A New Five-Year Plan for the U.S. Human Genome Project

Francis Collins and David Galas*

The U.S. Human Genome Project is part of an international effort to develop genetic and physical maps and determine the DNA sequence of the human genome and the genomes of several model organisms. Thanks to advances in technology and a tightly focused effort, the project is on track with respect to its initial 5-year goals. Because 3 years have elapsed since these goals were set, and because a much more sophisticated and detailed understanding of what needs to be done and how to do it is now available, the goals have been refined and extended to cover the first 8 years (through September 1998) of the 15-year genome initiative.

In 1990, the Human Genome programs of the National Institutes of Health (NIH) and the Department of Energy (DOE) developed a joint research plan with specific goals for the first 5 years [fiscal year (FY) 1991–95] of the U.S. Human Genome Project (*1*). It has served as a valuable guide for both the research community and the agencies' administrative staff in developing and executing the genome project and assessing its progress for the past 3 years. Great strides have been made toward the achievement of the initial set of goals, particularly with respect to constructing detailed human genetic maps, improving physical maps of the human genome and the genomes of certain model organisms, developing improved technology for DNA sequencing and information handling, and defining the most urgent set of ethical, legal, and social issues associated with the acquisition and use of large amounts of genetic information.

Progress toward achieving the first set of goals for the genome project appears to be on schedule or, in some instances, even ahead of schedule. Furthermore, technological improvements that could not have been anticipated in 1990 have in some areas changed the scope of the project and allowed more ambitious approaches. Earlier this year, it was therefore decided to update and extend the initial goals to address the scope of genome research beyond the completion of the original 5-year plan. A major purpose of revising the plan is to inform and provide a new guide to all participants in the genome project about the project's goals. To obtain the advice needed to develop the extended goals, NIH and DOE held a series of meetings with a large number of scientists and other interested scholars and representatives of the public, including many who previously had not been direct participants in the genome project. Reports of all these meetings are available from the Office of Communications of the National Center for Human Genome Research (NCHGR) and the Human Genome Management Information System of DOE (*2*, *3*). Finally, a group of representative advisors from NIH and DOE drafted a set of new, extended goals for presentation to the National Advisory Council for Human Genome Research of NIH and the Health and Environmental Research Advisory Committee of DOE. These bodies have approved this document as a statement of their advice to the two agencies, and the following represents the goals for FYs 1994–98 (1 October 1993 to 30 September 1998).

General Principles

Several general observations underlie the specific goals (Fig. 1) described here. The first observation is that successful development of new technology for genomic and genetic research has been essential to the achievements of the project to date and will continue to be critical in the future. It was clearly recognized, both in the 1988 National Research Council (NRC) report (*4*) and in the first NIH-DOE plan, that attainment of the ambitious goals originally set for the genome project would require significant technological advances in all areas, such as mapping, sequencing, informatics, and gene identification. As the genome project has proceeded, progress along a broad range of technological fronts has been conspicuous. Among the most notable of these developments have been (i) new types of genetic markers, such as microsatellites, that can be assayed by polymerase chain reaction (PCR); (ii) improved vector systems for cloning large DNA fragments and better experimental strategies and computational methods for assembling those clones into large, overlapping sets (contigs) that compose useful physical maps; (iii) the definition of the sequence tagged site (STS) (*5*) as a common unit of physical mapping; and (iv) improved technology and automation for DNA sequencing. Further substantial improvements in technology are needed in all areas of genome research, especially in DNA sequencing, if the project is to stay on schedule and meet the demanding goals that are being set.

