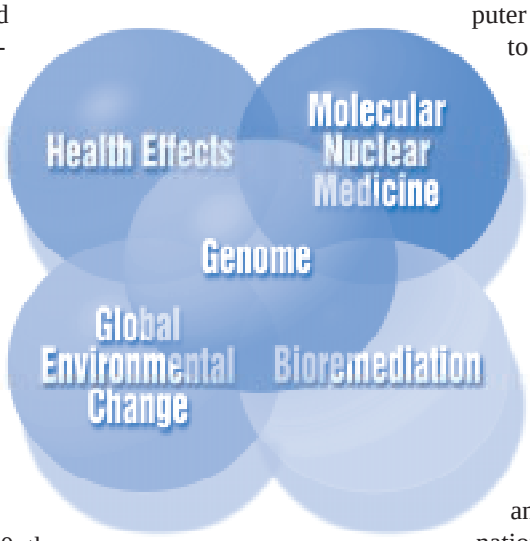
Introduction

genome (jē′nōm), n. all the genetic material in the chromosomes of an organism.

Scientific and technical terms are defined in the Glossary, p. 101. More historical details and other information appear in the Appendices beginning on p. 71.

Now completing its first decade, the Human Genome Program of the U.S. Department of Energy (DOE) is the longest-running federally funded program to analyze the genetic material—the genome— that determines an individual's characteristics at the most fundamental level. Part of the Biological and Environmental Research (BER) Program sponsored by the DOE Office of Biological and Environmental Research (OBER*), the genome program is a major component of the larger U.S. Human Genome Project.



Since October 1990, the project has been supported jointly by DOE and the National Institutes of Health (NIH) National Human Genome Research Institute (formerly National Center for Human Genome Research). Together, the DOE and NIH components make up the world's largest centrally coordinated biology research project ever undertaken.

The U.S. Human Genome Project is a 15-year endeavor to characterize the human genome by improving existing human genetic maps, constructing physical maps of entire chromosomes, and ultimately determining a complete sequence of the deoxyribonucleic acid (DNA) subunits. Parallel studies are being carried out on selected model organisms to facilitate interpretation of human gene function.

*In 1997 the Office of Health and Environmental Research (OHER) was renamed Office of Biological and Environmental Research (OBER).

The ultimate goal of the U.S. project is to identify the estimated 70,000 to 100,000 human genes and render them accessible for future biological study. A complete human DNA sequence will provide physicians and researchers in many biological disciplines with an extraordinary resource: an "encyclopedia" of human biology obtainable by computer and available to all.

For 50 years, programs in the DOE Office of Biological and Environmental Research have crossed traditional research boundaries in seeking new solutions to energy-related biological and environmental challenges (see Appendix F, p. 95, and **http://www.er.doe.gov/production/ober/ober.html***).*

Obtaining the complete sequence by 2005 will require a highly coordinated and focused international effort generating advances in biological methodology; instrumentation (particularly automation); and computer-based methods for collecting, storing, managing, and analyzing the rapidly growing body of data.

Project Origins

The potential value of detailed genetic information was recognized early; until recently, however, obtaining this information was far beyond the capabilities of biomedical research. DOE OBER and its two predecessor agencies—the Atomic Energy Commission and the Energy Research and Development Administration—had long sponsored genetic research in both microbial and higher systems. These studies included explorations into population genetics; genome structure, maintenance, replication, damage, and repair; and the consequences of genetic mutations. These traditional DOE activities evolved naturally into the Human Genome Program.

OBER's mission is described more fully in the Program Management section (p. 59) of this report.

By 1985, progress in genetic and DNA technologies led to serious discussions in the scientific community about initiating a major project to analyze the structure of the human genome. After concluding that a DNA sequence would offer the most useful approach for detecting inherited mutations, DOE in 1986 announced its Human Genome Initiative. The initiative emphasized development of resources and technologies for genome mapping, sequencing, computation, and infrastructure support that would culminate in a complete sequence of the human genome.

The National Research Council issued a report in 1988 recommending a dedicated research budget of $200 million annually for 15 years to determine the sequence of the 3 billion chemical subunits (base pairs) in the human genome and to map and identify all human genes.

To launch the nation's Human Genome Project, Congress appropriated funds to DOE and also to NIH, which had long supported research in genetics and molecular biology as an integral part of its mission to improve the health of all Americans. Other federal agencies and foundations outside the Human Genome Project also contribute to genome research, and many other countries are making important contributions through their own genome research projects.

Coordinated Efforts

In 1988 DOE and NIH signed a Memorandum of Understanding in which the agencies agreed to work together, coordinate technical research and activities, and share results. The two agencies assumed a joint systematic approach toward establishing goals to satisfy both short- and long-term project needs.

Early guidelines projected three 5-year phases, for which the first plan was presented to Congress in 1990. The 1990

Anticipated Benefits of Genome Research

Predictions of biology as "the science of the 21st century" have been made by observers as diverse as Microsoft's Bill Gates and U.S. President Bill Clinton. Already revolutionizing biology, genome research has spawned a burgeoning biotechnology industry and is providing a vital thrust to the increasing productivity and pervasiveness of the life sciences.

Technology and resources promoted by the Human Genome Project already have had profound impacts on biomedical research and promise to revolutionize biological research and clinical medicine. Increasingly detailed genome maps have aided researchers seeking genes associated with dozens of genetic conditions, including myotonic dystrophy, fragile X syndrome, neurofibromatosis types 1 and 2, a kind of inherited colon cancer, Alzheimer's disease, and familial breast cancer.

Current and potential applications of genome research will address national needs in molecular medicine, waste control and environmental cleanup, biotechnology, energy sources, and risk assessment.

Molecular Medicine

On the horizon is a new era of molecular medicine characterized less by treating symptoms and more by looking to the most fundamental causes of disease. Rapid and more specific diagnostic tests will make possible earlier treatment of countless maladies. Medical researchers also will be able to devise novel therapeutic regimens based on new classes of drugs, immunotherapy techniques, avoidance of environmental conditions that may trigger disease, and possible augmentation or even replacement of defective genes through gene therapy.

Microbial Genomes

In 1994, taking advantage of new capabilities developed by the genome project, DOE formulated the Microbial Genome Initiative to sequence the genomes of bacteria useful in the areas of energy production, environmental remediation, toxic waste reduction, and industrial processing. In the resulting Microbial Genome Project, six microbes that live under extreme conditions of temperature and pressure have been sequenced completely as

plan emphasized the creation of chromosome maps, software, and automated technologies to enable sequencing.

By 1993, unexpectedly rapid progress in chromosome mapping required updating the goals [*Science* **262**, 43–46 (October 1, 1993)], which now project through 1998 (see p. 5). This plan is being revised again in anticipation of the approaching high-throughput sequencing phase of the project. Last year marked an early transition to this phase as many more genome sequencing projects were funded. The second and third phases of the project will optimize resources, refine sequencing strategies, and, finally, completely determine the sequence of all base pairs in the genome.

Another area of DOE and NIH cooperation is in exploring the ethical, legal, and social issues (ELSI) arising from increased availability of genetic data and growing genetic-testing capabilities. The two agencies established a joint working group to confront these ELSI challenges and have cosponsored joint projects and workshops.

DOE Genome Program

A general overview follows of recent progress made in the DOE Human Genome Program. Refer to the timeline (pp. ii–iii) for other achievements toward U.S. goals, including contributions made outside DOE.

Physical maps

For DOE, an early goal was to develop chromosome physical maps, which involves reconstructing the order of cloned DNA fragments to represent their specific originating chromosomes. (A set of such cloned fragments is called a library.) Critical to this effort were the libraries of individual human chromosomes of August 1997. Structural studies are under way to learn what is unique about the proteins of these organisms—the ultimate aim being to use the microbes and their enzymes for such practical purposes as waste control and environmental cleanup.

Biotechnology

The potential for commercial development presents U.S. industry with a wealth of opportunities. Sales of biotechnology products are projected to exceed $20 billion by the year 2000. The genome project already has stimulated significant investment by large corporations and prompted the creation of new biotechnology companies hoping to capitalize on the far-reaching implications of its research.

