committees and industrial scientists have received numerous grants from both NIH and DOE. It is expected that industrial involvement will increase as the project proceeds, especially during the phase of large-scale sequencing.

Transfer of the technology into medical applications will be facilitated where necessary, but will also occur naturally. Many of the scientists supported by the

NCHGR human genome program are physicians or work closely with physicians who are involved in patient care.

The various institutes of the NIH all support research on diseases that result from genetic variation and a variety of mechanisms will be used to assure that information is transferred efficiently from the NCHGR to these institutes. A coordinating committee has already been established for this purpose. The NCHGR will be particularly alert to the need to stimulate preparation of reagents for use in the diagnosis and treatment of rare genetic diseases when such reagents may not be commercially viable.

5 YEAR GOAL: Enhance the already close working relationships with industry.

Encourage and facilitate the transfer of technologies and of medically important information to the medical community.

Implementation

Administration of Research

Management Systems

A number of management devices have been put into place to facilitate the administration of the human genome program. Both NIH and DOE have advisory and coordinating committees that provide overall planning and advice. These advisory committees have established a number of working groups to study in detail specific aspects of the program. Such working groups will be created as needed and will terminate as their work is complete.

A newsletter available to all interested parties has been started by the DOE. In the future, this newsletter will be published jointly by NIH and DOE. DOE and NIH will also make available an electronic bulletin board for rapid dissemination of information. Finally, an administrative database will be set up to capture information on genome research worldwide and enable us to track progress towards program goals. Cooperation between both agencies is planned for all these projects.

Role of Research Centers

Attaining the goals of the Human Genome Initiative will require research programs of varying magnitude and complexity. This nation's pre-eminent achievement in biomedical research is rooted in the traditional decentralized system of support for projects initiated by single investigators and small groups of scientists. Projects of this scope will contribute significantly to the

U.S. human genome program and will continue to represent a substantial fraction of the overall funding.

Many of the scientific projects envisioned as part of the genome program, however, can only be addressed by teams of investigators from various disciplines working collaboratively. For the research to proceed efficiently, sharing and exchange of equipment and other resources will be necessary, as well as coordinated collection and analysis of data.

The DOE has already established three genome centers within its National Laboratories. These centers are multidisciplinary and each has chosen specific human chromosomes as the focus for an intense and comprehensive effort of physical mapping. The centers also are engaged in the development of sequencing technology, informatics research, and studies of other novel technologies expected to contribute to the genome project. Each center has developed extensive collaborative networks with academia and with industry.

Starting in FY 1990, the NIH human genome program will also provide funding for the establishment and support of genome research centers. These centers will consist of groups of investigators from diverse disciplines who wish to come together to tackle a project they could not accomplish otherwise. Both academic institutions and industrial organizations will be eligible to compete for research center grants.

Research centers will be expected to address a major task of the genome project. For example, some centers may focus on the development of the complete physical map of a human chromosome, others on the sequencing of the complete genome of a model organism, or on the development and application of a particular technology. Research centers can provide a stable environment for large, long-term projects and facilitate the recruitment of new investigators into the program. Such centers also provide an excellent vehicle for collaborations with investigators who are not part of the center and with industry. It is expected that research centers will stimulate the coalescence of individual research efforts into collaborative networks.

In the next five years NIH expects to establish, depending on their size, ten to twenty research centers in various academic, non-profit and for-profit institutions. The funding of three centers is anticipated in fiscal year 1990. DOE may also establish additional centers in the National Laboratories.

Construction

Many of the institutions that have the strongest genome research programs suffer from lack of laboratory space available for expansion of these projects. The space limitation is both in the amount and kind of facilities needed for the specialized interdisciplinary activities involved, such as robotics and instrument development tied closely to biological activities. This is a serious problem that could delay progress at an optimal rate. Both NIH and DOE advisors have strongly urged the agencies to seek authority and funds for construction.

Workshops and Meetings

Because they will tie all research components together, research centers will in some sense represent the administrative centerpiece of the genome project. However, centers alone will not be sufficient. Another important administrative device is the organization of a variety of workshops and meetings designed to facilitate collaboration, assess the state of the art in a particular area and determine what actions are needed next.

A number of such workshops have been held in the past two years, with excellent success. For example, human chromosome-based workshops have been held for chromosomes 11,16, and X. Workshops on the human genetic map and on physical mapping databases have also taken place. In November, 1989, DOE held the first contractor/grantee workshop for investigators supported by its human genome program. In addition, meetings to begin organizing work on some of the model systems are underway for the mouse, drosophila, and yeast.

Some of these workshops were sponsored jointly by NIH and DOE, with contributions in several cases from the Howard Hughes Medical Institute. All workshops included participants from around the world. In each case there were a number of positive outcomes. Along with the exchange of information between investigators, mechanisms for the exchange of materials were established, collaborations were initiated, and agreements were reached on more uniform approaches among the different laboratories. As a result, greater efficiency on all the projects is anticipated.

Meetings of scientists with common research objectives will play an increasingly important role in terms of the coordination of the genome project. It is anticipated that mapping and sequencing groups will eventually coalesce around individual human chromosomes or around a model organism.

This appears at present to be the most logical and effective way to organize the project. However, such organization must evolve based on the commitment and efforts of individual scientists and cannot be centrally dictated at this early stage. Eventually, groups that assume leadership in each of these areas will emerge. In order to encourage coalescence, annual meetings of particular groups will be required in many cases. Both DOE and NIH will continue to organize and promote such workshops.

Role of NIH and DOE

The National Institutes of Health has a natural interest in the Human Genome Initiative in view of its long history of supporting research in genetics and molecular biology as an integral part of its mission to improve the health of all Americans. The Department of Energy has a long-standing program of genetic research directed at improving the ability to assess the effects of radiation and energy-related chemicals on human health. In recognition of these complementary interests, NIH and DOE have agreed to coordinate their individual genome activities.

National Institutes of Health

The human genome program of the National Institutes of Health was formally established after Congress appropriated earmarked funds to NIH in fiscal year 1988 to conduct research on mapping and sequencing of the human genome.

In October 1988, the Office of Human Genome Research was established to plan and coordinate NIH genome activities in cooperation with other federal agencies, industry, academia and international groups. As of October 1, 1989, the office became an independent funding unit within the NIH with authority to award grants and contracts and was renamed the National Center for Human Genome Research (NCHGR).

To provide ongoing advice from scientific experts and industry representatives, NIH established a permanent Program Advisory Committee on the Human Genome (PACHG) and, because virtually all of the institutes of NIH are involved in research that interacts with the human genome program, an internal NIH Coordinating Committee on the Human Genome also was formed.

While most of the research supported by the NIH genome program will take place at academic, non-profit, or for-profit institutions across the country, relevant intramural studies also will be considered for funding under the program.

U.S. Department Of Energy

The genome program of the Department of Energy started in fiscal year 1987 on a small scale and received earmarked funds for the first time in the fiscal year 1988 appropriation.

DOE's genome activities are represented mainly by multidisciplinary programs under way at three National Laboratories: Lawrence Berkeley Laboratory; Los Alamos National Laboratory; and Lawrence Livermore National Laboratory. Additional projects are supported at other National Laboratories, at universities, and in the private sector.

Oversight of DOE human genome activities is provided by the Health and Environmental Research Advisory Committee (HERAC). The Office of Health and Environmental Research (OHER), assisted by a steering committee representing the three National Laboratories and extramural grantees, manages the program and administers grants and contracts.

Memorandum of Understanding

Mechanisms for the coordination of human genome activities between DOE and NIH are specified in a 1988 Memorandum of Understanding. A joint advisory subcommittee was established to monitor and coordinate programs. Furthermore, there is extensive formal and informal interagency contact between program administrators. Panels convened by DOE or NIH to review genome research proposals, to assist in program coordination or to provide advice, are attended by representatives of both agencies, and regular joint workshops and meetings on genome-related issues are held.

The NIH and the DOE have had an excellent working relationship with regard to the human genome program in the past and expect that this relationship will become even closer and more useful in the future. The establishment of a joint informatics task force, a joint working group on ethical, legal, and social issues, and a joint mapping working group, in addition to the joint advisory subcommittee called for in the MOU, attest to the close cooperation. Additional joint working groups will be established

as needed. Each of these groups will provide information and advice to the parent advisory committees of both agencies.

In August, 1989, a group of NIH and DOE advisors met together with selected other experts to develop a joint plan for the genome project for the next five years. This plan was approved by the advisory committees of both agencies. Each agency will implement its genome program according to this overall scheme. Because of the success of the joint planning exercise and the need for frequent updates, the two agencies will repeat this process at regular intervals to assure continued close coordination.

Although there are areas of overlapping interest between DOE and NIH, there are also clear areas of distinction, based on the respective agency's interests and strengths. The following highlights major similarities and differences.

- Both agencies are interested in the physical map of the human genome as well as in the development of sequencing technology. These are very large objectives that will require strong efforts by both agencies. Since the methods for achieving this goal are still under development and rapidly changing, multiple parallel approaches are necessary and desirable. To this area the DOE brings extensive expertise in technology development and computer sciences, while NIH brings a wealth of biological and medical expertise. The combination of these approaches through collaboration is most likely to lead to the best results.

- The NIH genome program includes the study of the genomes of model organisms. These studies will contribute to technology development and to the interpretation of the human DNA sequence.

- NIH is devoting considerable effort to the completion of the human genetic map, which makes an important contribution to physical mapping and sequencing, and constitutes a valuable tool for scientists seeking to identify and eventually isolate disease genes.