A second general observation concerns an evolution in the levels of biological organization at which genomic research will likely function over the next few years. Initially, attention was focused on the chromosome as the basic unit of genome analysis. Large-scale mapping efforts, in particular, were directed at the construction of chromosome maps. The sophisticated genetic linkage maps now available and the detailed physical maps that are being produced are clear measures of the success of that approach. However, other units of study for the Human Genome Project will also have increasing usefulness in the future. Therefore, further mapping efforts directed at both larger and smaller targets should be encouraged. At one end of the scale, "whole genome" mapping efforts, in which the entire genome is efficiently analyzed, have become feasible with developments in PCR applications and robotics. These approaches generally produce relatively low-resolution maps with current technology. At the other end of the scale, increasing attention needs to be paid to detailed mapping, sequencing, and annotation of regions on the order of one to a few megabases in size. Although small in comparison with the whole genome, a megabase is still large in comparison with the capabilities of conventional molecular genetic analysis. Thus, development of efficient technology for approaching detailed analysis of several-megabase sections of the genome will provide a useful bridge between conventional genetics and genomics, and provide a foundation for innovation from which future methods for analysis of larger regions may arise.

Third, a goal for identifying genes within maps and sequences, implicit in the original plan, has now been made explicit. The progress already made on the original goals, combined with promising new approaches to gene identification, allow this element of genome analysis to be given greater visibility. This increased emphasis on gene identification will greatly enrich the maps that are produced.

It must also be noted that, as in the original 5-year plan, these goals assume a funding level for the U.S. Human Genome Project of $200 million annually, adjusted for inflation. As the detailed cost analysis for the first 5-year plan was performed in

F. Collins is the director of the National Center for Human Genome Research, National Institutes of Health, Bethesda, MD 20892.
D. Galas was associate director, Office of Health and Environmental Research, Department of Energy, Washington, DC 20585.

* Present address: Darwin Molecular, 2405 Carillon Point, Kirkland, WA 98033.

1991, a cost of living increase must be added for all years beyond FY 1991. This funding level has not yet been achieved (Table 1).

International Aspects

The Human Genome Project is truly international in scope, as the original planners envisioned it. Its success to date has been possible because of major contributions from many countries and the extensive sharing of information and resources. It is hoped and anticipated that this spirit of international cooperation and sharing will continue. This coordination has been achieved largely by scientist-to-scientist interaction, facilitated by the Human Genome Organization (HUGO), which has taken on responsibility for some aspects of the management of the international chromosome workshops in particular. These workshops have served to encourage collaboration and the sharing of information and resources and to facilitate the expeditious completion of chromosome maps.

Several notable individual international collaborations have marked the genome project so far. One is the United States–United Kingdom collaboration on the sequencing of the *Caenorhabditis elegans* genome. Scientists at the Los Alamos National Laboratory are collaborating with Australian colleagues to develop a physical map of chromosome 16, and investigators at the Lawrence Livermore National Laboratory are working with Japanese scientists on a high-resolution physical map of chromosome 21. Other joint efforts include the collaboration between NIH and the Centre d'Etudo du Polymorphisme Humain (CEPH) on the genetic map of the human genome and the Whitehead/Massachusetts Institute of Technology–Généthon collaboration on the whole-genome approach to the human physical map. These are but examples of the myriad interrelationships that have formed, generally spontaneously, among participating scientists.

Specific Goals

Genetic map. The 2- to 5-cM human genetic map of highly informative markers called for in the original goals is expected to be completed on time. However, improvements to make the map more useful and accessible will still be needed. If the field develops as predicted, there will be an increasing demand for technology that allows the nonexpert to type families rapidly for medical research purposes. In addition, to study complex genetic diseases, there is a need to be able to easily test large numbers of individuals for many markers simultaneously. In the long run, polymorphic markers that can be screened in a more automated fashion, and methods of gene mapping that obviate the need for a standard set of polymorphic markers are also desirable.

Table 1. The budget for the Human Genome Project for NIH and DOE (in millions of dollars). Budgets for 1994 and 1995 have not yet been determined.

