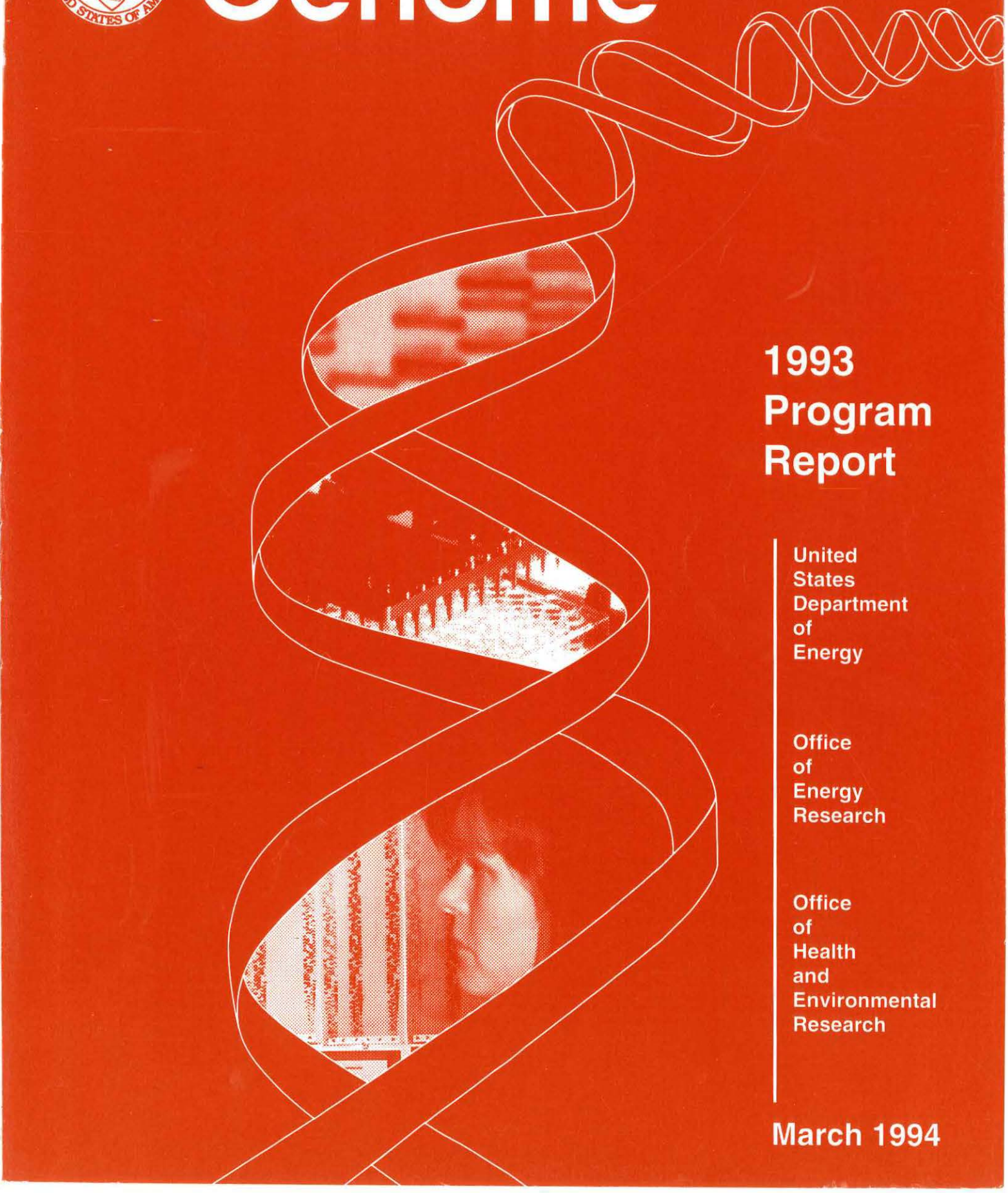
DOE-ER-0611P

Human Genome



1993 Program Report

United States Department of Energy

Office of Energy Research

Office of Health and Environmental Research

March 1994

Please address queries on this publication to:

Human Genome Program
U.S. Department of Energy
Office of Health and Environmental Research
ER-72 GTN
Washington, DC 20585
301/903-6488, Fax: 301/903-8521
Internet: *genome@er.doe.gov*

Human Genome Management Information System
Oak Ridge National Laboratory
P.O. Box 2008
Oak Ridge, TN 37831-6050
615/576-6669, Fax: 615/574-9888
BITNET: *bkq@ornlstc*
Internet: *bkq@ornl.gov*

This report has been reproduced directly from the best obtainable copy.

Available to DOE and DOE contractors from the Office of Scientific and Technical Information; P.O. Box 62; Oak Ridge, TN 37831. Price information: 615/576-8401.

Available to the public from the National Technical Information Service; U.S. Department of Commerce; 5285 Port Royal Road; Springfield, VA 22161.

Printed with soy ink on recycled paper

DOE/ER-0611P

Human Genome 1993 Program Report

Date Published: March 1994

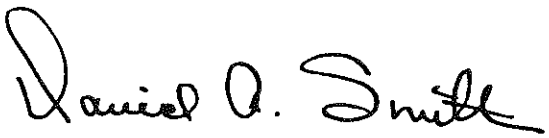
U.S. Department of Energy
Office of Energy Research
Office of Health and Environmental Research
Washington, D.C. 20585

Preface

The purpose of this report is to update the *Human Genome 1991–92 Program Report* (DOE/ER-0544P, published June 1992) and provide new information on the DOE genome program to researchers, program managers, other government agencies, and the interested public. This FY 1993 supplement includes abstracts of 60 new or renewed projects and listings of 112 continuing and 28 completed projects. These two reports, taken together, present the most complete published view of the DOE Human Genome Program through FY 1993.

Research is progressing rapidly toward the 15-year goals of mapping and sequencing the DNA of each of the 24 different human chromosomes. The DOE human genome centers at Lawrence Berkeley Laboratory, Lawrence Livermore National Laboratory, and Los Alamos National Laboratory serve as the focus of much of the multidisciplinary research, and their efforts are complemented by those at other DOE-supported laboratories and more than 80 universities, research organizations, and foreign institutions.

DOE welcomes general or scientific inquiries concerning its Human Genome Program. Announcements soliciting research applications appear early in the year in the *Federal Register*, *Science*, and other publications. The deadline for formal applications is generally midsummer for awards to be made the next fiscal year, and submission of preproposals in areas of potential interest is strongly encouraged. Further information may be obtained by contacting the program office (301/903-6488, Fax: -8521, Internet: *genome@ er.doe.gov*).



David A. Smith, Director
Health Effects and Life Sciences Research Division
Office of Health and Environmental Research
Office of Energy Research
U.S. Department of Energy

Contents

Human Genome 1993 Program Report

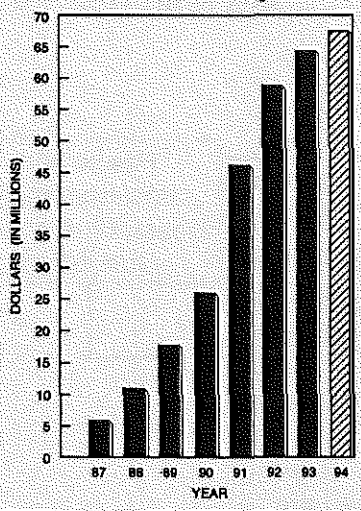
Program Management

The Human Genome Program is a major component of the genetics research programs maintained by the DOE Office of Health and Environmental Research (OHER). A part of the DOE Office of Energy Research, OHER was directed by David J. Galas from April 1990 to August 1993. Aristides Patrinos, Director of the OHER Environmental Sciences Division, is Acting Associate Director of OHER, and John C. Wooley is Deputy Associate Director. The Human Genome Program is administered through the Health Effects and Life Sciences Research Division, directed by David A. Smith.

Resource Allocation

In FY 1993 DOE program expenditures were over $64 million and included $2.3 million in capital equipment expenditures. The table below categorizes the 1993 distribution of funds. The presidential (proposed) budget for the program in FY 1994 is nearly $68 million.

Expenditures and FY 1994 Presidential Budget for the DOE Human Genome Program



Types of Institutions Conducting DOE-Sponsored Genome Research

8	DOE national laboratories
4	Other federal organizations
47	Academic sites
15	Private-sector institutions
15	Companies, including SBIR*
11	Foreign institutions (8 Russian, 2 Canadian, 1 Australian)

*Small Business Innovation Research

Human Genome Program Funds Distribution in FY 1993 (in $K)
(Commitments as of December 1993)

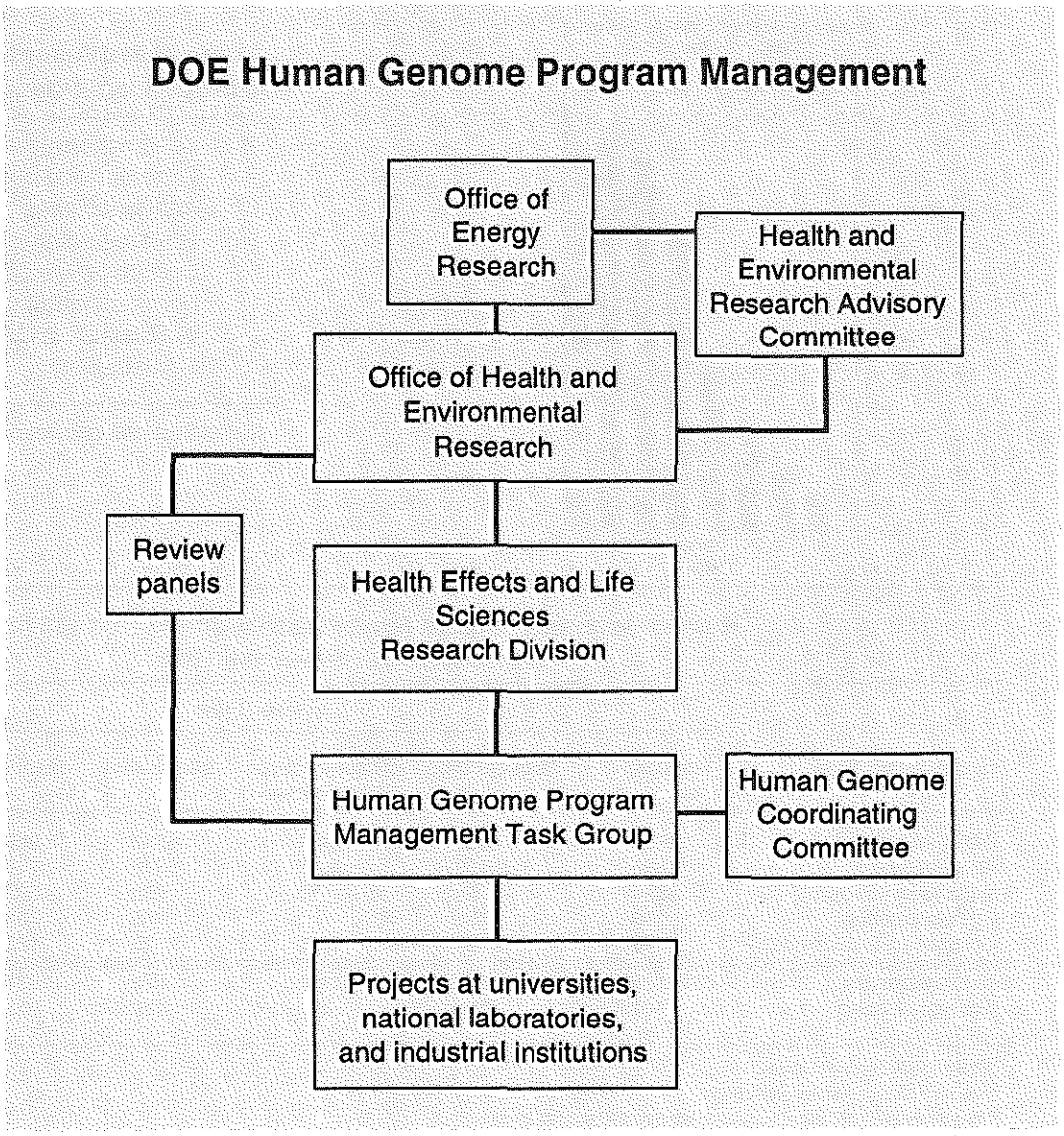
	Project Type						
Organization Type	**Mapping**	**Sequencing**	**Instrumentation Development**	**Informatics**	**ELSI[1]**	**Totals**	**[Percent of Total]**
DOE laboratories	19634	4426	8841	6593	366	39860	63.8
Academic sites	3836	2140	4672	5278	746	16672	26.7
Institutions (nonprofit)	2390	0	50	653	622	3715	5.9
NIH laboratories	190	0	0	85	35	310	0.5
Companies, including SBIR[2]	75	575	1075	70	102	1897	3.0
All organizations	26125	7141	14638	12679	1871	62454[3]	100
[Percent of Total]	[41.8]	[11.4]	[23.4]	[20.3]	[3.0]		[100]

[1] Ethical, Legal, and Social Issues.
[2] Includes $1757 thousand in SBIR grants.
[3] Total allocation of $64 million less capital equipment funds of $2.3 million.

Program Management

Human Genome Coordinating Committee and Human Genome Program Management Task Group

Program coordination is the responsibility of two groups (see chart below): the Human Genome Coordinating Committee (HGCC) and the Human Genome Program Management Task Group (see box for member lists of both groups, p. 3). HGCC, composed of contractor/grantee scientists, provides OHER with external expertise in mapping and sequencing, development and implementation of new technologies, and informatics. HGCC responsibilities include assisting OHER with overall coordination of DOE-funded genome research; facilitating the development and dissemination of novel genome technologies; ensuring proper management and sharing of data and samples; cooperating with other national and international efforts; and helping OHER establish ad hoc task



groups to analyze specific research areas such as ethical, legal, and social issues. HGCC also discusses informatics requirements of DOE programs; mapping and sequencing technologies; use of the mouse as a model organism; costs of resource distribution; use of flow-sorting facilities; research progress; and long-term goals.

The Human Genome Program Management Task Group administers the genome program and its evolving needs and reports to the OHER Director. The task group arranges for periodic workshops and coordinates peer review of research proposals, administration of awards, and collaboration with all concerned agencies and organizations.

Human Genome Coordinating Committee†

Chair: David A. Smith, Office of Health and Environmental Research, DOE

Elbert W. Branscomb, Biologist and Principal Investigator, Human Genome Center Informatics, Lawrence Livermore National Laboratory

Charles R. Cantor, Director, Center for Advanced Research in Biotechnology, Boston University

Anthony V. Carrano, Director, Human Genome Center and Associate Director, Biology and Biotechnology Research Program, Lawrence Livermore National Laboratory

C. Thomas Caskey, Director, Institute for Molecular Genetics, Baylor College of Medicine

Raymond F. Gesteland, Professor and Cochair, Department of Human Genetics, University of Utah; Investigator, Howard Hughes Medical Institute Laboratory for Genetic Studies at the Eccles Institute, University of Utah

Leroy E. Hood, William Gates III Professor of Biomedical Sciences and Chair, Department of Molecular Biotechnology, University of Washington; Director, National Science Foundation Science and Technology Center

David T. Kingsbury, Director, Genome Data Base, Johns Hopkins University

Robert K. Moyzis, Director, Center for Human Genome Studies, Los Alamos National Laboratory

Jasper Rine,* Director, Human Genome Center, Lawrence Berkeley Laboratory

Lloyd M. Smith, Assistant Professor, Analytical Division, Department of Chemistry, University of Wisconsin, Madison

HGCC Executive Officer: **Sylvia J. Spengler**, Human Genome Center, Lawrence Berkeley Laboratory

DOE Human Genome Program Management Task Group†

Member	Specialty
David A. Smith, Chair	Molecular biology
Benjamin J. Barnhart	Genetics/Radiation biology
Daniel W. Drell	Biology/ELSI
Gerald Goldstein	Physical science/ Instrumentation
Aristides Patrinos	Physical science
Robert J. Robbins	Bioinformatics
Murray Schulman	Radiation biology
Jay Snoddy	Molecular biology/ Informatics
Marvin Stodolsky	Molecular biology
David G. Thomassen	Biochemistry
John C. Wooley	Computational Biology

†All members of the DOE Human Genome Program Management Task Group are ex-officio members of HGCC.

*In January 1994 Mohandas Narla became Acting Director of the Lawrence Berkeley Laboratory Human Genome Center and a member of HGCC.

Program Management

Training: Human Genome Distinguished Postdoctoral Fellowships

OHER established the Human Genome Distinguished Postdoctoral Research Program in 1990 to support research on projects related to the DOE Human Genome Program. Current fellowship recipients and their host institutions are listed below. Fellows have the opportunity to participate in advanced genome-related research, interact with outstanding professionals, and become familiar with major programmatic issues. These fellowships complement the Alexander Hollaender Distinguished Postdoctoral Fellowships, which provide support in all areas of OHER-sponsored research. Both postdoctoral programs are administered by the Oak Ridge Institute for Science and Education, which is a university consortium and DOE contractor. For additional information, contact Linda Holmes at 615/576-4805, Fax: -0202.

DOE Human Genome Distinguished Postdoctoral Fellows

1992

Michael Smith (*Salk Institute of Biological Studies*)

Julia Parrish (*Baylor College of Medicine*)

David Lever (*Duke University*)

Rhett Affleck (*Los Alamos National Laboratory*)

Janet Warrington (*University of California, San Francisco/Lawrence Berkeley Laboratory*)

William Bruno (*Los Alamos National Laboratory*)

1993

Jeffrey Elbert (*Oak Ridge National Laboratory*)

John Kececioglu (*University of California, Davis*)

Mark Neff (*Lawrence Berkeley Laboratory*)

DOE Human Genome Research Centers

Established by DOE to foster collaborations by teams of investigators from various disciplines, the three DOE human genome research centers [located at Lawrence Berkeley Laboratory (LBL), Lawrence Livermore National Laboratory (LLNL), and Los Alamos National Laboratory (LANL)] address such major tasks as genetic and physical mapping, DNA sequencing, informatics related to mapping and sequencing, and technology development. Contracts to these centers are funded annually and peer reviewed through site visits every 2 to 3 years. In FY 1993, $30.365 million (48.8%) of the DOE OHER genome program budget was devoted to the centers, with LLNL receiving $9.713 million, LBL $8.599 million, and LANL $12.053 million.

Centers also play a key role in promoting distribution of genome research technology and resources through outside collaborations, public access to laboratory databases, and "visitor laboratories" at which visiting scientists and pre- and postdoctoral students can apply genome center expertise and technology to their own research. In addition, the centers strive to make available their resources—including biologicals, software, databases, instrumentation, and training opportunities—to the entire genome research community.

All center investigators are encouraged to engage in active collaborations with the private sector and transfer their resources and technologies for commercial development. Activities include the transfer of vectors, primers, and software to industry and the further development of instrumentation with industrial partners. Short descriptions and recent accomplishments of each center follow.

Lawrence Berkeley Laboratory

Established in 1988, the LBL Human Genome Center is developing research and analytical technology to speed genome mapping and sequencing and decrease costs. Research at the LBL center focuses on mapping and sequencing chromosome 21, which is the smallest human chromosome and contains an estimated 900 genes. Identified chromosome 21 genes include those implicated in Down's syndrome and familial Alzheimer's disease. A major goal of the center is to use low-density maps to catalyze construction of high-density genetic and physical maps, gene maps, and sequence. Toward this goal the center is sequencing a biologically interesting 3- to 4-Mb region of chromosome 21 (including the Down's syndrome region) and creating high-density genetic maps of that chromosome. Center investigators are aided by groups specializing in robotics, instrumentation automation, and informatics development. The current focus at LBL is to develop technology to sequence 10 Mb or more per year at $0.50 per base with more than 99.9% accuracy. In 1993 Jasper Rine was the center's director, and Sylvia Spengler was the deputy director. In January 1994 Mohandas Narla became acting director of the Human Genome Center.

Recent Accomplishments

- An 80-kb insert from a P1 clone from the Down's region was sequenced using the DOG-tag strategy that generates distance, orientation, and gene-size resolution. Over 150 P1 clones covering the Down's region have been selected and mapped to yeast artificial chromosomes (YACs). These P1s will provide substrates for the DOG-tag strategy.

- Novel mapping resources were developed for chromosome 21, including 60 mapped cDNA clones, over 40 dinucleotide repeat–based genetic markers, a YAC contig map, a set of somatic cell radiation hybrids (with David Cox, Stanford University), a cytogenetic map of the long arm enriched by fluorescence in situ hybridization (FISH) with 280 YACs and cosmids, and 188 cosmids forming 21 multiple YAC contigs and 4 single YAC contigs.

- Instruments supporting automated mapping and sequencing were developed, including an automated colony picker and imaging station with supporting hardware for gel casting and an expandable high-throughput polymerase chain reaction (PCR) apparatus.

- Companion software produced includes 21Bdb, a database variant of ACEDB (the *Caenorhabditis elegans* database) containing physical mapping and sequence data for chromosome 21; a sequence-analysis package; database management tools, including ERDRAW, SDT, QST, and OPM Editor; and an automated generator of user interfaces that changes the interface in response to changes in the underlying database.

Contact: Mohandas Narla, Acting Director (510/486-7029, Fax: -6746; Human Genome Center; LBL; Bldg. 74, Rm. 157; 1 Cyclotron Road; Berkeley, CA 94720).

Lawrence Livermore National Laboratory

The Human Genome Center at LLNL was established in 1990 as an outgrowth of studies on the identification and characterization of human DNA repair genes, specifically on chromosome 19. Major goals include new cloning, mapping, instrumentation, informatics, and sequencing technologies focused on the assembly, closure, and characterization of a high-resolution ordered clone map of human chromosome 19. The final high-resolution map will consist of cosmid contigs with YACs, bacterial artificial chromosomes (BACs), and P1 artificial chromosomes (PACs), as well as an *Eco*R1 restriction map for the minimal spanning set of cosmids. The physical map is being aligned with genetic maps of chromosome 19. Other goals are to isolate, map, and sequence chromosome 19 cDNAs, with emphasis on full-length clones; construct (with LANL) National Laboratory Gene Library Project (NLGLP) chromosome-specific lambda and cosmid libraries for distribution; and develop new cloning, mapping, and sequencing technologies. Anthony V. Carrano is the center's director.