Energy Sources

Biotechnology, fueled by insights reaped from the genome project, will play a significant role in improving the use of fossil-based resources. Increased energy demands, projected over the next 50 years, require strategies to circumvent the many problems associated with today's dominant energy technologies. Biotechnology promises to help address these needs by providing cleaner means for the bioconversion of raw materials to refined products. In addition, there is the possibility of developing entirely new biomass-based energy sources. Having the genomic sequence of the methane-producing microorganism *Methanococcus jannaschii*, for example, will enable researchers to explore the process of methanogenesis in more detail and could lead to cheaper production of fuel-grade methane.

Risk Assessment

Understanding the human genome will have an enormous impact on the ability to assess risks posed to individuals by environmental exposure to toxic agents. Scientists know that genetic differences make some people more susceptible—and others more resistant—to such agents. Far more work must be done to determine the genetic basis of such variability. This knowledge will directly address DOE's long-term mission to understand the effects of low-level exposures to radiation and other energy-related agents, especially in terms of cancer risk.

produced at Los Alamos National Laboratory (LANL) and Lawrence Livermore National Laboratory (LLNL). These libraries allowed the huge task of mapping and sequencing the entire 3 billion bases in the human genome to be broken down into 24 much smaller single-chromosome units. Availability of the libraries has enabled the participation of many laboratories worldwide. Some three generations of clone libraries with improving characteristics have been produced and widely distributed. In the DOE-supported projects, DNA clones representing chromosomes 16, 19, and 22 have been ordered (mapped) and are now providing material needed for large-scale sequencing.

Sequencing

Toward the goal of greatly increasing the speed and decreasing the cost of DNA sequencing, DOE has supported improvements in standard technologies and has pioneered support for revolutionary sequencing systems. Marked improvements have been made in reagents, enzymes, and raw data quality. Such novel approaches as sequencing by hybridization (using DNA "chips") and mass spectrometry have already found important, previously unanticipated applications outside the Human Genome Project.

Joint Genome Institute

In early 1997, the human genome centers at Lawrence Berkeley National Laboratory, LANL, and LLNL began collaborating in the Joint Genome Institute (JGI), within which high-throughput sequencing will be implemented [see p. 26 and *Human Genome News* **8**(2), 1–2]. The initial JGI focus will be on sequencing areas of high biological interest on several chromosomes, including human chromosomes 5, 16, and 19. Establishment of JGI represents a major transition in the DOE Human Genome Program.

Previously, most goals were pursued by small- to medium-sized teams, with modest multisite collaborations. The JGI will house high-throughput implementations of successful technologies that will be run with increasingly stringent process- and quality-control systems.

In addition, a small component aimed at understanding how genes function in the body—a field known as functional genomics—has been established and will grow as sequencing targets are met. High-throughput functional genomics represents a new era in human biology, one which will have profound implications for solving biological problems.

Informatics

In preparation for the production-sequencing phase, many algorithms for interpreting DNA sequence have been developed, and an increasing number have become available as services over the Internet. Last year, the GRAIL (for Gene Recognition and Analysis Internet Link) and GenQuest servers, developed and maintained at Oak Ridge National Laboratory, processed an average of almost 40 million bases of sequence each month.

As technology improves and data accumulates exponentially, continued progress in the Human Genome Project will depend increasingly on the development of sophisticated computational tools and resources to manage and interpret the information. The ease with which researchers can access and use the data will provide a measure of the project's success. Critical to this success is the creation of interoperable databases and other computing and informatics tools to collect, organize, and interpret thousands of DNA clones.

For additional information on the DOE genome programs, refer to Research Highlights, p. 9; Research Narratives, p. 25; this report's *Part 2, 1996 Research Abstracts*; and the Web site (*http://www.ornl.gov/hgmis*).

Five-Year Research Goals of the U.S. Human Genome Project

October 1, 1993, to September 30, 1998 (FY 1994 through FY 1998)*

Major events in the U.S. Human Genome Project, including progress made toward these goals, are charted in a timeline on pp. ii–iii.

Genetic Mapping

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

Physical Mapping

- Complete a sequence tagged site (STS) map of the human genome at a resolution of 100 kb.

DNA Sequencing

- Develop efficient approaches to sequencing one- to several-megabase regions of DNA of high biological interest.
- Develop technology for high-throughput sequencing, focusing on systems integration of all steps from template preparation to data analysis.
- Build up a sequencing capacity to allow sequencing at 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.

Gene Identification

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

Technology Development

- Substantially expand support of innovative technological developments as well as improvements in current technology for DNA sequencing and for meeting the needs of the Human Genome Project as a whole.

Model Organisms

- Finish an STS map of the mouse genome at a 300-kb resolution.
- Finish the sequence of the *Escherichia coli* and *Saccharomyces cerevisiae* genomes by 1998 or earlier.
- Continue sequencing *Caenorhabditis elegans* and *Drosophila melanogaster* genomes with the aim of bringing *C. elegans* to near completion by 1998.
- Sequence selected segments of mouse DNA side by side with corresponding human DNA in areas of high biological interest.

Informatics

- 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.
- Consolidate, distribute, and continue to develop effective software for large-scale genome projects.
- Continue to develop tools for comparing and interpreting genome information.

Ethical, Legal, and Social Implications

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

Training

- Continue to encourage training of scientists in interdisciplinary sciences related to genome research.

Technology Transfer

- Encourage and enhance technology transfer both into and out of centers of genome research.

Outreach

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

*Original 1990 goals were revised in 1993 due to rapid progress. A second revision was being developed at press time.

Evolution of a Vision:
Genome Project Origins,

In an interview at a DNA sequencing conference in Hilton Head, South Carolina, David Smith, a founder and former Director of the DOE Human Genome Program, recalled the establishment of this country's first human genome project. The impressive early achievements and spin-off benefits, he noted, offer more than mere vindication for project founders. They also provide a tantalizing glimpse into the future where, he observed, "scientists will be empowered to study biology and make connections in ways undreamt of before."*

The DOE Human Genome Program began as a natural outgrowth of the agency's long-term mission to develop better technologies for measuring health effects, particularly induced mutations. As Smith explained it, "DOE had been supporting mutation studies in Japan, where no heritable mutations could be detected in the offspring of populations exposed to the atomic blasts at Hiroshima and Nagasaki. The program really grew out of a need to characterize DNA differences between parents and children more efficiently. DOE led the development of many mutation tests, and we were interested in developing even more sensitive detection methods. Mortimer Mendelsohn of Lawrence Livermore National Laboratory, a member of the International Commission for Protection Against Environmental Mutagens and Carcinogens, and I decided to hold a workshop to discuss DNA-based methods (see Human Genome Project chronology, p. ii).

"Ray White (University of Utah) organized the meeting, which took place in Alta, Utah, in December 1984. It was a small meeting but very stimulating intellectually. We concluded the obvious—that if you really wanted to use DNA-based technologies, you had to come up with more efficient ways to characterize the DNA of much larger regions of the genome. And the ultimate sensitivity would be the capability to compare the complete DNA sequences of parents and their offspring."

Project Begins

Smith recalled reaction to the first public statement that DOE was starting a program with the aim of sequencing the human genome. "I announced it at the Cold Spring Harbor meeting in May 1986, and there was a big hullabaloo." After a year-long review, a National Academy of Sciences National Research Council panel endorsed the project and the basic strategy proposed. Smith pointed out that NIH and others were also having discussions on the feasibility of sequencing the human genome. "Once NIH got interested, many more people became involved. DOE and NIH signed a Memorandum of Understanding in October 1988 to coordinate our activities aimed at characterizing the human genome." But, he observed, it wasn't all smooth sailing. The nascent project had many detractors.

> ***"Genomics has come of age, and it is opening the door to entirely new approaches to biology."***

Responding to Critics

Many scientists, prominent biologists among them, thought having the sequence would be a misuse of scarce resources. Smith, laughing now, recalls one scientist complaining, "Even if I had the sequence, I wouldn't know what to do with it." Other critics worried that the genome project would siphon shrinking research funds away from individual investigator-initiated research projects. Smith takes the opposite view. "In fact, individual investigators can do things they would never be able to do otherwise. We're beginning to see that demonstrated at this meeting. For the first time, we're finding people exploring systematic ways of looking at gene function in organisms. The genome project opens up enormous new research fields to be mined. Cottage-industry biologists won't need a lot of robots, but they will have to be computer literate to put the information all together."