Fiscal year	NIH	DOE	Total	1991 Projection of Needs
1991	87.4	47.4	134.8	135.1
1992	104.8	61.4	166.2	169.2
1993	106.1	64.5	170.6	218.9
1994				246.8
1995				259.9

Goals

(i) Complete the 2- to 5-cM map by 1995.
(ii) Develop technology for rapid genotyping.
(iii) Develop markers that are easier to use.
(iv) Develop new mapping technologies.

Physical map. An STS-based physical map of the human genome is expected to be available in the next 2 to 3 years, with some areas mapped in more detail than others and an average interval between markers of about 300 kb. However, such a map will not likely be sufficiently detailed to provide a substrate for sequencing or to be optimally useful to scientists searching for disease genes. The original goal of a physical map with STS markers at intervals of 100 kb remains realistic and useful and would serve both sequencers and mappers. Using widely available methods, a molecular biologist can isolate a gene that is within 100 kb of a mapped marker, and a sequencer can use such a map as the basis for preparing the DNA for sequencing. To the extent that they do not introduce statistical bias, the use of STSs with added value (such as those derived from polymorphic markers or genes) is encouraged because such markers add to the usefulness of the map.

Goal

(i) Complete an STS map of the human genome at a resolution of 100 kb.

Physical maps of greater than 100-kb resolution are needed for DNA sequencing, for the purpose of finding genes and for other biological purposes. Although a variety of options are being explored for creating such maps, the optimal approach is by no means clear. There is a need to develop new strategies for high-resolution physical mapping as well as new cloning systems that are well integrated with advanced sequencing technology. Technology for sequencing is evolving rapidly. Therefore, preparation of sequence-ready sets of clones should be closely associated with an imminent intent to sequence.

There is a pressing need for clone libraries with improved stability and lower chimerism and other artifacts and a need for better technology for traveling from one STS to the next. A greater accessibility to clone libraries should also be encouraged.

DNA sequencing. Although the goal of sequencing DNA at a cost of $0.50 per base pair may be met by 1996 as originally projected, the rate at which DNA can be sequenced will not be sufficient for sequencing the whole human genome. Priority should be given during the next 5 years to increasing sequencing capacity by increasing the number of groups oriented toward large-scale production sequencing. Substantial new technology that will allow sequencing at higher rates and lower costs is also needed: evolutionary technology developed from improvements in current gel-based approaches and revolutionary technology developed on the basis of new principles. These developments will only occur if significantly greater financial resources can be invested in this area. It is estimated that an immediate investment of $100 million per year will be needed for sequencing technology alone, to allow the human genome to be sequenced by the year 2005.

Goals

(i) Develop efficient approaches to sequencing one- to several-megabase regions of DNA of high biological interest.
(ii) Develop technology for high throughput sequencing, focusing on systems integration of all steps from template preparation to data analysis.
(iii) Build up a sequencing capacity to a collective rate of 50 Mb per year by the end of the period. This rate should result in an aggregate of 80 Mb of DNA sequence completed by the end of FY 1998.

The standard model organisms should be sequenced as rapidly as possible, with *Escherichia coli* and *Saccharomyces cerevisiae* completed by 1998 or earlier and *C. elegans* nearing completion by 1998. It is often advantageous to sequence the corresponding regions of human and mouse DNA side by side in areas of high biological interest. The sequencing of full-length, mapped complementary DNA molecules is useful, especially if it is associated with technological innovation applicable to genomic sequencing.

The measurement of the cost of sequencing is complex and fraught with many uncertainties due to the diversity of approaches being used. However, we need to continue to reduce costs, as well as im-

prove our ability to assess the accuracy of the sequence produced. This latter point must be addressed in future sequencing efforts. Cost will be highly dependent on the level of accuracy achieved.