Recent Accomplishments

- Coverage has been achieved on an estimated 90% of chromosome 19 with about 800 contigs assembled from over 12,000 cosmids, with use of automated fluorescence-based fingerprinting.
- Rapid closure of gaps is progressing with the use of YACs, BACs, and PACs; 400 contigs representing about 60% of the chromosome have been regionally mapped to bands by FISH.
- More than 180 genes and polymorphic markers were localized on the contig map. FISH was used to localize over 800 cosmids to bands, including an ordered set of cosmids spaced an average of 1 Mb over the whole chromosome and 500 kb in selected bands.
- Over 166 kb from chromosome 19 repair-gene regions were sequenced.
- With several collaborators, organization was determined for a number of gene families, including carcinoembryonic, olfactory receptor, zinc finger, cytochrome P450, and fucosyl transferase.
- New vectors for cosmid and P1 cloning systems and novel *Alu* PCR primers for fingerprinting and forensic analysis were developed.
- Construction was completed for the NLGLP chromosome-specific lambda and cosmid libraries from sorted chromosomes 3, 7, and X; libraries for chromosomes 9, 12, 18, 19, 21, 22, and Y are also complete.
- Integrated mapping-analysis software was developed with interactive graphical display and linkage to local and national databases.
- A high-speed flow cytometer with sorting capabilities was completed for cell and chromosome purification.
- New instrumentation was developed for high-density filter production.

Research Centers: LLNL

Contacts: Linda Ashworth, Assistant to Center Director (510/422-5665, *ashworth1@llnl.gov*) or Anthony Carrano, Director (510/422-5698, Fax: /423-3110, *carrano1@llnl.gov*); Human Genome Center; LLNL; Biology and Biotechnology Research Program; 7000 East Avenue, L-452; P.O. Box 808; Livermore, CA 94551.

Graduate and postdoctoral research training is available through the Institute of Genetics and Genomics at LLNL (Harvey Mohrenweiser, 510/423-0534).

Los Alamos National Laboratory

The Center for Human Genome Studies at LANL was established in 1988. The LANL center's goals include assembly of complete high-resolution (0.1 Mb) maps of chromosome 16 and regions of chromosome 5, studies at the molecular level of chromosome structure and function, and isolation of selected genes of interest on chromosomes 5 and 16. Other goals are (1) short-term computational development and support for large-scale physical mapping and sequencing projects and long-term development of tools for storage, manipulation, and analysis of genome data; (2) development and application of new methods for physical mapping; (3) use of robotics in handling and storing DNA fragments; (4) construction of DNA libraries from flow-sorted chromosomes; (5) rapid, inexpensive, large-scale sequencing; and (6) studies of ethical, legal, and social issues arising from the increased availability and use of genome data. Technology transfer activities are progressing and include robotic instrumentation development, software licensing, library distribution, and rapid sequencing technology. Robert K. Moyzis is the center's director and Larry L. Deaven is the deputy director.

Recent Accomplishments

- Two maps have been constructed: (1) a high-resolution cosmid/YAC physical map of human chromosome 16 consisting of regionally localized contigs covering over 95% of the chromosome and (2) a framework sequence tagged site map of human chromosome 5 consisting of 300 new markers regionally assigned with a resolution of 4 Mb on 5q to 1 Mb on 5p.

- The unusual 3-D structure of telomeric DNA was determined (with Alex Rich, Massachusetts Institute of Technology); the human telomere represents the endpoint for genetic and physical maps and was identified and cloned at LANL in 1988.

- Highly conserved centromeric repetitive DNA regions, likely human centromere components, were identified and cloned.

- A novel complementary DNA (cDNA) library was constructed from mRNA obtained from a paternally encoded human pregnancy (hydatidiform mole).

- NLGLP chromosome-specific phage and cosmid libraries were constructed with LLNL (over 2500 DNA libraries were sent to research and production laboratories worldwide, including complete digest libraries for each human chromosome); also constructed were partial-digest phage or cosmid libraries for human chromosomes 4, 5, 6, 8, 10, 11, 13, 14, 16, 17, 20, X, and Y; and complete digest low-chimeric YAC libraries for human chromosomes 5, 9, 16, and 21.

- Flow-cytometry techniques were developed to detect single DNA molecules, resulting in a Cooperative Research and Development Agreement (called CRADA) with Life Technologies, Inc., for codevelopment of a rapid DNA-sequencing technology.

- A robot was constructed for high-density cosmid/YAC array replication and distribution.

Research Centers: LANL

- SIGMA (System for Integrated Genome Map Assembly), an X windows–based software tool, was developed for creating, editing, and viewing integrated genome maps.
- cDNA-Inform, a collection of public and private sequences and software for automatic searches and comparisons, was produced.

Contact: Lynn Clark, Technical Coordinator (505/667-9376, Fax: -2891; *moyzis@flovax.lanl.gov*); Center for Human Genome Studies; LANL; Life Sciences Division, MS M886; Los Alamos, NM 87545.

Abstracts of DOE-Funded Research

The abstracts in this section (beginning on p. 19) were contributed by DOE Human Genome Program grantees and contractors. Names of principal investigators are in bold print. Telephone and telefax numbers and electronic mail address are for the first investigator named unless noted otherwise. Descriptions of the research categories follow, and a table of contents of categorized projects with their principal investigators begins on the next page. An index of all investigators named in the abstracts is given at the end of this report.

Descriptions of Research Category Listings

- **Projects New in FY 1993:** Projects that were reviewed by the fall 1992 panel and funded in the genome program are represented by complete abstracts (including investigators, affiliations, and contact information).
- **Projects Renewed in FY 1993:** FY 1992 final-year projects that were renewed for FY 1993 are also represented by complete abstracts.
- **Projects Continuing into FY 1993:** Projects for which funding was initiated prior to FY 1993 and continues in FY 1993 are represented by titles, investigators, affiliations, and contact information. Complete abstracts are in the DOE *Human Genome 1991–92 Program Report* (please consult its index beginning on p. 243).
- **Projects Completed in FY 1993:** Titles of projects that terminated in the fall of 1992 are listed with their investigators.

Project Categories and Principal Investigators*

***Thirteen new projects (designated by an asterisk) are funded through small emergency grants to Russian scientists following December 1992 site reviews by David Galas (formerly OHER), Raymond Gesteland (University of Utah), and Elbert Branscomb (LLNL).**

Principal investigators of the research projects described by the abstracts in this section are listed here under their respective subject categories.

RESOURCE DEVELOPMENT

New Projects

Projects Continuing into FY 1993

PHYSICAL AND GENETIC MAPPING

New Projects

Project Categories and Principal Investigators

ETHICAL, LEGAL, AND SOCIAL ISSUES

Project Categories and Principal Investigators

INFRASTRUCTURE

Renewed Projects

Projects Continuing into FY 1993

SMALL BUSINESS INNOVATION RESEARCH (SBIR)

SBIR Phase I

Projects New in FY 1993

Projects Continuing into FY 1993

SBIR Phase II

Projects Continuing into FY 1993

SBIR Phases I and II

Projects Continuing into FY 1993

COMPLETED PROJECTS

Projects in this category are either completed or no longer receiving support through the DOE Human Genome Program.

Resource Development

Physical and Genetic Mapping

Mapping Instrumentation

Sequencing Technologies

Informatics

Ethical, Legal, and Social Issues

Infrastucture

Project Categories and Principal Investigators

SBIR

Projects New in FY 1993

***Thirteen new projects (designated by an asterisk) are funded through small emergency grants to Russian scientists following December 1992 site reviews by David Galas (formerly OHER), Raymond Gesteland (University of Utah), and Elbert Branscomb (LLNL).**

*Toward Cloning Human Chromosome 19 in Yeast Artificial Chromosomes

Inga P. Arman, Alexander B. Devin, and Svetlana P. Legchilina
Laboratory of Molecular Yeast Genetics; Institute of Molecular Genetics; Russian Academy of Sciences; Moscow 123182, Russia
+7-095/196-5625, Fax: -0221, Internet: *img@glass.apc.org*

We are collaborating on a project that has focused the efforts of several Russian laboratories on cloning, mapping, and sequencing the coding regions of human chromosome 19. We have constructed a partial yeast artificial chromosome (YAC) library from a hybrid (human/hamster) cell line that retains only human chromosome 19 and are screening the library for human DNA fragments. For now, our laboratory is primarily concerned with optimizing YAC library storage.

Long DNA fragments cloned in YACs are often used as a source of starting material for further restriction analysis and sequencing. However, maintaining cloned human DNA in yeast cells is well known to be a cause of human sequence variation. In some cases this variation, sometimes referred to as structure instability, is quite noticeable. Minimizing variation by proper YAC maintenance is therefore highly desirable.

One way to reduce the frequency of inaccurate sequences in a YAC library is to introduce certain changes in the genotype of recipient cells. Mutations of genes RAD52 and RAD1 have been found to cause reduction in chimerism and recombinational rearrangements in some YACs. We have begun a systematic study to determine the relationship of cloned DNA and host genotype. We isolated and characterized mutations in several nuclear spontaneous rho-mutability (SRM) genes that mediate maintenance of various redundant and optional genetic structures in yeast cells. The first step of this project will be to analyze YAC maintenance (i.e., structure variation and mitotic stability of YACs from our library) in mutant SRM cells.

*Refining the Map Location of 5q31-5q33 Deletion with Known Molecular Markers

Maxim L. Filipenko and **Elena I. Yantsen**
Human Genome Group; Department of Molecular Biology; Institute of Bioorganic Chemistry; Siberian Branch of the Russian Academy of Sciences; Novosibirsk 630090, Russia
+7-3832/351-667, Fax: -665, Internet: *max@modul.bioch.nsk.su*

A useful approach for constructing physical maps of defined chromosomal regions is to use a hybrid cell line containing human chromosomes with a deletion in the DNA region of interest. DNA from such cells may be used for selection of a restricted number of cosmid clones to enable easier mapping of a particular chromosome.

We prepared a hybrid cell line containing human chromosomes having a deletion that maps to 5q31-5q33 by G-banding. Our goal is to identify molecular markers in the deletion region and use them to screen yeast artificial chromosome (YAC) and cosmid libraries for corresponding clones. These markers and clones will provide useful starting points for developing physical maps of ordered clones.

The work will be performed in two steps: (1) from available sequence data, polymerase chain reaction primers will be designed and used to test for the presence of these markers in the 5q31-5q35 region; and (2) markers mapping positively to this region will be used to screen the YAC and cosmid libraries.

This work is being carried out as part of the Chromosome 5 project of the Russian National Human Genome Program.

Projects New in FY 1993

A Novel Bacteriophage P1–Derived Electroporation-Based Vector for the Construction of Large-Insert Recombinant DNA Libraries

Jeff Gingrich, Mark Batzer, Jeffrey Garnes, and **Anthony Carrano**
Human Genome Center; Lawrence Livermore National Laboratory; Livermore, CA 94550
510/423-8145, Fax: -3608, Internet: *gingrich1@llnl.gov*

A number of different cloning systems are currently being used to generate contiguous physical maps of individual chromosomes; these systems include yeast artificial chromosomes (YACs), cosmids, and bacteriophage P1. Each of these traditional cloning vehicles is either limited in the ability to propagate large DNA fragments (e.g., 40 kb for cosmids and 90 kb for P1 phage) or contains a high proportion of chimeric DNA molecules (YACs). To mitigate these limitations, we have modified the bacteriophage P1 vector system for producing recombinant molecules by transformation into host cells through electroporation. We call this cloning system P1-derived artificial chromosomes (PACs).

We are presently constructing a fivefold redundant total human genomic library using the PAC cloning system. The PAC cloning vector offers many of the advantages of bacterial artificial chromosomes (BACs), which are derived from the bacterial F-factor and offer ease of manipulation in bacterial hosts as well as a very low frequency of chimeric clones. In addition, the PAC cloning system offers a selectable marker for recombinant clones (*SacB* gene product, levansucrase) using medium that contains sucrose for selection and has the ability to amplify clones using an isopropyl ß-D-thiogalactopyranoside (IPTG)-inducible high-copy-number origin of replication. The maximum size of PAC clones is not limited by a headful packaging mechanism (traditional P1 cloning systems) and can accommodate inserts up to 300 kb in length.

To isolate chromosome 19–specific PAC clones for closing the chromosome, we used a Biomek workstation to construct a set of high-density colony filters from a portion of the PAC library. The filters have been screened using a variety of chromosome 19–specific probes, including inter-*Alu* polymerase chain reaction (PCR) products, "degenerate oligo-primed" PCR (DOP-PCR) products, and a 37-bp repetitive element (pe670). We isolated a number of chromosome 19–specific clones, and six putative pe670-positive PAC clones have been identified. Assuming a random distribution of pe670 repeats on chromosome 19, screening a 0.25-fold redundant filter set should result in localizing 13 pe670-positive PACs. The isolation of six pe670-positive PACs probably results from the nonrandom distribution of the pe670 repeat along chromosome 19. PACs that contained pe670 were used as probes to screen chromosome 19-specific high-density cosmid filters. The hybridization of individual PAC clones to a number of chromosome 19–specific cosmids is being confirmed. These data demonstrate the potential of PAC clones for generating a contiguous chromosome 19 physical map.

*Identification and Mapping of DNA-Binding Proteins Along Genomic DNA by DNA-Protein Crosslinking

A. D. Mirzabekov, V. L. Karpov, O. V. Preobrazhenskaya, D. A. Papazenko, and I. V. Priporava
Engelhardt Institute of Molecular Biology; Russian Academy of Sciences; Moscow 117984, Russia
Fax: +7-095/135-1405, Internet: *amir@imb.msk.su*

Techniques such as chemical protection and enzymatic footprinting together with gel-retardation assays enable the mapping of proteins on genomic DNA but provide no information on bound proteins. We have developed the "protein-image" hybridization method, which allows us to identify the size of proteins associated with specific DNA sequences. This is accomplished either directly in vivo in whole-cell experiments by using uv-induced DNA-protein crosslinks or in isolated nuclei by chemical crosslinking.

A protein can be precisely localized on DNA by digestion of crosslinked DNA with restriction endonuclease and exonuclease III. Many sequences can be tested by repeatedly hybridizing the same blot with a number of probes. Successful development of genome programs will provide huge amounts of new sequences, including those of regulatory and structural regions that may be associated with specific proteins. Without much additional effort, mapping of proteins along genomic DNA can be directly coupled with the program of genome sequencing.

*Sequence-Specific Proteins Binding to the Repetitive Sequences of the Human Genome

Olga Igorevna Podgornaya, Olga Barmina, Tamara Smirnova, and Aleksey Mittenberg
Laboratory of Cell Biology; Department of Cell Cultures; Institute of Cytology; Russian Academy of Sciences; St. Petersburg 194064, Russia
+7-812/247-7450, Fax: -0341, Internet: *root@cell.spb.su*

Although repetitive sequences occupy a large part of the whole eukaryotic genome, their purpose is not well understood. We will attempt to elucidate their role by examining sequence-specific DNA-binding proteins (SSBPs). Methods and technology for finding SSBP were established by using human *Alu* retroposon-like short interspersed repeated sequences (SINEs) as a model. Two proteins with molecular weights of 40 and 80 kilodaltons (kDa) were found in HeLa nuclear extracts and partially purified. The 80-kDa protein (*Alu*-SSBP) was found to bind the sequence within the *Alu* repeat that is homologous to the T-antigen binding site of SV40; cell cycle control may be mediated via *Alu*-SSBP.

Repetitive sequences with simple structures might play a role in three-dimensional chromatin organization. The presence of SSBP that specifically binds with human satellite 3 (HS3) was shown using the developed model system and a gel shift assay with constituents from a partially solubilized nuclear matrix. HS3 has multiple binding sites for this specific SSBP, and binding influences the secondary structure of HS3 by bending the DNA. The molecular weight of HS3-SSBP is 80 kDa; in the most stable retarded complex it apparently exists as a dimer, and in solution under physiological conditions it assembles with other matrix proteins into complexes with molecular weights of about 800 kDa.

A human embryo λgt11 expression library was screened with labeled HS3 under suitable conditions for DNA-protein binding. The identified clone produced a fusion protein with strong affinity for HS3. Immunofluorescence with antibodies against this fusion protein demonstrated a characteristic intranuclear pattern in HeLa cells and on the isolated nuclear matrix. Future plans are to obtain the genes of the *Alu*-binding proteins and look for proteins to the other chromosome-specific satellites in the nuclear matrix.

*Protein-Binding DNA Sequences

O. L. Polanovsky, A. G. Stepchenko, and N. N. Luchina
Engelhardt Institute of Molecular Biology; Russian Academy of Sciences; Moscow 117984, Russia
Fax: +7-095/135-1405, Internet: *pol@imb.msk.su*

A random modification method was developed to determine a target site on DNA that would be recognized by specific DNA binding proteins such as the Oct-2B transcription factor. To this end we have used as a probe a random oligonucleotide of the following structure:

CCGGGAAGCTG*nnnnnnnn*GTGCTGCCTTCGAC*nnnnnnnn*CACGACGGGCC,

where *n* is any of the four bases A, G, C, or T.

DNA fragments containing the binding area were cloned and sequenced. We found two groups of sequences interacting with the conservative POU domain of octomer proteins, one group containing a common tetranucleotide T/CAAA and the other having the tetranucleotide TAAT. All tested probes

Projects New in FY 1993

differ in their affinity to the POU domain. The stability of DNA-protein complexes depends on the structure of core and flanking sequences. The affinity of a group to the POU domain depends particularly on the presence of the second half of the binding site (ATGC). Our results indicate that the Oct-2B protein interacts with canonical sequence and degenerated sequences. These data have greatly increased the number of potential targets for octamer proteins on DNA and changed our view on gene-expression regulation by these protein factors.

Strategies for Identification of Evolutionarily Conserved DNA Sequences

Lisa Stubbs and **Elbert Branscomb**[1]
Biology Division; Oak Ridge National Laboratory; Oak Ridge, TN 37831-8077
615/574-0848, Fax: -1283
[1]Human Genome Center; Biology and Biotechnology Research Program; Lawrence Livermore National Laboratory; Livermore, CA 94551

Protein-coding sequences occupy only a small fraction of the mammalian genome and generally exist as small, widely scattered segments interspersed between large stretches of noncoding DNA. Identification of all such gene sequences located throughout the human genome is, therefore, a technically challenging task. If mapping and sequencing data are to be correlated eventually with biological functions encoded throughout the human genome, rapid and efficient methods must be developed to identify such functionally critical DNA sequences.

The goal of this project is to exploit mouse-human genomic homologies and develop and test strategies for identifying, isolating, and analyzing these evolutionarily conserved sequences from large cloned segments of human DNA. Pilot studies, which focus on specially targeted regions of human chromosome 19, are being done in close collaboration with genome center staff at Lawrence Livermore National Laboratory (LLNL).

Human and mouse genes are, on the whole, very similar in sequence; nongene regions, by contrast, will generally vary greatly between two such highly divergent species. Our approach is based on the fact that DNA sequences encoding important biological functions are most likely to be conserved throughout evolution. We are currently exploring several different means for selectively cloning the most similar sequences in mouse and human DNA. The ultimate goal of these studies is to provide means by which large, cloned genomic regions can be scanned rapidly for the presence of genes and other functionally significant sequences.

These strategies focus on using conserved sequences in murine yeast artificial chromosome (YAC) or P1-based artificial chromosome clones to "trap" sequences from human cosmids that are similar to those found in the mouse. This project has been designed to capitalize on LLNL's collection of contiguous cosmid and YAC clones, which span with some gaps the length of human chromosome 19. Most cloned HSA19 genes and DNA markers have been localized to specific cosmid clones, but many more unidentified genes are expected to be scattered throughout each contig. Our current efforts are focused on a limited number of well-characterized regions known to be similar in humans and mice. Eventually we intend to apply these methods to identify genes along the length of HSA19.

*Development of New Reactive Oligonucleotide Derivatives for Sequence-Specific Fragmenting of Genomic DNA and Mapping Naturally Open Sequences in Cellular DNA

V. Vlassov, S. Gaidamakov, T. Ivanova, T. Abramova, and O. Gimautdinova
Institute of Bioorganic Chemistry; Novosibirsk 630090, Russia
+7-3832/353-162, Fax: -459, Internet: *vlasov@modul.bioch.nsk.su*

A number of oligonucleotides equipped with alkylating groups and groups that generate free radicals have been developed for sequence-specific modification and cleavage of large DNA fragments. Alkylating derivatives of oligonucleotides were used to investigate the interaction of oligonucleotides with single-stranded and double-stranded DNAs. Also, cellular DNA was found to have some naturally open sequences capable of binding oligonucleotides. The biological role of this phenomenon remains unknown.

We are now optimizing the structure of alkylating groups of oligonucleotide derivatives to develop efficient reagents capable of introducing rare cuts in large DNA fragments. To achieve an optimal geometry for modifying single- and double-stranded DNA, we plan to introduce linkers of different size and flexibility between the oligonucleotide moiety and the reactive nitrogen mustard residue. We are also trying to develop binary mixtures of oligonucleotide derivatives bearing fragments of potential reactive structures, which will be activated upon simultaneous binding of mixture components to the DNA target by forcing contact of the compounds. The first compounds of this type will be structures capable of binding metal ions, which can produce activated oxygen species.

By the end of this year we plan to develop new improved versions of alkylating derivatives of oligonucleotides. We hope to develop binary oligonucleotide reagents that may open new possibilities in designing specific DNA probes and artificial DNA-cleaving molecules.