- NIH and DOE will share responsibility for dealing with the ethical, legal, and social implications of the genome project.

- NIH is taking major responsibility for funding public databases, such as GenBank, that will be the ultimate repositories for information generated by the project. This is an area within the congressional mandate of the recently created National Center for Biotechnology

Information (NCBI) of the National Library of Medicine. The National Center for Human Genome Research will collaborate with NCBI in this area.

- Both DOE and NIH are vigorously pursuing the development of electronic notebooks, algorithms for analyzing DNA sequences and the development of physical mapping and other databases that are closely integrated with research protocols and required to accomplish the work.

- NIH will support a comprehensive training program for pre-doctoral, post-doctoral, and senior investigators, as well as a career development program. These programs will address current and future needs of genomics research.

- DOE will provide support for pre-doctoral and post-doctoral individuals training at the National Laboratories.

- The medical applications of genomic research will be a special strength of the NIH program. Although these are not part of the immediate objectives of the genome project, which is focused on producing basic mapping and sequencing information, most of the investigators supported by NIH are closely tied to the medical research community or are themselves conducting such studies along with their genome research.

- DOE, which has close ties with the physics, computation, and engineering communities, will emphasize the transfer of technology in these areas.

While the list of similarities and differences is instructive for showing the great diversity of activities that are included in the genome project, the key to the DOE-NIH relationship is the fact that both agencies are working from the same blueprint. Over the past two years a great deal of synergism has developed between the two agencies, with productive collaborations established between DOE laboratories, NIH supported investigators, and industry.

Role of other Federal Agencies

Because the information to be derived from mapping and sequencing the human genome will be of very broad interest and applicability, a number of other federal agencies are involved in funding and carrying out activities related to the Human Genome Initiative.

National Science Foundation

The National Science Foundation (NSF) is interested in the support of projects focused on the scientific infrastructure for genome-related activities. Specific NSF activities have included funding in FY 1989 of a science and technology center dedicated to new technologies for DNA and protein chemistry. NSF is also involved in development of new software and algorithms for database searching and development of special-purpose hardware to increase the speed of biological database searches. Recently, the NSF decided to start a program for mapping and sequencing the genome of the model plant system *Arabidopsis thaliana* in collaboration with NIH and other agencies. This system will be an excellent one for developing and testing technology. NSF representatives regularly attend the NIH advisory committee and DOE steering committee meetings as liaison members to assure coordination of the programs.

U.S. Department of Agriculture

A growing interest in mapping and sequencing the genomes of plants important to agriculture and forestry led the U.S. Department of Agriculture (USDA) to establish an Office of Genome Mapping after a planning conference in December 1988. A coordinating committee was formed to devise the goals and scope of USDA's plant genome efforts, which are planned to extend over 10 years at an estimated cost of $500 million. Plant genes that confer pest and disease resistance as well as drought tolerance, along with other gene systems of economic importance, will be selected for mapping and sequencing.

The USDA's Agricultural Research Service also has an active animal science division that is interested in genome research. This is expected to be a growing area within USDA. A liaison member from the USDA attends the NIH Program Advisory Committee meetings and the DOE Human Genome Steering Committee meetings, and NIH and DOE staff have attended the various USDA planning meetings. As the USDA program proceeds, closer ties will be established.

Howard Hughes Medical Institute

The Howard Hughes Medical Institute (HHMI) has played an important role in supporting research and databases of importance to the genome project. Both DOE and NIH have worked with HHMI to coordinate activities and a representative of HHMI has attended almost all functions sponsored by one or both of the agencies. HHMI has been able to identify a role for itself in areas that are difficult for federal agencies to support, such as the critical funding provided to help the Human Genome Organization (HUGO) get started (also see below).

International Collaboration

The Human Genome Initiative is not limited to the United States. Many countries are interested in participating in the project and all are interested in the outcome. Programs with funding are currently underway in the United Kingdom (UK), Italy, and the Soviet Union. Funding is expected in the near future from the Commission of the European Community (EC), France, and Japan. However, all these programs are small compared to the U.S. program and are currently in the early stages of organization.

An association of interested scientists from around the world has been formed and incorporated as the Human Genome Organization. This organization plans to develop a number of activities to assist with the international coordination of the various national programs.

While NIH and DOE support HUGO and believe it could be most helpful as a facilitator, international interaction is already proceeding well. Individual investigators have formed numerous collaborations across national lines, almost all genome meetings are international in scope, and the staff responsible for the management of the various national programs have established good lines of communication.

For example, NIH staff has been represented at meetings of the EC Working Party and planning meetings in the UK. Both NIH and DOE representatives have attended planning meetings in Italy, Spain, the USSR, and Japan. The EC Working Party and the Medical Research Council in the UK, as well as Canada, have sent representatives to the NIH Program Advisory Committee meetings.

NIH and DOE have a policy of welcoming international collaboration in the basic research aspects of the human genome project. Because it is desirable to encourage other countries to contribute financially to this project, the agencies have decided that they will, in general, not fund a foreign research project unless it will make a unique contribution that cannot readily be duplicated in the United States. The agencies will, however, fund joint research projects between the United States and another country if there is also joint funding from the other country.

There are many opportunities where international collaboration could enhance progress on the Human Genome Initiative. Currently, the United States is in a leadership position with respect to scientific accomplishment and organization of the genome program. However, as other nations organize and initiate their programs, the United States will stand to gain by international collaboration as much as the other countries involved.

Budget

The budget for the two agencies follows:

	FY 1988 Actual	FY 1989 Actual	FY 1990 Estimate	FY 1991 Estimate
NIH	17.2	28.2	59.5	108.0
DOE*	10.7	18.5	27.9	47.9
Total	27.9	46.7	87.4	155.9

*does not include salaries and expenses of DOE employees devoted to this effort

The original cost estimates by the National Academy of Sciences and the Office of Technology Assessment were that a level of funding of approximately $200 million per year for about 15 years would be needed to complete the human genome project. No effort has been made at this early stage to revise or update these figures. Fiscal years 1988 through 1990 have been a period of getting organized and getting research underway. The five-year goals proposed in this document are for the period FY 1991 through FY 1995 and assume that a funding level of $200 million per year with inflationary increases can be reached rapidly. Only at this level will the critical mass of people and research projects be achieved that can move the human genome program forward at an optimal rate.

Contributions to the program by other countries are welcome. However, such contributions should not be viewed as decreasing the need for a critical level of activity in the United States. Rather, they will shorten the time needed to complete the project. The funding levels originally recommended by the OTA and the NRC are required to provide optimal benefit to the American research enterprise and to American industry.

The need for money for new construction was identified in the NRC report, although no specific figure was given. NIH has estimated that a sum of $100 million over five years will be needed to make space available for the expansion of research center activities. DOE estimates that $21 million will be needed to construct additional space at its centers.

Since technology and costs are still changing rapidly, it is hazardous to assign precise costs to specific areas. However, an approximate breakdown into categories over the next five years is currently estimated as follows:

45 to 55%	for human studies and associated technology and computational development;
20 to 30%	for model systems and associated technology and computational development;
20 to 30%	for infrastructure, including research on and maintenance of public databases; research training; technology development not included in one of the other areas; ethics; conferences; and administration.

It would be counterproductive to fix a particular budget distribution at this time in such a rapidly moving field, where relative costs are also changing constantly. Flexibility will be essential so that unexpected opportunities can be pursued effectively. Every effort will be made to complete the project as economically as possible.

Appendixes

U.S. DEPARTMENT OF ENERGY
HEALTH AND ENVIRONMENTAL RESEARCH
ADVISORY COMMITTEE

Committee Chairman

Dr. Sheldon Wolff
Director
Laboratory of Radiobiology and
Environmental Health
University of California
San Francisco, CA 94143

Committee Members

Dr. Vera Alexander
Director
Institute of Marine Sciences
University of Alaska
Fairbanks, AK 99775-1080

Dr. Kelly H. Clifton
Department of Human Oncology
University of Wisconsin
K-4/312 Clinical Science Center
600 North Highland Avenue
Madison, WI 53792

Dr. G. Sam Hurst
Institute of Resonance
Ionization Spectroscopy
University of Tennessee
10521 Research Drive
Suite 300
Knoxville, TN 37932

Dr. Kenneth K. Kidd
Department of Human Genetics
Yale University School of Medicine
New Haven, CT 06510

Dr. Leonard S. Lerman
Department of Biology
Massachusetts Institute of
Technology
77 Massachusetts Avenue
Cambridge, MA 02139

Dr. Simon A. Levin
Section of Ecology and Systematics
Director, Center for
Environmental Research
Cornell University
Corson Hall
Ithaca, NY 14853

Dr. James R. Mahoney
Director
National Acid Precipitation
Assessment Program
722 Jackson Place, N.W.
Washington, D.C. 20503

Dr. J. Justin McCormick
Carcinogenesis Laboratory
B-620 Fee Hall
Michigan State University
East Lansing, MI 48824

Dr. Mortimer L. Mendelsohn
Associate Director for Biomedical and Environmental Research
Lawrence Livermore National Laboratory
P.O. Box 803
Livermore, CA 94550

Dr. Mary Lou Pardue
Department of Biology
Room 16-717
Massachusetts Institute of Technology
Cambridge, MA 02139

Dr. Theodore L. Phillips
Department of Radiation Oncology
University of California School of Medicine
San Francisco, CA 94143