Gene identification. Identification of all the genes in the human genome and in the genomes of certain model organisms is an implicit part of the Human Genome Project. Although the previous 5-year plan did not explicitly identify this activity with a specific goal, progress in mapping and in technology now makes it desirable to do so. With both genetic and physical maps of the human genome and the genomes of certain model organisms becoming available and large amounts of sequence data beginning to appear, it is important to develop better methods for identifying all the genes and incorporating all known genes onto the physical maps and the DNA sequences that are produced. This information will make the maps most useful to scientists studying the involvement of genes in health and disease. While many promising approaches are being explored, more development is needed in this area.

Goals

(i) Develop efficient methods of identifying genes and for placement of known genes on physical maps or sequenced DNA.

Technology development. The development of new and improved technology is vital to the genome project. Certain technologies, such as automation and robotics, cut across many areas of genome research and need particular attention. Cooperation in technology development should be encouraged where possible because it is likely to be more effective and efficient than competition and duplication. The technology developed must be expandable and exportable, the long-term goal being to create technology that will be available in many basic science laboratories and allow the efficient sequencing of other genomes. Technology development is costly and has not been sufficiently funded.

Goal

(i) Substantially expand support of innovative technological developments as well as improvements in current technology for DNA sequencing and to meet the needs of the Human Genome Project as a whole.

Model organisms. Excellent progress has been made on the mouse genetic map and the *Drosophila* physical map, as well as the sequencing of the DNA of *E. coli*, *S. cerevisiae*, and C. *elegans*. Many of the original goals for this area are likely to be exceeded. Completion of the mouse map and sequencing of all the selected model organism genomes continue to be high priorities. The current emphasis for sequencing of mouse DNA should be placed on the sequencing of selected regions of high biologic interest side by side with the corresponding human DNA.

Goals

(i) Finish an STS map of the mouse genome at 300-kb resolution.

(ii) Finish the sequence of the *E. coli* and *S. cerevisiae* genomes by 1998 or earlier.

(iii) Continue sequencing C. *elegans* and *Drosophila* genomes with the aim of bringing C. *elegans* to near completion by 1998.

(iv) Sequence selected segments of mouse DNA side by side with corresponding human DNA in areas of high biological interest.

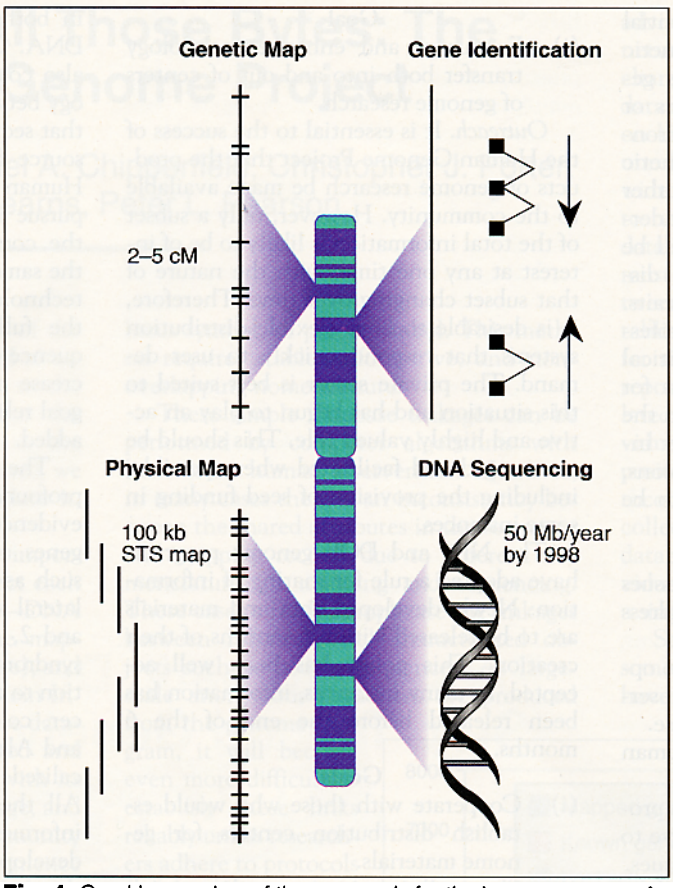

Fig. 1. Graphic overview of the new goals for the human genome. A 2- to 5-cM genetic map is expected to be completed by 1995 and a physical map with STS markers every 100 kb by 1998. Efficient methods for gene identification need to be developed and refined. The DNA sequencing goal of 50 Mb per year by 1998 includes all DNA, both human and model organisms, and assumes an exponential increase in sequencing capacity over time. Other important goals involving model organisms are not shown here, but are described in the text.