Projects Continuing into FY 1993

Monochromosomal Hybrids for the Analysis of the Human Genome

Raghbir S. Athwal
Fels Institute for Cancer Research and Molecular Biology; Temple University; Philadelphia, PA 19140
215/707-6931 or -4300, Fax: -4318

An Investigation of Gene Organization Within the Human Genome Utilizing cDNA Sequencing

E. Morton Bradbury and **Joe M. Gatewood**
Center for Human Genome Studies; Life Sciences Division; Los Alamos National Laboratory; Los Alamos, NM 87545
505/667-2690, Fax: /665-3024

Overcoming Genome-Mapping Bottlenecks

Charles R. Cantor
Center for Advanced Research in Biotechnology; Boston University; Boston, MA 02215
617/353-8500, Fax: -8501

Projects Continuing into FY 1993

Isolation of Chromosome-Specific cDNA Clones

Jan-Fang Cheng and Victor Boyartchuk
Human Genome Center; Cell and Molecular Biology Division; Lawrence Berkeley Laboratory; Berkeley, CA 94720
510/486-6549, Fax: -6816, Internet: *jfcheng@lbl.gov*

Chromosome Structure and Function

Larry L. Deaven, Evelyn Campbell, and Mary Campbell
Center for Human Genome Studies; Life Sciences Division; Los Alamos National Laboratory; Los Alamos, NM 87545
505/667-3114, Fax: /665-3024, Internet: *moyzis@flovax.lanl.gov*

Human Recombinant DNA Library

Larry L. Deaven, Jon L. Longmire, MaryKay McCormick,[1] Deborah L. Grady, and Robert K. Moyzis
Center for Human Genome Studies; Life Sciences Division; Los Alamos National Laboratory; Los Alamos, NM 87545
505/667-3114, Fax: /665-3024, Internet: *moyzis@flovax.lanl.gov*
[1]Massachusetts General Hospital; Charlestown, MA 02129

Gene Libraries for Each Human Chromosome: Construction and Distribution

Jeff Gingrich, Jeff Garnes, and Anthony V. Carrano
Human Genome Center; Biology and Biotechnology Research Program; Lawrence Livermore National Laboratory; Livermore, CA 94550
510/423-8145, Fax: -3608, Internet: *gingrich1@llnl.gov*

Molecular Cytogenetics and Computer-Assisted Microscopy

Joe Gray, Dan Pinkel, Wen-Lin Kuo, Damir Sudar, and Don Peters
Department of Laboratory Medicine; Division of Molecular Cytometry; University of California; San Francisco, CA 94143-0808
415/476-3461, Fax: -8218, Internet: *gray@lcaquips.ucsf.edu*

An Improved Method for Producing Radiation Hybrids Applied to Human Chromosome 19

Cynthia L. Jackson and Hon Fong L. Mark
Rhode Island Hospital and Brown University; Providence, RI 02903
401/277-4370, Fax: -8514

Chromosome Region-Specific Libraries for Human Genome Analysis

Fa-Ten Kao and Jing-Wei Yu[1]
Eleanor Roosevelt Institute and Department of Biochemistry, Biophysics, and Genetics; University of Colorado Health Sciences Center; Denver, CO 80262
303/333-4515, Fax: -8423
[1]Eleanor Roosevelt Institute; Denver, CO 80262

Human cDNA Mapping Using Fluorescence In Situ Hybridization

Julie R. Korenberg
Department of Pediatrics; Medical Genetics; Cedars-Sinai Medical Center; University of California; Los Angeles, CA 90048
310/855-6451, Fax: /967-0112

Construction of a Human Genome Library Composed of Multimegabase Acentric Chromosome Fragments

Michael J. Lane, Peter Hahn,[1] and **John Hozier**[2]
Departments of Medicine and Microbiology and [1]Department of Radiology; State University of New York-Health Science Center at Syracuse; Syracuse, NY 13210
315/464-5446, Fax: -8255
[2]Department of Medical Genetics; Florida Institute of Technology; Melbourne, FL 32901

The cDNA Genome: Strategies and Results with Particular Reference to Human Chromosome 19

Gregory G. Lennon, Harvey Mohrenweiser, and **Anthony V. Carrano**
Human Genome Center; Biology and Biotechnology Research Program; Lawrence Livermore National Laboratory; Livermore, CA 94551
510/422-5711, Fax: /423-3608, Internet: *greg@mendel.llnl.gov*

Identification and Characterization of Expressed Chromosomal Sequences

Christopher H. Martin, Carol A. Mayeda, and **Michael J. Palazzolo**
Human Genome Center; Cell and Molecular Biology Division; Lawrence Berkeley Laboratory; Berkeley, CA 94720
Martin and Palazzolo: 510/486-5909, Fax: -6816, Internet: *chrism@genome.lbl.gov* or *mjpalazzolo@lbl.gov*

Application of Flow-Sorted Chromosomes to the Construction of Human Chromosome–Specific Yeast Artificial Chromosomes

Linda Meincke, Mary Campbell, Evelyn Campbell, John Fawcett, **MaryKay McCormick,**[1] **Larry Deaven,**[2] and **Robert Moyzis**[2]
Life Sciences Division and [2]Center for Human Genome Studies; Los Alamos National Laboratory; Los Alamos, NM 87545
Deaven: 505/667-3114, Fax: /665-3024, Internet: *moyzis@flovax.lanl.gov*
[1]Massachusetts General Hospital; Charlestown, MA 02129

Chimera-Free, High-Copy-Number YAC Libraries and Efficient Methods of Analysis

Donald T. Moir
Collaborative Research, Inc.; Waltham, MA 02154
617/487-7979, Fax: -7960, Internet: *moir@crie.com*

Projects Continuing into FY 1993

Genome Organization and Function

Robert K. Moyzis, Julie Meyne, and Robert L. Ratliff
Center for Human Genome Studies; Life Sciences Division; Los Alamos National Laboratory; Los Alamos, NM 87545
505/667-3912, Fax: /665-3024, Internet: *moyzis@flovax.lanl.gov*

Isolation of cDNAs from the Human X Chromosome and Derivation of Related STSs

David L. Nelson
Institute for Molecular Genetics; Baylor College of Medicine; Houston, TX 77030-3498
713/798-3122, Fax: -5386, Internet: *nelson@bcm.tmc.edu*

Multiplex Mapping of Human cDNAs

William C. Nierman, Donna R. Maglott, and Scott Durkin
American *Type Culture* Collection; Rockville, MD 20852-1776
301/231-5559, Fax: /770-1848

Chromosomal Localization of Brain cDNAs

Mihael H. Polymeropoulos, Hong Xiao, and Carl R. Merrill
Neuroscience Center at St. Elizabeth's Hospital; National Institute of Mental Health; Washington, DC 20032
202/373-6077, Fax: -6087

Development of a Large-Scale Targeted Mutagenesis Program for Determining Organismal Function of Specific Human Genes

Mike Mucenski, Bill Lee, **Eugene Rinchik**,[1] and **Richard Woychik**
Biology Division; Oak Ridge National Laboratory; Oak Ridge, TN 37831-8077
615/574-0703, Fax: -1274
[1]Sarah Lawrence College; Bronxville, NY 10708

Sequence Tagged Sites for Human Chromosome 19 cDNAs

Michael J. Siciliano and Anthony V. Carrano[1]
Department of Molecular Genetics; University of Texas M.D. Anderson Cancer Center; Houston, TX 77030
713/792-2910, Fax: /794-4394
[1]Human Genome Center; Lawrence Livermore National Laboratory; Livermore, CA 94550

cDNA/STS Map of the Human Genome: Methods Development and Applications Using Brain cDNAs

James M. Sikela, Akbar S. Khan,[1] Arto K. Orpana, Andrea S. Wilcox, Janet A. Hopkins, and Tamara J. Stevens
Department of Pharmacology; University of Colorado Health Sciences Center; Denver, CO 80262
303/270-8637, Fax: -7097, Internet: *sikela_j%maul@vaxf.colorado.edu*
[1]Samuel Roberts Noble Foundation, Inc.; Ardmore, OK 73402

Chromosome-Specific cDNAs and Sequence Tagged Sites

Marcelo Bento Soares, Pierre Jelenc,[1] Stephen Brown, Maria de Fatima Bonaldo, and Agiris Efstratiadis[1]
Department of Psychiatry and [1]Department of Genetics and Development; Columbia University;
New York, NY 10032
212/960-2313, Fax: /795-5886

Development of Human Genomic Virus-Based Library of 150- to 200-kb Inserts

Jean-Michel H. Vos, Tian-Qiang Sun, and Sharon Michael
Department of Biochemistry and Biophysics and Lineberger Comprehensive Cancer Research Center; University of North Carolina; Chapel Hill, NC 27599-7295
919/966-6888, Fax: -3015, Internet: *vos@med.unc.edu*

Isolation of Specific Human Telomeric Clones by Homologous Recombination and YAC Rescue

Geoffrey Wahl and Linnea Brody
Gene Expression Laboratory; Salk Institute for Biological Studies; La Jolla, CA 92037
619/453-4100 Ext. 587, Fax: /455-1349

Development of an Embryonic Stem (ES) Cell–Based System for the In Vitro Generation of Germline Deletion Complexes Throughout the Mouse Genome

Richard Woychik and **Eugene Rinchik**[1]
Biology Division; Oak Ridge National Laboratory; Oak Ridge, TN 37831-8077
615/574-3965 or -3966, Fax: -1274
[1]Sarah Lawrence College; Bronxville, NY 10708

Physical and Genetic Mapping

Projects New in FY 1993

A Method for Heterozygous Carrier Screening Using an *E. coli* Mismatch Binding Protein, MutS; Application to the Cystic Fibrosis Gene

Alla Lishanski and Jasper Rine
Human Genome Center; Lawrence Berkeley Laboratory; Berkeley, CA 94720
510/486-7332, Fax: -6818, Internet: *allali@genome.lbl.gov*

An experimental strategy for detecting heterozygosity in genomic DNA has been developed based on a preferential binding of *Escherichia coli* MutS protein to DNA molecules containing mismatched bases. The binding was detected by a gel mobility–shift assay. This approach was tested using the most commonly occurring mutations within the cystic fibrosis gene (CFTR) as a model. Genomic DNA samples were amplified using 5′-end–labeled primers that bracket the site of the ΔF508 3-bp deletion in exon 10 of the CFTR gene. The renatured polymerase chain reaction (PCR) products from homozygotes produced homoduplexes, and the PCR products from heterozygotes produced heteroduplexes and homoduplexes (1:1). MutS protein bound more strongly to heteroduplexes corresponding to heterozygous carriers of ΔF508 and containing a CTT or GAA loop in one of the strands than to homoduplexes corresponding to homozygotes. The ability of MutS to detect heteroduplexes in PCR-amplified DNA extended to fragments ~500 bp in length. The method was also able to detect carriers of the point mutations in exon 11 of the CFTR gene by a preferential binding of MutS to single-base mismatches in PCR-amplified DNA.

Vertically Integrated Analysis of Human DNA

Maynard Olson
University of Washington; Seattle, WA 98195
206/685-7346, Fax: -7344, Internet: *mvo@u.washington.edu*

A systematic approach will be developed for the vertically integrated analysis of human DNA segments ranging in size up to a few megabase pairs. Vertical integration denotes sequential analysis of a genomic region at all levels of detail from low-resolution physical mapping to finished sequence. This project will emphasize the steps preceding DNA sequencing, particularly the hierarchical analysis of genome segments in somatic cell lines, yeast artificial chromosomes, cosmids, lambdas, plasmids, and filamentous-phage clones. The goal will be precise, high-resolution physical maps on which the ends of the inserts of small-insert clones, suitable for use as sequencing templates, will be placed exactly.

The project will focus on simple, modular, experimental strategies that are good candidates for high-level automation. The raw data will be captured by electronically imaging electrophoretic gels that have banding patterns predictable enough for completely automated interpretation. Early steps will be taken toward automation of these procedures on a scale that would allow the analysis of millions of clones.

Construction of a Genetic Map Across Chromosome 21

Elaine A. Ostrander
Fred Hutchinson Cancer Research Center; Seattle, WA 98104
206/667-5000, Fax: -6124

The goal of our group is to develop and implement technologies aimed at high-resolution mapping of individual human chromosomes. Our initial efforts have focused on developing strategies for identifying and characterizing polymorphic microsatellite repeat sequences; these repeats are extremely abundant in all mammalian genomes and usually have multiple, highly informative alleles. We intend to construct a high-density genetic map of chromosome 21 with polymorphic repeat markers spaced every 0.5 to 1 cM, based on simple sequence microsatellite repeats.

We have fully sequenced 43 $(CA)_n$ repeat-based markers prepared from chromosome 21 cosmids. After eliminating 2 previously identified markers and 5 containing *Alu* sequences, we localized 14 of the remaining markers to their specific locations on chromosome 21 and identified yeast artificial chromosomes containing the markers.

With computer-analysis programs prepared by Lawrence Berkeley Laboratory's engineering group, we have adapted the automated laser fluorescent (called A.L.F.) sequencer for use in fluorescence-tagged genetic mapping studies. The markers identified and mapped to date are fairly well distributed and represent a strong foundation on which to continue our studies.

A Physical and Genetic Map of Human Chromosome 21—A Prelude to Sequence: Overview

Jasper Rine, R. Blajez, Jan-Fang Cheng, Jeffrey Gingrich, S. R. Lowry, Elaine Ostrander,[1] Stewart Scherer, Sylvia Spengler, D. Scott, F. Shadravan, T. Torok, K. M. Wilson, and Y. Zhu
Human Genome Center; Lawrence Berkeley Laboratory; Berkeley, CA 94720
510/643-5592, Fax: /642-6420
[1]Fred Hutchinson Cancer Research Center; Seattle, WA 98104

Advances in physical and genetic maps have set the stage for the next biological goals of genome research: obtaining genomic sequence from substantial regions of the genome and preparing the genetic infrastructure for genotyping populations. Toward these ends we are preparing to sequence a 3- to 4-Mb region from a medically significant portion of chromosome 21 and are working toward saturating the chromosome with genetic markers. Lawrence Berkeley Laboratory (LBL) has designed, implemented, and tested on an 80,000-bp segment of genomic DNA a directed sequencing approach that produced completed DNA sequences at a high rate of throughput and with high accuracy.

In the evaluation of potential DNA sources for sequencing templates, the yeast artificial chromosomes (YACs) used in chromosome 21 sequence tagged sites (STSs) failed to exhibit obvious deletions. Thus, YACs appear to be poor choices for sequencing studies. The rapid success of the directed genome-sequencing strategy at LBL has focused our attention on the human P1 library of Shepherd and Sternberg. P1 clones can carry 100,000 inserted bases and are the best source of material that can serve directly as sequencing templates for our strategy. We have constructed pools 1000 clones deep from the Shepherd and Sternberg library and have screened these with STSs from the 22.2 to 22.3 region of chromosome 21. These P1s are now in the pipeline for large-scale directed DNA sequencing. This process involves constructing physical maps in which the distance (d) and orientation (o) of each gene-sized sequencing template (g) is known and tagged by sequence from each end. This procedure, known as dog tagging, offers the best-known method for large-scale genome sequencing and avoids the sequence-assembly headaches of random strategies.

We have used marker-selection techniques to isolate a large number of simple sequence repeats from chromosome 21 as a source of genetic markers. STSs produced from about 40 of these repeats have been assigned to the map, with a resolution of a few hundred kilobases. We have approximately doubled the density of genetic markers on this chromosome, making it the most densely marked human chromosome.

Our physical mapping efforts have focused on the distal third of the q arm. In this region we have used fluorescence in situ hybridization to map about 280 YACs plus cosmids and have constructed contig maps. This mapping has allowed us to detect and correct errors in the recently published map (Chumakov et al., 1992), including gene misplacement of as much as 2 Mb. Corrected maps for these regions will be presented.

We have developed methods for physical selection of cDNAs corresponding to mapped YACs and cosmids. We have mapped 21 new cDNAs to their respective locations on chromosome 21. By sequence analysis, each of these defines new genes and pioneer proteins. The cDNA effort is now focused on saturating the multimegabase target of the genomic-sequencing effort.

Projects New in FY 1993

All the programs involve close interaction among the center's biology, instrumentation and automation, and informatics groups. Major instrument development includes automation of steps in directed sequencing; a large-scale, extensible thermocycler; and a large-scale oligonucleotide developer. Software for sequence assembly and analysis is under development.

The capacity to produce genome information has outstripped the capacity of formal publication procedures to disseminate the information to the community. To help close the gap between producers and consumers, all of our unpublished cDNA sequences have been deposited in the cDNA Inform database (Los Alamos National Laboratory), all genetic markers in Genome Data Base, and the sequences from which they are derived in GenBank®. In addition, we are establishing a public database at LBL that will serve as an open notebook for chromosome 21 mapping data and sequence information from P1 clones. Our mapping data enter this database directly, and the sequence enters as each 3-kb dog unit is complete.

Chromosome Mapping by Fluorescent In Situ Hybridization to Interphase Nuclei

Barbara Trask and **Ger van den Engh**
Department of Molecular Biotechnology; School of Medicine; University of Washington; Seattle, WA 98195
206/685-7347, Fax -7354, Internet: *trask@fishnet.mbt.washington.edu*

This project aims to develop a new, efficient approach for high-resolution chromosome mapping by in situ hybridization. The approach is based on the observation that the folding of DNA in the interphase nucleus can be described by a random walk model. This model provides the theoretical basis for using the average distance observed between hybridization sites in the interphase nucleus to estimate distance between two markers measured along the DNA strand. We have developed a rapid graphical method that makes possible the accumulation of thousands of distance measurements per day and have used this approach to confirm the published map of a 4-Mb region of chromosome 4. We plan to (1) further improve data collection methods and (2) develop software for calculating the most probable probe order and relative distance from a set of interphase distance measurements. We will develop similar graphical procedures for efficiently mapping probes to metaphase chromosomes and combine these techniques to build selected chromosome region maps with average density of 100 kb.

*A Chromosome 13 Mapping Project Based on the Los Alamos Cosmid Library

Nick K. Yankovsky, B. I. Kapanadze, V. M. Brodjansky, and G. E. Sulimora
Laboratory of Genome Analysis; Institute of General Genetics; Russian Academy of Sciences; Moscow 117809, Russia
+7-095/135-4307, Fax: -1289, Internet: *bion@glas.apc.org*

The main goal of this project is to contribute to contig mapping of human chromosome 13 with cosmids and YACs as elements. The starting point is production of microsatellite markers from cosmids already mapped to the 13q14 region. Sources of clones are an s*Cos* cosmid chromosome 13 library from Los Alamos National Laboratory (LANL) and a similar Lorist-based library from the Imperial Cancer Research Fund (ICRF) in the United Kingdom. Restriction patterns for more than 150 cosmids have been established with one or two restriction enzymes, and a supporting database has been created to support contig assembly.

Chromosome 13–specific clones are found by hybridization with a nylon-gridded library of all the cosmid clones. A total human genome YAC library has been obtained from ICRF; pools have been formed, and screening with chromosome 13–specific probes has begun. Some equipment for this project was supplied by the Russian Human Genome Organization board.

Projects Continuing into FY 1993

New Strategies for Closure of the Chromosome 19 Contig Map

Mark Batzer and Anthony V. Carrano
Human Genome Center; Biology and Biotechnology Research Program; Lawrence Livermore National Laboratory; Livermore, CA 94551
510/422-5721, Fax: /423-3608, Internet: *batzer2@llnl.gov*

DNA Sequence Mapping by Fluorescence In Situ Hybridization

Brigette Brandriff, Laurie Gordon, Anne Bergmann, Mari Christensen, Anne Fertitta, and **Anthony Carrano**
Human Genome Center; Biology and Biotechnology Research Program: Lawrence Livermore National Laboratory; Livermore, CA 94551
510/423-0758, Fax: -3608, Internet: *brandriff@llnl.gov*

Core Facility for Support of Chromosome 19 Physical Mapping

Anthony V. Carrano, Anne Olsen, Mark Batzer, Jane Lamerdin, and Linda K. Ashworth
Human Genome Center; Biology and Biotechnology Research Program; Lawrence Livermore National Laboratory; Livermore, CA 94550
510/422-5698, Internet: *carrano1@llnl.gov*

Mapping and Ordered Cloning of Human X Chromosome

C. Thomas Caskey and **David L. Nelson**
Institute for Molecular Genetics; Baylor College of Medicine; Houston, TX 77030-3498
Nelson: 713/798-3122, Fax: -5386, Internet: *nelson@bcm.tmc.edu*

Massive Isolation and Contig Building of Chromosome-Specific YAC Clones

Jan-Fang Cheng and Julia Nikolic
Human Genome Center; Cell and Molecular Biology Division; Lawrence Berkeley Laboratory; Berkeley, CA 94720
510/486-6549, Fax: -6816, Internet: *jfcheng@lbl.gov*

Physical and Transcription Mapping of Human Chromosome 11

Glen A. Evans, David McElligott, Steven Clark, Suzanne Clancy, Licia Selleri, Michael Smith, Merl Hoekstra, and Gary Hermanson
Molecular Genetics Laboratory; Salk Institute for Biological Studies; San Diego, CA 92186-5800
619/453-4100 Ext. 279, Fax: /558-9513, Internet: *gevans@salk-sd2.sdsc.edu*

A Clone-Limited STS Strategy for Physical Mapping

Christopher H. Martin, Carol A. Mayeda, and **Michael J. Palazzolo**
Human Genome Center; Cell and Molecular Biology Division; Lawrence Berkeley Laboratory; Berkeley, CA 94720
Martin and Palazzolo: 510/486-5909, Fax: -6816, Internet: *chrism@genome.lbl.gov* or *mjpalazzolo@lbl.gov*

Projects Continuing into FY 1993

Interdigitation of the Genetic and Physical/Cosmid Contig Maps of Human Chromosome 19

Harvey W. Mohrenweiser, Elbert Branscomb, and Anthony V. Carrano
Human Genome Center; Biology and Biotechnology Research Program; Lawrence Livermore National Laboratory; Livermore, CA 94551
510/423-0534, Fax: /422-2282, Internet: *harvey@cea.llnl.gov*

Physical Mapping of Human Chromosome 16

N. A. Doggett, C. E. Hildebrand, M. K. McCormick,[2] L. L. Deaven,[1] D. F. Callen,[3] G. R. Sutherland,[3] K. Okumura,[4] D. C. Ward,[5] and **R. K. Moyzis**[1]
Life Sciences Division and [1]Center for Human Genome Studies; Los Alamos National Laboratory; Los Alamos, NM 87545
Moyzis: 505/667-3912, Fax: /665-3024, Internet: *moyzis@flovax.lanl.gov*
[2]Massachusetts General Hospital; Charlestown, MA 02129
[3]Department of Cytogenetics and Molecular Genetics; Adelaide Children's Hospital; North Adelaide, South Australia 5006, Australia
[4]Juntendo University School of Medicine; Department of Immunology; Tokyo 113, Japan
[5]Department of Human Genetics; Yale University School of Medicine; New Haven, CT 06510

Assembly, Closure, and Characterization of a Chromosome 19 Contig Map

Anne S. Olsen, Emilio Garcia, Linda Ashworth, Alex Copeland, and Anthony V. Carrano
Human Genome Center; Biology and Biotechnology Research Program; Lawrence Livermore National Laboratory; Livermore, CA 94551
510/423-4927, Fax: -3608, Internet: *olsen@ecor1.llnl.gov*

Developing a Physical Map of Human Chromosome 22

Melvin I. Simon, Bruce Birren, and Hiroaki Shizuya
Biology Division; California Institute of Technology; Pasadena, CA 91125
818/356-3944, Fax: /796-7066

Physical Structure of Human Chromosome 21

Cassandra L. Smith, Denan Wang,[1] Kaoru Yoshida,[1] Jesus Sainz,[2] Carita Fockler,[1] and Meire Bremer[1]
Center for Advanced Research in Biotechnology; Boston University; Boston, MA 02215
617/353-2800, Fax: -5929, Internet: *cls@buenga.bu.edu*
[1]Division of Chemical Biodynamics; Lawrence Berkeley Laboratory; Berkeley, CA 94720
[2]Cedars-Sinai Medical Center; Los Angeles, CA 90048

Generating a Comparative Physical Map of Mouse Chromosome 7

Lisa J. Stubbs, **Eugene Rinchik**,[1] and Estela Generoso
Biology Division; Oak Ridge National Laboratory; Oak Ridge, TN 37831-8077
615/574-0848 or -0864, Fax: -1283
[1]Sarah Lawrence College; Bronxville, NY 10708

Correlation of Physical and Genetic Maps of Human Chromosome 16

David F. Callen, Sinoula Apostolou, Elizabeth Baker, Liang Z. Chen, Helen Kozman, Sharon A. Lane, Julie Nancarrow, Hilary A. Phillips, Yang Shen, Andrew D. Thompson, Scott A. Whitmore, Norman A. Doggett,[1] Raymond L. Stallings,[1] C. Edgar Hildebrand,[1] John C. Mulley, Robert I. Richards, and **Grant R. Sutherland**
Department of Cytogenetics and Molecular Genetics; Adelaide Children's Hospital; North Adelaide, South Australia 5006, Australia
Sutherland: + 61/8-204-7333 or -7284, Fax: -7384 or -7342
[1]Center for Human Genome Studies; Life Sciences Division; Los Alamos National Laboratory; Los Alamos, NM 87545

Mapping Instrumentation

Projects New in FY 1993

*Electrotransformation for Introducing DNA Into Industrial Bacilli

Alexandre S. Boitsov
Department of Biophysics; St. Petersburg State Technical University; St. Petersburg 195251, Russia
+7-812/552-7964, Fax: -6086, Internet: *boitsov@bio.stu.spb.su*

This group is involved in research in two major fields, bacilli genetics and genetic engineering. We have recently begun exploring electrotransformation (ET) as a method for introducing DNA into industrial bacilli. In these projects physics and biology researchers from different St. Petersburg institutes gathered together in ECOGENE, a science technology company.