Dr. Richard C. Reba
Director, Division of Nuclear Medicine
The George Washington University Medical Center
The University Hospital
901 Twenty-third Street, N.W.
Washington, D.C. 20037

Dr. Paul G. Risser
Vice President for Research
University of New Mexico
Scholes Hall 108
Albuquerque, NM 87131

Dr. Kenneth Rothman
P.O. Box 57
Chestnut Hill, MA 02167

Dr. Robert E. Sievers
Department of Chemistry
University of Colorado
P.O. Box 215
Boulder, CO 80309-0215

Dr. Ignacio Tinoco
Department of Chemistry
University of California
430 Latimer Hall
Berkeley, CA 94720

Dr. Audrey V. Wegst
Diagnostic Technology Consultants, Inc.
4747 Troost Avenue
Kansas City, MO 64110

Dr. Harel Weinstein
Chairman, Department of Physiology and Biophysics
Mt. Sinai School of Medicine
Annenberg 2186
1 Gustave L. Levy Place
New York, NY 10029

Department of Energy Staff

George D. Duda
Executive Director

Jean M. Hummer
Executive Secretary

NATIONAL INSTITUTES OF HEALTH
PROGRAM ADVISORY COMMITTEE ON THE HUMAN GENOME

Chairman

Norton D. Zinder, Ph.D.
John D. Rockefeller, Jr. Professor
The Rockefeller University
1230 York Avenue
New York, New York 10021-6399

Executive Secretary

Elke Jordan, Ph.D.
Deputy Director
Center for Human Genome Research
National Institutes of Health
Building 1, Room 201
Bethesda, MD 20892

Members

Bruce M. Alberts, Ph.D.
Chairman
Department of Biochemistry and Biophysics
University of California San Francisco, Box 0448
San Francisco, CA 94143-0448

David Botstein, Ph.D.
Vice President
Genentech, Inc.
460 Point San Bruno Boulevard
South San Francisco, CA 94080

Jaime G. Carbonell, Ph.D.
Associate Professor
Computer Science Department
Carnegie-Mellon University
Wean Hall, Room 4212
Pittsburgh, PA 15213

Joseph L. Goldstein, M.D.
Chairman
Dept. of Molecular Genetics
University of Texas
Southwestern Medical Center
5323 Harry Hines Boulevard
Dallas, TX 75235-9046

Leroy E. Hood, M.D., Ph.D.
Chairman
Division of Biology, 147-75
California Institute of Technology
1201 East California Boulevard
Pasadena, CA 91125

Victor A. McKusick, M.D.
University Professor
Division of Medical Genetics
Johns Hopkins Hospital
600 N. Wolfe St., Blalock 1007
Baltimore, MD 21205

Maynard V. Olson, Ph.D.
Professor
Department of Genetics
Washington University School
of Medicine
P.O. Box 8031
4566 Scott Avenue
Saint Louis, MO 63110

Mark L. Pearson, Ph.D.
Director, Molecular Biology
Central Research & Development
Department
E. I. Du Pont de Nemours and
Company
P.O. Box 80328
Wilmington, DE 19880-0328

Cecil B. Pickett, Ph.D.
Executive Director of Research
Merck Frosst Centre for
Therapeutic Research
16711 Trans Canada Highway
Kirkland, PQ H9H 3L1
Canada

Phillip A. Sharp, Ph.D.
Professor and Director
Center for Cancer Research
40 Ames Street, Room E17-529B
Massachusetts Institute of
Technology
Cambridge, MA 02139

Nancy S. Wexler, Ph.D.
President
Hereditary Disease Foundation
Associate Professor
Dept. of Neurology & Psychiatry
Columbia Presbyterian Medical
Center
722 West 168th Street, Box 58
New York, NY 10032

JOINT DOE-NIH SUBCOMMITTEE ON THE HUMAN GENOME

Co-Chairs

Sheldon Wolff, Ph.D.
Laboratory of Radiobiology
and Environmental Health
LR 012
University of California
3rd and Parnassus Avenue
San Francisco, CA 94143-0750

Norton D. Zinder, Ph.D.
John D. Rockefeller, Jr. Professor
The Rockefeller University
1230 York Avenue
New York, New York 10021-6399

Members

David Botstein, Ph.D
Vice President
Genentech, Inc.
460 Point San Bruno Boulevard
South San Francisco, CA 94080

Charles Cantor, Ph.D.
Department of Genetics and
Development
College of Physicians & Surgeons
Columbia University
New York, NY 10032

Jaime G. Carbonell, Ph.D.
Associate Professor
Computer Science Department
Carnegie-Mellon University
Wean Hall, Room 4212
Pittsburgh, PA 15213

Anthony Carrano, Ph.D.
Lawrence Livermore National
Laboratory
Biomedical and Environmental
Science Division
P.O. Box 5507
Livermore, CA 94550

Leonard Lerman, Ph.D.
Department of Biology
Building 56-743
Massachusetts Institute of
Technology
Cambridge, MA 02139

Robert Moyzis, Ph.D.
Center for Human Genome Studies
MS M886
Los Alamos National Laboratory
Los Alamos, NM 87545

MaryLou Pardue, Ph.D.
Department of Biology
Room 16-717
Massachusetts Institute of
Technology
Cambridge, MA 02139

Maynard V. Olson, Ph.D.
Professor
Department of Genetics
Washington University School
of Medicine
P.O. Box 8031
4566 Scott Avenue
Saint Louis, MO 63110

Mark L. Pearson, Ph.D.
Director, Molecular Biology
Central Research & Development
Department
E. I. Du Pont de Nemours and
Company
P.O. Box 80328
Wilmington, DE 19880-0328

Nancy S. Wexler, Ph.D.
President
Hereditary Disease Foundation
Associate Professor
Dept. of Neurology & Psychiatry
Columbia Presbyterian Medical
Center
722 West 168th Street, Box 58
New York, NY 10032

MEMORANDUM OF UNDERSTANDING

BETWEEN THE

UNITED STATES DEPARTMENT OF ENERGY

AND THE

NATIONAL INSTITUTES OF HEALTH

TO COORDINATE RESEARCH AND TECHNICAL ACTIVITIES

RELATED TO THE HUMAN GENOME

I. Introduction

The National Institutes of Health (NIH), Department of Health and Human Services, and the United States Department of Energy (DOE) agree to foster interagency cooperation that will enhance the human genome research capabilities of both agencies.

DOE and NIH are the Federal Agencies primarily responsible for supporting research relating to the human genome. There has been considerable discussion in the scientific community over the past two years about the need for a coordinated long-term project to map and sequence the human genome. While NIH and DOE have informally coordinated such research efforts, the increasing complexity and scope of the project require a more formal mechanism. The purpose of this Memorandum of Understanding (MOU) is to provide for the formal coordination of the activities of DOE and NIH, and to provide for interfaces with relevant activities both within and outside the United States. The MOU also provides a mechanism by which NIH and DOE can jointly obtain outside advice regarding the human genome project.

II. Definition

For the purposes of this MOU, human genome research encompasses efforts to develop and apply technologies for the large-scale mapping, sequencing and analysis of the human genome. It includes the development of shared centralized facilities such as repositories for cloned DNA fragments, databases, and data centers to collect and distribute the large amounts of information generated on the project.

III. Goals

The goals of the project include: completion of a high-resolution genetic map of the human genome; completion of a series of complementary physical maps of increasing resolution; acquisition of a collection of ordered DNA clones encompassing the entire genome; determination of the complete nucleotide sequence of a reference genome; location of all the genes; and development of the tools to use the above information for a variety of biological and medical applications. Parallel studies in model organisms will be required in order to achieve a full understanding of the human genome.

IV. Management and Program Guidelines

A. Establishment of a joint advisory subcommittee chosen from the members of the DOE Health and Environmental Research Advisory Committee and the NIH Program Advisory Committee on the Human Genome.

The joint subcommittee will receive charges jointly prepared by NIH and DOE and communicated to their appropriate parent advisory committees. The joint subcommittee shall be co-chaired by representatives from the DOE and NIH committees. The joint subcommittee shall meet quarterly in order to advise and review the relevant activities of the two agencies. Subcommittee reports will be delivered through the two parent advisory committees to appropriate senior officials of NIH and DOE.

B. Establishment of an Interagency Working Group (IAWG) on genome research between DOE and NIH. The IAWG will be co-chaired by NIH and DOE and will meet at least on a quarterly basis to explore the need for and the feasibility of initiating a variety of cooperative

and complementary programs and projects in order to advance knowledge in human genome research. The IAWG will also provide oversight of activities carried out under this MOU. In addition to the chairpersons, the IAWG will consist of an equal number of full members from DOE and NIH. Additional *ad hoc* members may be added for temporary assignments by either agency with prior concurrence of the chairpersons.

C. Continued coordination with other Federal agencies, with outside scientific groups, both national and international, and with private organizations involved in the genome project.

D. Continued joint participation and sponsorship of meetings and workshops for the purposes of planning and review of technical progress including an annual symposium to review progress in the science, to identify areas of need, and to address general policy questions.

E. Development of synchronous calendars for the agencies' research award cycles.

F. Concurrent funding and management of selected programs in human genome research that require utilization of unique NIH or DOE facilities.

G. Maintenance of regularly scheduled joint program staff meetings to exchange program information and plans.

H. Promotion of the sharing of technological advances and relevant biological materials (probes, cell lines, etc.) among investigators supported by both agencies. Assurance that relevant data are rapidly placed in appropriate databases and that relevant biological materials are rapidly placed in appropriate repositories.