The genome project also is providing enabling technologies essential to the future of the emerging biotechnology industry, catalyzing its tremendous growth. According to Smith, the technologies are capable of more than elucidating the human genome. "We're developing an infrastructure for future research. These technologies will allow us to efficiently characterize any of the organisms out there that pertain to various DOE missions, with such applications as better fuels from biomass, bioremediation, and waste control. They also will lead to a greater understanding of global cycles, such as the carbon cycle, and the identification of potential biological interventions. Look at the ocean; an amazing number of microbes are in there, but we don't know how to use them to influence cycles to control some of the harmful things that might be happening. Up to now, biotechnology has been nearly all health oriented, but applications of genome research to modern biology really go beyond health. That's one of the things motivating our program to try to develop some of these other biotechnological applications."

Responding to criticism about not researching gene function early in the project, Smith reasserted that the purpose of the Human Genome Project is to build technologies and resources that will enable researchers to learn about biology in a much

*The Seventh International Genome Sequencing and Analysis Conference, September 1995.

Present and Future Challenges, Far-Reaching Benefits

more efficient way. “The genome budget is devoted to very specific goals, and we make sure that projects contribute toward reaching them.”

International Scope

Smith credited the international community with contributing to many project successes. “The initial planning was for a U.S. project, but the outcome, of course, is that it is truly international, and we would not be nearly as far as we are today without those contributions. Also, there’s been a fair amount of money from private companies, and support from the Muscular Dystrophy Association in France and The Wellcome Trust in the United Kingdom has been extremely important.”

Technology Advances

While noting enormous advances across the board, Smith cited automation progress and observed that tremendously powerful robots and automated processes are changing the way molecular biology is done. “A lot of novel technologies probably won’t be useful for initial sequencing but will be very valuable for comparing sequences of different people and for polymorphism studies. One of the most gratifying recent successes is the DNA polymerase engineering project. Researchers made a fairly simple change, but it resulted in a thermosequenase that may answer a lot of problems, reduce the cost of sequencing, and give us better data.”

Progress in genome research requires the use of maturing technologies in other fields. “The combination of technologies that are coming together has been fortuitous; for example, advances in informatics and data-handling technologies have had a tremendous impact on the genome project. We would be in deep trouble if they were at a less-mature stage of development. They have been an important DOE focus.”

ELSI

Smith described tangible progress toward goals associated with programs on the ethical, legal, and social issues (ELSI) related to data produced by the genome project. “ELSI programs have done a lot to educate the thinkers, and this has produced a higher level of discourse in the country about these issues. DOE is spending a large fraction of its ELSI money on informing special populations who can reach others. Educating judges has been especially well received because they realize the potential impact of DNA technology on the courts.”

According to Smith, more people and groups need to be involved in ELSI matters. “We have some ELSI products: the DOE-NIH Joint ELSI Working Group has an insurance task force report, and a DOE ELSI grantee has produced draft privacy legislation. Now it’s time for others to come and translate ELSI efforts into policy. Perhaps the new National Bioethics Advisory Commission can do some of this.”

New Model for Biological Research

Smith spoke of a changing paradigm guiding DOE-supported biology. “Some years ago, the central idea or dogma in molecular biology research was that information in DNA directs RNA, and RNA directs proteins. Today, I think there is a new paradigm to guide us: Sequence implies structure, and structure implies function. The word ‘implies’ in our new paradigm means there are rules,” continued Smith, “but these are rules we don’t understand today. With the aid of structural information, algorithms, and computers, we will be able to relate sequence to structure and eventually relate structure to function. Our effort focuses on developing the technologies and tools that will allow us to do this efficiently.”

“That’s how I think about what we do at DOE,” he said. “We’re working a lot on technology and projects aimed at human and microbial genome sequencing. For understanding sequence implications, we are making major, increasing investments in synchrotrons, synchrotron user facilities, neutron user facilities, and big nuclear magnetic resonance machines. These are all aimed at rapid structure determination.” Smith explained that now we are seeing the beginnings of the biotechnology revolution implied by the sequence-to-structure-to-function paradigm. “If you really understand the relationship between sequence and function, you can begin to design sequences for particular purposes. We don’t yet know that much about the world around us, but there are capabilities out there in the biological world, and if we can understand them, we can put those capabilities to use.”

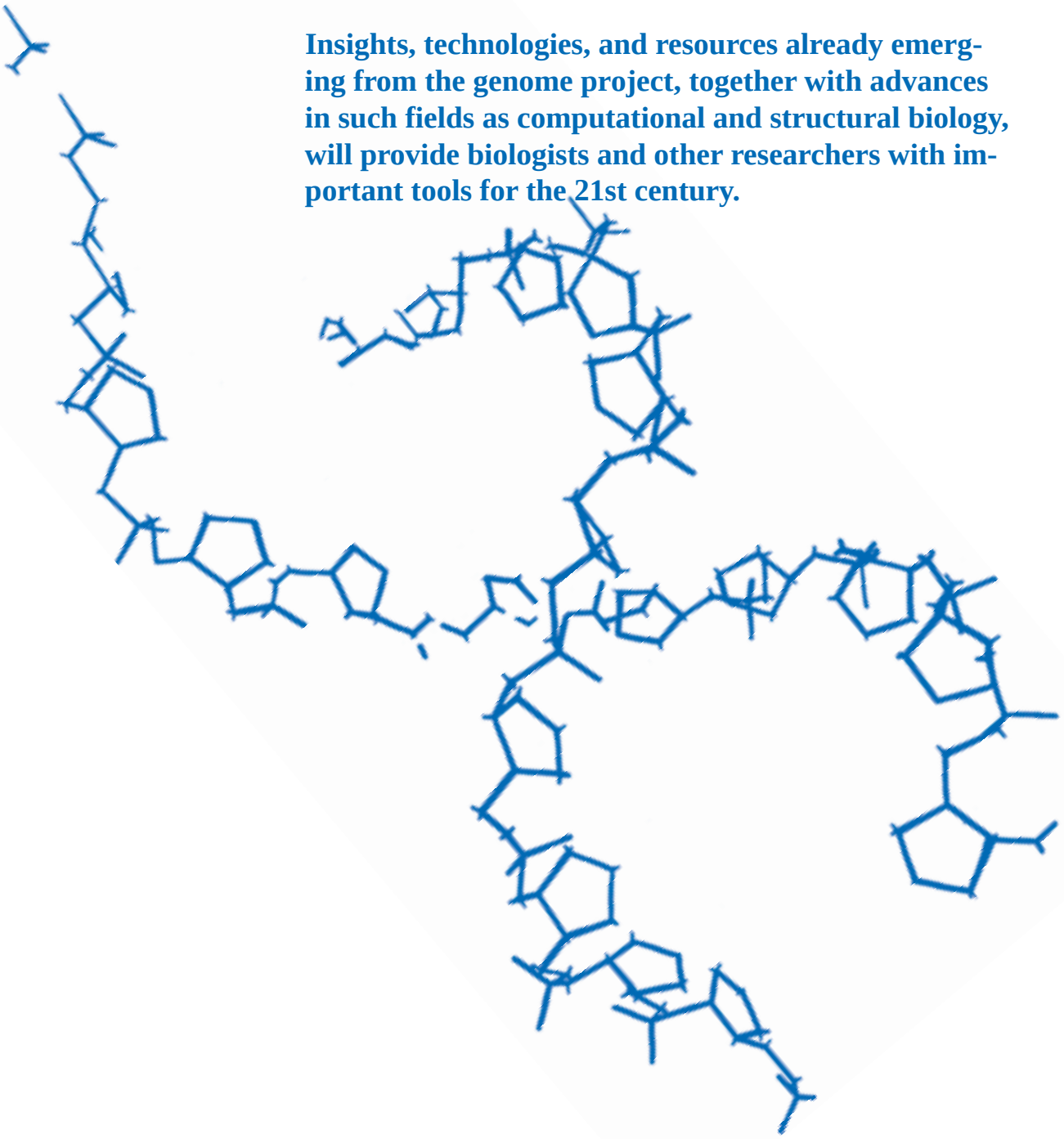
“Comparative genomics,” he continued, “will teach us a tremendous amount about human evolution. The current phylogenetic tree is based on ribosomal RNA sequences, but when we have determined whole genomic sequences of different microbes, they will probably give us different ideas about relationships among archaebacteria, eukaryotes, and prokaryotes.”