We are attempting to create a system of highly efficient but gentle methods for increasing cell-membrane permeability and for introducing biomolecules (DNA, protein) into cells via an electric field. These methods appear to have a number of advantages over traditional techniques in that the metal electrodes do not come into contact with the cell suspension; an extremely high intensity of the electric field (400 kV/cm or more) can be achieved with pulse duration of 10 ns and more; and cell survival, electrostimulation efficiency of the microbiological processes, and ET with biomolecules are increased.

The project is based on experiments revealing the unexpected roles of electric field intensity for cell-wall permeability and the dependence of pulse shape on ET efficiency. These observations became apparent through use of a specially constructed apparatus in which electric pulse parameters were independent of cell suspension and pulse shapes could vary.

This project consists of (1) development of electronic equipment and (2) a theoretical study of the biophysics process. We are developing techniques for introducing very large DNA molecules into *Escherichia coli* cells that cannot be efficiently transformed by the classical method of electroporation. Our immediate goals are the following:

- Elucidate the principal pattern of *E. coli* ET with plasmids up to 50 kb.
- Optimize the ET protocol for *E. coli* cells with plasmids of 100 to 330 kb.
- Experiment with bacilli, yeasts, and mycoplasma to finish the development of apparatus with optimal pulse shape.
- Complete an ET theoretical model.

*Development of Intracellular Flow-Karyotype Analysis

Andrei I. Poletaev, Sergei I. Stepanov,[1] Valeri V. Zenin,[2] Nikolay Aksenov,[2] Tatijana V. Nasedkina,[3] and Yuri V. Kravazky[4]
Engelhardt Institute of Molecular Biology; Russian Academy of Sciences; Moscow 117984, Russia
+7-095/135-9824, Fax: -1405 or /938-2187, Internet: *polet@imb.msk.su*
[1]St. Petersburg Institute of Nuclear Physics and [2]Institute of Cytology; Russian Academy of Sciences; St. Petersburg, Russia
[3]Institute of Molecular Biology; Russian Academy of Sciences
[4]Physico-Technical Institute; Moscow, Russia

Intracellular flow karyotyping appears to be a feasible and beneficial method for analyzing karyotype aberrations from individual cells using flow cytogenetics. This technology might be especially useful for various studies of karyotype instability and tumorigenesis.

Groups headed by Scott Cram (Los Alamos National Laboratory) and Andrei Poletaev (Russian Academy of Sciences) are collaborating to achieve the following six goals. The Russian group is carrying out research described in all six goals; American investigators will concentrate on the last three.

1. Optimize the technology of hydrodynamic destruction of mitotic cells by capillary-flow high-gradient devices.
2. Develop alternative methods (particularly ultrasonic disintegration) of cell membrane destruction.
3. Improve methods for preparing human cells for analysis while maintaining stable chromosome staining inside the cells.
4. Adapt the method for modern serial flow-cytometer systems.
5. Develop new algorithms and computer programs for data interpretation.
6. Conduct pilot research using different human cell line models to investigate the method's parameters.

First, we will establish the main principles of intracellular analysis technology and build two instruments, one for general study and the other for making improvements to the method. Second, we will experiment with human cells and improve equipment and methods. Last, we intend to adapt the new technology for modern serial flow cytometers.

Atomic Force Microscopy of Biochemically Tagged DNA

Matthew N. Murray, Helen Hansma,[2] D. Frank Ogletree,[1] William F. Kolbe, Sylvia Spengler, Cassandra Smith,[3] Charles Cantor,[3] and **Miguel Salmeron**[1]
Human Genome Center and [1]Material Science Division; Lawrence Berkeley Laboratory; Berkeley, CA 94720
Salmeron: 510/486-6230, Fax: -4995, Internet: *salmeron@lbl.gov*
[2]Department of Physics; University of California; Santa Barbara, CA 93106
[3]Center for Advanced Research in Biotechnology; Boston University; Boston, MA 02215

Small DNA fragments of known length were made using the polymerase chain reaction. These fragments had biotin molecules (vitamin H) covalently attached to each end and were then labeled with streptavidin. This tetrameric complex was expected to bind up to four DNA molecules via their attached biotin molecules. The DNA was then imaged with atomic force microscopy (AFM). As expected theoretically, images revealed the protein at the end of the DNA strands as well as the presence of dimers, trimers, and tetramers of DNA bound to a single protein. Imaging time was about 1 min.

With these results, we believe we have shown that AFM does have sufficient resolution to map DNA. In its simplest form, mapping involves measuring the physical distance between two points of DNA. In this experiment we have demonstrated the ability of AFM to perform this task by attaching a large protein marker to genetically engineered pieces of human DNA and using AFM to locate the marker and measure the known length from the protein to the other end of the DNA.

Flow Karyotyping and Flow Instrumentation Development

Ger van den Engh and **Barbara Trask**
Department of Molecular Biotechnology; School of Medicine; University of Washington; Seattle, WA 98195
206/685-7345, Fax: -7301, Internet: *engh@fishnet.mbt.washington.edu*

The purpose of this project is to develop means and methods for flow karyotyping and chromosome sorting. Analytical flow karyotyping is being applied to a variety of areas related to genomic research and medical diagnostics. Chromosomes purified by flow sorting are used for the production of clones or DNA libraries amplified by polymerase chain reaction (PCR).

Projects New in FY 1993

The project is a continuation of the principal investigators' work at Lawrence Livermore National Laboratory. We have extensively explored chromosome analysis by flow cytometry and have improved techniques of chromosome preparation and staining. We have built flow instruments that accurately measure and sort chromosomes with high efficiency. We have produced software that facilitates analysis and comparison of human flow karyotypes and optimizes sort purity and throughput. As a result of these developments, quantitative DNA measurement of human chromosomes from clinical peripheral blood cultures and established cell lines has become a straightforward and reproducible technique that can be applied to a variety of genetic studies. Examples are the description of normal chromosome heteromorphism, quantification of deletion size in contiguous gene syndromes, and routine monitoring of somatic cell hybrids. Our developments in high-speed sorting technology have led to chromosome-enriched libraries (e.g., the DOE gene library project). Our techniques have been adopted by other laboratories, and the instrumentation, which has been licensed to industry, will soon be available commercially.

This work should increase the availability of this technique to the genetics and genomics research community. Experiments will provide quantitative information on normal and abnormal chromosomes, understanding of interactions of DNA-binding dyes and chromatin, deletion maps of particular chromosome regions to facilitate directed disease-gene mapping, improved purification of chromosome types, and simplification of flow technology for export to other research institutions. In addition, techniques will be developed for mapping probes to chromosomes sorted onto filters, producing DNA or RNA sequence libraries by PCR amplification of small numbers of sorted chromosomes or cells, and identifying and handling single bacteria carrying transfected sequences.

Projects Continuing into FY 1993

Technology Development for Large-Scale Physical Mapping

Tony J. Beugelsdijk, Patricia A. Medvick, Robert M. Hollen, and Randy S. Roberts
Los Alamos National Laboratory; Los Alamos, NM 87545
505/667-3169, Fax: /665-3911

Advanced Flow Cytometry Technique Development

James H. Jett, John C. Martin, and **Mark E. Wilder**
Center for Human Genome Studies; Life Sciences Division; Los Alamos National Laboratory; Los Alamos, NM 87545
505/667-3843, Fax: /665-3024, Internet: *jett@flovax.lanl.gov*

DNA Separation by Pulsed-Field Capillary Electrophoresis

Barry L. Karger
Barnett Institute; Northeastern University; Boston, MA 02115
617/437-2867, Fax: -2855

Image Acquisition and Analysis

William F. Kolbe, Joseph E. Katz, and **Joseph M. Jaklevic**
Human Genome Center and Engineering Division; Lawrence Berkeley Laboratory; Berkeley, CA 94720
510/486-7199, Fax: -5857, Internet: *wfkolbe@lbl.gov*

Automated Methods for Large-Scale Physical Mapping

Patricia A. Medvick, Robert M. Hollen, Tony J. Beugelsdijk, Randy S. Roberts, David M. Trimmer, Leonard A. Stovall, and Mark A. Kozubel
Los Alamos National Laboratory; Los Alamos, NM 87545
505/667-2676, Fax: /665-3911, Internet: *pm@lanl.gov*

Robotics and Automation

Donald C. Uber, Joseph M. Jaklevic, and Edward H. Theil
Human Genome Center and Engineering Division; Lawrence Berkeley Laboratory; University of California; Berkeley, CA 94720
510/486-6378, Fax: -6816

Sequencing Technologies

Projects New in FY 1993

Sequencing by Hybridization: Development of an Efficient Large-Scale Methodology

Radomir Crkvenjakov
Center for Mechanistic Biology and Biotechnology; Argonne National Laboratory; Argonne, IL 60439-4833
708/252-3161, Fax: -3387, Internet: *crk@everest.anl.gov*

We proposed DNA sequencing by hybridization (SBH) in 1987. Steady progress in research and theory, including the sequencing of an unknown short (343-bp) DNA by this method, opens the way for rapid development and laboratory-scale implementation of the SBH approach. To achieve our research objective of developing potential daily SBH rates of up to 1 Mb per laboratory, we are exploiting SBH Format 1, in which DNA samples arrayed on a surface are sequentially interrogated by oligonucleotide probes.

This strategy is based on the development of a high-throughput line for simultaneous production of hybridization scores on hundreds of thousands of 1- to 2-kb clones. DNA sample preparation and dense offprinting on filters, hybridization, and imaging are highly parallelized and streamlined for easy automation. A throughput capacity of 1 million scores/d is projected for the next year, increasing to 10 million/d in the near future.

Three levels of sequencing information can be obtained depending on the numbers of probes scored per clone in an experiment. Mapping and identification using clone sequence signatures can be achieved with relatively few (50 to 200) probes. Positioning and identifying genome structural elements (partial sequencing) requires more-extensive hybridizations. Complete sequencing by SBH requires data from several thousand probes, either on three to five related genomes or, in the case of single genomes, in combination with single-pass gel sequencing of one genome equivalent.

In an orderly progression toward complete sequencing, we have almost completed the development of SBH for the first group of applications. Typing of 20,000 cDNA clones from human brain with 60 to 110 probes led to grouping them into at least 5000 gene clusters, revealing the abundance structure of the libraries used. A model experiment on known clones simulated the cosmid-sized DNA subclone library of ten equivalents. This experiment demonstrated that SBH data from 110 probes can lead to clone clustering so that the entire DNA is represented in a one- to two-equivalent set of clones drawn from the clusters. This can reduce the redundancy of gel sequencing by five- to tenfold. The principle of partial sequencing was demonstrated by identifying the gamma-actin cDNA cluster only on the basis of its hybridization scores.

Intermediate-term goals are to (1) prepare sequence-ready maps of 1- to 2-kb subclones of human cosmids or bacterial artificial chromosomes and of several related bacterial genomes; (2) identify partially sequenced cDNAs in previously sequenced libraries to avoid redundancy in gene discovery and efficiently provide cDNAs from as-yet-unknown genes for complete sequencing; and (3) starting from the above maps, combine hybridization data from 3000 probes and single-pass gel sequencing to obtain very accurate finished sequence at a scale of 5 to 20 Mb/year.

Coupling Sequencing by Hybridization with Gel Sequencing for Inexpensive Analysis of Genes and Genomes

Radoje Drmanac, Snezana Drmanac, and Ivan Labat
Integral Genetics Group; Center for Mechanistic Biology and Biotechnology; Argonne National Laboratory; Argonne, IL 60439
708/252-3175, Fax: -3387, Internet: *rade@everest.bim.anl.gov*

Since 1987 when we conceived sequencing by hybridization (SBH), we have developed several procedures and concepts that enable immediate use of the method as well as future "chip"-based technologies. In particular, hybridization conditions were defined and proved by correct sequencing

of 343 bp in a blind test. Solutions for inexpensive, large-scale genome analysis with state-of-the-art technologies are represented by (1) partial sequencing or fine structural (and sequence-ready) mapping with 100 to 1000 probes and (2) full sequencing by integrating the incomplete gel and SBH data from single or several similar genomes. A basis for genome sequencing without subcloning is provided by Format 1 (an array of DNA samples) or Format 2 (an array of probes) sequencing chips based on microbeads, and by a recently proposed combination of the two formats. Format 3 (combinatorial chip) involves ligation of arrayed probes and probes in solution.

To implement Format 1, we have developed a data-production line with the present capacity of 1 million clone-probe measurements/d. A high-throughput polymerase chain reaction (PCR) procedure is established using BioOvens. Biomek 1000 is adapted to spot 31,000 DNA samples on a 6- by 9-in. filter. This dot density provides 50 Mb of DNA per membrane, ready for fine mapping and sequencing. Development of a hybridization machine with a capacity of 24 filters is in progress. The PhosphorImager is used to collect data from 33P-labeled probes and our image-analysis program to report dot intensities. Priorities for upgrading current facilities toward a capacity of 10 million scores/d are an automated setting of 10,000 PCR reactions/d, labeling of 100 probes/d, and robotized retrieval of selected subsets of clones.

By the described setup, 20,000 cDNA clones from a brain library (M. B. Soares, Columbia University) have been hybridized with 256 probes. About 13,000 groups or single clones have been recognized by our clustering program. Screening provides a rational choice of clones for gene mapping and full sequencing. The method's simplicity allows inexpensive screening of millions of cDNAs from dozens of tissues. Our first target is 100,000 clones from the brain library. To demonstrate sequence-ready mapping (ordering of shotgun clones), 1100 M13 subclones from a cosmid (B. Koop, University of Victoria, Canada) have been hybridized with 250 probes, and screening a shotgun library of the 2-Mb genome of the archebacteria *Pyrococcus furiosus* (F. Robb, University of Maryland, Baltimore) has been started.

The next target is a proof of the proposed inexpensive sequencing scheme, which requires 3000 probes and targeted single-pass gel sequences with as much as 20% errors. A further advancement would be comparative sequencing of 4 similar bacterial genomes. Megabase sequencing based on reading 14-mers through ligation of back-to-back hybridized 7-mers will be investigated in parallel.

Sequencing By Hybridization With Oligonucleotide Matrices (SHOM)

Andrei Mirzabekov, Yuri Lysov, Eduard Kraindlin, Gennadi M. Ershov, and Vladimir Florentiev
Engelhardt Institute of Molecular Biology; 117984 Moscow, Russia
+7-095/135-2311, Fax -1405, Internet: *amir@imb.msk.su*

Sequencing by hybridization with oligonucleotide matrix (SHOM) by this research group has led to the development of sequencing "microchips." These microchips consist of glass plate covered with polyacrylamide gel squares (about 30 x 30 μm) that are 20 μm thick and contain certain chemically immobilized octanucleotides. Hybridization with DNA fragments can discriminate among perfect duplexes, duplexes containing single internal mismatches, and major parts of duplexes containing terminal mismatches. Developed procedures have been used successfully in model experiments to sequence a heptadecanucleotide and localize a single base change in three other heptadecanucleotides. A theory has been developed to describe DNA hybridization with gel-immobilized oligonucleotides. The theory predicts the apparent thermostability of duplexes and the thermostability dependence on concentration of immobilized oligonucleotides, gel thickness, and washing time [K. Khrapko et al., *DNA Sequence* 1, 375–88 (1991); M. Livshits et al., "Dissociation of DNA Duplexes with Gel-Immobilized Oligonucleotides" (in preparation)].

Projects New in FY 1993

Prototype automatic sequencing equipment has been created and tested. This equipment consists of a thermostated plate and a fluorescent microscope with a charged-coupled-display (CCD) camera connected to a computer for measuring hybridization of fluorescently labeled DNA with immobilized oligonucleotides. Software has also been developed for image analysis of the pattern of hybridized octamers and for DNA sequence reconstitution.

A "continuous stacking hybridization" approach, which makes the efficiency of a matrix of immobilized octanucleotides as high as the efficiency of a tridecanucleotide matrix, has been suggested. The procedure is based on additional rounds of hybridizing the octamer matrix with chosen fluorescently labeled pentanucleotides and unlabeled DNA. Computer simulations have shown that several rounds of continuous stacking hybridization of DNA with an octanucleotide matrix in the presence of a mixture of preselected pentanucleotides imitates hybridization with a tridecanucleotide matrix and thus can be effectively used to sequence DNA that is several thousand nucleotides long [Yu, Lysov et al., "DNA Sequencing by Hybridization to Oligonucleotide Matrix: Calculation of Continuous Stacking Hybridization Efficiency" (in preparation)]. The use of gel to immobilize oligonucleotides provides the important possibility of increasing the capacity of matrices for immobilized oligos and equalizing the thermostability of G+C– and A+T–rich duplexes. Our future efforts will be concentrated on optimizing conditions, materials, equipment, and software so that SHOM can internally sequence millions of bases per day of near-nonrepetitive DNA several thousand nucleotides long.

*Development of a Simple and Rapid Technique for Sequencing DNA Fragments

Oleg I. Serpinsky, Galina F. Sivolobova, Galina V. Kochneva, Ilnur H. Urmanov, Victor N. Krasnikh, and Yura A. Gorbunov
Institute of Molecular Biology; Koltsovo, Novosibirsk Region 633159, Russia
+7-3832/647-887, Fax: /328-831

A limiting factor in Sanger sequencing is the preparation of DNA templates for carrying out polymerase chain reaction. The goal of this project is to develop a simple and timesaving technique for preparing DNA samples. Our project includes the following steps:

- Construct a specialized transposable genetic element [Tn5s2 (on Tn5 basis)] containing the (1) NPTII gene for selecting tagged DNA plasmids after insertion mutagenesis, (2) IS1 gene for generating the deletion variants of DNA plasmids in which Tn5s2 will be inserted, (3) original ribosomal S12 gene of *Escherichia coli* as a genetic marker to allow selection of *E. coli* clones bearing deletion DNA plasmids, and (4) fragment of M13 bacteriophage DNA allowing single-stranded DNA (ssDNA) to be obtained.
- Choose or create *E. coli* strains necessary for transposition and prepare deletion variants of plasmid DNA and their single-stranded forms.
- Improve ssDNA-isolation methods and determine sequences of some DNA fragments (using the transposon Tn5s2 as an example).

We believe this technique will not require DNA subcloning in specialized vectors if a plasmid without kanamycin resistance is used. This technique may be easily automated and thus increase the efficiency of Sanger sequencing.

Large-Scale DNA Sequencing with a Primer Library

F. William Studier and **John J. Dunn**
Biology Department; Brookhaven National Laboratory; Upton, NY 11973
516/282-3390 or -3012, Fax: -3407
Internet: *studier@genome1.bio.bnl.gov* or *dunn@genome1.bio.bnl.gov*

Our aim is to develop a DNA sequencing capacity that can contribute significantly to the goal of sequencing the human genome within the next 10 years. We found that strings of three contiguous hexanucleotides (hexamers) can prime sequencing reactions specifically on templates at least as large as cosmid DNAs (40 kb) if the template DNA is saturated with a single-stranded DNA-binding protein. Most of the 4096 possible hexamers seem to participate effectively in such priming reactions, and the initial success rate of 60 to 90% compares favorably with conventional priming. The ability to prime sequencing reactions from a hexamer library may allow sequencing by primer walking on multiple templates as fast as sequencing reactions can be assembled. As the next steps toward realizing this potential, our immediate goals are to (1) integrate triple-hexamer priming chemistry with four-color fluorescent labeling and detection, (2) implement capillary electrophoresis with a replaceable matrix for rapid readout of many sequencing reactions in parallel, and (3) maximize priming effectiveness by learning more about the factors that affect it. Our ultimate aim is to develop a fully automated machine capable of producing hundreds of thousands of base pairs of finished DNA sequence per day.

Novel Separation and Detection Methods for Gene Mapping and DNA Sequencing

Project Renewed in FY 1993

Edward S. Yeung
Department of Chemistry; Iowa State University; Ames, IA 50011
515/294-8062, Fax: -0266, Internet: *yeung@ameslab.gov*, BITNET: *yeung@alisuvax*

Electrophoresis is one of the most powerful proven techniques available for gene mapping and sequencing. The number of possible resolution elements indicates that separation efficiencies and information content in two-dimensional gels easily outperform other techniques. Recently, electrophoresis in capillary tubes has shown potential for extended size range in sequencing runs and for substantially increased speed. The major problem is in detecting the separated components.

In conventional electrophoresis, a tag is introduced to allow measurement by absorption, fluorescence, or radiography. At best, only semiquantitative results are obtained because of unreliable chemistry and difficulties in probing a two-dimensional spot, which can be distorted. Staining can also affect component migration and lead to sequencing errors.

We propose to develop novel separation, detection, and imaging techniques for real-time monitoring in electrophoresis. Emphasis will be on schemes that allow multiplexing and on methods that do not require specialized fluorescent or radioactive tags. These techniques will be used for substantially increasing the speed, reliability, and sensitivity in gene mapping and DNA sequencing applications, both in slab gels and in capillary gels.