I. Promotion of coordination and exchange of data with other countries.

J. Advance sharing of public policy statements relevant to human genome research.

V. Administration

A. Public Information Coordination: Subject to the Freedom of Information Act (5 U.S.C. 552), decisions on disclosure of information to the public regarding projects and programs implemented under the Memorandum of Understanding will be made following consultation between DOE and NIH representatives.

B. Intellectual Property: Specific provisions concerning the disposition of rights in intellectual property will be included in any interagency agreement under this Memorandum of Understanding.

C. Amendment and Termination: This Memorandum of Understanding may be modified or amended by written agreement between NIH and DOE and terminated by mutual agreement of DOE and NIH or by either party upon 90-day written notice to the other.

D. Effective Date: This Memorandum of Understanding is effective when signed by both parties.

James B. Wyngaarden
James B. Wyngaarden
Director
National Institutes of Health

Sept 30, 1988
Date

Robert O. Hunter, Jr.
Robert O. Hunter, Jr.
Director
Office of Energy Research
U. S. Department of Energy

October 7, 1988
Date

ADDITIONAL PARTICIPANTS IN THE DEVELOPMENT OF THE FIVE-YEAR PLAN

Dr. C. Thomas Caskey
Institute for Molecular Genetics
Baylor College of Medicine
One Baylor Plaza, T809
Houston, TX 77030

Dr. Francis S. Collins
University of Michigan
4708 Medical Science II
Ann Arbor, MI 48109-0618

Dr. Uta Francke
Howard Hughes Medical Institute
Beckman Center
Stanford University Medical Center
Stanford, CA 94305-5428

Dr. Eric Lander
Whitehead Institute for
Biomedical Research
9 Cambridge Center
Cambridge, MA 02142

Dr. Shirley Tilghman
Molecular Biology Department
Princeton University
Princeton, NJ 08544

Dr. Raymond White
University of Utah
Department of Human Genetics
Salt Lake City, UT 84132

JOINT INFORMATICS TASK FORCE PROPOSAL

Background

What is genome informatics? Genome informatics is a scientific discipline that encompasses all aspects of genome information acquisition, processing, storage, distribution, analysis, and interpretation. This activity combines the tools and techniques of mathematics, computer science, and biology to produce a variety of molecular maps of genomes, including DNA and protein sequences, with the aim of understanding the biological significance of such data.

The large amount and extraordinary complexity of mapping and sequencing data to be generated by the human genome project requires the development of a coordinated national program to make the information and analysis tools from this project freely available to the widest possible range of scientists and physicians in the most useful, timely, and cost-effective fashion.

Recommendation

Members of the Database Working Group of the NIH Program Advisory Committee for the Human Genome Project, the Informatics Task Force of the DOE Human Genome Steering Committee and the GenBank Advisory Panel of the National Institute of General Medical Sciences, plus representatives of the National Science Foundation and the Howard Hughes Medical Institute, met in November, 1989, to discuss how this might best be accomplished. They recommend the establishment by the NIH and DOE genome advisory committees of a national Joint Informatics Task Force (JITF) that would supersede the current separate genomics informatics working groups.

Mission of the JITF

The mission of the JITF should be to identify user needs, to set genome informatics goals, to establish research and development priorities, to enhance the effectiveness of computational solutions to genome informatics problems, and to make funding recommendations to the NIH and DOE human genome committees in both technical and policy areas relating to:

- o genome database structures, management, and services;
- o informatics tool development, including algorithms, software, and hardware for organization and analysis;
- o data exchange standards;
- o electronic networks for collection and distribution of genome information;
- o training and education of informatics personnel;
- o coordination of genome informatics activities among laboratories, agencies and nations.

Membership and Representation on the JITF

The membership of the JITF, made up of recognized experts in the fields of molecular biology and computer science from academia, government, and industry, should be selected and approved by the DOE and NIH parent committees. The members should be chosen particularly for their breadth of understanding of the biological and computational issues facing the human genome project. The role of the working group would be to provide general wisdom about computational issues, and to provide access to particular knowledge about specific areas (eg., chip design, genetic linkage analysis, object-oriented databases) by constituting special subcommittees. Although some effort should be made toward balancing the expertise of working group members, the major focus should be on breadth, perspective, and practical experience with related projects.

A standing working group of 8-10 members with overlapping and rotating terms and a chairman reporting to the parent NIH and DOE committees is recommended. *Ad hoc* advisory panels will be convened by the JITF to deal with specific technical and policy issues. The JITF should represent the United States interest in dealings with international informatics groups, such as the Human Genome Organization (HUGO). Meetings of the JITF should be quarterly, or more frequently if requested by the parent committees. Funding for the JITF activities should be shared by the NIH and DOE. Representatives of other government (eg. NSF, USDA, FDA) and private (eg. HHMI, HUGO) agencies with responsibilities and activities in the area of genomic informatics should be invited to attend as liaison members of the JITF.

Near-term Goals

The immediate goals of the JITF will be to facilitate the implementation of the objectives of the five-year plan for the human genome project, namely:

- to support the rapid acquisition, database management, and public dissemination of genetic maps and DNA and protein sequence information;
- to develop effective software and database designs to support large-scale mapping and sequencing projects;
- to create database tools that provide easy access to up-to-date physical mapping, genetic mapping, chromosome mapping, and sequence information that also allows ready comparison of data among these datasets.

Future Goals

We outline here our current perceptions of some of the informatics issues to be considered by the JITF regarding database connectivity, informatics coordination, and networking. With the goals of expediting biological research by facilitating the sharing of data and software, the JITF must address issues of standardization of nomenclature, data exchange, and communication network protocols as well as database schema integration.

Connections Between Databases

For each of the 24 unique human chromosomes, the DNA sequence constitutes a unique line of connectivity to which other physical, structural, and genetic data are referenced. Datasets collinear with the nucleotide sequence include:

- the physical maps
- ordered clone maps
- the genetic map
- the cytogenetic map
- polymorphisms
- experimentally identified genes
- protein-coding sequences
- links to protein structure databases

- o links to maps of other species
- o links to the bibliographic, patent, and other factual databases.

The requirements of the genome database will stress the limits of present day informatics technology. Some of these challenges are specific to genome informatics, while others are more broadly applicable. In addition to the current generation of sequence databases, a variety of other useful, more specialized biological databases are being developed, which should be connected logically to the sequence databases. These include sequence database subsets concerned with protein superfamilies having common structural motifs and DNA regulatory sequences, for example, as well as supersets describing biosynthetic pathways. Recent discussions of the value of sequence tagged sites (STSs) in defining physical maps with greater precision, in facilitating the direct comparison of mapping results using different techniques, and ultimately in connecting physical maps with DNA sequence, further illustrate the need to develop such connections between the emerging databases.

Genome databases will be essential information resources for a diverse set of user communities with diverse requirements. Biochemists will use the sequence and protein data to plan novel genetically engineered organisms that can produce scarce, medically significant human proteins and hormones. Medical practitioners and researchers will use the genetic and sequence data to discover, diagnose, understand, and treat the many human genetic diseases that exist. Detailed knowledge of metabolic pathways may make it possible to design treatments for many diseases that arise from over- or underactivity in a normal pathway, such as atherosclerosis, and autoimmune disorders such as arthritis. Evolutionary biologists will use data on homology relationships among sequences to understand the mechanisms of protein evolution, which will in turn lead to insights in protein function that can be translated into protein design methods for medicine and industry. This list of applications is necessarily incomplete, in part because many of the applications of such a database cannot be foreseen, and perhaps not even imagined at this time. Nonetheless, it is clear that the human genome information resource will be an invaluable asset to many areas of society for generations to come.

Coordination of the Informatics Effort

The human genome project will involve the development of a variety of databases and analytical software tools in a variety of laboratories. To a large extent these systems will be developed locally to support the experimental projects that require them, but there is a recognized need to share these resources wherever feasible, thereby increasing the effectiveness of the national effort.

Data

Coordination of data refers to measures that can be adopted to facilitate the exchange of data between groups and the integration of databases. Such measures include developing conventions for representing data at several difference levels. The levels are:

- o database schema
- o nomenclature
- o exchange format
- o database management systems (DBMS).

Exchange of data is the simplest task, and requires the establishment of conventions for the first three levels. Database integration is more demanding, and imposes the additional fourth level. We discuss each level in detail below:

Database Schema Level: The design of a database specifies what kinds of objects are represented, what kinds of attributes they can have and how they can be related to each other. For any problem many alternative designs are possible, and incompatibilities in the way different databases represent similar concepts can pose obstacles to exchange and integration of data. In order to minimize incompatibilities and maximize the reuse of database design effort across databases with similar representational problems, the JITF should consider establishing guidelines for database designs.

Nomenclature Level: Biological nomenclature is important because the ability to relate information in different databases ultimately depends on being able to determine what names each system uses for the same objects (see the DOE-NIH Nomenclature for Physical Mapping Report, for example). While traditional biological nomenclature can be irregular and even inconsistent, we expect that in the future, "official" naming systems will arise for various classes of objects, and these systems will most likely be the responsibility of the groups that maintain databases for such objects. The aim is to have an "official" nomenclature that is stable, maintained by a well-defined social entity, and is readily available in computer readable form. These nomenclatures can then be identified for use as standards in the construction of other databases.