Feeling good about progress over the previous 5 years, Smith summed it up succinctly: “Genomics has come of age, and it is opening the door to entirely new approaches to biology.”

David Smith retired at the end of January 1996. Taking responsibility for the DOE Human Genome Program is Aristides Patrinos, who is also Associate Director of the DOE Office of Biological and Environmental Research. Marvin Frazier is Director of the Health Effects and Life Sciences Research Division, which manages the Human Genome Program.

Looking to the Future

Insights, technologies, and resources already emerging from the genome project, together with advances in such fields as computational and structural biology, will provide biologists and other researchers with important tools for the 21st century.



Transitioning to large-scale sequencing

The early years of the Human Genome Program have been remarkably successful. Critical resources and infrastructures have been established, and technologies have been developed for producing several useful types of chromosomal maps. These gains are supporting the project's transition to the large-scale sequencing phase. Some highlights and trends in the U.S. Department of Energy's (DOE) Human Genome Program after FY 1993 are presented in this section.

Clone Resources for Mapping, Sequencing, and Gene Hunting

The demands of large chromosomal mapping and sequencing efforts have necessitated the development of several different types of clone collections (called libraries) carrying human DNA. Three generations of DOE-developed libraries are being distributed to research teams in the United States and abroad. In these libraries, human DNA segments of various lengths are maintained in bacterial cells.

NLGLP Libraries

The first two generations are chromosome-specific libraries carrying small inserts of human DNA (15,000 to 40,000 base pairs). As part of the National Laboratory Gene Library Project (NLGLP) begun in 1983, these libraries were prepared at Los Alamos National Laboratory (LANL) and Lawrence Livermore National Laboratory (LLNL) using DOE flow-sorting technology to separate individual chromosomes. Library availability has allowed the very difficult whole-genome tasks to be divided into 24 more manageable single-chromosome projects that could be pursued at separate research centers. Completed in 1994, NLGLP libraries have provided critical resources to genome researchers worldwide (*http://www-bio.llnl.gov/genome/html/cosmid.html*). Very high resolution chromosome maps based principally on NLGLP libraries were published in 1995 for chromosomes 16 and 19. These are described in detail in the Research Narratives section of this report (see LLNL, p. 27, and LANL, p. 35).

PACs and BACs

The third generation of clone resources supporting chromosome mapping is composed of P1 artificial chromosome (PAC) and bacterial artificial chromosome (BAC) libraries. A prototype PAC library was produced by the team of Leon Rosner (then at DuPont) many years ago, but more efficient production began with improvements introduced by the DOE-supported teams headed by Melvin Simon at Caltech (BACs) and Pieter de Jong at Roswell Park (PACs).

In contrast to cosmids, BACs and PACs provide a more uniform representation of the human genome, and the greater length of their inserts (90,000 to

DOE Genome Research Web Site
http://www.ornl.gov/hgmis/research.html

Research Narratives

Separate narratives, beginning on p. 25, contain detailed descriptions of research programs and accomplishments at these major DOE genome research facilities.

- Lawrence Livermore National Laboratory
- Los Alamos National Laboratory
- Lawrence Berkeley National Laboratory
- University of Washington Genome Sequencing Laboratory
- Genome Database
- National Center for Genome Resources

Research Abstracts

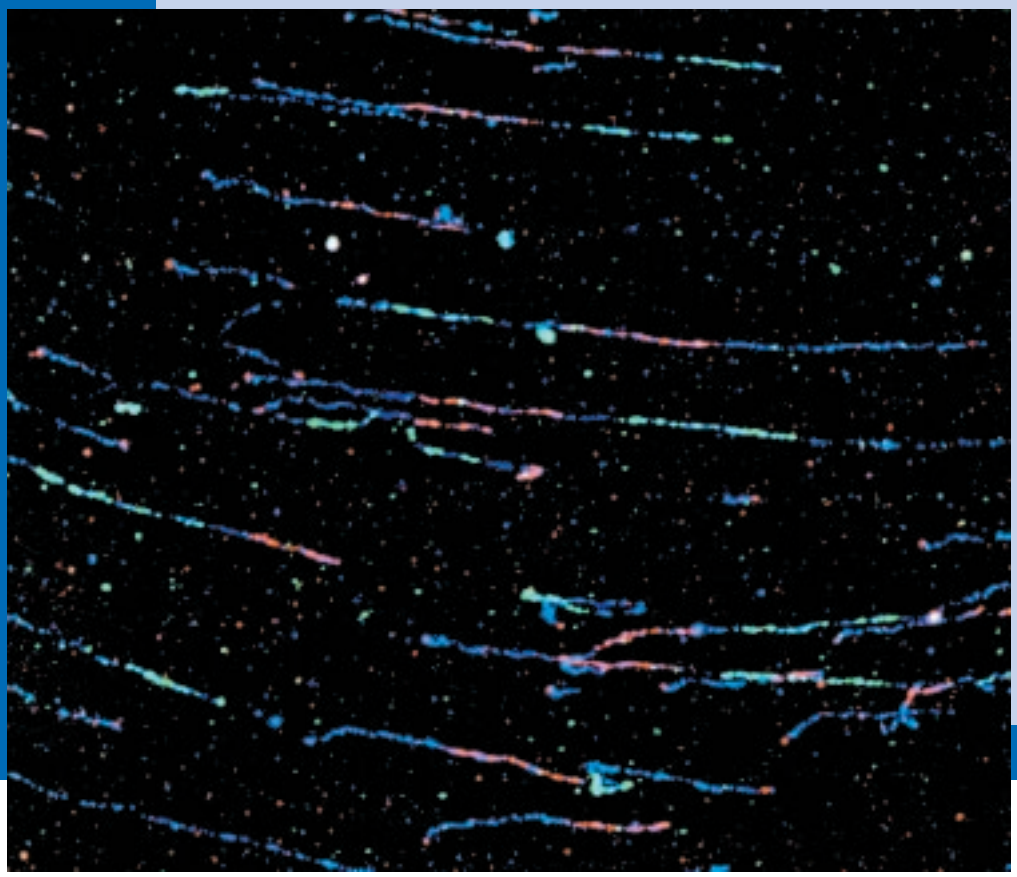
Descriptions of individual research projects at other institutions are given in *Part 2, 1996 Research Abstracts*.

300,000 base pairs) facilitates both mapping and sequencing. Their usefulness was illustrated dramatically in 1993 when the first breast cancer–susceptibility gene (*BRCA*1) was found in a BAC clone after other types of resources had failed. The next year, with major support from NIH, de Jong's PACs contributed to the isolation of the second human breast cancer–susceptibility gene (*BRCA*2).

Mapping

The assembly of ordered, overlapping sets (contigs) of high-quality clones has long been considered an essential step toward human genome sequencing. Because the clones have been mapped to precise genomic locations, DNA sequences obtained from them can be located on the chromosomes with minimal uncertainty.

The large insert size of BACs and PACs allows researchers to visually map them on chromosomes by using fluorescence in situ hybridization (FISH) technology (see photomicrograph below). These mapped BACs and PACs represent very valuable resources for the cytogeneticist exploring chromosomal abnormalities. Two major medical genetics resources have been developed: (1) The Resource for Molecular Cytogenetics at the University of California, San Francisco, in collaboration with the Lawrence Berkeley National Laboratory (LBNL) team led by Joe Gray (*http://rmc-www.lbl.gov*) and (2) The Total Human Genome BAC-PAC Resource at Cedars-Sinai Medical Center, Los Angeles, developed by Julie Korenberg's laboratory (see map, p. 12, and Web site, *http://www.csmc.edu/genetics/korenberg/korenberg.html*).