Projects Continuing into FY 1993

Sequencing by Hybridization: Methods to Generate Large Arrays of Oligonucleotides

Thomas M. Brennan
Engineering Division; Lawrence Berkeley Laboratory; Berkeley, CA 94720
On site at Stanford University; Palo Alto, CA 94301
415/725-7423, Fax: -1534, Internet: *brennan@sumex-aim.stanford.edu*

Detection of Luminescence from Lanthanide Ions as Labels for DNA Sequencing

Gilbert M. Brown, Robert S. Foote, K. Bruce Jacobson, Frank W. Larimer, Roswitha S. Ramsey, Richard A. Sachleben, and Richard P. Woychik
Oak Ridge National Laboratory; Oak Ridge, TN 37831-6119
615/576-2756, Fax: -5235

Vacuum Ultraviolet Ionizer Mass Spectrometer for Genome Sequencing

C. H. Winston Chen, Marvin G. Payne,[1] and **K. Bruce Jacobson**
Health and Safety Research Division; Oak Ridge National Laboratory; Oak Ridge, TN 37831
615/574-5895, Fax: /576-2115
[1]Department of Physics; Georgia Southern University; Statesboro, GA 30460

Development of a Fully Integrated Technology to Facilitate Sequencing the Human Genome

George Church
Department of Genetics; Harvard University; Boston, MA 02115
617/732-7562, Fax: -7663, Internet: *church@rascal.bwh.harvard.edu*

Sequencing by Hybridization

Radomir Crkvenjakov and **Radoje Drmanac**
Biological and Medical Research Division; Argonne National Laboratory; Argonne, IL 60439-4833
708/252-3161 or -3175, Fax: -3387, Internet: *crkve@mcs.anl.gov*

Genomic Instrumentation Development: Detection Systems for Film and High-Speed Gel-Less Methods

Jack B. Davidson and Robert S. Foote[1]
Instrumentation and Controls Division; [1]University of Tennessee Graduate School of Biomedical Sciences and Biology Division; Oak Ridge National Laboratory; Oak Ridge, TN 37831-6010
615/574-5599, Fax: -4058

Single-Molecule Detection Using Charge-Coupled Device Array Technology

M. Bonner Denton and **Richard Keller,**[1]
Department of Chemistry; University of Arizona; Tucson, AZ 85721
602/621-8246, Fax: -8272, Internet: *mbdenton@ccit.arizona.edu*
[1]Chemical and Laser Sciences Division; Los Alamos National Laboratory; Los Alamos, NM 87545

Multicolumn Gel Electrophoresis and Laser-Induced Fluorescence Detection for DNA Sequencing at 64,000 Bases/Hour

Norman J. Dovichi
Department of Chemistry; University of Alberta; Edmonton, Alberta, Canada T6G 2G2
403/492-2845, Fax: -8231, Internet: *norm_dovichi@dept.chem.ualberta.ca*

Rapid Preparation of DNA for Automated Sequencing

John J. Dunn and **F. William Studier**
Biology Department; Brookhaven National Laboratory; Upton, NY 11973
512/282-3012 or -3390, Fax: -3407, Internet: *dunn@genome1.bio.bnl.gov*

Using Scanning Tunneling Microscopy to Sequence the Human Genome

Thomas L. Ferrell, Robert J. Warmack, David P. Allison, K. Bruce Jacobson, Gilbert M. Brown, and Thomas G. Thundat
Oak Ridge National Laboratory; Oak Ridge, TN 37831-6123
Warmack: 615/574-6215, Fax: -6210, BITNET: *rjw@ornlstc*

DNA Sequence Analysis by Solid-Phase Hybridization

Robert S. Foote,[1] Richard A. Sachleben,[2] and K. Bruce Jacobson[1]
University of Tennessee Graduate School of Biomedical Sciences; [1]Biology Division and [2]Chemistry Division; Oak Ridge National Laboratory; Oak Ridge, TN 37831-8077
615/574-0801, Fax: -1274

Advanced Sequencing Technology

Raymond F. Gesteland and Robert Weiss
Department of Human Genetics; University of Utah; Salt Lake City, UT 84112
801/581-5190, Fax: /585-3910, Internet: *rayg@genetcs.med.utah.edu*

Megabase Sequencing of Human Immune Receptor Loci

Leroy E. Hood
Department of Molecular Biotechnology; University of Washington; Seattle, WA 98195
206/685-7367, Fax: -7301

DNA Sequencing Using Stable Isotopes

K. Bruce Jacobson, Heinrich F. Arlinghaus,[1] Gilbert M. Brown,[2] Robert S. Foote, Frank W. Larimer, Richard A. Sachleben,[2] Norbert Thonnard,[1] and Richard P. Woychik
Biology Division and [2]Chemistry Division; Oak Ridge National Laboratory; Oak Ridge, TN 37831-8077
615/574-1204, Fax: -1274, BITNET: *bru@ornl.stc*
[1]Atom Sciences, Inc.; Oak Ridge, TN 37830

Projects Continuing into FY 1993

Advanced Detectors for Mass Spectrometry

Joseph M. Jaklevic, W. Henry Benner, and Joseph Katz
Human Genome Center and Engineering Division; Lawrence Berkeley Laboratory; Berkeley, CA 94720
510/486-5647, Fax: -5857, Internet: *jmjaklevic@lbl.gov,* BITNET: *jmj@lbl*

Rapid DNA Sequencing Based on Fluorescence Detection of Single Molecules

James H. Jett, Richard A. Keller, John C. Martin, and **E. Brooks Shera**
Center for Human Genome Studies; Los Alamos National Laboratory; Los Alamos, NM 87545
Keller: 505/667-3018, Fax: /665-3024

Transposon-Based Genomic Sequencing

Christopher H. Martin, Michael Strathmann, Carol A. Mayeda, and **Michael J. Palazzolo**
Human Genome Center; Cell and Molecular Biology Division; Lawrence Berkeley Laboratory; Berkeley, CA 94720
Martin and Palazzolo: 510/486-5909, Fax: -6816, Internet: *chrism@genome.lbl.gov* or *mjpalazzolo@lbl.gov*

Ultrasensitive Fluorescence Detection of DNA

Richard A. Mathies, Mark A. Quesada, Hays S. Rye,[1] Xiaohua Huang, Jiun W. Chen, and **Alexander N. Glazer**[1]
Departments of Chemistry and [1]Molecular and Cell Biology; University of California; Berkeley, CA 94720
510/642-4192, Fax: -3599

Preparation of Oligonucleotide Arrays for Hybridization Studies

Michael C. Pirrung
Department of Chemistry; Duke University; Durham, NC 27708-0346
919/660-1556, Fax: -1591

Thioredoxin-Gene 5 Protein Interactions: Processivity of Bacteriophage T7 DNA Polymerase

Jeff Himawan, Stanley Tabor, and **Charles C. Richardson**
Department of Biological Chemistry and Molecular Pharmacology; Harvard Medical School; Boston, MA 02115
617/432-3129, Fax: -3362

Improvement and Automation of Ligation-Mediated Genomic Sequencing

Arthur D. Riggs and Gerd P. Pfeifer
Department of Biology; Beckman Research Institute of the City of Hope; Duarte, CA 91010
818/301-8352, Fax: /358-7703

A High-Speed Automated DNA Sequencer

Lloyd M. Smith
Department of Chemistry; University of Wisconsin; Madison, WI 53706
608/263-2594, Fax: /262-0453, Internet: *smith@bert.wisc.edu*

High-Speed DNA Sequence Analysis by Matrix-Assisted Laser Desorption Mass Spectrometry

Lloyd M. Smith and **Brian Chait**[1]
Department of Chemistry; University of Wisconsin; Madison, WI 53706
608/263-2594, Fax: /262-0453, Internet: *smith@bert.wisc.edu*
[1]Rockefeller University; New York, NY 10021

Automation of the Front End of DNA Sequencing

Lloyd M. Smith and **David Mead**[1]
Department of Chemistry; University of Wisconsin; Madison, WI 53706
608/263-2594, Fax: /262-0453, Internet: *smith@bert.wisc.edu*
[1]Chimerx; Madison, WI 53704

Ion Cyclotron Resonance-Mass Spectroscopy of DNA Molecular Ions

Richard D. Smith, Charles G. Edmonds, and Joseph A. Loo
Chemical Sciences Department; Pacific Northwest Laboratory; Richland, WA 99352
509/376-0723 or -5665, Fax: -0418

Large-Scale DNA Sequencing with a Primer Library

F. William Studier and **John J. Dunn**
Biology Department; Brookhaven National Laboratory; Upton, NY 11973
516/282-3390, Fax: -3407

Time-of-Flight Mass Spectrometry of DNA for Rapid Sequence Determination

Peter Williams and **Neal Woodbury**
Department of Chemistry; Arizona State University; Tempe, AZ 85287-1604
602/965-4107, Fax: -2747

Informatics

Projects New in FY 1993

*Method for Direct Sequencing of Diploid Genomes on Oligonucleotide Arrays: Theoretical Analysis and Computer Modeling

Alexander B. Chetverin, A. R. Rubinov, M. S. Gelfand, S. A. Spirin,[1] M. E. Ivanov,[2] R. F. Nakipov, and O. I. Razgulyaev
Viral RNA Biochemistry Group; Institute of Protein Research and [1]Institute of Mathematical Problems in Biology; Russian Academy of Sciences; 142292 Pushchino, Moscow Region, Russia
[2]A. N. Belozersky Institute of Physical and Chemical Biology; Moscow State University; Moscow, Russia
Internet: *chetveri@cgi.nks.sn*

Projects to sequence several large genomes, particularly the human genome, make the automation of nucleic acid sequencing one of the most important problems of biochemistry and molecular biology. Recently several groups in Great Britain, Yugoslavia, Russia, and the United States suggested a new, easily automatable method for DNA sequencing by hybridization (SBH) with oligonucleotides, compiling the complete list of oligonucleotides of a fixed length (k-tuples) that occur in a DNA fragment (fragment vocabulary) and subsequently reconstructing the fragment sequence by linking maximally overlapping k-tuples from the vocabulary.

In its current form SBH does not accomplish complete automation of genome sequencing, since the need still exists for random cloning of millions of fragments. To overcome this problem, A. B. Chetverin and F. R. Kramer recently suggested a method for direct sequencing of large (including diploid) genomes on oligonucleotide arrays.

Unlike existing strategies for total genome sequencing, the suggested approach would not require preliminary fragment cloning and chromosome mapping and could be completely automated. Substantial reductions in time and cost would result; preliminary estimates show that the cost of human genome sequencing could be reduced 10- to 100-fold. Furthermore, total sequencing of an individual diploid genome can be considered with this method, as opposed to sequencing a haploid set of fragments arbitrarily compiled from different individual genomes. One of the stages (i.e., fragment sorting and pool reconstruction, see below) can be used for direct sequencing of the total pool of cellular RNA, with obvious advantages for genetics, developmental studies, and medicine.

In this project we plan to develop, implement, and test computer algorithms to solve the problem of reconstructing fragment sequences in a pool. This is important for the following reasons. First, computer processing of biochemical data is an integral and essential part of the method. Second, practical implementation of the method should be preceded by intensive computer modeling and theoretical analysis to determine such optimal biochemical parameters as oligonucleotide size, mean and maximal values of fragment length and pool size measured as their combined length, and signal-to-noise (s/n) ratio in hybridization procedures. Such a preliminary analysis would substantially increase reliability and decrease the cost of collecting biochemical data. Finally, some arising mathematical problems are novel, and they have independent value for discrete mathematics and computer science.

Accurate Restoration of DNA Sequences

Gary A. Churchill
Biometrics Unit; Cornell University; Ithaca, NY 14853-7801
607/255-5488, Fax: -4698, Internet: *gary@amanita.biom.cornell.edu*

This project will develop statistical methods and algorithms to detect potential errors in DNA sequence data and compute summary measures of sequence quality. Methods will be based on a stochastic model of sequencing errors that will apply to any technology that generates overlapping linear sequence fragments, including gel and capillary electrophoresis. New high-speed sequencing

technologies are expected to be developed that may produce errors in raw fragment sequences at a much higher rate then existing methods. The problem of assembling overlapping sequence fragments will also be addressed.

Our long-term goal is to develop an efficient and fully automated quality-control procedure that should be an integral part of any large-scale sequencing project. This automated procedure should eventually replace much expensive human effort expended in rechecking raw sequence data. Model-based statistical methods can help improve the efficiency of existing sequencing technology (for example, by allowing for longer gel runs) and will be an essential component of any system that can produce high-quality finished sequence at the target rate of 1 Mb/d.

Computer-Aided Genome Map Assembly with SIGMA (System for Integrated Genome Map Assembly)

Michael J. Cinkosky, Michael A. Bridgers, William M. Barber, Mohamad Ijadi, and James W. Fickett
Theoretical Biology and Biophysics Group; Los Alamos National Laboratory; Los Alamos, NM 87545
505/665-0840, Fax: -3493, Internet: *michael@t10.lanl.gov*

SIGMA (System for Integrated Genome Map Assembly), a recently released graphical genome map editor, supports the following:

- graphical, mouse-based genome map editing;
- integration of data from many different types of physical and linkage experiments and at all appropriate resolution levels, from banded ideograms to restriction fragments;
- creation of multiple "views" on a single map;
- creation by users of new classes of map objects on demand; and
- workgroup map building through the use of client-server database management system technology.

In addition, SIGMA enables users to store map-based data as part of the map itself. This feature:

- keeps underlying data as part of the map, allowing users to know the real support for any given map;
- automatically evaluates the map against underlying data, pointing out places where the two disagree; and
- creates a platform for automatic map assembly algorithms being developed by many groups worldwide.

The software, documentation, and a number of sample maps in SIGMA format, including current Genome Data Base maps, are available by anonymous FTP from *atlas.lanl.gov.* Additional information may be obtained by sending a message containing only the word **sigma-info** to *bioserve@t10.lanl.gov.*

Informatics for the Sequencing by Hybridization Project

Aleksandar Milosavljevic and **Radomir Crkvenjakov**
Center for Mechanistic Biology and Biotechnology; Argonne National Laboratory; Argonne, IL 60439-4833
708/252-3161, Fax: -3387, Internet: *crk@everest.anl.gov*

Methods for the design and analysis of massive hybridization experiments have been developed on a solid theoretical basis. Algorithmic information theory and minimal length encoding are being used to design methods for comparing partial sequence data and databases of sequenced DNA. Prelimi-

Projects New in FY 1993

nary experiments led to the first identification of 55 gamma-actin cDNA clones based on their hybridizations to 110 heptamer probes. In addition, clustering methods are being developed and applied to discover groups of similar clones in cDNA and genomic libraries through comparison of their hybridization signatures. Clone clustering revealed transcriptional structure in human brain cDNA libraries. A preliminary experiment demonstrated the potential of the clustering method to reduce by 10 times the sequencing effort needed to cover a 12-kb segment of human genomic DNA. Mutual information and other concepts from information theory are being applied to interpret measurements optimally and approach design of massive hybridization experiments.

A relational database that contains a complete record of massive hybridization experiments has been developed using Sybase client-server technology. A complete suite of programs for experiment design, entry of experimental data, and data analysis has been built into the UNIX C-shell in order to facilitate the writing of C-shell scripts for these functions. The programs are written in C++ and interfaced with the database by using Sybase db-library. A new level of experiment design and debugging is being developed to facilitate a complete experiment design in the computer; this design will then be automatically converted into robotic instructions for such routine laboratory operations as microtiter plate manipulation and dot-blot filter printing. Preliminary research is being performed using the Quintus Prolog system to interface the database of hybridization experiments with databases of annotated DNA sequences. The logic programming level will enable the rapid design of computational experiments as well as a uniform data representation level, both of which are necessary for machine discovery of biologically relevant patterns.

Sequencing by Hybridization Algorithms and Computational Tools

Radoje Drmanac, Ivan Labat, and Nick Stavropoulos
Integral Genetics Group; Center for Mechanistic Biology and Biotechnology; Argonne National Laboratory; Argonne, IL 60439
708/252-3175, Fax: -3387, Internet: *rade@everest.bim.anl.gov*

Sequencing by hybridization (SBH) requires sophisticated computational procedures for data acquisition and evaluation and for DNA screening, mapping, and sequencing applications. We have been developing algorithms and programs and performing simulations to prove many SBH possibilities other than the straight sequencing of short DNA fragments. For example, we demonstrated a 10- to 50-fold increase of SBH efficiency by using overlapped and similar sequences in the assembly process. Furthermore, we showed that partial sequences obtained by 100 to 1000 probes are sufficient for gene identification and recognition of overlapped and similar sequences.

Recently we have started to produce large sets of real hybridization data that define practical requirements and serve as a final check for the necessary programs. Several computational tools have been developed to enable use of the data. The programs are based on heuristic rules and resemble expert systems resistant to common experimental imprecision. All the programs use hybridization intensities without conversion to 0/1 (binary) form.

Acquisition of hybridization data from filters containing 31,000 dots (1 dot/mm^2) is the first step. An image-analysis program (DOTS) has been developed (J. Jarvis and R. Drmanac) that automatically defines filter position and reports hybridization intensity for each dot. Programs for evaluating and normalizing data (SCORES), identifying groups of similar clones (CLUSTERS), and ordering clones (CORD) are in the final phase of development. The programs are written in C for UNIX platforms with an X-Windows interface. Through SCORES and CLUSTERS, 20,000 cDNA clones have been sorted into 13,000 groups. Further, an algorithm has been developed for matching hybridization signatures with known sequences by simulating the expected probe scores for known sequences. The CORD program defines contigs of 1- to 2-kb clones hybridized by 200 probes and provides sequence-ready maps. The maps allow clone selection for complete sequencing by less then 2 reads/bp. In various simulation experiments, CORD has shown a tolerance to more hybridization errors than observed in our experiments and to the high abundance of *Alu* repeats found in human sequences.

Key programs remain to be developed for assembling sequence by (1) combining single-pass gel sequences having as much as a 20% error rate with hybridization data from 3000 probes and (2) integrating hybridization data from similar sequences. A program is also needed for using partial sequence data to identify genes and other genome elements.

*Computer Analysis of Functional Regions of the Human Genome

Nikolay A. Kolchanov
Laboratory of Theoretical Molecular Genetics; Institute of Cytology and Genetics; Siberian Branch of the Russian Academy of Sciences; Novosibirsk 630090, Russia
+7-3832/353-335, Fax: /356-558, Internet: *kol@cgi.nsk.su*

This work is directed toward computer analysis of the structural-functional organization and evolution of functional regions of human and mammalian genomes. Attention will be focused on genes coding for proteins, 5′ regulatory regions, 3′ flanking regions, and repeat sequences.

We plan to study the distribution of short oligonucleotides adjacent to transcribed regions. Our specific aim is to identify statistically significant oligonucleotide patterns in the transcription starting region of RNA polymerase II. Particular emphasis will be on oligonucleotide distribution that can be approximated to linear trends within promoter regions (200 to 400 bp).

Evolution of the distribution pattern of short nucleotides adjacent to transcription start regions (200 to 300 bp) will also be studied. For this purpose, we intend to perform computer modeling of primate and rodent genes. Simulation results will be compared with real data, and the most adequate evolutionary models of promoter regions in neutral and adaptive variants will be chosen. A computer method will be designed to recognize starting points of transcription by RNA polymerase in eukaryotic genes. Linear-discriminant analysis of many features relevant to oligonucleotide distribution in the transcription start region will be performed for this purpose.

Contextual features of genes coding for proteins will be analyzed. Analysis will take into account the exon-intron structure in these genes. We intend to identify contextual features specific to each given functional region by analyzing exons, introns, and their boundaries containing donor and acceptor splicing sites. These features will be used to develop computer methods for recognizing exons, introns, and whole genes in nucleotide sequences. Methods will include traditional approaches based on discriminant analysis and methods of classification theory and recursive contextual systems.

Computer analysis will be performed to determine molecular mechanisms of mutation emergence in the coding parts of genes in human and other vertebrate genomes. Our goal is to determine the role of polynucleotide context and estimate the contribution of different mutagenesis mechanisms to mutation emergence. Mechanisms to be studied include template chain dislocation, gene conversion, heteroduplex repair, and effects of specific and nonspecific signals of polynucleotide context on mutation emergence.

Polyadenylation sites in the 3′ ends of genes transcribed by RNA polymerase II will be analyzed. These sites, key signals in 3′ flanking regions of a given group of genes, determine features of 3′-end processing of pre-mRNAs. We intend to analyze specific features of nucleotide context to determine the location of polyadenylation sites in genomic DNA. Based on these results, a method will be developed for recognizing polyadenylation sites in sequences of human and other vertebrate genomes. We intend to identify specific contextual features of polyadenylation sites that determine the magnitude of their functional activities. Methods will be developed for estimating functional activities of given sites derived from sequences.

We will also study the insertion-site contextual features of *Alu* repeats in the primate genome. These features include homology and complementarity between *Alu* sequences and insertion regions as well as distribution patterns of short oligonucleotides within insertion sites of *Alu* repeats. The number of potential insertion points of *Alu* into human and primate genomes will be estimated.

Projects New in FY 1993

21Bdb: A Database for Human DNA Sequence Information

Suzanna Lewis, John McCarthy, Edward Theil, Arun Aggarwal, Donn Davy, Sam Pitluck, and Michael Palazzolo
Human Genome Computing Group; Lawrence Berkeley Laboratory; Berkeley, CA 94720
McCarthy: 510/486-5307, Fax: -4004, Internet: *jlmccarthy@lbl.gov*

21Bdb is a variant of ACEDB, a suite of database and display software originally developed by Richard Durbin and Jean Thierry-Mieg to meet the needs of the *Caenorhabditis elegans* project. 21Bdb includes all the functionality of ACEDB and extends those capabilities to meet new requirements of the Lawrence Berkeley Laboratory (LBL) chromosome 21 sequencing project. 21Bdb is being used to maintain and provide information for laboratory personnel and the chromosome 21 research community.

Three aspects of ACEDB that have been especially useful for LBL are schema design, data presentation, and collaboration. ACEDB makes relatively easy the continuous refinement of the database schema to match ongoing research needs and permit timely responses to rapidly changing laboratory requirements. Second, ACEDB already includes numerous graphical displays for genomic data and an independent simple graphics library that is completely portable across many platforms. This feature enables us to formulate quickly our own customized data displays, as we have already done for flyDB (developed for the *Drosophila* physical mapping project). Finally, LBL staff have established a very productive ongoing collaboration with the original developers of ACEDB to extend ACEDB via the Internet. Many LBL enhancements have already been incorporated into the standard ACEDB distribution. ACEDB will also be used by the new Sanger Sequencing Center in Cambridge, England, thus making available more opportunities for collaboration on human genome sequencing tools.