Data Exchange Format Level: Standard formats for representing data in flat files are essential for the exchange of data between databases. A more complex approach involves using a formal language to describe file formats, and including the description with the data. The receiver then needs

software to turn the format description into a program which can be used to read the data -- ASN.1 (ISO 8824) is one language for specifying abstract syntax. By describing abstract data objects first, the description can be applied equally well to flat file formats, networking data exchange, transaction formats, and database schemas.

DBMS Level: Integration of databases can occur in at least two ways: physical and virtual. Physical integration means that the data from disparate sources are shipped to a central repository and combined there into a single database. This is essentially a problem of data exchange, so if the three data exchange levels discussed above have been properly standardized, physical integration should present no additional difficulties. Virtual integration means the establishment of a software "front-end" or interface, which will provide the appearance of a single database, while in fact a number of databases, typically at difference network sites, may actually be accessed to answer queries. Virtual integration is tremendously facilitated by the use at all sites of the same database product, particularly if the product is designed to support such integration.

Software Coordination

The problems of software exchange and integration are considerably more difficult than the corresponding problems on the data side. There are a number of reasons for this. First is the problem of the social context: software exists within a community of users of varying degrees of sophistication, including usually at least one "hacker" without whom the system is frequently unusable. It is very hard to share a system without sharing the hacker. The effort involved in making a system exportable to other users, including the development of clear user's manuals, installation guides, bug reporting and maintenance procedures, on-line help and easy-to-use interfaces, represents a huge additional cost increment beyond the minimal development cost necessary for in-house use. Laboratories whose primary focus is doing genome research cannot be expected to take up the burden of software production as well. Corporations might, if there were a sufficient market, but genome research may be in the position of "orphan drugs" in this regard.

A second issue is the problem of hardware and operating system (OS) compatibility. Numerous debates in the working group failed to produce a consensus on recommendations for hardware or OS standards. Until such a standard does emerge, various software engineering techniques can be applied within projects to minimize the difficulties associated with porting. These include careful

modularization of hardware and OS dependencies, the use of common languages, such as FORTRAN or C, and the provision of programmatic interfaces to all significant "back-end" functionality, so that a new interface can easily be pasted onto a different machine.

At present, mapping and pilot sequencing efforts in the human genome program are quite individualistic, so the goal of adopting a completely general software package in not realistic. This means that multiple software systems will be developed. Since only five or six institutions currently combine both the biological and the computational expertise needed for the human genome program, these institutions should be supported at a level that will allow efficient progress in the development of parallel, hopefully complementary, informatics packages that could be implemented and beta-site tested by groups without this expertise.

Only after considerable experience is gained, and efforts at different locations are compared, will it be possible to select those software approaches that merit reworking to a point where they can be more broadly distributed and implemented at many sites. This unification will form the major challenge of the informatics component in the middle stage of the human genome project.

Tools for Analysis of Map and Sequence Data

Analysis tools are in fact needed at all stages of the genome project. For example, in the construction of contigs and restriction maps for clone fingerprint data, we already need more powerful methods of analyzing redundant, even conflicting, data from multiple overlapping clones. At present, the most powerful method for predicting the function of an unknown sequence is by comparison with all known sequences, and a sophisticated algorithm like FASTA with GenBank can now take on the order on an hour. However, as the database grows a hundred-fold or more in the later stages of the genome project, better algorithms and hardware will be needed for such comparisons. In addition, as the sequence data become abundant, understanding their biological significance will also require improved methods for predicting functional genes, regulatory motifs, chromosome structural features, etc. Thus the development of new and more powerful computational tools required for assimilation and comprehension of the data to be generated in the human genome program will be an integral part of the project. Accordingly, the JITF should consider mechanisms to promote awareness of the needed tools among mathematicians and computational scientists, and to ensure that new developments in this area are promptly known and widely available to the genome community.

Networking for the Human Genome Research Community

Computer literacy and utilization within the human genome community are not uniform. More effective use of existing networks for activities ranging from electronic mail to database access to systems development is a clear priority. Capabilities that are commonplace in some laboratories present major time and cost hurdles to other researchers interested in the results but not the technology. At this early stage of the human genome project, the JITF can provide an important benefit to the community by leading the way with information and support for individual labs that wish to exploit existing network and computing technologies. The value of networking to progress in the genome project will increase rapidly as the data itself grows--the more complete the dataset, the more useful it will become to a larger numbers of users. The initial effort made by a small group should be development of an easy-to-use guide that would indicate available options, realistic benefits, recommendations on hardware and software (minimal, alternate, easy-to-use), costs, personnel requirements, and other sources of detailed information. No effort should be made to change existing networks or add new capabilities, develop hardware, software, or even training material.

There are several levels of "connectivity" (i.e., participation) to the networks. Initials efforts should be focused on labs trying to climb the lower levels. Generalization becomes more difficult at the higher levels. A limited characterization of the different levels follows:

> **Level 1 - Electronic Mail and Usenet News:** This is the minimal networking level, which needs to be accessible to every researcher and funded as part of research costs. Researchers are connected via terminal (or emulator running on a PC or Mac with a modem) to a local computer supporting electronic mail and attached to one of the standard networks (arpanet, bitnet, nsfnet, etc.) Individuals can login to the local computer to use electronic mail, and can participate in the usenet news bulletin boards. Local support is now necessary for this access, but should be part of the nationwide genome support funding. A list of the major newsgroups, bulletinboards, etc., will be a part of the guide.
>
> **Level 2 - Remote Access to Databases:** Level 1, plus researchers can access remote databases but cannot update them without permission. Using no more than the Level 1 hardware, public databases are accessible with little more than an account number and vendor- or owner-supplied documentation. Biologist-friendly documentation available on connection to the databases should be developed as a part of the guide.

Level 3 - Local Database Capability: Level 2, plus local database capability. This provides the ability to maintain a local database, to share it with the research community, and to obtain copies of remote databases. (The local copy is just that--it's existence is independent of the original.)

Interactions Between Computer Scientists, Mathematicians, and Biologists

The computational problems arising from the genome initiative will not be solved merely by biologists describing their problems in biological terms and hoping that computer scientists learn to communicate effectively. Possible mechanisms that could stimulate such interactions follow:

Joint Research Proposals: Computer scientists and biologists should be encouraged to seek joint funding of specific projects relevant to the genome initiative. A major benefit to be derived from such a scheme would be the establishment of rapport between the two scientists so that continuing collaborations would be facilitated.

Training Courses: Computer scientists could spend one to three weeks receiving an intensive, in-depth exposition of the biology relevant to the genome initiative. A parallel course in which biologists learned the elements of computer science could provide a complementary service.

Small Workshops: Workshops that bring together computer scientists and biologists with appropriate interdisciplinary expertise and focused on specific issues, such as database management techniques or analytical methods, should be arranged. Ideally such workshops would be highly focused and small in size to ensure the maximum opportunity for information exchange and informal discussion, which so often spawns fruitful collaborations.

National or Regional Meetings: Biologists and computer scientists can provide summaries of their recent work, and most importantly, can demonstrate on-line their latest software developments. Such meetings would encourage the participation of postdoctoral fellows and graduate students from both disciplines by ensuring the availability of direct funding from the meeting for such participants.

Advanced Interdisciplinary Training: In the long term, scientists will be needed who are educated in both biology and mathematics to effectively pursue research in genome informatics. This requires that students majoring in either biology or the mathematical sciences obtain minors in the

other discipline. Such degree programs will have to be designed and promoted, and regular interdisciplinary courses (and faculty positions) created. Special funding programs should be established that would encourage computer scientists, either at the graduate or postgraduate level, to undertake research projects in biological labs, and vice versa, to encourage biologists to undertake further training in a computer science lab. A program of Senior Fellowships could be initiated to encourage interdisciplinary sabbatical visits. Such cross-training will be essential if a new generation of computer-literate biologists and biologically astute computer scientists is to emerge.

Other Items for JITF Consideration

Several policy issues that require JITF input were recognized by the members of the working group. These overlap with the charges to the parent committees and need to be developed with their guidance.

Examples of such issues are:

- What level of funding should NIH and DOE provide for the informatics component of the genome project? How should this vary with time, and with total level of funding for the genome project as a whole?

- How can the networking efforts of the NIH and DOE be effectively integrated with those of the NSF, which has already taken the lead in providing high-throughput networking capabilities to the academic community?

- How can private industry and the foundations most effectively access the academic and government databases and bulletin boards? What is their "fair-share" component of the cost of these services?

- What is the role of the private sector in developing and disseminating informatics capabilities?

- What are the policies regarding data deposition in public databases going to be? Who should have access - and when?

- How can the most effective software tools be made most widely available? Who should be responsible for this service?

- Who will assume responsibility for service and maintenance of genomic databases?

- How can we best interact with the international scientific community with whom we both collaborate and compete?