FISH Mapping on DNA Fibers. The fluorescence microscope reveals several individual cloned DNA fibers from yeast artificial chromosomes (YACs, in blue) after molecular combing to attach and stretch the DNA molecules across a glass microscope slide. Also shown are the locations of two P1 clones, labeled green and red, mapped onto the YAC fibers using FISH. Digital imaging technology can be used to assemble physical maps of chromosomes with a resolution of about 3 to 5 kilobases. ***[Source: Joe Gray, University of California, San Francisco]***

Coordinated Mapping and Sequencing

A simple strategy was proposed in 1996 for choosing BACs or PACs to elongate sequenced regions most efficiently [*Nature* **381**, 364–66 (1996)]. The first step is to develop a BAC end sequence database, with each entry having the BAC clone name and the sequences of its human insert ends. In toto, the source BACs should represent a 15- to 20-fold coverage of the human genome. Then for any BAC or chromosomal region sequenced, a comparison against the database will return a list of BACs (or PACs) that overlap it. Optimal choices for the next BACs (or PACs) to be sequenced can then be made, entailing minimal everlap (and therefore minimal redundancy of sequencing).

Two pilot BAC-PAC end-sequencing projects were initiated in September of 1996 to explore feasibility, optimize technologies, establish quality controls, and design the necessary informatics infrastructure. Particular benefits are anticipated for small laboratories that will not have to maintain large libraries of clones and can avoid preliminary contig mapping (see abstracts of Glen Evans; Julie Korenberg; Mark Adams, Leroy Hood, and Melvin Simon; and Pieter de Jong in Part 2 of this report).

Updated information on BAC-PAC resources can be found on the Web (*http://www.ornl.gov/meetings/bacpac/95bac.html*). [See Appendix C: Human Subjects Guidelines, p. 77 or *http://www.ornl.gov/hgmis/archive/nchgrdoe.html* for DOE-NIH guidelines on using DNA from human subjects for large-scale sequencing.]

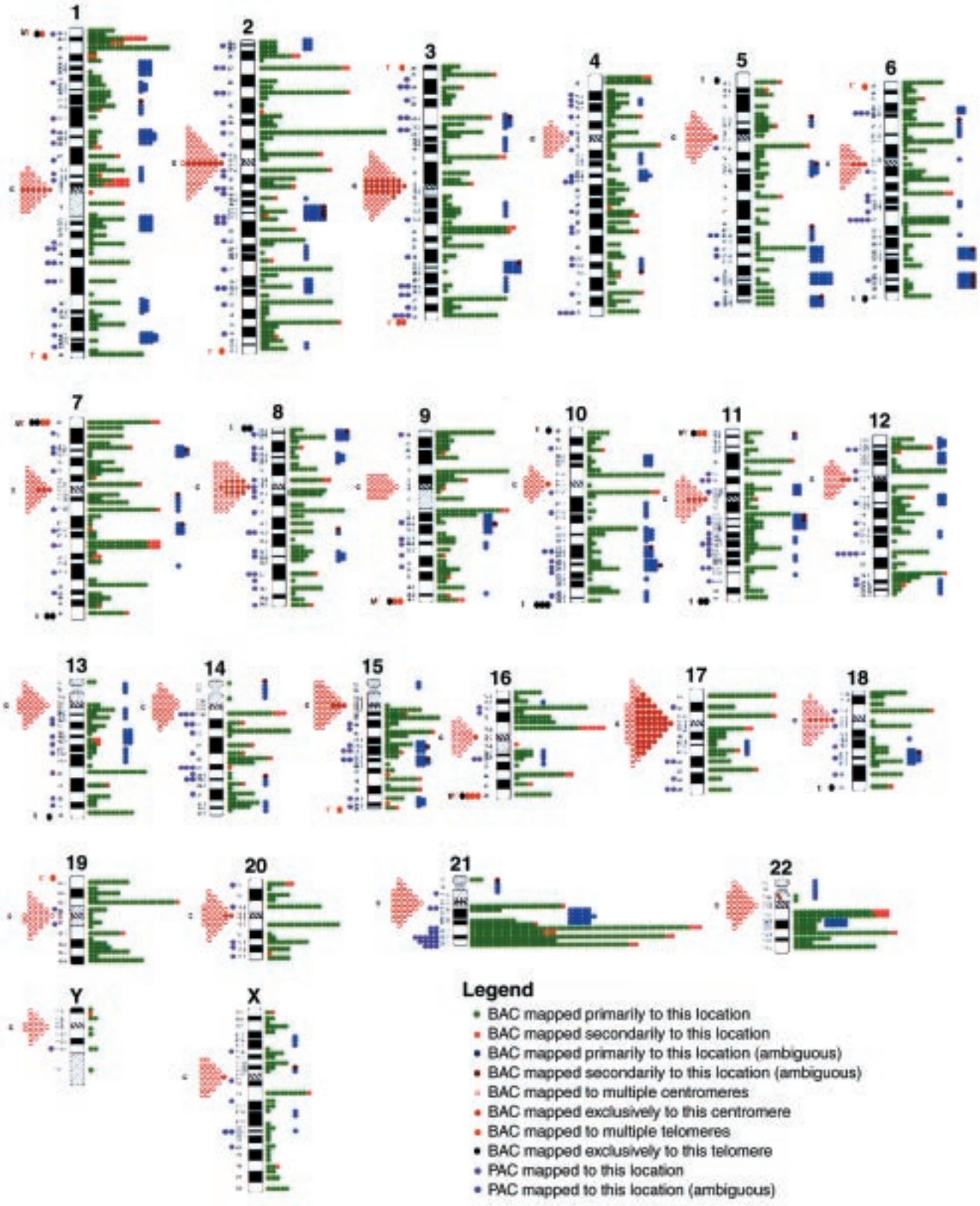
cDNA Libraries

In 1990, DOE initiated projects to enrich the developing chromosome contig maps with markers for genes. Although the protein-encoding messenger RNAs are good representatives of their source genes, they are unstable and must be converted to complementary DNAs (cDNAs) for practical applications. These conversions are tricky, and artifacts are introduced easily. The team led by Bento Soares (University of Iowa) has optimized the steps and continues to produce cDNA libraries of the highest quality. At LLNL, individual cDNA clones are put into standard arrays and then distributed worldwide for characterization by the international IMAGE (for Integrated Molecular Analysis of Gene Expression) Consortium (see box, p. 13).

Initially supported under a DOE cDNA initiative, Craig Venter's team (now at The Institute for Genomic Research) greatly improved technologies for reading sequences from cDNA ends (expressed sequence tags, called ESTs). Together with complementary analysis software, ESTs were shown to be a valuable resource for categorizing cDNAs and providing the first clues to the functions of the genes from which they are derived. This fast EST approach has attracted millions of dollars in commercial investment. Mapping the cDNA onto a chromosome can identify the location of its corresponding gene. Many laboratories worldwide are contributing to the continuing task of mapping the estimated 70,000 to 100,000 human genes.

HAECs

All the previously described DNA clones are maintained in bacterial host cells. However, for unknown reasons, some regions of the human genome appear to be unclonable or unstable in bacteria. The team led by Jean-Michel Vos (University of North Carolina, Chapel Hill) has developed a human artificial episomal chromosome (HAEC) system based on the Epstein-Barr virus that may be useful for coverage of these especially difficult regions. In the broader biomedical community, HAECs also show promise for use in gene therapy.



BAC-PAC Map. *The Total Human Genome BAC-PAC Resource represents an important tool for understanding the genes responsible for human development and disease (*http://www.csmc.edu/genetics/korenberg/korenberg.html*). The Resource, consisting of more than 5000 BAC and PAC clones, covers every human chromosome band and 25% of the entire human genome. Each color dot represents a single BAC or PAC clone mapped by FISH to a specific chromosome band represented in black and white. The clones, which are stable and useful for sequencing, have been integrated with the genetic and physical chromosome maps.* *[Source: Julie Korenberg, Cedars-Sinai Medical Center]*

Resources for Gene Discovery

Hunting for disease genes is not a specific goal of the DOE Human Genome Program. However, DOE-supported libraries sent to researchers worldwide have facilitated gene hunts by many research teams. DOE libraries have played a role in the discovery of genes for cystic fibrosis, the most common lethal inherited disease in Caucasians; Huntington's disease, a progressive lethal neurological disorder; Batten's disease, the most prevalent neurodegenerative childhood disease; two forms of dwarfism; Fanconi anemia, a rare disease characterized by skeletal abnormalities and a predisposition to cancer; myotonic dystrophy, the most common adult form of muscular dystrophy; a rare inherited form of breast cancer; and polycystic kidney disease, which affects an estimated 500,000 people in the United States at a healthcare cost of over $1 billion per year.