LBL is exploiting a directed sequencing technique on its chromosome 21 project. The implications for laboratory data management are significant. By definition, a directed strategy requires that biologists know the complete heritage of each DNA sequence and its position in relationship to other DNA sequences. This knowledge simplifies and makes more tractable the sequence-assembly process. The database must record all subclones derived from each P1 clone, the P1 subclone map, the transposon-insertion map of each subclone, descriptions of all transposon-inserted priming sites derived from each subclone, and sequencing status and results for every priming site. Recorded data are available both graphically and computationally.

This physical map and chromosome 21 sequence data generated by this project will be available to the community through the 21Bdb database. These data include not only the LBL P1 physical map but also corresponding linkages to the Genethon yeast artificial chromosome map via sequence tagged sites derived at both laboratories. The database also incorporates high-resolution maps of individual P1s made before sequencing and the sequence data itself. Collaborative work on providing public access is under way. Emphasis is on a graphical presentation that looks and feels natural for biologists.

Multiple Alignment and Homolog Sequence Database Compilation

Hwa A. Lim
Florida State University; Tallahassee, FL 32306
904/644-1010, Fax: -0098, Internet: *hlim@scri.fsu.edu*

In the last few years, especially with the prolific growth of sequence data since the Human Genome Project began, the problem of sequence alignment has been a main research subject in theoretical and computational molecular biology. Although different algorithms have been forwarded, they are all based on very simple evolutionary models. The underlying evolutionary mechanisms are not

studied in detail and therefore are not adequately considered in existing alignment algorithms. Most efforts have been concentrated on accuracy and speed improvement, but the objective function has never been considered carefully.

The first goal of this project is to develop an alignment procedure capable of perceiving and considering features of evolutionary pattern for query sequence samples. To achieve this, the following approaches will be pursued: (1) alignment of many sequence families contained in databases, (2) study of various alignment parameter values on the basis of available three-dimensional structure alignments and evolutionary models, and (3) alignment of computer-simulated sequences. This study should also improve understanding of mechanisms of sequence evolution. The statistically based approach of diagonal fragment analysis will be used as a basic alignment technique.

The second goal is to compile probable sequence families in a homolog sequence database useful in various applications. To this end, a massively parallel machine such as the Connection Machine allows investigators to perform the task within a reasonable time.

The third goal of this project is to apply databases developed for human genome sequence investigations. This will include investigation of human multigene families and comparative analysis of human sequences with other species, leading to a better characterization of their functional roles.

Construction of an Integrated Database to Support Genetic Sequence Analysis

Ross Overbeek and **Patrick Gillevet**[1]
Mathematics and Computer Science Division; Argonne National Laboratory; Argonne, IL 60439
708/252-7856, Fax: -5986, Internet: *overbeek@mcs.anl.gov*
[1]National Center for Human Genome Research; National Institutes of Health; Bethesda, MD 20855
301/402-2540 or -2534, Fax: -2120, Internet: *gillevet@uranus.nchgr.nih.gov*

This 3-year project will attempt to create an integrated database to support comparative analyses of genomes. The database will initially include data from GenBank®, Swiss Protein Data Bank, Swiss Enzymes Data Bank, EcoSeq Database, ProSite Dictionary of Protein Sites and Patterns, the compilation of compounds distributed by Peter Karp, a representation of the more significant metabolic pathways, genetic maps for a number of bacterial genomes, and additional data relating to the specific sequencing project at Harvard University. This first version will be extended as rapidly as feasible to include data relating to the physical structure of proteins (largely from the Brookhaven Protein Data Bank), manually and automatically generated alignments, phylogeny (most notably the tree and supporting data distributed by the Ribosomal Database Project), and other curated databases such as CarBank and Genome Data Base.

The immediate goal is to compile a database of information about *Mycoplasma capricolum*, and the long-range goal is to identify functional components of the organism, relating genes to their roles in pathways and explicating regulatory mechanisms for pathways. These sweeping goals require the establishment of a framework for comparative analysis.

Although the initial use of the system will be for comparative analysis of microorganisms and their genetic properties, we plan to develop and distribute a tool that will include data on all organisms and have widespread usefulness within the community. This tool will emerge from connecting (1) an X-Windows–based system devoted to integration of analysis tools and (2) the Argonne GenoBase project focused on integrating existing databases.

Since the advent of the genome project, substantial advances have been made in the availability of genomic data. The number of databases for retention of DNA sequence is increasing, and specialized databases for peptides, enzymes, motifs, alignments, metabolic pathways, and two-dimensional protein gels are also now available. Quality and variety will continue to improve rapidly, necessitating a system to integrate the growing collection of heterogeneous databases.

Projects New in FY 1993

Chromosome 21 Physical Mapping and Analysis

Stewart Scherer
Human Genome Center; Lawrence Berkeley Laboratory; Berkeley, CA 94720
510/486-4856 or -5468, Fax: -6816, Internet: *stew@genome.lbl.gov* or *stew@lenti.med.umn.edu*

Large-scale sequencing of human DNA will be initiated in the next few years. Although moderate-resolution physical maps of human chromosomes are now being assembled with yeast artificial chromosome (YAC) clones, these clones clearly will not be useful sources of templates for DNA sequencing. To provide such a source, the Lawrence Berkeley Laboratory Human Genome Center is constructing a contig of P1 clones from chromosome 21q22.3 that will also be useful in efforts to saturate the region for cDNAs and genetic markers. We focused on this gene-rich 3-Mb region because it is known to be involved in Down's syndrome.

As part of this work, we are enhancing and correcting the existing chromosome 21 YAC map based on sequence tagged sites. We will attempt to understand why certain YAC clones are unstable and whether these regions are better propagated in bacterial hosts.

Analysis of P1-sized genomic sequences with existing computer software is ill-suited to the high-throughout DNA sequencing anticipated in the genome project. We have developed a series of algorithms for rapid comparison of new sequences with statistical models of underlying genome structure. The output is presented graphically so the user is directed rapidly to regions of unusual sequence organization. We intend to integrate our own novel algorithms into a single package with sequence-analysis procedures developed by others. This package will highlight interesting features detected by these programs and provide a graphical overview of the largest contemplated sequences.

*DBEMP: Data Base on Enzymes and Metabolic Pathways

Evgenij E. Selkov
Laboratory for Mathematical Simulation of Multienzyme Systems; Institute of Theoretical and Experimental Biophysics; Russian Academy of Sciences; 142292 Pushchino, Moscow Region, Russia

The ability to relate sequence data to a comprehensive representation of functional gene roles and coded proteins is highly desirable. Data describing metabolic pathways and their regulation will also be vital in creating a framework for studying disorders that result from disruptions of specific metabolic systems. Since the early 1980s, the research group under E. Selkov has been encoding quantitative data from more than 14,000 journal articles into the DataBase on Enzymes and Metabolic Pathways (DBEMP). This database contains by far the most extensive data set relating to enzymes and metabolism and provides a critical functional context for the emerging volume of sequence data.

During 1993, the Russian team achieved the following.

- The number of encoded journal articles increased by 30% from about 10,000 to over 14,000.
- A CD-ROM version of DBEMP was developed and will be available to both the research and commercial communities in early 1994.
- A set of 902 metabolic maps was compiled, a selection of which will soon be available to the research community via Internet (World Wide Web). The maps cover most essential pathways of *Escherichia coli, Saccharomyces cerevisiae*, mollicutes, mammalia, and higher plants.
- Software is under development to allow users to enter their own data into DBEMP. The software will be distributed worldwide to accelerate the encoding of relevant data.
- Parsers have been built to perform consistency checks of all fields of all DBEMP records. Spell checkers and format checkers are also in place.

The collaborating team at Argonne National Laboratory (ANL) has incorporated selections from DBEMP into the ANL Genobase, a system of integrated biological databases that allows users to query numerous data repositories about genetic sequences and proteins. As a result, 192 selected metabolic pathways have been integrated with the Swiss Protein Data Bank, portions of the EMBL Nucleic Acid Data Bank (including sequences for *E. coli* and five other organisms), the Enzyme Data Bank, the Blocks database, the ECO2DB data bank, and the Prosite motif-pattern data bank.

New Approaches to Recognizing Functional Domains in Biological Sequences

Gary D. Stormo
Department of Molecular, Cellular, and Developmental Biology; University of Colorado; Boulder, CO 80309-0347
303/492-1476, Fax: -7744, Internet: *stormo@boulder.colorado.edu*

Problems in identifying coding regions and other important functional domains in genomic DNA sequences will be approached using a combination of dynamic programming and neural network methods. Dynamic programming returns optimal partitioning of sequences into regions of different classes, given a particular weighting of evidence for those classes. The neural network is used to find the weights that maximize the performance of dynamic programming predictions. Dynamic programming can also be used to obtain suboptimal sequence partitioning, which can be very effective in assessing prediction reliability in different regions and in providing alternative partitioning models in cases where a high degree of reliability is not achieved. Combining an optimization procedure like dynamic programming with a machine-learning procedure like neural networks should be applicable to a wide range of problems beyond those being studied in this project.

Human Genome Center Informatics Group

Edward H. Theil, Arun Aggarwal, Donn Davy, Suzanna Lewis, Victor Markowitz, John McCarthy, Sam Pitluck, Eugene Veklerov, and Manfred Zorn
Human Genome Center; Lawrence Berkeley Laboratory; Berkeley, CA 94720
510/486-7501, Fax: -5936, Internet: *ehtheil@lbl.gov*

The Informatics Group of the Lawrence Berkeley Laboratory (LBL) Human Genome Center develops software to address problems related to the electronic capture, representation, and organization of data generated by the genome laboratories. The group works with biologists and engineers at the center to provide many forms of computer assistance, including custom software, access to external databases, and tools for portability of data. In addition, the group is concerned with longer-range problems of sequence and clone assembly, database design, and tools for data management.

The Flydb database has been developed to represent physical map information for the model organism *Drosophila melanogaster.* Flydb, which can display physical maps and in situ images, contains a contig-assembly algorithm based on the clone-limited sequence tagged site (STS) mapping strategy used at LBL. Another database, 21Bdb, displays physical maps of chromosome 21 based on STS markers and includes yeast artificial chromosomes, polymorphic repeats, and P1 clones known to contain these markers. 21Bdb also allows maps to be manipulated by moving STSs from one relative ordering to another with "click and drag." These databases support the full functionality associated with ACEDB (*Caenorhabditis elegans* database) as well as the ability to retrieve digitized images used in the mapping process.

Other projects include a sequence-assembly system based on the directed strategy used at LBL (a collaboration with Baylor College of Medicine), new ways to visualize the results of sequence analysis, and high-level tools to model laboratory protocols as part of data flow.

Projects New in FY 1993

Software for Sequence Assembly Based on the Directed Approach

Eugene Veklerov, Suzanna Lewis, **Christopher Martin**, Sam Pitluck, and Edward Theil
Human Genome Computing Group; Lawrence Berkeley Laboratory; Berkeley, CA 94720
Veklerov: 510/486-7532, Fax: -6816, Internet: *veklerov@cse.lbl.gov*
Martin: 510/486-5654, Fax: -6816, Internet: *chrism@guyana.lbl.gov*

Existing software packages do not fully support the directed DNA sequencing strategy in use at Lawrence Berkeley Laboratory (LBL). Specifically, they are inadequate in the following areas:

Algorithms: The assembly algorithms were originally designed for the shotgun strategy and not to take advantage of all the information available to biologists using a directed strategy. Algorithms that properly use this information can overcome performance difficulties when the sequences become very long or when repeated sequences cause ambiguities.

Data Model: The sequencing strategy developed at LBL relies on a hierarchy of maps of increasingly higher resolution. The various pieces of sequencing software must be able to incorporate all these maps into a comprehensive data model.

User Interface: The large volume of data generated by large-scale sequencing requires that all data be available in a simple graphical form. The most time-consuming operations should be fully automated while still allowing the biologist to override automatic procedures.

We have written several programs that alleviate some difficulties in applying the Staden xdap package to our strategy. These programs perform several disjoint functions, including:

- graphical display of the xdap alignment algorithm output;
- assembly of 3- to 4-kb fragments into a P1 clone; and
- location of inconsistent gel-reading positions in the consensus line.

The programs will be incorporated into new, much more flexible software designed to remedy some of the inadequacies of existing packages. Because of Smalltalk's fast production of prototypes and superior data-modeling capabilities, we are using it to implement the system in a collaboration with Charles Lawrence's group at Baylor College of Medicine.

Using Metadata To Automatically Generate User Interfaces for Genomic Databases

Manfred D. Zorn
Information and Computing Sciences Division; Lawrence Berkeley Laboratory; Berkeley, CA 94720
510/486-5041, Fax: -4004, Internet: *mdzorn@lbl.gov*, BITNET: *mdzorn@lbl*

The Human Genome Project has a growing need to manage and distribute information. Databases for this purpose are often cumbersome for biologists to use or require extensive effort to build friendlier user interfaces. Adaptations of database structure to the changing needs of an evolving research area lead to costly modifications of user-interface applications.

We are developing software for the automatic generation of graphical, user-friendlier, forms-based user interfaces from high-level database definitions. An extended-entity-relationship (EER) model captures real-world objects and defines the underlying database. The EER schema, which constitutes part of the metadata, is used to create a user-interface object model that is stored in a configuration file. A generic user-interface application reads in the configuration file to produce a user interface for a particular database. The object definition in the configuration file defines not only the elements in the user interface but also an internal self-describing data structure and mappers that specify the translation between the database and the user-interface formats. Procedures that access the database and retrieve information are created by specifying queries in an EER-based

query language for the objects in the configuration file. Thus the user interface and the connection to the underlying database are generated automatically, and database changes are easily propagated to create a modified user interface.

Biopoet: A System for Large-Scale Sequence Analysis

Manfred D. Zorn, Jane Macfarlane,[1] and Robert Armstrong[2]
Human Genome Center and [1]Information and Computing Sciences Division; Lawrence Berkeley Laboratory; Berkeley, CA 94720
510/486-5041, Fax: -4004, Internet: *mdzorn@lbl.gov*, BITNET: *mdzorn@lbl*
[2]Sandia National Laboratory; Livermore, CA 94550

In the past year, the dramatic increase in the rate of new-sequence generation has presented a major challenge for sequence analysis. Increasingly longer sequences are being analyzed as finished sequences become larger than 100 kb, and database size doubles almost every year for sequence-similarity searches. Sophisticated computing technology for tackling these problems already exists in faster machines, parallel processing, and distributed computing. However, optimal access requires detailed knowledge of particular resources.

POET, the Parallel Object-Oriented Environment and Toolkit, is modeled after the X11 toolkit and enables both high- and low-level control of computational methods. The object-oriented programming paradigm offers data encapsulation and methods for hiding implementation details to present a unified object view to the user. Existing software can be adapted to exploit the power of parallel processing. Thus sequence analysis can be performed transparently to the user in reasonable time where POET divides either the query sequence or the database into multiple pieces to run on parallel computers or on a number of workstations in a distributed environment.

We are developing BioPOET, a prototype system that integrates sequence analysis into a friendly user interface and performs comparisons of large sequences. The user interface, developed in ParcPlace Smalltalk (produced by ParcPlace Systems) allows parameter specification for several analysis options and for launching the analysis program. A graphical display presents the results to the user.

Efficient Algorithms and Data Structures in Support of DNA Mapping and Sequence Analysis

Eugene Lawler and **Daniel Gusfield**[1]
Electronics Research Laboratory; University of California; Berkeley, CA 94720
510/642-4019, Fax: -5775, Internet: *lawler@arpa.berkeley.edu*
[1]Division of Computer Science; University of California; Davis, CA 95616
916/752-7131, Fax: -4767, Internet: *gusfield@cs.ucdavis.edu*

The objective of this project is to identify computational problems of fundamental importance to molecular biologists engaged in the Human Genome Project, devise new algorithmic approaches for solving these problems, program and test the algorithms that are developed, and make useful computer code available to the biology community. An educational component of this project is the training of Ph.D.'s in computer science who will be qualified to take up careers in computational biology.

Nearly all our research concerns the design and adaptation of data structures and algorithms for solving problems in sequence analysis or "stringology." This includes problems in string alignment and matching, local similarity search, restriction site mapping, clone ordering, and fragment assembly. Our emphasis is on finding solutions that are programmable, useful, and effective, as well as elegant and theoretically satisfying.

Projects Renewed in FY 1993

Projects Renewed in FY 1993

Building on prior results, we plan the following lines of investigation.

1. *Local Similarity Search:* Adaptation of Chang-Lawler filtering technique for approximate pattern matching. *O(kn)* dynamic programming algorithm when *k* is prespecified bound on number of errors. Dynamic programming for nonlinear scoring functions.
2. *Detection of Random Repeats and Palindromes:* Improvement of suffix-tree and other algorithms for random repeats, approximate palindromes, contiguous tandem repeats, etc.
3. *Comparison of Alignments:* Measure of similarity of two alignments. Dynamic programming table analysis to generate most dissimilar optimal alignments. Alignment comparison and its relation to parametric analysis carried out by PARAL.
4. *Multiple String Alignment:* Bounded-error heuristics for alternative scoring functions. Application of multiple common substring computation. Multiple alignments and consensus strings related to phylogenetic tree.
5. *Clone Ordering:* Dynamic programming for generation of least-cost agreement with probe data. Adaptation of traveling salesman algorithms.
6. *Sequencing by Hybridization:* Information theoretic analysis of hybridization-array design. Pooling of oligos and/or clones. Application of graph algorithms.
7. *Fast Fourier Transform (FFT):* Combinatorial interpretation of FFT algorithm when used to generate match counts. Match counts as filter for matching algorithms. FFT as subprocedure in other algorithms.
8. *Evolutionary Reconstruction Under High-Order Mutations:* Algorithms to find the least-cost reconstruction of a set of sequences where high-order mutations such as inversions, repetitions, and recombinations are permitted in addition to point mutations.

Foundations for a Syntactic Pattern Recognition System for Genomic DNA Sequences

David B. Searls
Department of Genetics; University of Pennsylvania School of Medicine; Philadelphia, PA 19104-6145
215/573-3107, Fax: -3111, Internet: *dsearls@cbil.humgen.upenn.edu*

The goal of this work is to extend, refine, and apply the principal investigator's research to linguistic analysis of biological sequences. A software system will be created to perform sophisticated pattern-recognition and related functions at abstraction and expression levels beyond current *general-purpose* pattern-matching systems for biological sequences; it will also perform with more-uniform language, environment, and graphical user interface and with greater flexibility, extensibility, embeddability, and ability to incorporate other algorithms than possible with current *special-purpose* analytic software. Specific aims are:

1. *Extended development of the graphical user interface and visualization tools.* A current dynamic parse-visualization tool will be enhanced and supplemented with static data-visualization routines for high-level iconic depiction of parse results. A graphical interface will be implemented to support interactive grammar development and refinement in a rapid-prototyping mode.
2. *Development of embeddability "hooks" for incorporation of and by other algorithms.* The system will be made into a platform for applying other algorithms in a hierarchical fashion; focusing them on regions of interest; providing a uniform environment for input, output, and parameter management; and assembling results into the grammar's structural model. The grammar system will be made embeddable in other platforms where appropriate.

3. *Incorporation of advanced parser technology and application to eukaryotic gene parsing.* Current developments in areas such as island and probabilistic parsing will be embedded in the system, driven by the specific practical problem of efficiently recognizing protein-coding eukaryotic genes. Current statistical and heuristic gene-finding algorithms will be adapted to grammatical expression to allow for greater flexibility and contextually structured application.
4. *Extension of input formats accepted and header information processed by the parser.* For graphical depiction and high-level parsing, the current GenBank® flat-file entry parser will be extended to handle a variety of other formats and extract additional information from features tables. Facilities will also be developed for transparent connection to relational databases and ASN.1-formatted data streams.
5. *Extension of the grammar system to encompass protein sequence at multiple levels.* The parser will be extended to accept single-letter protein code as the primary sequence for describing motifs. Longer-term goals include the development of secondary structure grammars and the potential description of tertiary structures using coordinate grammars.
6. *Collaborations aimed at specific biological and computational problems.* To drive system development farther in biologically relevant directions, collaborations for grammar development will be undertaken with biologists and for parser development with computational biologists. A facility will be provided for remote access to the parser.
7. *Distribution and promotion of software and associated libraries.* Periodic software releases will be accompanied by full documentation and a reasonable level of support, particularly in developing new grammars. Grammars for use with the parser or other programs will be maintained in a central, publicly accessible repository of biological feature specifications.

Projects Continuing into FY 1993

Computational Support for the Human Genome Center: Statistical and Mathematical Analysis, Data Processing, and Databasing

Elbert Branscomb, Tom Slezak, David Nelson, and Anthony V. Carrano
Human Genome Center; Biology and Biotechnology Research Program; Lawrence Livermore National Laboratory; Livermore, CA 94551
510/422-5681, Fax: /423-3608, Internet: *elbert@alu.llnl.gov*

GnomeView: A Graphical Interface to the Human Genome

Richard J. Douthart, Joanne E. Pelkey, and David A. Thurman
Life Sciences Center; Pacific Northwest Laboratory; Richland, WA 99352
509/375-2653, Fax: -3649, Internet: *dick@gnome.pnl.gov*

Robust Contig Construction

Michael Cinkosky, Randall Dougherty, **Vance Faber**, Mark Goldberg,[1] Mark Mundt, Robert Pecherer, Doug Sorenson, and **David Torney**
Theoretical Biology and Biophysics Group; Los Alamos National Laboratory; Los Alamos, NM 87545
Torney: 505/667-7510, Fax: /665-3493, Internet: *dct@life.lanl.gov*
[1]Rensselaer Polytechnic Institute; Troy, NY 12181

Projects Continuing into FY 1993

HGIR: Information Management for a Growing Map

James W. Fickett, Michael J. Cinkosky, Michael A. Bridgers, Henry T. Brown, Christian Burks, Philip E. Hempfner, Tran N. Lai, Debra Nelson,[1] Robert M. Pecherer, Doug Sorenson, Peichen H. Sgro, Robert D. Sutherland, Charles D. Troup, and Bonnie C. Yantis
Theoretical Biology and Biophysics Group; Los Alamos National Laboratory; Los Alamos, NM 87545
505/665-5340, Fax: -3493, Internet: *jwf@life.lanl.gov*
[1]Department of Human Genetics; University of Utah; Salt Lake City, UT 84112

Identification of Genes in Anonymous DNA Sequences

Christopher A. Fields and **Carol A. Soderlund**[1]
The Institute for Genomic Research; Gaithersburg, MD 20878
301/869-9056, Fax: -9423
[1]Sanger Center; Cambridge, U.K.