Contributors to the Informatics Report

Benjamin Barnhart	Department of Energy
George Bell	Los Alamos National Laboratory
Dennis Benson	National Library of Medicine
David Botstein	Genentech, Inc.
Elbert Branscomb	Lawrence Livermore National Laboratory
Christian Burks	Los Alamos National Laboratory
Charles Cantor	Lawrence Berkeley Laboratory
Jaime Carbonell	Carnegie-Mellon University
Jim Cassatt	National Institute of General Medical Sciences
Elson Chen	Genentech, Inc.
John Devereaux	University of Wisconsin
Helen Donis-Keller	Washington University
Irene Eckstrand	National Institute of General Medical Sciences
Nat Goodman	Codd and Date, Inc.
Mark Guyer	National Center for Human Genome Research
Greg Hamm	Rutgers University
Diane Hinton	Howard Hughes Medical Institute
Elke Jordan	National Center for Human Genome Research
Eric Lander	Massachusetts Institute of Technology
Stanley Letovsky	Carnegie-Mellon University
David Lipman	National Library of Medicine
Donna Maglott	American Type Culture Collection
Thomas Marr	Cold Spring Harbor Laboratory
Eugene Myers	University of Arizona
Vicky Nichols	Chemical Abstracts Services
Frank Olken	Lawrence Berkeley Laboratory
Jim Ostell	National Library of Medicine
Ross Overbeek	Argonne National Laboratory
Mark Pearson	E.I. Du Pont de Nemours and Company
Peter Pearson	Johns Hopkins University
Robert Pecherer	Los Alamos National Laboratory
Jane Peterson	National Center for Human Genome Research
Robert Robbins	National Science Foundation
Richard Roberts	Cold Spring Harbor Laboratory
Henry Schaeffer	North Carolina State University
Jeff Schmaltz	Department of Energy
Dieter Soll	Yale University
Sylvia Spengler	Lawrence Berkeley Laboratory
Marvin Stodolsky	Department of Energy
Ed Theil	Lawrence Berkeley Laboratory
Michael Waterman	University of Southern California
John Wooley	National Science Foundation

REPORT OF THE WORKING GROUP ON ETHICAL, LEGAL, AND SOCIAL ISSUES RELATED TO MAPPING AND SEQUENCING THE HUMAN GENOME

The plan to map and sequence the human genome has profound implications for the alleviation of human suffering due to genetic disease. Genes directly causing or predisposing to human disease will be placed on the map for all to investigate. Additionally, normal genes which may be involved in the pathways leading to the development of new treatments will be captured and fundamental biological lessons in genetic regulation and functioning will be learned through the Human Genome Initiative.[1]

Any scientific endeavor of this magnitude must be developed in concert with a plan to ensure that the public has access to the benefits in improved health care, which should be a result of the research. It is also imperative to protect individuals and society from possible hazards which may be a consequence of our improved ability to detect and predict hereditary illness. The use of genetic information, for good or ill, has long been an issue in our society. But the quantity and complexity of genetic information that should become available requires that special precautions be taken.

Accordingly, the National Center for Human Genome Research is giving high priority to the development of a program to address the ethical, legal, and social implications of the Human Genome Initiative. This plan will attempt to anticipate the impact of the Human Genome Initiative and address what protections need to be in place so that the information generated can be of maximum benefit to individuals and society.

Although initially the Human Genome Initiative will produce information that will lead to the detection and diagnosis of genetic disease, the long-range goal will go beyond this to providing improved treatment, prevention, and ultimately cure. The interim phase, before adequate treatment is available, is the one in which the most deleterious consequences can occur, such as discrimination against gene carriers, loss of employment or insurance, stigmatization, untoward psychological reactions and attention. Once effective treatment is available for an illness, most of these

---

[1]The Human Genome Initiative is discussed in detail in the National Academy of Science's 1988 report, ***Mapping and Sequencing the Human Genome*** and the Office of Technology Assessment's 1988 report, ***Mapping Our Genes--The Genome Projects: How Big, How Fast?***

problems disappear. As the fruits of the Human Genome Initiative are realized, there will be an increased need for improved professional and public education to take advantage of the information gained.

In responding to the desires of the scientific community to understand the social, ethical, and legal implications of research on the human genome, the Office of Human Genome Research developed a program announcement, which appeared in the March 3, 1989 NIH Guide to Grants and Contracts. Applications were requested to address questions such as: (1) What are the concerns to society and to individuals?; (2) What questions in the areas of ethics and law need to be addressed?; (3) What can be learned from precedents?; (4) What are the policy alternatives and the pros and cons of each?; and (5) How can we inform and involve the public?

At its January 1989 meeting, the Program Advisory Committee on the Human Genome established the working group on ethics to develop a plan for this component of the human genome program. After considerable informal discussion within the group and with other scholars in ethics, law, and related fields over subsequent months, the working group had its first formal meeting on September 14-15, 1989. A roster of the members is attached.

At this meeting, the working group began to define and develop a plan of activities to address the ethical, legal, and social issues arising out of the application of knowledge gained as a result of the Human Genome Initiative. Representatives of the National Science Foundation (NSF) and the National Endowment for the Humanities were invited to present their grant programs for research on ethics, science, and society, and the working group noted that there was considerable opportunity for collaboration with these agencies, taking advantage of their expertise and experience in managing grants in this field.

The working group agreed that the purpose of the ethics component of the human genome program should be to:

- anticipate and address the implications for individuals and society of mapping and sequencing the human genome;
- examine the ethical, legal, and social consequences of mapping and sequencing the human genome;
- stimulate public discussion of the issues; and
- develop policy options that would assure that the information is used for the benefit of individuals and society.

The working group was strongly supportive of a program that would anticipate problems before they arise and develop suggestions for dealing with them that would forestall adverse effects. The approach to accomplishing these objectives should be several fold:

- to stimulate research on the issues through grants;
- to refine the research agenda through workshops, commissioned papers, and invited lectures on specific topics selected by the working group;
- to solicit public input from the community-at-large through town meetings and public testimony;
- to support the development of educational materials for all levels; and
- to encourage international collaboration in this area.

Stimulate Research

The working group is eager to encourage investigators in the research community to explore the wide range of issues pertinent to the human genome program. Outcomes of this research may be used to develop educational programs, policy recommendations or possible legislative recommendations.

In discussing the ethical, legal, and social consequences of the Human Genome Initiative, the working group deemed the following topics to be of particular importance and will strongly encourage research in the following areas.

1. Fairness in the use of genetic information with respect to:
   - insurance (acquisition and maintenance of health, life, disability, catastrophic, long-term care, and automobile insurance coverage)
   - employment (equal access)
   - the criminal justice system
   - the education system
   - adoptions
   - the military
   - any other areas to be identified

2. The impact of knowledge of genetic variation on the individual, including issues of:

   - o stigmatization
   - o ostracism
   - o labelling
   - o individual psychological responses, including impact on self image

3. Privacy and confidentiality of genetic information regarding:

   - o ownership and control of genetic information
   - o consent issues

4. The impact of the Human Genome Initiative on genetic counseling in the following areas:.

   - o prenatal testing
   - o pre-symptomatic testing
   - o carrier status testing, especially for very common disorders such as cystic fibrosis
   - o testing when there is no therapeutic remedy available, such as for Huntington's disease
   - o counseling and testing for polygenic disorders
   - o population screening versus testing

5. Reproductive decisions influenced by genetic information:

   - o effect of genetic information on options available
   - o use of genetic information in the decision-making process

6. Issues raised by the introduction of genetics into mainstream medical practice:

   - o qualifications and continuing education of all appropriate medical and allied health personnel
   - o standards and quality control
   - o education of patients
   - o education of the general public

7. Uses and misuses of genetics in the past and the relevance to the current situation, e.g.:

   - o the eugenics movement in the U.S. and abroad
   - o problems arising from screening for sickle-cell trait and other recent examples in which screening or testing sometimes achieved unintended and unwanted outcomes
   - o the misuse of behavioral genetics to advance eugenics or prejudicial stereotypes

8. Questions raised by the commercialization of the products from the Human Genome Initiative in the following areas:

   - o intellectual property rights (patents, copyrights, and trade secrets)
   - o property rights
   - o impact on scientific collaboration and candor
   - o accessibility of data and materials

9. Conceptual and philosophical implications of the Human Genome Initiative on:

   - o the concept of human responsibility
   - o the issue of free will versus determinism
   - o the concept of genetic disease, particularly in view of the high rate of human genetic variability and the large numbers of people who will be found to have genetic vulnerabilities

Most of this research can best be accomplished through the support of scholarly research and conferences. The working group recommended that support for conferences be limited to those that are highly focussed and produce a specific product, such as recommendations or policy options. The types of research to be supported should be varied and involve many of the disciplines traditional to the humanities. General surveys for purposes of information gathering are not recommended at this time.

Refine the Research Agenda

The working group is intentionally small so that others with specific necessary expertise can be recruited to join the effort as needed. To accomplish its task, the working group plans to invite individuals from a variety of disciplines to help refine

the research and policy agenda. This activity will include small workshops, commissioned papers, and invited lectures by knowledgeable individuals. In an effort to gather needed information in a timely manner, the working group will convene two to three times annually to collect information and discuss how this new knowledge will be integrated into a plan to refine the research agenda and propose future action.

Initial plans for the first workshop are underway. The format of a focus group is envisioned. Participants will include prominent individuals from various occupations and professions on which the Human Genome Initiative will have an impact such as, insurance companies, industry, labor unions, geneticists, "consumers" of genetic information and services, constitutional law, newsmedia, and the arts. The intent is to invite individuals who may not have been actively involved in the Human Genome Initiative or genetic research or services, but who can view the issues from a fresh perspective.

Participants will be provided background materials compiled by members of the working group and will be encouraged to discuss, on the basis of their experience and expertise, the most salient ethical, legal, and social repercussions of the plan to map and sequence the human genome and suggest areas of research, policy development, or legislation that they feel should be in place. From these discussions, the working group will formulate specific recommendations to bring before the advisory committee.