The team led by Fa-Ten Kao (Eleanor Roosevelt Institute) has microdissected several chromosomes and made derivative clone libraries broadly available to disease-gene hunters. This resource played a critical role in isolating the gene responsible for some 15% of colon cancers.

Of Mice and Humans: The Value of Comparative Analyses

A remaining challenge is to recognize and discriminate all the functional constituents of a gene, particularly regulatory components not represented within cDNAs, and to predict what each gene may actually do in human biology. Comparing human and mouse sequences is an exceptionally powerful way to identify homologous genes and regulatory elements that have been substantially conserved during evolution.

Researchers led by Leroy Hood (University of Washington, Seattle) have analyzed more than 1 million bases of sequence from T-cell receptor (TCR)

To IMAGE the Human Gene Map

Since 1993, the Integrated Molecular Analysis of Gene Expression (IMAGE) Consortium has played a major role in the development of a human gene map. Founding members of the IMAGE Consortium are Bento Soares (Columbia University, now at University of Iowa), Gregory Lennon (LLNL), Mihael Polymeropoulos (National Institutes of Health's National Institute of Mental Health), and Charles Auffrey (Généthon, in France). Because cDNA molecules represent coding (expressed-gene) areas of the genome, sets of cloned cDNAs are a valuable resource to the gene-mapping community. The cDNA libraries representing different tissues have many members in common. Thus, good coordination among participating laboratories can minimize redundant work. The international IMAGE Consortium laboratories fulfill this role by developing and arraying cDNA clones for worldwide use. *[http://www-bio.llnl.gov/bbrp/image/image.html]*

From the IMAGE cDNA clones, researchers at the Washington University (St. Louis) Sequencing Center determine ESTs with support from Merck, Inc. The data, which are used in gene localization, are then entered into public databases. More than 10,000 chromosomal assignments have been entered into Genome Database (*http://www.gdb.org*). Including replica copies, over 3 million clones have been distributed, probably representing about 50,000 distinct human genes.

The IMAGE infrastructure is being used in two additional programs. At LLNL, the IMAGE laboratory arrays mouse cDNA libraries produced by Soares for the Washington University Mouse EST project (*http://genome.wustl.edu/est/mouse_esthmpg.html*) with sequencing sponsored by the Howard Hughes Medical Institute. Additional clone libraries are being used in a collaborative sequencing project sponsored by the NIH National Cancer Institute as part of the Cancer Genome Anatomy Project to identify and fully sequence genes implicated in major cancers (*http://www.ncbi.nlm.nih.gov/ncicgap*).

chromosome regions of both human and mouse genomes. Many subtle functional elements can be recognized only by comparing human and mouse sequences. TCRs play a major role in immunity and autoimmune disease, and insights into their mechanisms may one day help treat or even prevent such diseases as arthritis, diabetes, and multiple sclerosis (possibly even AIDS).

Comparative analysis is also used to model human genetic diseases. Given sequence information, researchers can produce targeted mutations in the mouse as a rapid and economical route to elucidating gene function. Such studies continue to be used effectively at Oak Ridge National Laboratory (ORNL).

DNA Sequencing

From the beginning of the genome project, DOE's DNA sequencing-technology program has supported both improvements to established methodologies and innovative higher-risk strategies. The first major sequencing project, a test bed for incremental improvements, culminated with elucidation of the highly complex TCR region (described above) by a team led by Hood.

A novel "directed" sequencing strategy initiated at LBNL in 1993 provides a potential alternative approach that can include automation as a core design feature. In this approach, every sequencing template is first mapped to its original position on a chromosome (resolution, 30 bases). The advantages of this method include a large reduction in the number of sequencing reactions needed and in the sequence-assembly steps that follow. To date, this directed strategy has achieved significant results with simpler, less repetitive nonhuman sequences, particularly in the NIH-funded *Drosophila* genome program. The system also is in use at the Stanford Human Genome Center and Mercator Genetics, Inc.

The preparation of DNA clones for sequencing involves several biochemical processing steps that require different solution environments. At the Whitehead Institute, Trevor Hawkins has improved systems for reversible binding of DNA molecules to magnetic beads that are compatible with complete robotic management. The second-generation Sequatron fits on a tabletop with a single robotic arm moving sample trays between servicing stations. This very compact system, supported by sophisticated software, may be ideal for laboratories with limited or costly floor space.

Fluorescent tags are critical components of conventional automated sequencing approaches. The team of Richard Mathies and Alexander Glazer (University of California, Berkeley) has made a series of improvements in fluorescence systems that have decreased DNA input needs and markedly increased the quality of raw data, thereby supporting longer useful reads of DNA sequence.

Complementary improvements in enzymology have been achieved by the team of Charles Richardson and Stanley Tabor (Harvard Medical School). Current widely used procedures for automated DNA sequencing involve cycling between high and low temperatures. The Harvard researchers used information about the three-dimensional structure of polymerases (enzymes needed for DNA replication) and how they function to engineer an improved Taq polymerase. ThermoSequenase, which is now produced commercially as part of the ThermoSequenase kit, reduces the amount of expensive sequencing reagents required and supports popular cycle-sequencing protocols.

The application of higher electrical fields in gel electrophoresis separation of DNA fragments can increase sequencing speed and efficiency. Conventional thick gels cannot adequately dissipate the additional heat produced, however. Two promising routes to "thinness" are ultrathin slab gels and

capillary systems. An ultrathin gel system was developed by Lloyd Smith (University of Wisconsin, Madison) and licensed for commercial development.

The replacement of gels by pumpable solutions of long polymers is making capillary array electrophoresis (CAE) potentially practical for DNA sequencing. The first CAE system for DNA was demonstrated by the team of Barry Karger (Northeastern University). In 1995, Karger and Norman Dovichi (University of Alberta, Canada) separately identified CAE conditions under which DNA sequencing reads could be extended usefully up to the 1000-base range. Another CAE system, developed by Edward Yeung (Iowa State University), has been licensed for commercial production (see box, p. 23). Mathies has developed a system in which a confocal microscope displays DNA bands. Application of this system to the sizing of larger DNA fragments binding multiple fluors allows single-molecule detection.

Replacing the gel-separation step with mass spectroscopy (MS) is another promising approach for rapid DNA sequencing. MS uses differences in mass-to-charge ratios to separate ionized atoms or molecules. Early efforts at MS sequencing were plagued by chemical reactivity during the "launching" phase of matrix-assisted laser desorption ionization (MALDI). MALDI badly degraded the DNA sample input. However, the degradation chemistry was elucidated in Smith's laboratory, leading to improvements. At ORNL, the team of Chung-Hsuan Chen has performed extensive trials of alternative matrices and has achieved significant improvements that now support sequence reads up to 100 DNA bases. The system is undergoing trials for DNA diagnostic applications.

The most revolutionary sequencing technology is being pursued by the team of Richard Keller and James Jett at LANL. Their goal is to read out sequence from single DNA molecules, work that builds on LANL's expertise in flow cytometry. The strand to be sequenced is labeled first with fluors that distinguish the four DNA subunits and is then suspended in a flow stream. An exonuclease cleaves the subunits, which flow past an interrogating laser system that reports the subunits' identities. All system constituents are operational but limited by the low subunit release rates of commercially available exonucleases. A current developmental focus is on identifying more active exonucleases.

Synthetic DNA strands in the 15- to 30-base range (oligomers) play essential roles in DNA sequencing; in sample-preparation steps for the polymerase chain reaction, which copies DNA strands millions of times; and in DNA-based diagnostics. The cost of custom oligomer synthesis once was a limiting factor in many research projects. A more economical, highly parallel oligomer synthesis technology was developed by Thomas Brennan at Stanford University (see last bullet, p. 22, for further details).