BISP: VLSI Solutions to Sequence-Comparison Problems

Tim Hunkapiller, Leroy Hood, Ed Chen,[1] and Michael Waterman[2]
Department of Molecular Biotechnology; University of Washington; Seattle, WA 98195
206/685-7365, Fax: -7302, Internet: *tim@mudhoney.mbt.washington.edu*
[1]Jet Propulsion Laboratory; Pasadena, CA 91109
[2]University of Southern California; Los Angeles, CA 90089

Efficient Identification and Analysis of Low- and Medium-Frequency Repeats

Jerzy Jurka, Aleksandar Milosavljevic,[1] Jolanta Walichiewicz, and Sherman Yang
Linus Pauling Institute of Science and Medicine; Palo Alto, CA 94306
415/327-4064, Fax: -8564, Internet: *jurek@jmullins.stanford.edu*
[1]Biological/Medical Research Division; Argonne National Laboratory; Argonne, IL 60439-4833

A Human Genome Database

David Kingsbury, Ken Fasman, and Peter L. Pearson
Genome Data Base; Johns Hopkins University School of Medicine; Baltimore, MD 21205
410/955-7058, Fax: /614-0434, Internet: *dkingsbu@gdb.org*

Genome Assembly Manager

Charles B. Lawrence, Eugene W. Myers,[1] and Sandra Honda
Department of Cell Biology; Baylor College of Medicine; Houston, TX 77030-3498
713/798-6226, Fax: /790-1275, Internet: *chas@mbir.bcm.tmc.edu*
[1]Department of Computer Science; University of Arizona; Tucson, AZ 85721

Laboratory Information Management System (LIMS) for Megabase Sequencing

Victor M. Markowitz
Data Management Group and Human Genome Center; Information and Computing Sciences Division; Lawrence Berkeley Laboratory; Berkeley, CA 94720
510/486-6835, Fax: -4004, Internet: *v_markowitz@lbl.gov*

Database Tools Development

Victor M. Markowitz,[1,2] Arie Shoshani,[1] and Ernest Szeto[1]
[1]Data Management Group and [2]Human Genome Center; Information and Computing Sciences Division; Lawrence Berkeley Laboratory; Berkeley, CA 94720
510/486-6835, Fax: -4004, Internet: *v_markowitz@lbl.gov*

A Computer System for Access to Distributed Genome Mapping Data

Thomas G. Marr and **Andrew Reiner**
Cold Spring Harbor Laboratory; Cold Spring Harbor, NY 11724
516/367-8393, Fax: -8416, Internet: *marr@cshl.org*

Applying Machine Learning Techniques to DNA Sequence Analysis

Jude W. Shavlik, Michiel O. Noordewier,[1] Geoffrey Towell, Mark Craven, Andrew Whitsitt, Kevin Cherkauer, and Lorien Pratt[1]
Department of Computer Science; University of Wisconsin; Madison, WI 53706
608/262-7784, Fax: -9777, Internet: *shavlik@cs.wisc.edu*
[1]Department of Computer Science; Rutgers University; New Brunswick, NJ 08903

Computational Analysis and Support for Extensive Physical Mapping of Genomes

Tom Blackwell, David Balding, Frederic Fairfield, Jim Fickett, Catherine Macken, Karen Schenk, **David Torney**, Burton Wendroff, and Clive Whittaker
Los Alamos National Laboratory; Los Alamos, NM 87545
Torney: 505/667-7510, Fax: /665-3493, Internet: *dct@life.lanl.gov*

Informatics Support for Mapping in Mouse-Human Homology Regions

Edward Uberbacher, Richard Mural,[1] Eugene Rinchik,[2] and Richard Woychik[1]
Engineering Physics and Mathematics Division and [1]Biology Division; Oak Ridge National Laboratory; Oak Ridge, TN 37831-6364
615/574-6134, Fax: -7860, Internet: *ube@ornl.gov* or *ube@ubersun.epm.ornl.gov*
[2]Sarah Lawrence College; Bronxville, NY 10708

An Intelligent System for High-Speed DNA Sequence Pattern Analysis and Interpretation

Edward Uberbacher, Richard Mural,[1] Ralph Einstein, and **Reinhold Mann**
Engineering Physics and Mathematics Division and [1]Biology Division; Oak Ridge National Laboratory; Oak Ridge, TN 37831-6364
615/574-6134, Fax: -7860, Internet: *ube@ornl.gov* or *ube@ubersun.epm.ornl.gov*

Ethical, Legal, and Social Issues (ELSI) Related to Data Produced in the Human Genome Project

Projects New in FY 1993

Guidelines for Protecting Privacy of Information Stored in Genetic Data Banks

George J. Annas and Leonard H. Glantz
Law, Medicine, and Ethics Program; Boston University School of Public Health; Boston, MA 02118
617/638-4626, Fax: -5299

This 16-month project will answer the following questions to develop proposed policies and laws for safeguarding personal genomic information stored in "genetic data banks":

1. In what ways are genomic and medical information similar and different?
2. Under what circumstances should private entities or public agencies be permitted legally to obtain DNA samples from individuals?
3. When is consent required for the storage of DNA samples and genomic data? Is informed consent possible for such storage when the material might someday reveal personal health-related information about individuals and their genetic relatives?
   - Who "owns" genomic information? Are genomic records different from medical records because they always include information about others? Does the genetic data bank have a special duty to notify individuals of new information that can be obtained from their stored genetic data? Must genetic relatives be notifed that they have or are at risk for developing a serious condition?
   - Who can have access to the stored DNA and genomic information and for what purposes? Should the subject have access and destruction rights, and under what circumstances?
   - Should time limits be set for storage of DNA and genomic information? How should such limits be set?

To answer these questions, we will review the legal literature, case law, and statutes on privacy with an emphasis on issues specifically related to highly personal information stored in computers. We will identify policies and laws that are most analogous to genetic data-banking concerns and adapt them to special privacy issues raised by the use and storage of genetic data. The product will be a monograph on genetic privacy, including proposed guidelines (with explanations and justifications) that could be adopted as laws or regulations governing the collection, storage, and use of genomic information and DNA samples. The guidelines will be presented at national meetings and submitted for publication in the medical and legal literature.

Genome Technology and Its Implications: A Hands-On Workshop for Educators

Diane Baker and **Paula Gregory**[1]
Human Genome Center; University of Michigan; Ann Arbor, MI 48109-0674
313/747-2738, Fax: /764-4133
[1]National Center for Human Genome Research; Bethesda, MD

The goal of this project is to develop and conduct workshops for instructing high school science teachers in the latest developments in human genetic research and its ethical and social implications. The workshops are designed to help teachers incorporate this information into their everyday classroom curriculum.

Two 5-day continuing-education workshops will be conducted for 25 participants at the Human Genome Center. Educators will be trained in the pedagogy of molecular genetics and in methods for bringing the latest in biotechnology and its implications to the attention of their students. The workshops will be open to teachers throughout the Great Lakes states, in cooperation with the

education subcommittee of the Great Lakes Regional Genetics Group. These workshops will not only introduce teachers to biotechnology but will also promote partnerships among education specialists, clinical geneticists, research scientists, and teachers.

HUGO International Yearbook: Genetics, Ethics, Law, and Society (GELS)

Alex Capron and **Bartha Knoppers**[1]
Pacific Center for Health Policy and Ethics; University of Southern California; Los Angeles, CA 90089-0071
213/740-2557, Fax: -5502, Internet: *acapron@law.usc.edu*
[1]Faculty of Law; University of Montreal; Quebec, Canada H3C 3JF

This project has the goal of planning over a 9-month period how to create inexpensively and efficiently a resource of information and analysis under the aegis of the Human Genome Organization (HUGO) Ethics Committee. This resource will consist of a data bank and an annually published compilation, the *HUGO International Yearbook: Genetics, Ethics, Law, and Society (GELS)*. The yearbook will enable people around the world in science, medicine, government, law, and ethics to stay abreast of pertinent statutes, regulations, guidelines, and analysis produced in other nations and by international organizations. Keeping track of developments is difficult because work on ethical, legal, and social implications (ELSI) of human genome research is not only international but also interprofessional and interdisciplinary.

Besides a compilation of documents presented in an organized fashion, the yearbook will contain original material analyzing ELSI information by topic and describing national, regional, and international trends. Thus, persons or groups interested in comparative information on topics of concern to genome research efforts—such as accessibility and confidentiality of genetic information, policies on patenting gene sequences, and implications for employment and insurance—will have a ready source of governmental policies and rulings to allow informed appraisal of such policies in light of developments in other fields and other countries. Presentation of this material should also enrich thinking in the field by facilitating the discovery of connections (in principles, objectives, and practical methods) among ELSI topics that are not always analyzed together.

The yearbook will be planned by a 12-person committee consisting of internationally representative ELSI experts and specialists in bibliographic and documentation methodologies. The committee will include representatives from Europe and North America and at least one person with firsthand knowledge about developments in Central and South America, Asia and India, Australia and Oceania, and Africa. The project is supported by DOE and the Canadian Genome and Technology Program.

The Human Genome: Science and the Social Consequences; Interactive Exhibits and Programs on Genetics and the Human Genome

Charles C. Carlson
The Exploratorium; San Francisco, CA 94123
415/563-7337, Fax: /561-0307, Internet: *charliec@exploratorium.edu*

The Exploratorium is continuing a long-term commitment to increase public awareness of the Human Genome Project, the basic science of genetics, and the implications for society. The comprehensive multiyear program has produced five interactive exhibits on DNA and the highly successful September 1992 genetics and biotechnology symposium entitled *Winding Your Way Through DNA*, which was cohosted with the University of California, San Francisco. Other components of the plan include a DNA extraction demonstration, 13 new interactive exhibits about genetics, and the Genetics Pathway for teachers and students to address the basic science of genetics. A lecture series focusing specifically on ethical, legal, and social issues will serve as both

Projects New in FY 1993

a public forum and a means of creating video records for a freestanding exhibit about ethical considerations. These exhibits, demonstrations, and programmatic activities will involve collaborations with regional biotechnology firms, universities, and museums.

Important concepts addressed by exhibits and programs are DNA as the molecule of heredity; mutation; variation; and the relationship between genes and proteins. Focus is on the Human Genome Project as an example of how basic research can result in practical applications and raise important social and ethical questions. This effort will produce a major museum section on the Human Genome Project and genetics that will serve as a model for other museums nationwide. It will be viewed by more than 625,000 visitors annually and used by more than 500 teachers attending training programs.

The Human Genome Project: Information Management, Access, and Regulation—Educational Materials for High School Biology

Joseph D. McInerney and Lynda Micikas
Biological Sciences Curriculum Study; Colorado College; Colorado Springs, CO 80903
719/578-1136, Fax -9126, Internet: *jmcinerney@cc.colorado.edu*

The Biological Sciences Curriculum Study (BSCS) will produce and distribute free of charge to all 50,000 U.S. high school biology teachers an instructional module and software titled *The Human Genome Project: Information Management, Access, and Regulation.* BSCS will use the process of curriculum development that it has refined in the 35 years since the organization's inception. The module will provide about 25 pages of background information for teachers on information technology as it relates to the Human Genome Project and on ethical, legal, and social issues related to genomic databases. Five days of classroom instruction will involve students directly in manipulation of hypothetical mapping and sequence databases; online searches of the Genome Data Base; and classroom analyses of the ethical, legal, and social issues.

An eight-person advisory committee will oversee the 22-month project. This committee will have expertise in clinical genetics, molecular biology, genome databases, classroom teaching, ethics, and law. The education committees of the American Society of Human Genetics, National Society of Genetic Counselors, and Council of Regional Networks for Genetic Services will review the project materials at two critical junctures, at no cost to the project.

Pilot Senior Research Fellowship Program: Bioethical Issues in Molecular Genetics

Declan Murphy and **Claudette Cyr Friedman**
Library of Congress; Washington, DC 20540
Friedman: 202/707-1513, Fax: -1714

As we approach the end of the twentieth century, the pace of scientific discovery and technological innovation is so rapid that society confronts an almost continuous stream of difficult new philosophical questions regarding the social consequences of its own inventiveness. These questions are particularly acute in research into molecular genetics and new therapeutic possibilities for treating genetic-based disorders. Critics and proponents have expressed concern about the nature and direction of this new research, social implications of new genetic technologies, and potential disruptions to traditional definitions of our own humanity. Conflicting arguments and perspectives involving an array of emotionally charged individual issues make it all the more difficult to grasp how traditional philosophy will change as a result of these scientific advances.

Rigorous new scholarship must be generated to examine philosophical issues surrounding uses and potential misuses of molecular genetics research. Rethinking philosophical paradigms in approaching these difficult questions will greatly benefit practicing scientists, end users of these new technologies, policymakers who must regulate them (often with too little information), and society as a whole.

Given that the Human Genome Project has mandated a portion of its funding for examining the ethical implications of this work, DOE and the Library of Congress (LC) are sponsoring a 1-year Senior Research Fellowship in Bioethical Issues in Molecular Genetics. The fellow's task will be to (1) conduct a critical review of existing bioethics literature on molecular genetics and (2) generate new frameworks for analyzing ethical questions and public-policy implications arising from research on the human genome. The fellow will produce both oral and written presentations of his or her work for DOE, LC, Congress, the greater scientific community, and the general public.

DNA Banking and DNA Data Banking: Legal, Ethical, and Public Policy Issues

Philip R. Reilly
Shriver Center; Waltham, MA 02254
617/642-0230, Fax: /893-5340

The research objectives of this project are to focus on issues of genetic privacy in critically analyzing selected aspects of DNA banking and DNA data banking. Such banking presently takes place in four sectors: (1) DNA banking by the military to assist in identifying human remains; (2) state-based DNA forensic data banking to assist in resolving violent crimes, identifying missing persons, and analyzing crime patterns; (3) academically based repositories housed in the laboratories of scientists who are studying particular genetic disorders; and (4) commercially based repositories that offer DNA banking services to researchers and individuals. The primary goals of our research are to learn how informational privacy and related issues are handled, especially in computer security; analyze current approaches; and suggest guidelines for the future.

Although we will study all four sectors, our primary focus will be on academically and commercially based DNA data banking, which we believe carry the broadest potential for harm to the privacy interests of individuals. These banks remain essentially unregulated, unlike the military, which has implemented strict rules limiting access to stored DNA information (Weedn, 1992, 1991). Similarly, state forensic laboratories operate under statutory mandates that typically include generalized directives on the need for confidentiality (Armed Forces Institute of Pathology/American Registry of Pathology, course materials).

Social Science Concepts and Studies of Privacy: A Comprehensive Inventory and Analysis for Considering Privacy, Confidentiality, and Access Issues in the Use of Genetic Tests and Applications of Genetic Data

Alan F. Westin
Center for Social and Legal Research; Hackensack, NJ 07601
201/996-1154, Fax: -1883, America Online: *alanrp@aol.com*

The objectives of this project are to completely update social science theoretical and empirical work on privacy, confidentiality, and individual rights of access to personal records since the publication of Westin's *Privacy and Freedom* in 1967; to refine the concepts and operative dynamics first formulated by Westin in light of research and social developments since 1967; and to relate this updated research and reconceptualization to privacy, confidentiality, and access issues likely to arise from the use of genetic data generated by the Human Genome Project, including the uses of genetic information in computerized data banks.

Projects New in FY 1993

This research will also examine the development, implementation, and effectiveness of legal and organizational privacy-protection measures in the United States since the late 1980s; current debates about updating such measures to reflect scientific, technological, social, and cultural changes in the early 1990s; and implications for selection of social, organizational, and legal policies to deal with genetic testing and the application of genetic data in the late 1990s and early 21st century.

In addition to an overall project monograph and five to seven topical area reports in hard copy, the project will produce a set of computerized resources on disk, CD-Rom, or online. These resources will include the monograph and area reports, an annotated master bibliography, a master classification system for privacy issues relating to genetic information uses, a directory of scholars working on privacy aspects of genetic data applications, and a set of selected and classified abstracts of major social science works dealing with privacy aspects of genetic data uses.

Resources on the Ethical, Legal, and Social Implications of the Human Genome Project

Michael S. Yesley
Los Alamos National Laboratory; Los Alamos, NM 87545
505/667-3766, Fax: /665-4424, Internet: *msy@lanl.gov*

A comprehensive bibliographic database and collection of publications related to the ethical, legal, and social implications (ELSI) of the Human Genome Project have been developed at the General Law Library of Los Alamos National Laboratory (LANL). Over 5600 books and articles have been catalogued, and copies of most of these materials have been obtained for the library collection.

Yesley and his staff have also prepared a second printed edition of *ELSI Bibliography: Ethical Legal and Social Implications of the Human Genome Project*, extracted from the computer database. This edition has been distributed to several thousand interested individuals and medical libraries. It provides a current and comprehensive resource for identifying publications on 17 major topics, e.g., counseling, discrimination, eugenics, DNA fingerprinting, patents, privacy, reproduction, screening, and therapy. This is not a complete inventory of the topics (keywords) that can be searched in the underlying computer database. Yesley will provide custom searches of the database to researchers interested in other topics. Also, the database may be made accessible online, enabling researchers to perform their own searches.

The ELSI bibliography is drawn from the computer database, which can be searched and sorted on a variety of parameters and combinations of parameters. For example, a search can be limited to any topic or name, sorted chronologically or alphabetically by author, or limited to certain periods or publications. The results of custom searches can be sent by e-mail, fax, or U.S. mail to the requestors. Researchers are also invited to use the extensive collection of ELSI books and articles at the LANL General Law Library.

Adjudicating Genetic Testing and Gene Therapy Litigation: Factors Influencing the Privacy Law and Bioethics Attitudes and Opinions of Trial and Appellate Judges Managing Lawsuits Related to Scientific and Technological Advances Emanating from the Human Genome Project

Franklin M. Zweig
Center for Health Policy Research; George Washington University; Washington, DC 20052
202/296-6922, Fax: -0025 or /785-0114

The George Washington University Center for Health Policy Research is conducting a 12-month research and education project in genetic science to assist federal and state courts in adjudicating lawsuits involving genetic testing and gene therapy. The project's goal is to produce a desk book to assist judges' management of such information under the rubric "novel scientific evidence." To aid in desk book production, a midproject workshop will brief judges and other court-affiliated personnel about basic science and research horizons related to the Human Genome Project and will produce two documents. The National Conference of Metropolitan Courts; National Association of Court Managers; State Justice Institute; National Institute of Justice, SEARCH, Inc.; Federal Bureau of Investigation; and Association of Academic Health Centers will collaborate in the project.

1993 Conferences

Maximizing the Return From Genome Research: A Conference Concerning the Effect of Patent Scope on Commercial Incentives

May 20–23, 1993, at Concord, New Hampshire
Thomas G. Field; Franklin Pierce Law Center; University of New Hampshire; Concord, NH
603/228-1541, Fax: /224-3342

Teaching Ethics in the Biomedical and Biological Sciences: A Workshop for Research Faculty

June 12–16, 1993, at Bar Harbor, Maine
Ed Golub; Pacific Center for Ethics and Applied Biology; San Diego, CA
619/625-0734, Fax: /793-3632

Between Design and Choice: The Social Shaping of Genetic and Reproductive Technologies

April 2–4, 1993, at Ithaca, New York
Sheila Jasanoff; Cornell University; Ithaca, NY
607/255-6049, Fax: -6044

Projects Continuing into FY 1993

"The Secret of Life"

Paula Apsell and **Graham Chedd**
WGBH Educational Foundation; Boston, MA 02134
Chedd: 617/492-2777 Ext 4352, Fax: /787-5781

Predicting Future Disease: Issues in the Development, Application, and Use of Tests for Genetic Disorders

Ruth E. Bulger and Jane E. Fullarton
National Academy of Sciences; Institute of Medicine; Washington, DC 20418
Fullarton: 202/334-3913, Fax: -2031

Human Genetics Education for Middle and Secondary Science Teachers

Debra L. Collins, R. Neil Schimke, and **Linda Segebrecht**[1]
Department of Medical Genetics; University of Kansas Medical Center; Kansas City, KS 66160-7318
913/588-6043, Fax: -3995, Internet: *ukanvm.cc.ukans.edu*
[1]Science Pioneers; Kansas City, MO 64110

Pathways to Genetic Screening: Patient Knowledge—Patient Practices

Troy Duster and Diane Beeson[1]
Institute for the Study of Social Change; University of California; Berkeley, CA 94720
510/642-0813, Fax: -8674
[1]Department of Sociology; California State University; Hayward, CA 94542

Studies of Genetic Discrimination

Marvin Natowicz
Division of Medical Genetics; Shriver Center; Waltham, MA 02254
617/642-0176, Fax: /894-9968

"Medicine at the Crossroads"

George Page and **Stefan Moore**
WNET/Thirteen; New York, NY 10019
212/560-2767, Fax: /582-3297

Impact of Technology Derived from the Human Genome Project on Genetic Testing, Screening, and Counseling: Cultural, Ethical, and Legal Issues

Ralph W. Trottier, Lee A. Crandall,[1] David Phoenix,[2] Mwalimu Imara,[2] and Ray E. Moseley[1]
Department of Pharmacology and Toxicology and [2]Department of Community Health and Preventive Medicine; Morehouse School of Medicine; Atlanta, GA 30310
404/752-1711, Fax: /755-7318
[1]Department of Community Health and Family Medicine; University of Florida; Gainesville, FL 32610

BIOSCI Electronic Newsgroup Network for the Biological Sciences

Project Renewed in FY 1993

Peter Arzberger
Division of Biological Instrumentation and Resources; National Science Foundation; Arlington, VA 22230
703/306-1469, Fax: -0356, Internet: *parzberg@nsf.gov*

The BIOSCI network is an electronic forum for distributing information and discussing issues relevant to biologists. The network is divided into a number of specialized biological fields or interests called "newsgroups," to which individual researchers can subscribe. In addition to communication in specialized biology fields, newsgroups are dedicated to various aspects of the Human Genome Project; general information; experimental methods; announcements of scientific meetings, funding, and job opportunities; and journal contents/abstracts. By submitting items via computer linkup (modem or local area network connection), contributors can communicate with the entire newsgroup without knowing individual contact information. Transmission is immediate, and BIOSCI includes several types of items that journals do not.

This project will actively promote newsgroup use through demonstrations, training sessions, and assistance with software installation. More newsgroups will be included in underserved biological areas, and message archiving and retrieval by e-mail or file transfer protocol will be facilitated. Plans are also under way for prominent scientists to serve as moderators, a possible first step in the evolution of a few selected newsgroups into electronic journals. A database of all newsgroups dedicated to the biological sciences will be maintained, along with a directory of electronic mail addresses of biologists.