Solicit Public Input

The working group unanimously agreed that a critical component of its mission is to inform the general public (in the broadest sense) about the Human Genome Initiative and to solicit from them their questions and concerns about human genome research.

The town meeting format was considered appropriate for soliciting public input. However, to be effective such meetings must be carefully planned, taking into consideration the need to reach a broad cross-section of the public, and factors such as site, selection of participants, and wide publicity. A meeting of this type is tentatively planned for early 1991, or the end of the first year of this plan.

Support of Education

The human genome program should include a strong educational component involving both formal and informal education targeted to all educational levels.

It is suggested that NIH collaborate with NSF to develop model curricula that would be appropriate for the following groups: students at all levels, the newsmedia, medical practitioners, genetic counselors, scientists, teachers, and groups targeted for genetic services. Because NSF has experience in curriculum development, the working group believes that co-funding of appropriate NSF programs would be an efficient way for NIH to accomplish its goals in this area. In addition, a program of individual postdoctoral fellowships, such as those funded in the scientific components of the human genome project, are recommended for support of individuals who have doctoral degrees in biomedicine and want to pursue studies in the ethical, legal, or social aspects of human genome research or vice versa.

Additional activities that should be pursued are:

- short courses in ethical, legal, and social aspects of human genome research for scientists; and

- short courses in genomics for scholars from the humanities who want to do research on the ethical, legal and social implications of the genome project.

International Collaboration

The working group supports the concept of international collaboration in this area under guidelines similar to those for biomedical research on the human genome. Collaborative projects should be supported by funds from all the participants in the collaboration. The Human Genome Organization (HUGO) could play an obvious role in this area, which would be welcomed.

The Human Genome Initiative will have a profound impact on the lives of people in all countries, including those without genome research programs. Ideally, representatives from all interested countries should participate in considering the issues that will arise. An international organization, such as UNESCO, could facilitate cooperation in this area.

Diseases and the suffering they cause respect no geographical boundaries. The sharing of results from the Human Genome Initiative across geographical barriers must be encouraged. Although differences exist cross culturally in the use of genetic information, the working group hopes there are also sufficient similarities so that its efforts can be useful to all.

Contributors to the Report on Ethical, Legal, and Social Issues Related to Mapping and Sequencing the Human Genome

Nancy S. Wexler, Ph.D.
Hereditary Disease Foundation
and
Department of Neurology and Psychiatry
College of Physicians and Surgeons
Columbia University
722 West 168th Street, Box 58
New York, NY 10032

Jonathan R. Beckwith, Ph.D.
Department of Microbiology
and Molecular Genetics
Harvard Medical School
200 Longwood Avenue
Boston, MA 02115

Robert Cook-Deegan, M.D.
7717 Goodfellow Way
Derwood, MD 2085

Patricia King, J.D.
Georgetown University
Law Center
600 New Jersey Ave., N.W.
Washington, D.C. 20001

Victor A. McKusick, M.D.
Division of Medical Genetics
Johns Hopkins Hospital
600 North Wolfe Street, Blalock 1007
Baltimore, MD 21205

Robert F. Murray Jr., M.D.
Department of Pediatrics,
Medicine, Oncology, and Genetics
Box 75
Howard University College of Medicine
Washington, D.C. 20050

Thomas H. Murray, Ph.D.
Center for Biomedical Ethics
Case Western Reserve University
2119 Abington Road
Cleveland, OH 44106

Contributors to the Report on Ethical, Legal, and Social Issues

Jonathan R. Beckwith is a bacterial geneticist interested in genetic screening. For more than 20 years, he has been interested in and concerned about the long-range implications of genetics and behavior and genetics and intelligence quotient.

Robert Cook-Deegan is a clinician whose interest in genetics dates back to his research on Alzheimer's Disease. While on an Office for Technology Assessment fellowship, he prepared two reports on human gene therapy and public policy related to the human genome project. Dr. Cook-Deegan is currently writing a book on how the Human Genome Initiative got started in the United States.

Patricia King is an attorney and academician whose legal career has been in civil rights law. She has served on the National Committee for Protection of Human Subjects, the Recombinant DNA Advisory Committee, and the Presidential Committee on Ethics. Ms. King is a Fellow of the Hastings Center and is interested in genetics and how it affects minorities.

Victor A. McKusick is an internationally recognized geneticist who has been active in human genetics research for over 40 years. More than any other person, he has been responsible over the years for collecting data on inherited diseases. Since 1973, he has collected and coordinated data on the human gene map, which in 1988 included 2,000 genes.

Robert F. Murray, Jr. is a clinical researcher who directs a clinical genetics program in sickle-cell disease. He became involved initially with the ethical aspects of screening for sickle-cell disease. Dr. Murray is concerned about individuals who want to plan their destiny based on new technologies emanating from genetics research.

Thomas H. Murray is a social psychologist who has written extensively about genetic screening in the work place. He has undertaken fellowships with an emphasis on humanities at Yale University and The Hastings Center. Dr. Murray was recently elected a Fellow of The Hastings Center and is currently co-authoring a publication with a geneticist for the British Medical Association.

Nancy S. Wexler is a clinical psychologist and a researcher. Her mother died of Huntington's Disease and she is a potential consumer of the information generated by the Human Genome Initiative. Many of her current efforts are to get individuals, interest groups, and the federal government to anticipate how information generated from the Human Genome Initiative can be used maximally to benefit the individual.

JOINT MAPPING WORKING GROUP

David Botstein, Ph.D.
Vice President
Genentech, Inc.
460 Point San Bruno Blvd.
South San Francisco, CA 94080

Anthony Carrano, Ph.D.
Lawrence Livermore
National Laboratory
Biomedical and Environmental
Science Division
P.O. Box 5507
Livermore, CA 94550

C. Thomas Caskey, M.D.
Professor and Director
Institute for Molecular Genetics
Baylor College of Medicine
One Baylor Plaza, T809
Houston, TX 77030

David R. Cox, M.D., Ph.D.
Associate Professor
Department of Psychiatry
Langley Porter Institute
Box F-0984
University of California
San Francisco
San Francisco, CA 94143

Robert Moyzis, Ph.D.
Center for Human Genome Studies
MS M886
Los Alamos National Laboratory
Los Alamos, NM 87545

Maynard V. Olson, Ph.D.
Professor
Department of Genetics
Washington University
School of Medicine
P.O. Box 8031
4566 Scott Avenue
St. Louis, MO 63110

JOINT INFORMATICS TASK FORCE

Chairman

Dr. Dieter Soll
Dept. of Molecular Biophysics
and Biochemistry
Yale University
P.O. Box 6666
260 Whitney Ave.
New Haven, CT. 06511

Dr. George Bell
Group T-10
MS K710
Los Alamos National Laboratory
Los Alamos, NM 87545

Dr. David Botstein
Genentech, Inc.
460 Point San Bruno Blvd.
South San Francisco, CA 94080

Dr. Elbert Branscomb
Lawrence Livermore National
Laboratory
Biomedical Science Division
P.O. Box 5507, L-452
Livermore, CA 94550

Dr. John Devereaux
Genetics Computer Group, Inc.
575 Science Drive
Suite B
Madison, WI 53711

Technical Coordinator

Mr. Gregory Hamm
Molecular Biology
Computer Laboratory
Waksman/CABM
P.O. Box 759
Rutgers University
Piscataway, NJ 08855

Dr. Eric Lander
Whitehead Institute
9 Cambridge Center
Cambridge, MA 02142

Dr. Thomas G. Marr
Cold Spring Harbor Laboratory
P.O. Box 100
Cold Spring Harbor, NY 11724

Mr. Frank Olken
Lawrence Berkeley Laboratory
1 Cyclotron Road
M/S 50B-3238
Berkeley, CA 94720

Dr. Ross Overbeek
Mathematics and Computer
Science Division
Argonne National Lab
9700 S. Cass Avenue
Argonne, IL 60439

Dr. Nathan Goodman
32 Kennard Road
Brookline, MA 02146

Dr. Mark Pearson
E.I. Du Pont de Nemours & Co.
Central Research & Development
Experimental Station
Building 328, Room 251
P.O. Box 80328
Wilmington, DE 19880-0328

Dr. Sylvia Spengler
Human Genome Center
Lawrence Berkeley Laboratory
459 Donner
Berkeley, CA 94720

Dr. Mike Waterman
University of Southern California
Department of Mathematics
University Park
Los Angeles, CA 90089-1113

JOINT WORKING GROUP ON ETHICAL, LEGAL, AND SOCIAL ISSUES

Chair
Nancy S. Wexler, Ph.D.
Hereditary Disease Foundation
and
Department of Neurology and Psychiatry
College of Physicians and Surgeons
Columbia University
722 West 168th Street, Box 58
New York, NY 10032

Jonathan R. Beckwith, Ph.D.
Department of Microbiology
and Molecular Genetics
Harvard Medical School
200 Longwood Avenue
Boston, MA 02115

Patricia King, J.D.
Georgetown University Law Center
600 New Jersey Ave., N.W.
Washington, D.C. 20001

Victor A. McKusick, M.D.
Division of Medical Genetics
Johns Hopkins Hospital
600 North Wolfe Street, Blalock 1007
Baltimore, MD 21205

Robert F. Murray Jr., M.D.
Department of Pediatrics,
Medicine, Oncology, and Genetics
Box 75
Howard University College of Medicine
Washington, D.C. 20050

Thomas H. Murray, Ph.D.
Center for Biomedical Ethics
Case Western Reserve University
2119 Abington Road
Cleveland, OH 44106

SCIENTIFIC GOALS OF THE U.S. HUMAN GENOME PROJECT

1. Mapping and Sequencing the Human Genome

Genetic Map

5 Year Goal: Complete a fully connected human genetic map with markers spaced an average of 2 to 5 centimorgans apart. Identify each marker by a sequence-tagged site (STS).