The sequencing by hybridization (SBH) technology provides information only on short stretches of DNA in a single trial (interrogation), but thousands of low-cost interrogations can be performed in parallel. SBH is very useful for rapid classification of short DNAs such as cDNAs, very low cost DNA resequencing, and detection of DNA sequence differences (polymorphisms) over short regions. The team of Radomir Crkvenjakov and Radoje Drmanac invented one format of SBH while in Yugoslavia, made substantial improvements at Argonne National Laboratory (ANL), and later started Hyseq Inc. to commercialize these technologies. At ANL, another implementation, SBH on matrices (SHOM) of gels, holds promise for high-accuracy sequence proofreading and diverse DNA diagnostics. The ANL team, led by Andrei Mirzabekov, collaborates

with the Englehardt Institute in Moscow, where SHOM was demonstrated initially.

Informatics: Data Collection and Analysis

Explosive growth of information and the challenges of acquiring, representing, and providing access to data pose continuing monumental tasks for the large public databases. Over the last 3 years, the Genome Database (GDB), the major international repository of human genome mapping data, has made extensive changes culminating in the enhanced representation of genomic maps and gene information in GDB V6.0. Major issues for the Genome Sequence DataBase (GSDB), established in 1994, are to capture and annotate the sequence data and to represent it in a form capable of supporting complex, ad hoc queries. Both GDB and GSDB have been restructured recently to handle the increasing flood of data and make it more useful for downstream biology (see Research Narratives, GDB, p. 49, and GSDB, p. 55. *[http://www.gdb.org* and *http://www.ncgr.org/gsdb]*

Victor Markowitz, formerly of LBNL, has developed a suite of database tools allowing substantial modifications of underlying data structures while the biologists' query tools remain stable. *[http://gizmo.lbl.gov/DM_TOOLS/DMTools.html]*

The Genome Annotation Consortium (based at ORNL) was initiated in 1997 to be a modular, distributed informatics facility for analyzing and processing (e.g., annotating) genome-scale sequence data.

The many improvements in World Wide Web software now enable maps to be downloaded simply by using a browser with accessory software provided by GDB. Computers sift stretches of DNA sequence for patterns that identify such biologically important features as protein-coding regions (exons), regulatory areas, and RNA splice sites. Other computer tools are used to compare a new sequence (i.e., a putative gene) against all other database entries, retrieve any homologous sequences that already have been entered, and indicate the degree of similarity.

The Gene Recognition and Analysis Internet Link (GRAIL) at ORNL localizes genes and other biologically important sequence features (see box, p. 17).

Another analytical service that returns informative, annotated data is MAGPIE, provided through ANL by Terry Gaasterland. MAGPIE is designed to reside locally at the site of a genome project and actively carry out analysis of genome sequence data as it is generated, with automated continued reevaluation as search databases grow (*http://www.mcs.anl.gov/home/gaasterl/magpie.html*). Once an automated functional overview has been established, it remains to pinpoint the organisms' exact metabolic pathways and establish how they interact. To this end, the WIT (What is There) system, which succeeds PUMA, supports the construction of metabolic pathways. Such constructions or models are based on sequence data, the clearly established biochemistry of specific organisms, and an understanding of the interdependencies of biochemical mechanisms. WIT, which was developed by Evgenij Selkov and Ross Overbeek at ANL, offers a particularly valuable tool for testing current hypotheses about microbial biology. *[http://www.cme.msu.edu/WIT]*

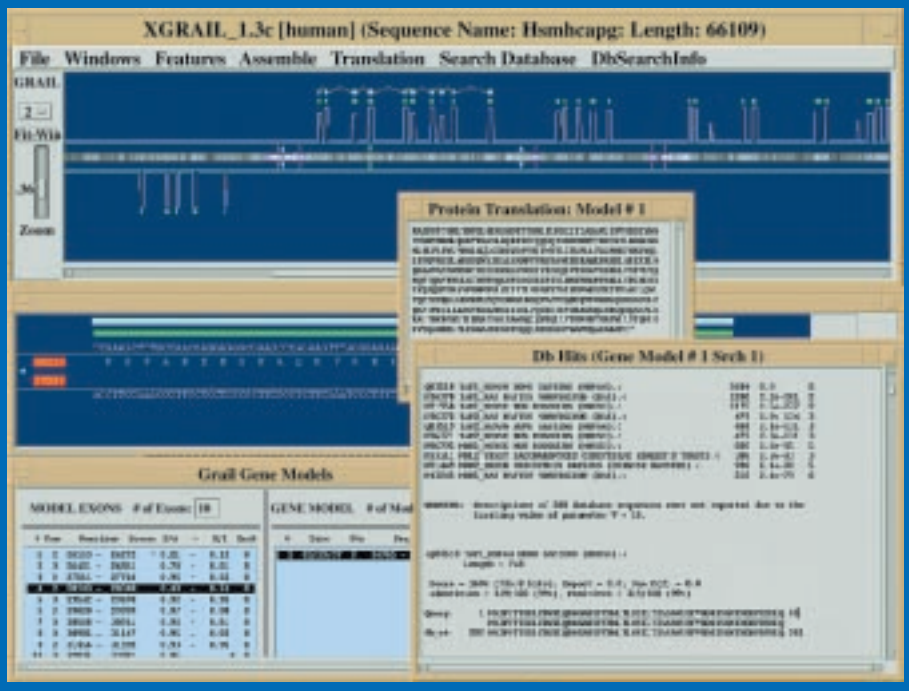
Researchers at the University of Colorado have developed another approach for predicting coding regions in genomic DNA, combining multiple types of evidence into a single scoring function, and returning both optimal and ranked suboptimal solutions. The approach is robust to substitution errors but sensitive to frameshift errors. The group is now exploring methods for predicting other classes of sequence regions, especially promoters. *[*software

GRAIL and GenQuest

In 1996 the Gene Recognition and Analysis Internet Link (GRAIL) processed nearly 40 million bases of sequence per month, making it the most widely used "gene-finding" system available. Developed at Oak Ridge National Laboratory (ORNL) by a team led by Ed Uberbacher, GRAIL uses artificial intelligence and machine learning to discover complex relationships in sequence data. The genQuest server, also at ORNL, compares information generated by GRAIL with data in protein, DNA, and motif databases to add further value to annotation of DNA sequences.

GRAIL's latest version (1.3) combines a Motif Graphical Client with improved sensitivity and splice-site recognition, better performance in AT-rich regions, new analysis systems for model organisms, and frameshift detection. This system can be used on a wide variety of UNIX platforms, including Sun, DEC, and SGI. The many ways to access GRAIL include a command line sockets client that permits remote program calls to all basic GRAIL-genQuest analysis services, thus allowing convenient integration of GRAIL results into automated analysis pipelines.

Contact GRAIL staff through the Web site at *http://compbio.ornl.gov* or at *GRAILMAIL@ornl.gov* for e-mail and ftp access.



The figure above shows the GRAIL analysis of part of the human major histocompatibility locus, which carries genes responsible for cellular immunity. Included in this analysis are potential exons (gene-coding regions), gene models, CpG islands (areas rich in bases C and G found in most mammalian genes), and repetitive DNA elements. ***[Source: Richard Mural, ORNL]***

and information: *http://beagle.colorado.edu/~eesnyder/GeneParser.html]*

The Baylor College of Medicine (BCM) Search Launcher improves user access to the wide variety of database-search tools available on the Web. Search Launcher features a single point of entry for related searches, the addition of hypertext links to results returned by remote servers, and a batch client. *[http://gc.bcm.tmc.edu:8088/search-launcher/launcher.html]*

FASTA-SWAP, also from the BCM group, is a new pattern-search tool for databases that improves sensitivity and specificity to help detect related sequences. BEAUTY, an enhanced version of the BLAST database-search program, improves access to information about the functions of matched sequences and incorporates additional hypertext links. Graphical displays allow correlation of hit positions with annotated domain positions. Future plans include providing access to information from and direct links to other databases, including organism-specific databases.