Projects Continuing into FY 1993

The Human Genome Distinguished Postdoctoral Fellowships

Linda Holmes and **Alfred Wohlpart**
Science and Engineering Education Division; Oak Ridge Associated Universities; Oak Ridge, TN 37831-0117
615/576-3192, Fax: -0202

Support of Human Genome Program Proposal Reviews

Amanda Lumley and **James Wright**
Training and Management System Division; Oak Ridge Associated Universities; Oak Ridge, TN 37831-0117
615/576-6752, Fax: -0202, Internet: *lumleya%orau2@cunyvm.cuny.edu*

Human Genome Management Information System

Betty K. Mansfield, Anne E. Adamson, Denise K. Casey, K. Alicia Davidson, Sheryl A. Martin, Donna B. Stinnett, **John S. Wassom**, Judy M. Wyrick, and Laura N. Yust
Health Sciences Research Division; Oak Ridge National Laboratory; Oak Ridge, TN 37831-6050
615/576-6669, Fax: /574-9888, Internet: *bkq@ornl.gov*, BITNET: *bkq@ornlstc*

Human Genome Coordinating Committee Administration

Sylvia Spengler
Human Genome Center; Lawrence Berkeley Laboratory; Berkeley, CA 94720
510/486-4879, Fax: -5717, Internet: *sylviaj@violet.berkeley.edu* or *ux5.lbl.gov*

Projects Continuing into FY 1993

Assistance for Ethical, Legal, and Social Issues Projects

Michael S. Yesley
Center for Human Genome Studies; Los Alamos National Laboratory; Los Alamos, NM 87545
505/667-3766, Fax: /665-4424, Internet: *msy@lanl.gov*

SBIR awards are designed to stimulate commercialization of new technology for the benefit of both private and public sectors. The DOE SBIR program consists of three phases:

- **Phase I:** Awards for up to 6 months and $75,000 for a firm to explore the scientific and technical merit and feasibility of a research idea.
- **Phase II:** Awards for up to 2 years and $500,000 to expand on Phase I results and pursue further development. Only Phase I awardees are eligible for Phase II, which is the principal research and development effort.
- **Phase III:** Private or non-SBIR federal funding to commercialize Phase II results. No Phase III projects are listed here.

SBIR Phase I

Phase I Projects New in FY 1993

High-Performance Searching and Pattern Recognition for Human Genome Databases

Douglas J. Eadline
Paralogic, Inc.; Bethlehem, PA 18015
215/861-6960, Fax: -8247, Internet: *deadline@plogic.com*

The goal of the Human Genome Project is to determine the sequence of most or all 3 billion nucleotide bases. Resulting genome database size will require new technologies that provide cost-effective and time-effective searching, interpretation, and analysis. One very promising approach to this problem is to (1) apply grammars expressed in the Prolog language ("linguistic analysis") to the "genetic language" expressed by DNA and (2) search in a concurrent or parallel fashion. The recent emergence of parallel computers and parallel programming tools has created the opportunity for rapid and cost-effective linguistic analysis of DNA sequences. This project seeks to determine the feasibility of combining grammar-based searching, Paralogic *n-parallel* **PROLOG™** and parallel computers to search genome databases efficiently and quickly.

A High-Spatial-Resolution Spectrograph for DNA Sequencing

Cathy D. Newman
CHROMEX, Inc.; Albuquerque, NM 87107
505/344-6270, Fax: -6095

The Human Genome Project, if it is to be successful, will require much more rapid DNA sequencing. Present technology is estimated to be 100 to 500 times too slow to permit sequencing of the entire human genome by 2005. The ABI 373A DNA sequencer, the most successful technology currently employed for large-scale DNA sequencing, involves real-time, four-color, fluorescent line imaging during the electrophoresis process. In our estimation, this basic technology can be scaled to significantly higher levels and is likely to remain the method of choice for future DNA sequencing. Indeed, a number of research groups have been developing faster fluorescent DNA sequencers. Although great strides have been made in reducing the time of electrophoresis by 10- to 20-fold, future gains in throughput will depend primarily on increased multiplexing. High-throughput instruments will require a detection system with hundreds of spatially resolved channels.

CHROMEX proposes to test a completely new concept in high-spatial-resolution spectroscopy. Performance will be achieved by using a radically different approach to the spectrograph design. Such a spectrograph, when combined with a charge-coupled device array detector, will produce a system capable of rapidly recording spectral signatures from more than 300 spatial locations.

Phase I Projects New in FY 1993

Nonradioactive Detection Systems Based on Enzyme-Fragment Complementation

Peter Richterich
Collaborative Research, Inc.; Waltham, MA 02154
617/487-7979, Fax: /891-5062, Internet: *peter@cric.com*

Assays based on DNA hybridization are very commonly used in molecular biology, clinical diagnostics, and the Human Genome Project. However, efficiency and sensitivity of hybridizations are often limited because of the nonspecific binding of probe molecules or detection-system components to target DNA or membrane supports.

In this project, detection methods will be developed for extremely sensitive, reliable, background-free detection in hybridization-based assays. The system will be based on two neighboring hybridization probes, each labeled with one subunit of alkaline phosphatase. Individual subunits of alkaline phosphatase are inactive, and nonspecific binding of hybridization probes to target DNA or membrane support will therefore not lead to background noise. When two probes bind next to each other on the target DNA, the two enzyme subunits can combine and form the active dimeric enzyme, which can be detected with chemiluminescent or colorimetric substrates. Increased hybridization specificity results from the necessity for two probes to hybridize next to one another.

In Phase I of this work, the feasibility and potential of such binary detection systems will be demonstrated. Conditions for generation, purification, storage, and use of subunit-labeled oligonucleotide probes will be defined and optimized. Finally, the binary detection systems will be used for chemiluminescent multiplex sequencing.

Separation Media for DNA Sequencing

David S. Soane and Herbert H. Hooper
Soane Technologies, Inc.; Hayward, CA 94545
510/293-1850, Fax: -1860

High-speed high-throughput DNA sequencing methods often rely on electrophoretic separation in very narrow geometries such as microcapillaries and ultrathin slabs. As channel dimensions decrease, the separation medium becomes a critical limiting factor in the speed, accuracy, and reproducibility of DNA fractionations. Conventional in situ casting of gels is not well suited for confined geometries, and the current approach of using noncrosslinked, entangled polymer solutions has several performance deficiencies, including those of resolution and reproducibility. This project involves a completely novel separation-media concept that combines the desirable performance attributes of crosslinked networks, the loading/unloading feature of dilute polymer solutions, and the user convenience of precast gels. The new approach will use discrete "smart" gel particles that form a free-flowing pseudo network under appropriate conditions and can be readily injected into and flushed from narrow capillaries. The feasibility of using this technology for DNA separation by capillary electrophoresis will be investigated in this project.

Interactive DNA Sequence Processing for a Microcomputer

Wayne Dettloff and **Holt Anderson**
Advanced Technology Applications, Inc.; Research Triangle Park, NC 27709
919/248-1800, Fax: -1455

DNA and biosequences are being identified faster than they can be compared and analyzed, and present techniques for rigorously searching a large database are costly and time-consuming. Under all plausible growth projections, the problem will soon become overwhelming.

A custom, very large scale integrated (VLSI) circuit has been designed that could be used for high-speed comparison, analysis, and interpretation of DNA and protein sequences. In a single pass with a statistically rigorous criterion, the systolic array can scrutinize each biosequence segment in a database to determine its homology to an input pattern. Phase I research includes designing and prototyping a printed circuit board for an IBM-compatible AT personal computer; this low-cost circuit board uses a full custom VLSI integrated circuit. Software is being developed to enable the board to interface with commonly used public-domain software packages (e.g., BLAST3), and the performance of the system in actual laboratory settings is being evaluated against current techniques. In Phase II, an interactive analysis platform will be developed on the basis of evaluation results obtained in Phase I.

Low-Cost Massively Parallel Neurocomputing for Pattern Recognition in Macromolecular Sequences

John R. Hartman
Computational Biosciences, Inc.; Ann Arbor, MI 48106
313/426-9050, Fax: -5311, Internet: *john@cbi.com*

Connectionist (neural network) approaches to pattern recognition and analysis have attracted great interest recently because of their flexibility, ability to learn by example, and ability to "self-organize" to reveal hidden patterns and relationships in the input data set. Neural networks are inherently parallel computational structures and thus potentially excellent candidates for implementation on massively parallel computers. Particularly in the training stage, serial implementations of connectionist models are often limited either by network size or by the number of practical training trials. The explosion in the size of macromolecular sequence databases (such as GenBank®) has created a need for pattern-analysis solutions with superior cost-performance characteristics.

Phase I will implement efficient parallel algorithms for all critical components of a basic multilayer, feed-forward neural network model. These components include a "linear combiner" for the multiplication of connection-weight matrices with input (or error) vectors, a sigmoidal activation function to introduce nonlinearity in neuron behavior, and a complete parallel implementation of the back-propagation (generalized Delta rule) training algorithm. Once implemented, the network will be evaluated against alternative parameterizations in a prototype DNA sequence-pattern-recognition task.

Electrophoretic Separation of DNA Fragments in Ultrathin Planar-Format Linear Polyacrylamide

Michael T. MacDonell and **Darlene B. Roszak**
Ransom Hill Bioscience, Inc.; Ramona, CA 92065
619/789-9483, Fax: -6902

Linear or uncross-linked polyacrylamide has been used successfully in capillary electrophoresis to separate nucleic acids. Typical acrylamide concentrations for those applications are 3 to 14% (w/v), with consistencies ranging from almost liquid to moderately viscous. Its relatively fluid nature at typical concentrations and the absence of cross links have caused linear polyacrylamide in planar (slab) gel electrophoresis to be overlooked.

Projects Continuing into FY 1993

We have described an application of ultrathin (100 μm) high-viscosity slabs of linear polyacrylamide to planar electrophoresis of nucleic acid fragments [M. T. MacDonell and D. B. Roszak, *Gene Anal. Tech. Appl.* **10**, 10–15 (1993)]. The approach is rapid-end-yield, high-resolution separation of nucleic acid fragments in linear polyacrylamide supports. The mobility of DNA fragments of various lengths in a range of linear polymer concentrations is compared with the mobility for conventional cross-linked gels.

The reptative migration of larger DNA fragments in linear polymers is predictable from models of cross-linked acrylamide and agarose, but the migration of smaller fragments is not entirely consistent with the Ogston model. Relative mobilities for very small DNA fragments are about half those predicted by the Ogston regime.

The tendency of smaller fragments to deviate from predicted mobilities benefits the user because the overall effect is not unlike that of a field-strength gradient. An additional benefit of using water-soluble linear polymers is that quantitative recovery of DNA fragments from these gels requires only that the band be excised and dropped into buffer. The useful range of linear polymer concentrations in planar sequencing appears to be 20 to 35% (w/v), appropriate for the electrophoretic separation of DNA fragments ranging from 50 to about 2500 bp in length.

An Acoustic Plate Mode DNA Biosensor

Douglas J. McAllister
BIODE, Inc.; Cape Elizabeth, ME 04107
207/883-1492, Fax: -1482

This project is developing a new solid-state biosensor technology for biological measurements. The technology is based on merging two relatively different technologies: nucleic acid probe (NAP) DNA hybridization technology with acoustic plate mode (APM) microsensors. Recent advances in the design and operation of APM devices have demonstrated a DNA hybridization sensor principle with excellent sensitivity (nanogram/milliliter), selectivity, and temperature stability when used with a model probe system.

This project is the first attempt at direct electronic in situ sensing of a diagnostically significant DNA gene sequence that codes for the abundant late matrix antigen of *Cytomegalovirus*. The new DNA biosensor operates in a continuous and therefore homogeneous mode and yet is expected to equal or exceed the sensitivity limits of existing dot-blot technologies. Thus, the biosensor is expected over time to respond in situ to the target DNA. In contrast, the current technology yields a discrete endpoint result. The biosensor response will be an electronic signal, allowing numerous data-handling options.

The Phase I project consists of several components: (1) construction of an optimized dual delay line APM sensing element, (2) design and construction of a dual oscillator measurement system for the optimized APM, (3) optimization of the chemistry employed to attach NAP covalently, (4) evaluation of sensor performance in detecting in situ hybridization of target DNA, and (5) completion of the sensor characterization.

Phase II Projects Continuing into FY 1993

Increased Speed in DNA Sequencing by Utilizing LARIS and SIRIS to Localize Multiple Stable Isotope–Labeled Fragments

Heinrich F. Arlinghaus
Atom Sciences, Inc.; Oak Ridge, TN 37830
615/483-1113, Fax: -3316

Site-Specific Endonucleases for Human Genome Mapping

George Golumbeski, Kimberly Knoche, Susanne Selman, Jim Hartnett, and Lydia Hung
Promega Corporation; Madison, WI 53711
608/274-4330, Fax: /277-2516

Current large-scale genome mapping methodology suffers from a lack of tools for generating specific DNA fragments in the megabase-size range. To address this need, Promega Corporation conducted several Phase I studies. These studies examined the feasibility of developing a set of site-specific endonucleases capable of generating DNA fragments in the 2- to 100-Mb-size range in a single step. Phase I demonstrated that I-*Ppo*I, the group I intron-encoded endonuclease, has excellent potential for use in general molecular biological and human genome research. This potential stems from I-*Ppo*I's ability to be expressed and purified at high yield, its stability and activity under a variety of reaction conditions, and its highly efficient single-site cleavage of genomic DNA embedded in agarose.

Phase II is designed to develop I-*Ppo*I for commercialization and broaden its potential for human genome mapping and analysis of other complex genomes. To accomplish these goals, we have systematically examined I-*Ppo*I's ability to tolerate single base substitutions within its recognition site. In addition, through cross-linking studies and crystallographic analysis, we are investigating the structure of I-*Ppo*I when bound to its recognition site. Using this information, we propose to: (1) identify structural modifications and reaction conditions that enhance I-*Ppo*I's rare-cutting capabilities by relaxing the enzyme's specificity; (2) isolate I-*Ppo*I proteins with mutations that enable the nuclease to cleave at altered recognition sequences; (3) further extend the cutting capabilities of the nuclease by combining approaches 1 and 2; and (4) test the ability of the native, mutant, and modified enzymes to cut human and other complex genomic DNA in agarose.

Thus, our Phase II work should provide a set of conditions and modified I-*Ppo*I enzymes with a range of useful cutting frequencies for complex genome-mapping applications. In addition, this work will provide a systematic structure-function analysis of a novel type of DNA-protein interaction. These results should lead to successful commercialization efforts that will focus on both immediate complex genomic research applications and on the longer-term goal of incorporating I-*Ppo*I into genomic mapping instrumentation.

High-Performance DNA and Protein Sequence Analysis on a Low-Cost Parallel-Processor Array

John R. Hartman and David L. Solomon
Computational Biosciences, Inc.; Ann Arbor, MI 48106
313/426-9050, Fax: -5311

Projects Continuing into FY 1993

Rapid, High-Throughput DNA Sequencing Using Confocal Fluorescence Imaging of Capillary Arrays

David L. Barker and **Jay Flatley**
Molecular Dynamics; Sunnyvale, CA 94086
408/773-1222, Fax: -8343

The goals of the Human Genome Project require DNA sequencing techniques that can increase throughput by an order of magnitude or more. Current automated fluorescence sequencing instruments require 5 to 10 h for a single gel run and accommodate a maximum of 36 reaction sets. Capillary gels can be run at higher voltages to yield separation of 300 to 500 bases in 1 to 2 h, but fluorescence-detection methods are limited to analyzing 1 capillary at a time.

Confocal fluorescence imaging can produce great sensitivity to detection and has the geometrical advantage of delivering excitation light through the same lens that collects fluorescence emission. R. A. Mathies and his colleagues have shown that a single scanning-confocal-fluorescence detector can detect DNA fragments in a parallel array of capillaries. Phase I was directed toward testing the feasibility of this technology for high-throughput DNA sequencing, including automated methods for loading and analyzing up to 96 simultaneous reaction sets in 1 to 2 h.

Phase II will identify and test various dye chemistries, lasers, base-coding methods, and software to identify solutions that would be successful in a commercial instrument. Low-viscosity gel matrices will also be tested, and a capillary-array filling and flushing system will be designed and built. A working prototype capillary-array sequencer will be built for testing in a high-throughput laboratory.

Spatially Defined Oligonucleotide Arrays

Stephen P. A. Fodor
Affymetrix; Santa Clara, CA 95051
408/481-3400, Fax: -0422, Internet: *steve_fodor@qmgates.affymetrix.com*

In Phase I, spatially defined arrays of oligonucleotide probes were constructed to study the feasibility of DNA sequencing by hybridization. Newly developed techniques in light-directed polymer synthesis were used to construct high-density oligonucleotide arrays, explore kinetic and solvent-related parameters of target hybridization, and read the hybridization positions by epifluorescence microscopy. Specific combinatorial synthesis strategies were designed to address experimental issues of parallel hybridization.

Phase II research will improve the basic technology by developing advanced instrumentation, including high-speed detection systems; upgrading the image-analysis software to handle larger data sets; and formulating algorithms for the design of application-specific arrays of probes. Completion of this work will lead to sequencing instrumentation that could provide order-of-magnitude improvements in DNA sequencing productivity and generate new technologies for exploring genetic diversity for diagnostic applications.

Chemiluminescent Multiprimed DNA Sequencing

Chris S. Martin, Corinne E. M. Oleson, and **Irena Bronstein**
Tropix, Inc.; Bedford, MA 01730
617/271-0045, Fax: /275-8581

The objective of this study is to improve the performance efficiency of a nonisotopic detection method for DNA sequencing. After electrophoretic separation, transfer to a nylon membrane, and incubation with a streptavidin-alkaline phosphatase conjugate, DNA sequencing-reaction products labeled with biotinylated primers are imaged by utilizing a chemiluminescent detection procedure that incorporates 1,2-dioxetane substrates for alkaline phosphatase. Upon dephosphorylation, these enzymatic substrates decompose and emit light.

In Phase I, the feasibility of a multiprime strategy for increasing the amount of DNA sequence information from a single membrane was assessed. Multiple sets of DNA sequencing reactions were loaded into a single set of gel lanes, and following electrophoretic separation and DNA transfer to nylon membrane, each set of sequencing reactions was individually detected. Multiple DNA sequencing primers were used, each bearing a unique ligand label, including biotin, digoxigenen, 2,4-dinitrophenyl, or fluorescein. Each set of reactions was detected with a ligand-specific alkaline phosphatase conjugate. The performance of various labels and enzyme conjugates was evaluated, and an optimum procedure was developed for rapid sequential detection of each set of labeled fragments.

Phase II will further optimize the procedure for sequential detection of each set of labeled fragments and will incorporate the newly developed protocols and individual reagents into a DNA-sequencing kit with multiple hapten-labeled primers. A complementary chemiluminescent detection kit containing hapten-specific enzyme conjugates will be produced for sequential identification of overlapping polymerase chain reaction (PCR) products. The creation of a specific membrane optimized for chemiluminescence will be investigated.

In addition, an apparatus for automating the blot-development steps will increase throughput and permit system scaleup. Techniques will also be developed for interfacing the multiplex-labeling DNA-sequencing procedures with PCR amplification and single-stranded DNA template purification. Protocols for these techniques will be incorporated into the research kits and will greatly expand their usefulness.

Projects Completed in FY 1993*

*Projects in this section have been completed or did not receive support through the DOE Human Genome Program in FY 1993. Page numbers refer to the 1991–92 program report (PR).

Resource Development

A Bacteriophage T4 In Vitro DNA Packaging System to Clone Long DNA Molecules (PR, p. 93)

Venigalla B. Rao, Vishakha Thakhar, and Lindsay W. Black

Synthetic Endonucleases (PR, p. 96)

Betsy M. Sutherland and Gary A. Epling

Expressed Sequence Tags (ESTs) from Human Brain cDNAs for Genome Mapping (PR, p. 97)

J. Craig Venter, Mark Adams, Mark Dubnick, Chris Fields, Jenny Kelley, Anthony Kerlavage, Ruben Moreno, and James Nagle

New Hosts and Vectors for Genome Cloning (PR, p. 98)

Philip A. Youdarian and Philip Greener

Physical and Genetic Mapping

Cloning and Characterization of Human Chromosome 21 YACs (PR, p. 102)

Jeffrey C. Gingrich and Steven Lowry

Mapping Instrumentation

Field-Flow Fractionation of Chromosomes and DNA (PR, p. 112)

J. Calvin Giddings

High-Resolution DNA Mapping by Scanning Transmission Electron Microscopy (PR, p. 112)

James F. Hainfeld

Automating the Analysis of Dot-Blot Hybridizations (PR, p. 113)

Joseph Jaklevic, Tony Hansen, William Kolbe, Linda Sindelar, Edward Theil, and Donald Uber

Cloning-Independent Mapping Technology for Genomic Fidelity, Contig Linking, cDNA Site Analysis, and Gene Detection (PR, p. 115)

Leonard Lerman

Quantitation in Electrophoresis Based on Lasers (PR, p. 117)

Edward S. Yeung

Sequencing Technologies

Informatics

Ethical, Legal, and Social Issues

Infrastructure

SBIR

Phase I

Phase II

Index to Principal and Coinvestigators Listed in Abstracts

Index

Index

Index

ANL*	Argonne National Laboratory, Argonne, IL
BNL*	Brookhaven National Laboratory, Upton, NY
DOE	Department of Energy
GDB*†	Genome Data Base
HERAC*	Health and Environmental Research Advisory Committee
HGCC*	Human Genome Coordinating Committee
HGMIS*†	Human Genome Management Information System (ORNL)
LANL*	Los Alamos National Laboratory, Los Alamos, NM
LBL*	Lawrence Berkeley Laboratory, Berkeley, CA
LLNL*	Lawrence Livermore National Laboratory, Livermore, CA
NCHGR†	National Center for Human Genome Research
NIH†	National Institutes of Health, Bethesda, MD
NLGLP*	National Laboratory Gene Library Project (LANL, LLNL)
OHER*	Office of Health and Environmental Research
ORNL*	Oak Ridge National Laboratory, Oak Ridge, TN
PNL*	Pacific Northwest Laboratory, Richland, WA
SBIR	Small Business Innovative Research

* Denotes U.S. Department of Energy organizations or funding recipients.
† Denotes U.S. Department of Health and Human Services organizations or funding recipients.

UNITED STATES DEPARTMENT OF ENERGY
ER-72
WASHINGTON, DC 20585

OFFICIAL BUSINESS

FIRST-CLASS MAIL
POSTAGE & FEES PAID
U.S. DEPT. OF ENERGY

PERMIT NO. G20