Physical Map

5 Year Goal: Assemble STS maps of all human chromosomes with the goal of having markers spaced at approximately 100,000 base-pair intervals.

Generate overlapping sets of cloned DNA or closely spaced unambiguously ordered markers with continuity over lengths of 2 million base pairs for large parts of the human genome.

DNA Sequencing

5 Year Goal: Improve current methods and/or develop new methods for DNA sequencing that will allow large-scale sequencing of DNA at a cost of $0.50 per base pair.

Determine the sequence of an aggregate of 10 million base pairs of human DNA in large continuous stretches in the course of technology development and validation.

2. Model Organisms

5 Year Goal: Prepare a genetic map of the mouse genome based on DNA markers. Start physical mapping on one or two chromosomes.

Sequence an aggregate of about 20 million base pairs of DNA from a variety of model organisms, focusing on stretches that are one million base pairs long, in the course of the development and validation of new and/or improved DNA sequencing technology.

3. Informatics: Data Collection and Analysis

5 Year Goal: Develop effective software and database designs to support large-scale mapping and sequencing projects.

Create database tools that provide easy access to up-to-date physical mapping, genetic mapping, chromosome mapping, and sequencing information and allow ready comparison of the data in these several data sets.

Develop algorithms and analytical tools that can be used in the interpretation of genomic information.

4. Ethical, Legal and Social Considerations

5 Year Goal: Develop programs addressed at understanding the ethical, legal, and social implications of the human genome project.

Identify and define the major issues and develop initial policy options to address them.

5. Research Training

5 Year Goal: Support research training of pre- and post-doctoral fellows starting in FY 1990. Increase the numbers of trainees supported until a steady state of about 600 per year is reached by the fifth year.

Examine the need for other types of research training in the next year.

6. Technology Development

5 Year Goal: Support innovative and high-risk technological developments as well as improvements in current technology to meet the needs of the genome project as a whole.

7. Technology Transfer

5 Year Goal: Enhance the already close working relationships with industry.

Encourage and facilitate the transfer of technologies and of medically important information to the medical community.

GLOSSARY

Atomic force microscopy: A variation of Scanning Tunneling Microscopy that involves measurements of forces at the atomic level.

Base pair: Two nucleotides (adenosine and thymidine or guanosine and cytidine) held together by weak bonds. Two strands of DNA are held together in the shape of a double helix by the bonds between base pairs.

Capillary gel: A very thin capillary tube filled with a semi-solid gel and used in electrophoretic separation of molecules.

Centimorgan: A unit of measure of recombination frequency. One centimorgan is equal to a 1 percent chance that a genetic locus will be separated from another marker due to recombination in a single generation.

Chromosome: A rod-like structure found in the cell nucleus and containing the genes. Chromosomes are composed of DNA and proteins. They can be seen in the light microscope during certain stages of cell division.

Chromosomes 1-22: Human chromosomes are ordered and named according to size, the largest being chromosome 1 and the smallest chromosome 22.

Contigs: Groups of overlapping clones representing a continuous region of DNA.

Cytological mapping: Mapping of genes using DNA probes that bind to the chromosome at the site of the gene and are visible in a light microscope.

DNA (deoxyribonucleic acid): The molecule that encodes genetic information. DNA is a double-stranded molecule held together by weak bonds between pairs of nucleotides on opposite strands. There are four nucleotides in DNA: adenosine (A), guanosine (G), cytidine (C), and thymidine (T). In nature, base pairs form only between A and T and between G and C, thus the sequence of each single strand can be deduced from that of its partner.

DNA sequence: The order of base pairs whether in a stretch of DNA, a gene, a chromosome, or an entire genome.

DNA sequencing: Determining the sequence of the nucleotides in DNA.

Double helix: The shape in which two linear strands of DNA are bonded together.

Electrophoresis: A method of separating large molecules (such as DNA fragments or proteins) from a mixture of similar molecules. An electric current is passed through a medium containing the mixture, and each kind of molecule travels through the medium at a different rate, depending on its electrical charge and size. Separation is based on these differences.

***Escherichia coli (E. coli)*:** A common intestinal bacterium geneticists have used for many studies.

Gene: The fundamental physical and functional unit of heredity. A gene is an ordered sequence of nucleotides located in a particular position on a particular chromosome.

Gene mapping: Determining the relative locations of different genes on chromosomes.

Genetic code: The sequence of nucleotides, coded in triplets along the mRNA, that determines the sequence of amino acids in protein synthesis. The DNA sequence of a gene can be used to predict the mRNA sequence, and the genetic code can in turn be used to predict the amino acid sequence.

Genetic linkage map: A map of the relative positions of genetic loci on a chromosome, determined on the basis of how often the loci are inherited together. Distance is measured in centimorgans.

Genome: All the genetic material in the chromosomes of a particular organism; its size is generally given as the total number of base pairs.

Genome projects: Research and technology development efforts aimed at mapping and sequencing some or all of the genome of human beings and other organisms.

Human Genome Initiative: An initiative whose goal is to map and sequence the human genome. The concept was first formally proposed in 1986.

Human Genome Project: The implementation of the concepts proposed as the Human Genome Initiative.

Human Genome Program: The individual programs, such as those at DOE and NIH, that make up the Human Genome Project.

Informatics: The study of the application of computer and statistical techniques to the management of information. In genome projects, informatics includes the development of methods to search databases quickly, to analyze DNA sequence and to determine DNA structure from DNA sequence data.

Marker: An identifiable physical location on a chromosome (e.g., restriction enzyme cutting site, gene, RFLP marker) whose inheritance can be monitored. Markers can be expressed regions of DNA (genes) or some segment of DNA with no known coding function but whose pattern of inheritance can be determined.

Mass spectroscopy: A method of determining chemical structure based on the mass of the molecule and derived fragments.

Messenger RNA (mRNA): A class of ribonucleic acid (RNA) whose role is to carry the genetic code from the chromosome to the ribosome, the site of protein synthesis.

Nucleotide: A subunit of DNA or RNA consisting of a nitrogenous base (adenine, guanine, thymine, or cytosine in DNA; adenine, guanine, uracil, or cytosine in RNA) a phosphate molecule, and a sugar molecule (deoxyribose in DNA and ribose in RNA). Thousands of nucleotides are linked to form the DNA or RNA molecule.

Physical map: A map of the locations of identifiable landmarks on DNA (e.g., restriction enzyme cutting sites, genes, RFLP markers). A physical map may also be a set of overlapping clones, called a contig. Distance is measured in base pairs. For the human genome, the lowest-resolution physical map is the banding patterns of the 24 different chromosomes; the highest-resolution map would be the complete nucleotide sequence of the chromosomes.

Polymerase chain reaction (PCR): An enzymatic reaction that precisely and rapidly amplifies a small segment of DNA millions of times or more. The reaction can start with one molecule of DNA.

Pulsed-field gel electrophoresis (PFGE): A type of gel electrophoresis in which pulses of current are applied to the sample at various angles, enabling scientists to separate and order by size extremely large segments of DNA.

Radiation hybrid: A somatic cell hybrid that contains pieces of human chromosomes generated by irradiation.

Recombinant DNA: The hybrid DNA produced in the laboratory by joining pieces of DNA from different sources.

Recombinant DNA technology: Techniques for cutting apart and splicing together pieces of DNA from different sources.

Recombination: The process by which portions of DNA are exchanged or deleted. Recombination occurs naturally between or within chromosomes, particularly during the formation of sperm and egg cells.

Restriction enzyme: An enzyme that recognizes a specific base sequence (usually four to six base pairs in length) in a double-stranded DNA molecule and cuts both strands of the DNA molecule at every place where this sequence appears.

Restriction enzyme cutting site: A specific nucleotide sequence of DNA at which a restriction enzyme cuts the DNA. Some sites occur frequently in DNA (e.g., every several hundred base pairs), others much less frequently (e.g., every 10,000 base pairs).

Restriction fragment length polymorphism (RFLP): The presence of two or more variants in the size of DNA fragments from a specific region of DNA that has been exposed to a particular restriction enzyme. These fragments differ in length because of an inherited variation in a restriction enzyme recognition site.

Scanning tunneling microscopy (STM): A very high-resolution imaging technique that is able to resolve material at atomic distances, opening the possibility of reading DNA sequence by microscopy.

Sequence-tagged site (STS): A short DNA sequence that uniquely identifies a mapped gene or other marker. The order and spacing of these sequences comprise an STS map.

Stable isotopes: Nonradioactive isotopes.

Technology transfer: The process of converting scientific knowledge into useful products.

X-chromosome: A sex chromosome. Normal human females have two X chromosomes in each cell, while normal males have one X and one Y chromosome in each cell.

X-ray imaging: High-resolution microscopy using an x-ray beam.

Y-chromosome: A sex chromosome. Normal human males carry one X chromosome and one Y chromosome in each cell.

Yeast artificial chromosome vectors (YAC): Plasmids that contain those portions of yeast chromosomal DNA needed for replication and maintenance and with which foreign DNA can be cloned. YAC vectors can accommodate foreign DNA fragments up to 1 million base pairs in size.