Informatics. In order to collect, organize, and interpret the large amounts of complex mapping and sequencing data produced by the Human Genome Project, appropriate algorithms, software, database tools, and operational infrastructure are required. The success of the genome project will depend, in large part, on the ease with which biologists can gain access to and use the information produced. Although considerable progress has been made in this area since the beginning of the genome project, there is a continuing need for improvements to stay current with evolving requirements. As the amount of information increases, the demand for it and the need for convenient access increase also. Thus, data management, data analysis, and data distribution remain major goals for the future.

Goals

(i) Continue to create, develop, and operate databases and database tools for easy access to data, including effective tools and standards for data exchange and links among databases.

(ii) Consolidate, distribute, and continue to develop effective software for large-scale genome projects.

(iii) Continue to develop tools for comparing and interpreting genome information.

Ethical, legal, and social implications (ELSI). The ELSI components of the Human Genome programs of NIH and DOE are strongly connected with genomic research so that policy discussions and recommendations are couched in the reality of the science. To date, the focus of the ELSI programs has been on the most immediate potential applications in society of genome research. Four areas were identified by advisers to the ELSI program for initial emphasis: privacy of genetic information, safe and effective introduction of genetic information in the clinical setting, fairness in the use of genetic information, and professional and public education. The program gives strong emphasis to understanding the ethnic, cultural, social, and psychological influences that must inform policy development and service delivery. Initial policy options for genetic family studies, clinical genetic services, and health care coverage have been developed, and reports on a range of urgent issues are expected by 1995.

As the genome project progresses, the need to prepare for even broader public impact becomes increasingly important. Poli-

cies are needed to anticipate the potential consequences of widespread use of genetic tests for common conditions, such as genetic predisposition to certain cancers or genetic susceptibility to certain environmental agents. In addition, as the genetic elements of behavioral and other nondisease-related traits are better understood, increased educational efforts will be needed to prevent stigmatization or discrimination on the basis of these traits. Continued emphasis on public and professional education at all levels will be critical to achieving these goals. Mechanisms for developing policy options that build on the current research portfolio and actively involve the public, the relevant professions, and the scientific community need to be developed.

Goals

(i) Continue to identify and define issues and develop policy options to address them.
(ii) Develop and disseminate policy options regarding genetic testing services with potential widespread use.
(iii) Foster greater acceptance of human genetic variation.
(iv) Enhance and expand public and professional education that is sensitive to sociocultural and psychological issues.

Training. There is a continuing need for individuals highly trained in the interdisciplinary sciences related to genome research. The original goal of suporting 600 trainees per year proved to be unattainable, because the capacity to train so many individuals in interdisciplinary sciences did not exist. However, now that a number of genome centers have been established, it is anticipated that training programs will expand. Although no numerical goal is specified, expansion of training activities should be encouraged, provided standards are kept high. Quality is more important than quantity.

Goal

(i) Continue to encourage training of scientists in interdisciplinary sciences related to genome research.

Technology transfer. Technology transfer is already occurring to a remarkable extent, as evidenced by the number of genome-related companies that are forming. Many interactions and collaborations have been established between genome researchers and the private sector. In addition to the need to transfer technology out of centers of genome research, there is also a need to increase the transfer of technology from other fields into the genome centers. Increased cooperation with industry, as well as continued cooperation between the agencies, is highly desirable. Care must be taken, however, to avoid conflicts of interest.