PROCRUSTES uses comparisons of the same gene of different species to delimit gene structure much more accurately. The product of a collaboration between Pavel Pevzner (University of Southern California) and two Russian researchers, PROCRUSTES is based on the spliced-alignment algorithm, which explores all possible exon assemblies and finds the multiexon structure that best fits a related protein. *[http://www-hto.usc.edu/software/procrustes]*

***The Ethical, Legal, and Social** Issues component of the DOE Human Genome Program supports projects to help judges understand the scientific validity of the genetics-based claims that are poised to flood the nation's courtrooms. Robert F. Orr (left) of the North Carolina Supreme Court and Francis X. Spina of the Massachusetts Appeals Court at the New England Regional Conference on the Courts and Genetics (July 1997) participate in a hands-on laboratory session. As a prelude to learning the fundamentals of DNA science and genetic testing, the judges are precipitating DNA (seen as streaks on the glass rod in the tube) from a solution containing the bacterium* Escherichia coli. *[Courts and Science On-Line Magazine:* http://www.ornl.gov/courts*]*

Ethical, Legal, and Social Issues (ELSI)

From the outset of the Human Genome Project, researchers recognized that the resulting increase in knowledge about human biology and personal genetic information would raise complex ethical and policy issues for individuals and society. Rapid worldwide progress in the project has heightened the urgency of this challenge.

Most observers agree that personal knowledge of genetic susceptibility can be expected to serve humankind well, opening the door to more accurate diagnoses, preventive intervention, intensified screening, lifestyle changes, and early and effective treatment. But such knowledge has another side, too: risk of anxiety, unwelcome changes in personal relationships, and the danger of stigmatization. Often, genetic tests can indicate possible future medical conditions far in advance of any symptoms or available therapies or treatments. If handled carelessly, genetic information could threaten an individual with discrimination by potential employers and insurers.

Other issues are perhaps less immediate than these personal concerns but no less challenging. How, for example, are products of the Human Genome Project to be patented and commercialized? How are the judicial, medical, and educational communities—not to mention the public at large—to be educated effectively about genetic research and its implications?

To confront these issues, the DOE and NIH ELSI programs jointly established an ELSI working group to coordinate policy and research between the two agencies. *[*An FY 1997 report evaluating the joint ELSI group is available on the Web (*http://www.ornl.gov/hgmis/archive/elsirept.html*).*]*

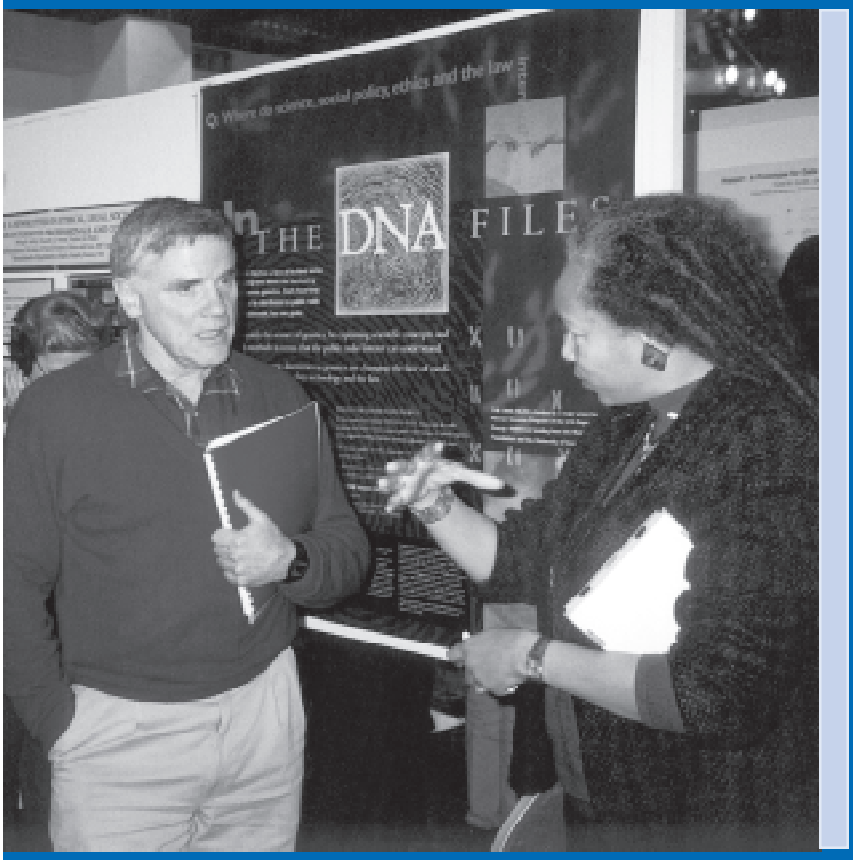
The DOE Human Genome Program has focused its ELSI efforts on education, privacy, and the fair use of genetic information (including ownership and commercialization); workplace issues, especially screening for susceptibilities to environmental agents; and implications of research findings regarding interactions among multiple genes and environmental influences.

A few highlights from the DOE ELSI portfolio for FY 1994 through FY 1997 are outlined below.

- Three high school curriculum modules developed by the Biological Sciences Curriculum Study (BSCS). *[http://www.bscs.org]*
- An educational program in Los Angeles to develop a culturally and linguistically appropriate genetics curriculum based on a BSCS module (see above) for Hispanic students and their families. *[http://vflylab.calstatela.edu/hgp]*
- A series of workshops to educate a core group of 1000 judges around the nation and a handbook with companion videotape to assist federal and state judges in understanding and assessing genetic evidence in an increasing number of civil and criminal cases (see photo above).

- Educational materials developed by the Science+Literacy for Health Project of the American Association for the Advancement of Science (AAAS) and targeted at or above the 6th- to 8th-grade reading levels. [AAAS: 202/326-6453; *Your Genes, Your Choices* booklet: *http://www.nextwave.org/ehr/books/index.html]*
- A program at the University of Chicago aimed at developing a knowledge base for physicians and nurses who will train other practitioners to introduce new genetic services.
- A series of radio programs (see photo at right) on the science and ethical issues of the genome project and a TV documentary program on ELSI issues. *[http://www.pbs.org]*
- *The Gene Letter*, a monthly online newsletter on ELSI issues for healthcare professionals and consumers. *[http://www.geneletter.org]*
- A congressional fellowship program in human genetics, administered through AAAS, for one annual fellowship for a mid-career geneticist. *[society@genetics.faseb.org]*
- The draft Genetic Privacy Act, prepared as a model for privacy legislation and covering the collection, analysis, storage, and use of DNA samples and the genetic information derived from them. *[http://www.ornl.gov/hgmis/resource/privacy/privacy1.html]*
- Privacy studies at the Center for Social and Legal Research, including an analysis of the effects of new genetic technologies on individuals and institutions.

For details on these and other projects, see ELSI Abstracts, p. 45, in Part 2 of this report. In addition to the specific projects listed in Part 2, the DOE program sponsors a number of conferences and workshops on ELSI topics.



Leroy Hood (left) of the University of Washington, Seattle, talks with Bari Scott at the 1996 DOE Human Genome Program Contractor-Grantee Workshop. Scott represented the Genome Radio Project (see text at left), which is supported by the Ethical, Legal, and Social Issues Program of the DOE Human Genome Program. (See the project's abstract in Part 2 of this report for more information.)

DOE ELSI Web Site

http://www.ornl.gov/hgmis/resource/elsi.html

Protection of Human Research Subjects

In 1996, President Clinton appointed the National Bioethics Advisory Commission to provide guidance on the ethical conduct of current and future biological and behavioral research, especially that related to genetics and the rights and welfare of human research subjects (*http://www.nih.gov/nbac/nbac.htm*).

Also in 1996, DOE and NIH issued a document providing investigators with guidance in the use of DNA from human subjects for large-scale sequencing projects (see Appendix C: Human Subjects Guidelines, p. 77). *[http://www.ornl.gov/hgmis/archive/nchgrdoe.html]*

Lawrence Livermore National Laboratory researcher Maria de Jesus, who designed software to automate DNA isolation. [Source: Linda Ashworth, LLNL]