Goal

(i) Encourage and enhance technology transfer both into and out of centers of genome research.

Outreach. It is essential to the success of the Human Genome Project that the products of genome research be made available to the community. However, only a subset of the total information is likely to be of interest at any one time, with the nature of that subset changing over time. Therefore, it is desirable to have flexible distribution systems that respond quickly to user demand. The private sector is best suited to this situation and has begun to play an active and highly valued role. This should be encouraged and facilitated where possible, including the provision of seed funding in some instances.

The NIH and DOE genome programs have adopted a rule for sharing of information: Newly developed data and materials are to be released within 6 months of their creation. This policy has been well accepted. In many instances, information has been released before the end of the 6 months.

Goals

(i) Cooperate with those who would establish distribution centers for genome materials.
(ii) Share all information and materials within 6 months of their development. The latter should be accomplished by submission of information to public databases or repositories, or both, where appropriate.

Conclusion

To date, the Human Genome Project has experienced gratifying success. However, enormous challenges remain. The technology that will lead to the sequencing of the entire human genome at reasonable cost must still be developed. Major support of research in this area is essential if the genome project is to succeed in the long run. The new goals described here are designed to address the long- and short-term needs of the project.

Although there is still debate about the need to sequence the entire genome, it is now more widely recognized that the DNA sequence will reveal a wealth of biological information that could not be obtained in other ways. The sequence so far obtained from model organisms has demonstrated the existence of a large number of genes not previously suspected. For example, almost half of the open reading frames identified in the genomic DNA of *C. elegans* appear to represent previously unidentified genes. Similar results have been observed in both *S. cerevisiae* and *E. coli* genomic DNA. Comparative sequence analysis has also confirmed the high degree of homology between genes across species. It is clear that sequence information represents a rich source for future investigation. Thus, the Human Genome Project must continue to pursue its original goal, namely, to obtain the complete human DNA sequence. At the same time, it is necessary to assure that technologies are developed that will allow the full interpretation of the DNA sequence once it is available. In order to increase emphasis on this area, an explicit goal related to gene identification has been added.

The genome project has already had a profound impact on biomedical research, as evidenced by the isolation of a number of genes associated with important diseases, such as Huntington's disease, amyotrophic lateral sclerosis, neurofibromatosis types 1 and 2, myotonic dystrophy, and fragile X syndrome. Genes that confer a predisposition to common diseases such as breast cancer, colon cancer, hypertension, diabetes, and Alzheimer's disease have also been localized to specific chromosomal regions. All these discoveries benefitted from the information, resources, and technologies developed by human genome research. As the genome project proceeds, many more exciting developments are expected including technology for studying the health effects of environmental agents; the ability to decipher the genomes of many other organisms, including countless microbes important to agriculture and the environment; as well as the identification of many more genes involved in disease. The technology and data produced by the genome project will provide a strong stimulus to broad areas of biological research and biotechnology. Exciting years lie ahead as the Human Genome Project moves toward its second set of 5-year goals.

REFERENCES AND NOTES

1. U.S. Department of Health and Human Services and Department of Energy, *Understanding Our Genetic Inheritance. The U.S. Human Genome Project: The First Five Years* (April 1990).
2. National Institutes of Health, National Center for Human Genome Research, Office of Communications, Bethesda, MD 20892. Phone, (301)402-0911; Fax, (301)402-4570.
3. U.S. Department of Energy, Human Genome Management Information System, Oak Ridge National Laboratory, PO Box 20008, Oak Ridge, TN 37831–6050. Phone, (615) 576-6669; Fax, (615) 574-9188.
4. National Research Council, Committee on Mapping and Sequencing the Human Genome, *Mapping and Sequencing the Human Genome* (National Academy Press, Washington, DC, 1988).
5. M.V. Olson, L. Hood, C. Cantor, D. Botstein, *Science* **245**, 143 (1989).